EUSKAL HERRIKO UNIBERTSITATEA

University of the Basque Country

PhD dissertation

Unsupervised Machine Translation

Mikel Artetxe

2020

EUSKAL HERRIKO UNIBERTSITATEA

University of the Basque Country

Unsupervised Machine Translation

PhD dissertation submitted by Mikel Artetxe under the supervision of Gorka Labaka and Eneko Agirre.

San Sebastian, June 2020.

Acknowledgments

Thank you / Eskerrik asko...

...lerro hauek irakurtzen ari zaren kuxkuxero horri, gutxirekin konformatu beharko zarelako.

...Ixakide guztiei eta, bereziki, Gorka eta Enekori, konplitzeko gutxienekoa izateaz gain sobera merezi duzuelako. Erbestetik bada ere ~~sekta~~ familia honen parte izaten jarraitzea espero dut!

...Dani, Sebastian, and the awesome Language Team at DeepMind.

...Holger, Kyunghyun and the rest of my past—and future—colleagues at FAIR.

This work was supported by a Facebook Fellowship and a FPU grant from MECD.

Abstract

The advent of neural sequence-to-sequence models has led to impressive progress in machine translation, with large improvements in standard benchmarks and the first solid claims of human parity in certain settings. Nevertheless, existing systems require strong supervision in the form of parallel corpora, typically consisting of several million sentence pairs. Such a requirement greatly departs from the way in which humans acquire language, and poses a major practical problem for the vast majority of low-resource language pairs.

The goal of this thesis is to remove the dependency on parallel data altogether, relying on nothing but monolingual corpora to train unsupervised machine translation systems. For that purpose, our approach first aligns separately trained word representations in different languages based on their structural similarity, and uses them to initialize either a neural or a statistical machine translation system, which is further trained through back-translation.

More concretely, having trained word embeddings in different languages from monolingual corpora, we learn a linear transformation to map them to a common space. The resulting cross-lingual embeddings can be used to translate at the word level by taking the nearest neighbor of each source word in the target language. Previous methods required a bilingual dictionary to learn such mapping, and worked by minimizing the distance between equivalent words in different languages through different optimization objectives. In this thesis, we propose a new mathematical framework that generalizes a substantial body of previous work, and design new variants that outperform them in standard benchmarks. In addition, we propose an iterative self-learning approach that alternates between the alignment learning and the dictionary induction in a bootstrapping fashion. By combining this procedure with an unsupervised initialization method, we are able to learn cross-lingual word embeddings in a completely unsupervised manner, while obtaining results that are comparable to those of existing supervised systems.

Having aligned the word embeddings in different languages, learning a fully fledged machine translation system requires generalizing from word level to text level translation. In this thesis, we explore two approaches to that end based on the two dominant paradigms in corpus-based machine translation: neural machine translation and phrase-based statistical machine translation. For the former, our proposed approach uses an attentional sequence-to-sequence model with a shared encoder and language specific decoders. We initialize the input layer of the encoder using cross-lingual word embeddings, and train the rest of the parameters in an unsupervised manner combining denoising autoencoding and on-the-fly back-translation. As for phrase-based statistical machine translation, we build an initial phrase-table by aligning n-gram embeddings, combine it with a language model and a distortion model, and further improve the resulting machine translation system through unsupervised tuning and iterative back-translation. Finally, we propose a method to combine both approaches by training two conventional neural machine translation systems in opposite direction through iterative back-translation, using the previous unsupervised statistical machine translation system for warmup.

While previous attempts at learning machine translations systems from monolingual corpora had strong limitations, our work—along with other contemporaneous developments—is the first to report positive results in standard, large-scale settings. For instance, our proposed system obtains 22.5 BLEU points in the well-known English-German WMT 2014 benchmark, outperforming the supervised shared task winner back in 2014 despite using the exact same monolingual data and none of the parallel data. Together with other parallel developments, the contributions made at this thesis establish the foundations of unsupervised machine translation, opening exciting opportunities for future research.

Note for non-Basque speaking readers

This dissertation is structured as a collection of articles. The introductory chapter is in Basque, whereas the conclusions and the articles themselves are in English. Non-Basque speaking readers are recommended to first read the Conclusions chapter to get an overview of the main contributions made at this thesis, followed by the papers in the appendix in their recommended reading order.

Contents

Abstract v

1 Sarrera 1
- 1.1 Motibazioa . . . 1
- 1.2 Helburuak eta ikerketa-lerroak . . . 5
- 1.3 Tesia osatzen duten artikuluak . . . 7
- 1.4 Tesitik kanpo utzitako artikuluak . . . 12
- 1.5 Oinarriak . . . 16
  - 1.5.1 Hitz-bektoreak . . . 17
  - 1.5.2 Itzulpen automatikoa . . . 22
- 1.6 Erlazionatutako lana . . . 33
  - 1.6.1 Hitz-bektoreen hizkuntza arteko lerrokatzea . . . 33
  - 1.6.2 Itzupen automatiko gainbegiratu gabea . . . 45

2 Conclusions 55

Glossary 61

Bibliography 69

A Appendix 101
- A.1 Artetxe et al. (EMNLP 2016) . . . 103
- A.2 Artetxe et al. (AAAI 2018) . . . 109
- A.3 Artetxe et al. (ACL 2017) . . . 117
- A.4 Artetxe et al. (ACL 2018) . . . 129
- A.5 Artetxe et al. (ICLR 2018) . . . 139
- A.6 Artetxe et al. (EMNLP 2018) . . . 151
- A.7 Artetxe et al. (ACL 2019a) . . . 163

A.8 Artetxe et al. (ACL 2019b) 173
A.9 Artetxe et al. (ACL 2020a) 179

1 Sarrera

Tesi-txosten hau artikulu-bilduma modura antolatuta dago. Tesia osatzen duten artikuluak A eranskinean aurki daitezke, eta kapitulu honetan egindako lanaren ikuspegi orokorra azalduko dugu. Lehendabizi, tesiaren gaia aurkeztu eta motibatuko dugu (1.1 atala). Jarraian, tesiaren helburuak eta bertan landutako ikerketa-lerroak azalduko ditugu (1.2 atala). Ondoren, tesia osatzen duten artikuluak aurkeztuko ditugu (1.3 atala), bai eta tesian zehar landuagatik txosten honetatik kanpo utzitakoak ere (1.4 atala). Horren ostean, tesiaren oinarri diren hitz-bektoreen eta itzulpen automatikoaren nondik norakoak azalduko ditugu (1.5 atala). Kapituluarekin amaitzeko, tesi honekin erlazionatutako literaturako lanak izango ditugu hizpide (1.6 atala). Azkenik, 2 kapituluan tesi honetatik ateratako ondorioak azalduko ditugu.

1.1 Motibazioa

1799ko uztailean Napoleonen soldaduek alde batean inskripzioak zituen harri bat aurkitu zuten Egiptoko Rosetta hiritik hurbil. Idazkera ezberdineko hiru atal bereiz zitezkeen bertan: egiptoar hieroglifikoa goiko aldean, egiptoar demotikoa tartekoan, eta greziera behekoan. Gerora jakingo zenez, k.a. II. mende hasieran Ptolomeo V.aren koroatzearen harira ateratako dekretu baten bertsio paraleloak ziren. Aurkikuntzaren garaian Antzinako Egiptoko idazkera bien ezagutza galdua zen mende luzez, eta hieroglifikoak ulertzeko ahalegin ugariak antzuak izan ziren ordura arte. Rosetta harria delakoak, baina, grezierazko bertsioarekin loturak egitea ahalbidetu zuen izen bereziak heldulekutzat hartuta, eta giltzarri izan zen hori, XIX. mendearen lehen zatian, hieroglifikoak deszifratu ahal izateko (Pope, 1999).

Ethnologue argitalpen ezagunaren arabera, 2019. urtean 7,111 hizkuntza zeuden bizirik munduan[1] (Eberhard et al., 2019a,b,c), eta poliglotetan poliglotenari ere egiptoar hieroglifikoak bere garaian bezain ulertezinak egingo zaizkio gehien-gehienak. Argitalpen

[1] https://www.ethnologue.com/guides/how-many-languages

beraren datuetan oinarrituz, mundu mailan ausaz aukeratutako bi pertsonek gutxienez hizkuntza komun bat hitz egiteko probabilitatea gehienez ere % 6,56koa dela estima dezakegu,[2] eta ama-hizkuntza bera izatekoa gehienez ere % 2,86koa. Bestela esanda, batez bestean ezinezkoa zaigu 35 pertsonatik 34rekin gure ama-hizkuntzan hitz egitea, eta 15etik 14rekin ezin gaitezke inongo hizkuntzatan komunikatu. Bistan da, beraz, zaindu beharreko altxor bat ez ezik, hizkuntza-aniztasuna komunikaziorako hesi bat ere badela geroz eta globalagoa den mundu honetan.

Hesi hori gainditzeko asmoz, itzulpen automatikoa hizkuntzaren prozesamenduaren eta, modu zabalagoan, adimen artifizialaren aplikazio entzutetsuenetariko bat izan da euren hastapen-hastapenetatik. Hasierako hurbilpenak erregeletan oinarritzen baziren ere, itzulpen automatiko modernoak corpusak ditu abiapuntu. Oinarrizko printzipioa egiptoar hieroglifikoak deszifratzea ahalbidetu zuen bera da: testu paraleloetatik—aldez aurretik pertsona batek eginiko itzulpenetatik alegia—itzulpen-patroiak ikastea modu gainbegiratuan. Urte luzez eredu estatistikoak erabili izan dira horretarako (Brown et al., 1990; Koehn et al., 2003), baina duela bospasei urtetik hona eredu neuronalak gailendu zaizkie (Sutskever et al., 2014; Bahdanau et al., 2015). Euren eskutik itzulpen automatikoak izugarrizko aurrerapausoak eman ditu, eta hainbat autore gizakiaren pareko emaitzak lortu dituztela baieztatzera iritsi dira (Hassan et al., 2018), domeinu eta hizkuntza-bikote jakinetan beti ere. Horren erakusgarri, WMT 2019ko ingelesa-alemana ataza partekatuan eskuzko ebaluatzaileek itzultzaile automatiko baten irteera hobetsi zuten itzultzaile profesionalen lanaren gainetik, modu estatistikoki esanguratsuan (Barrault et al., 2019).

Rosetta harriaren 100 lerroak,[3] baina, oso motz geratzen dira halako itzultzaile automatiko bat eraikitzeko. WMT 2019ko ataza partekatuan gizagaineko emaitzak eskuratu zituen sistema entrenatzeko, adibidez, 27,7 milioi esaldi paralelo erabili zituzten jatorrizko 38,8 milioiak iragazi ondoren (Ng et al., 2019). Meng et al. (2019) haratago joan ziren, 40 mila milioi esaldi paralelo erabili baitzituzten ingelesa-txinera itzultzaile automatiko bat entrenatzeko, ohiko corpus paraleloekiko hobekuntza nabarmenak eskuratuz. Zen-

[2] Estimazio hau egiteko $\sum_i \#L_i^2/N^2$ formula erabili dugu, non $\#L_i$ *Ethnologue*-ren arabera i. hiztun kopuru handiena duen hizkuntzaren hiztun kopuru totala den, eta $N = 7.713.468.205$ Nazio Batuen araberako 2019ko uztaileko munduko populazio totala (`https://population.un.org/wpp`). $\forall i > 200, \#L_i = \#L_{200}$ hartu dugu, *Ethnologue*-k hiztun kopuru handieneko 200 hizkuntzen datuak soilik eskaintzen baititu modu irekian (`https://www.ethnologue.com/guides/ethnologue200`). Horrenbestez, emaniko estimazioa goi-borne bat da, falta diren datuetarako balio posible altuenak erabiltzeaz gain $k > 1$ hizkuntza partekatzen dituzten bikoteak k aldiz zenbatzen baititu behin bakarrik beharrean. Estimazio zehatzak egiteko beharrezkoa da hizkuntza ezberdinak hitz egitearen arteko dependentziak modelatzea, eta *Ethnologue*-k ez du halako daturik eskaintzen. Pertsona batek hizkuntza jakin bakoitza hitz egitea gertaera independenteak direla suposatuz, % 6,23-6,39 tartera muga dezakegu portzentaia (behe-bornea eta goi-bornea kalkulatzeko falta diren datuetarako balio posible minimo eta maximoak hartuz, hurrenez hurren).

[3] Rosetta harriak 14 lerro ditu hieroglifikoz, 32 demotikoz eta 54 grezieraz. Horietako batzuk ez dira osorik mantendu, eta hieroglifikozko lerro batzuk ere falta dira.

baki horiek perspektiban jartze aldera, *Moby Dick* liburuak 9.104 esaldi baino ez ditu.[4] Horrenbestez, lehen sistema eraikitzeko halako 3 mila libururen parekoa erabili zuten, eta bigarren sistema eraikitzeko halako 4,4 milioi libururen parekoa, euren itzulpenekin batera. Egunean 8 ordu eskainiz, pertsona batek hurrenez hurren 48 eta 69.215 urte inguru beharko lituzke hori guztia irakurri ahal izateko bakarrik.[5]

Konparaketa zuzenak egitea zaila bada ere, bistan da gizakiok ez dugula halako gain-begiratze sendorik behar hizkuntza bat ikasteko. Arazo hori ez da itzulpen automatikora mugatzen, eta ikasketa sakonaren erronka handienetariko bat dela esan izan da (LeCun et al., 2015). Horren erakusgarri, AlphaGo sistemak Lee Sedol munduko Go txapeldunaren aurka erdietsitako garaipena lorpen gogoangarritzat jo izan bada ere, aipatzekoa da sistema horren azken bertsioa bere buruaren aurka 4,9 milioi partida jokatuz entrenatu zutela (Silver et al., 2017),[6] pertsona batek bere bizitzan zehar joka ditzakeenak baino askoz ere gehiago. Datu etiketatuen beharra arintze aldera, transferentzia-ikasketa (Sharif Razavian et al., 2014; Yosinski et al., 2014; Devlin et al., 2019) eta metaikasketa (Finn et al., 2017) moduko gaiek garrantzi handia hartu dute azkenaldian.

Lagin-efizientziaren berezko interesaz gain, itzulpen automatiko modernoak corpus paralelo handiekiko duen menpekotasuna arazo praktiko bat ere bada. Izan ere, corpus paraleloen iturri modura Europako Batasuna edo Nazio Batuak moduko erakundeek argitaratutako dokumentu itzuliak erabili izan dira urte luzez (Koehn, 2005; Rafalovitch et al., 2009; Eisele and Chen, 2010; Chen and Eisele, 2012; Ziemski et al., 2016), eta, berrikiago, *crawling* bidezko meatze-teknikak ere asko zabaldu dira (Esplà et al., 2019; Schwenk et al., 2019a,b). Frantsesa-ingelesaren kasuan, adibidez, Nazio Batuen corpus paraleloak 25,8 milioi esaldi paralelo biltzen ditu (Ziemski et al., 2016), eta ParaCrawl *crawling* corpusaren BiCleaner v6 bertsio iragaziak 73,4 milioi[7] (Esplà et al., 2019). Hizkuntza gehien-gehienen errealitatea, baina, oso bestelakoa da. Horren erakusgarri da Guzmán et al. (2019) lana, nepalera-ingelesa eta sinhala-ingelesa bikoteak lantzen dituena. Aipatzekoa da ez nepalera ez sinhala ez direla hizkuntza gutxituak inondik inora: Nepaleko hizkuntza nagusia da bata eta Sri Lankakoa bestea, biak dira ofizialak norbere herrialdean, eta 24,5 eta 17,3 milioi hiztun dituzte, txekierak, grezierak edo suedierak baino gehiago (Eberhard et al., 2019a,b). Bi kasuetan, baina, apenas bildu ahal izan zituzten milioi erdi esaldi paralelo, gehien-gehienak domeinu oso berezietakoak eta, horrenbestez, erabilgarritasun mugatukoak: nepalera-ingelesaren kasuan 495 mila

[4] Zenbaketa hori Gutenberg proiektuko ingelesezko testu lauko bertsioaren gainean egin dugu (`https://www.gutenberg.org/ebooks/2701`), *Moses*-en esaldi-banatzailea erabiliz (`https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl`)

[5] Estimazio horrek *Moby Dick*-en luzerako liburu bat eta bere itzulpena irakurtzeko 23na ordu behar direla suposatzen du, nobela horren `https://www.audible.com` dendako 10 audio-liburu salduenen batez besteko iraupena dena.

[6] Datu hori AlphaGo Zero bertsioari dagokio, Lee Sedol garaitu zuen AlphaGo Lee baino hobea dena. Azken horren xehetasunak ez dira argitaratu.

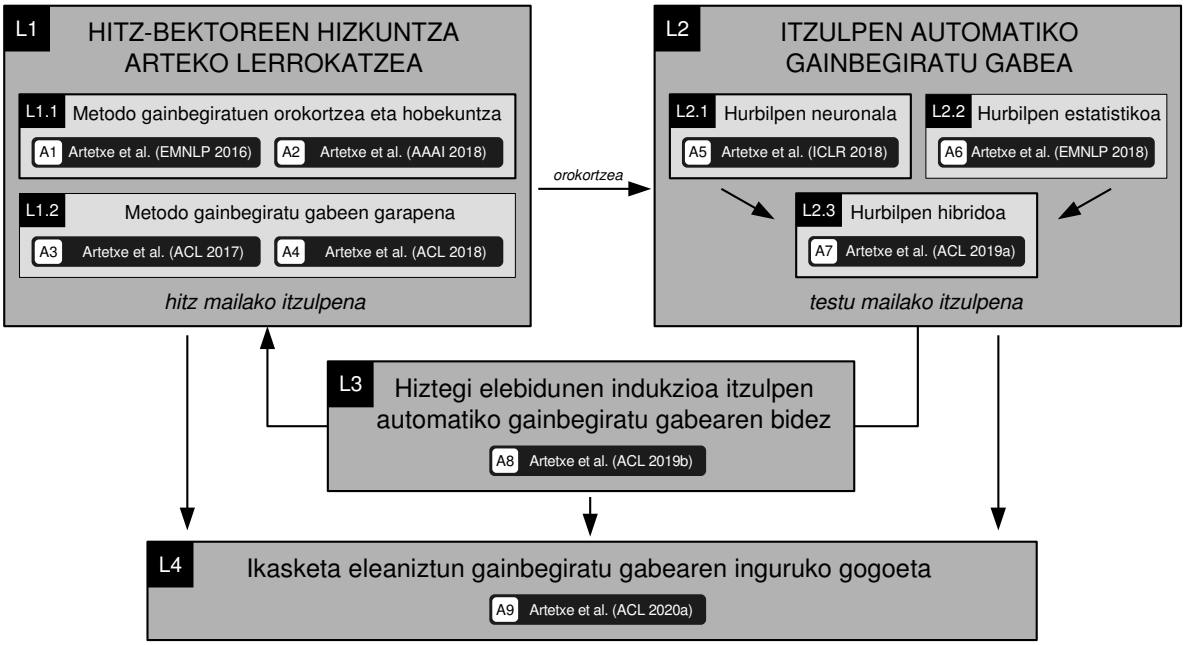
[7] `https://paracrawl.eu/`

segmentu software lokalizazioetatik, 62 mila esaldi Bibliatik, 33 mila *crawling* corpus iragazietatik eta 7 mila bestelako iturrietatik, eta sinhala-ingelesaren kasuan berriz 601 mila esaldi azpitituluetatik, 47 mila *crawling* corpus iragazietatik eta 46 mila segmentu software lokalizazioetatik. Bi kasuetan ere egungo itzulpen automatikoko teknikek emaitza kaskarrak ematen zituztela erakutsi zuten Guzmán et al. (2019) lanean, baliabide urriko hizkuntzak itzulpen automatikoaren erronka handienetariko bat direla ondorioztatuz.

Ezinbestekoa al da, baina, corpus paralelo bat izatea itzultzaile automatiko bat eraikitzeko? Noraino irits gintezke corpus elebakarrak soilik erabiliz? Posible ote da edozein hizkuntza itzultzen ikastea Rosetta harririk gabe? Galdera horiek bultzaturik, tesi honek itzulpen automatiko gainbegiratu gabearen inguruan ikertzea du helburu. Itzulpen automatiko gainbegiratu gabearen bereizgarria corpus elebakarrak soilik erabiltzea da, bi hizkuntzen arteko loturaren seinale espliziturik gabe. Deszifratze estatistikoaren inguruko ikerketa-lerroa alde batera utzita, orain arte landu gabeko gai bat zen hau, eta tesi honek paradigma berri honen bideragarritasuna frogatu eta bere oinarriak finkatzeko balio izan du. Berritzailea ez ezik, ikerketa-lerro hau interes handikoa ere bada, eta bere garrantzia justifikatze aldera honako bi arrazoiak nabarmen ditzakegu:

- **Berezko interes zientifikoa**. Zein puntutaraino da posible bi hizkuntzaren arteko baliokidetzak aurkitzea euren lagin independentetatik soilik abiatuta? Hizkuntzen izaera eta unibertsaltasunaren inguruan egin daitekeen galdera funtsezkoenetariko bat dugu hori, eta problema ireki eta esanguratsu bati dagokio bere baitan. Ikerketa-lerro hau galdera horri erantzuten ahalegintzen da hurbilpen enpiriko bat jarraituz, eredu konputazionaletan oinarritzen dena. Era berean, itzulpen automatiko gainbegiratu gabearen mugak aztertzea lagungarria izan daiteke oinarri dituen printzipioen mugak ere hobeto ezagutzeko, hipotesi distribuzionalarekin gertatzen den legez. Modu zabalagoan, ikerketa-lerro honek nola hizkuntza hala egungo eredu konputazionalen propietate eta barne funtzionamenduaren inguruko ezagutza zabaltzen lagun dezake.

- **Interes praktikoa**. Arestian aipatu bezala, corpus paraleloen gabezia arazo praktiko nabarmen bat da hizkuntza-bikote gehienentzat. Corpus paraleloekiko menpekotasun hori gainditzen duen neurrian, paradigma gainbegiratu gabeak bide berriak zabaltzen ditu, horrenbestez, kalitatezko itzulpen automatikoa hizkuntza-bikote gehiagotara iris dadin. Horrek ez du esan nahi, dena den, itzulpen automatiko gainbegiratu gabea baliabide urriko inguruneetarako hurbilpen egokiena denik nahitaez. Izan ere, praktikan ohikoa da nolabaiteko baliabide elebidunen bat izatea: Biblia bezalako corpus paraleloren bat, hiztegi txikiren bat... Nahiz eta ohiko itzultzaile automatiko bat entrenatzeko motz geratu, halako baliabideak ez lirateke, printzipioz, kaltegarriak izan beharko. Zentzu horretan, itzulpen automatiko gainbegiratu gabeak corpus elebakarren erabilera modu isolatuan aztertzeko ingurune bat ematen du, baliabide paraleloak barneratzeko gai diren



1.1 irudia: **Tesiaren eskema.** Bertan landuriko ikerketa-lerroak (1.2 atala) eta horietako bakoitzaren baitan argitaratutako artikuluak (1.3 atala) laburbiltzen ditu.

metodo erdigainbegiratuak garatzeko oinarritzat balio dezakeena etorkizun batean.

1.2 Helburuak eta ikerketa-lerroak

Tesi honen helburua corpus elebakarretatik soilik abiatuta itzultzaile automatikoak entrenatzeko metodo gainbegiratu gabeen inguruan ikertzea da. Horretarako jarraituriko hurbilpenak hiru urrats ditu: (i) corpus elebakarretatik hitz-bektoreak ikastea hizkuntza ezberdinentzat modu independentean, (ii) hizkuntza ezberdinetako hitz-bektoreak lerrokatzea euren antzekotasun estrukturalean oinarrituz, hizkuntza bateko hitzak itzultzeko erabil daitekeena dagozkien bektoreen beste hizkuntzako auzokide hurbilenak hartuz, eta (iii) hitz-bektore eleaniztun horietan oinarrituz itzultzaile automatikoak sortzea, hitz mailako itzulpenetik testu mailako itzulpenera orokortzea eskatzen duena. Lehenengo urratsak oinarri sendoak ditu dagoeneko Hizkuntzaren Prozesamenduaren arloan (ikus 1.5.1 atala), eta tesian zehar beste biak landu ditugu horrenbestez. Zehatzagoak izanez, 1.1 irudiak jaso bezala, tesian landuriko ikerketa-lerro nagusiak honakoak izan dira:

[**L1**] **Hitz-bektoreen hizkuntza arteko lerrokatzea.** Ikerketa-lerro honetan transformazio linealen bidez hizkuntza ezberdinetako hitz-bektoreak lerrokatzeko teknikak landu ditugu. Tesi honi ekin zitzaionean baziren horretarako hainbat metodo, baina guztiak ziren gainbegiratuak eta 5.000 sarrera inguruko entrenamendu-hiztegiak zerabiltzaten. Hori horrela, ikerketa-lerro honen baitan bi norabide nagusi

landu ditugu:

[**L1.1**] **Metodo gainbegiratuen orokortzea eta hobekuntza.** Puntu honen baitan hitz-bektoreak lerrokatzeko marko matematiko orokor bat landu dugu. Marko horrek hainbat parametro ditu, eta aurreko metodoak haien konfigurazio konkretuei dagozkie. Berrinterpretazio horri esker familia ezberdinetako metodoen arteko loturak egin ahal izan ditugu, eta euren portaera hobeto ulertu. Horrekin batera, aurreko emaitzak hobetzen dituzten aldaera berriak ere proposatu ditugu.

[**L1.2**] **Metodo gainbegiratu gabeen garapena.** Puntu honen baitan hitz-bektoreak lerrokatzeko aurreko metodoek beharrezkoa zuten gainbegirapena murriztu eta, azken buruan, erabat ezabatzeko metodoak landu ditugu. Horretarako bi teknika berri garatu ditugu: (i) autoikasketa iteratiboa, eta (ii) hizkuntza barneko antzekotasunean oinarritutako hasieraketa. Teknika biok inolako hiztegirik gabe hitz-bektoreak lerrokatzea ahalbidetzen dute, aurreko metodoek 5.000 sarrerako hiztegiekin lortzen zituzten pareko emaitzak lortuz.

[**L2**] **Itzulpen automatiko gainbegiratu gabea.** Ikerketa-lerro honetan hitz mailan itzultzeko baliagarri diren hitz-bektore lerrokatuetatik abiatuta, testu mailan itzultzeko gai diren itzultzaile automatikoak entrenatzeko metodo gainbegiratu gabeak landu ditugu. Itzulpen automatikoaren baitan dauden paradigma ezberdinekin bat etorriz (ikus 1.5.2 atala), horretarako hiru hurbilpen jorratu ditugu:

[**L2.1**] **Hurbilpen neuronala.** Lerro honen baitan kodetzaile-deskodetzaile arkitekturan oinarritutako itzultzaile automatiko neuronalak modu gainbegiratu gabean entrenatzeko metodo bat garatu dugu. Proposatutako metodoak lerrokatutako hitz-bektoreak baliatzen ditu kodetzailea hasieratzeko, eta neuronasare osoa entrenatu zarata murriztea eta atzeranzko itzulpena uztartuz.

[**L2.2**] **Hurbilpen estatistikoa.** Lerro honen baitan sintagmetan oinarritutako itzultzaile automatiko estatistikoak modu gainbegiratu gabean entrenatzeko teknikak landu ditugu. Horretarako jarraituriko hurbilpenak lau urrats nagusi ditu: (i) hitz-bektoreak orokortzea n-grama edo hitz-segiden bektoreak ikasteko, (ii) hizkuntza ezberdinetako n-gramen bektoreak lerrokatuz itzulpen-taula bat induzitzea, (iii) itzulpen-taula hori hizkuntza-eredu batekin konbinatzea hasierako itzultzaile automatiko estatistiko bat sortzeko, eta (iv) hasierako soluzio hori hobetzea doikuntza gainbegiratu gabearen eta atzeranzko itzulpen iteratiboaren bidez.

[**L2.3**] **Hurbilpen hibridoa.** Lerro honen baitan aurreko bi hurbilpenak uztartzeko metodo bat landu dugu. Ideia nagusia bi noranzkoetan dabilen itzultzaile neuronal bikote bat atzeranzko itzulpen iteratiboaren bidez entrenatzea da, prozedura hasieratzeko itzultzaile estatistiko gainbegiratu gabe bat erabiliz.

[L3] **Hiztegi elebidunen indukzioa itzulpen automatiko gainbegiratu gabearen bidez.** Arestian aipatu bezala, hitz-bektoreen lerrokatzea hitz mailako itzulpenak egiteko erabil daiteke, eta hiztegi elebidunen indukzioa izan da, hain justu ere, eurak ebaluatzeko ataza ohikoena. Era berean, itzulpen automatiko gainbegiratu gabeak hitz-bektoreen lerrokatzea du abiapuntu. Ikerketa-lerro honetan zikloa itxi eta itzulpen automatiko gainbegiratu gabea hiztegi indukziorako baliatzeko teknikak landu ditugu.

[L4] **Ikasketa eleaniztun gainbegiratu gabearen inguruko gogoeta.** Azken lerro honen baitan ikasketa eleaniztun gainbegiratu gabearen motibazio, bilakaera eta arazo metodologikoen inguruan hausnartu dugu, eta etorkizuneko erronkak identifikatu. Izan ere, azken urteotan pisu handia hartu du arlo horrek, eta beste faktore askoren artean tesi honetan eginiko ekarpenak ere giltzarri izan dira horretarako. Hori ikusirik, arloaren egungo egoeraz gogoeta egin eta ikerketa-komunitate zabalagoarekin honen inguruko eztabaida bat sustatu nahi izan dugu.

1.3 Tesia osatzen duten artikuluak

Atal honek tesia osatzen duten artikuluak aurkezten ditu. Artikuluak eurak A eranskinean aurki daitezke, eta jarraian lan bakoitzaren ikuspegi orokorra azaldu eta tesiaren testuinguru zabalagoan kokatuko ditugu. 1.1 irudiak artikulu hauen eta aurreko atalean azaldutako ikerketa-lerroen arteko lotura azaltzen du.

Artikuluak gomendatutako irakurketa-ordenan zerrendaturik daude. Antolaketa hau artikuluen edukiaren arabera egin da, 1.2 ataleko ikerketa-lerroen ordena logiko bera jarraituz. Hurrenkera hau bat dator, era berean, ordena kronologikoarekin, bi kasutan izan ezik: A2 artikulua A3 artikuluaren ondoren argitaratu zen, eta A4 artikulua A5 artikuluaren ondoren.

[A1] Artetxe et al. (EMNLP 2016)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Hitz-bektore modernoen hizkuntza arteko lerrokatzearen hastapenak Mikolov et al. (2013b) lanean kokatzen dira. Hizkuntza bateko hitz-bektoreak beste hizkuntza batekoekin lerrokatzeko transformazio lineal bat ikastea proposatu zuten bertan, hiztegi elebidun baten arabera bi hizkuntzetako hitz-bektoreen arteko distantzia euklidearren karratuen batura minimizatuz.

Artikulu honetan oinarrizko helburu-funtzio horren hiru hedapen aztertzen ditugu: ortogonalitatea, luzera-normalizazioa eta batezbestekoa zentratzea. Aldaera horiek inbariantza elebakarra bermatzeko, kosinu-antzekotasuna maximizatzeko eta kobariantza gurutzatua maximizatzeko balio dute, hurrenez hurren. Formulazio ezberdina erabiliagatik, Faruqui and Dyer (2014) eta Xing et al. (2015) lanetan proposatutako metodoak faktore horien arabera azal daitezkeela erakusten dugu, metodo biok Mikolov et al. (2013b) laneko oinarrizko helburu-funtzioaren aldaeratzat ikus daitezkeela erakutsiz. Proposatutako markoaren baitan, faktore bakoitzaren ekarpena argiago neurtzen dugu eta, hirurak uztartuz, ordura arte zeuden emaitzarik onenak gainditu.

[A2] Artetxe et al. (AAAI 2018)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Lan honek A1 artikuluko markoa zabaltzen du. Marko berriak luzera-normalizazioa eta batezbestekoa zentratzea aurreprozesu modura mantentzen ditu, baina, ortogonalitatea murriztapentzat tratatu beharrean, aurreko metodoak transformazio lineal zehatz batzuen konposaketa modura orokortzen ditu. Transformazio nagusia ortogonala da eta hizkuntzen arteko lerrokatzeaz arduratzen da. Horretaz gain hautazko beste transformazio batzuk daude: zuritzea, birpisaketa, deszuritzea eta dimentsionaltasun-murrizketa. Aurreko hainbat metodo (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Shigeto et al., 2015; Xing et al., 2015; Zhang et al., 2016; Smith et al., 2017) aurreprozesu eta hautazko transformazioen konfigurazio ezberdinei dagozkie. Deskonposaketa horrek aurreko metodoen arteko lotura berriak egin eta euren funtzionamendua hobeto ulertzen laguntzen du, alderantzizko erregresioa zergatik den mesedegarria azalduz, adibidez. Ikasitakoari esker, aurreko emaitzarik onenak gainditzen dituen aldaera berri bat ere proposatzen dugu.

[A3] Artetxe et al. (ACL 2017)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Artikulu honetan hitz-bektoreak lerrokatzeko beharrezko gainbegirapena modu nabarmenean arintzeko metodo bat proposatzen dugu, autoikasketa iteratiboan oinarritzen dena. Arestian aipatu bezala, hitz-bektoreak lerrokatzeko ohiko metodoek transformazio lineal bat ikasten dute hiztegi elebidun baten arabera. Era berean, lerrokatutako

hitz-bektoreak hiztegi elebidunak induzitzeko erabil daitezke, jatorrizko hizkuntzako hitz bakoitzaren helburuko hizkuntzako auzokide hurbilena hartuz. Proposatutako teknika prozedura hori modu iteratiboan errepikatzean oinarritzen da: hasierako hiztegi batetik abiatuta hitz-bektoreak lerrokatzen ditugu, lerrokatze horretan oinarrituz beste hiztegi bat induzitu, eta hiztegi berri hori hitz-bektoreak berriro ere lerrokatzeko erabili, urrats horiek prozesuak konbergitu arte errepikatuz. Prozedura horrek hasierako hiztegiarekiko independentea den optimizazio-helburu global baten optimo lokal batera konbergitzen du. Enpirikoki ere, proposatutako sistemak aurreko metodoek 5.000 sarrerako hiztegiekin eskuratzen zituzten pareko emaitzak lortzen ditu 25 sarrerako hiztegi txiki batetik soilik abiatuta. Era berean, hasierako hiztegitzat zenbaki-zerrenda bat erabilita ere pareko emaitzak lortzen ditugu. Honela, artikulu hau aitzindari izan zen baliabide paralelorik gabe eta heuristiko ahuletan soilik oinarrituz kalitatezko lerrokatzeak ikas zitezkeela erakusten.

[A4] Artetxe et al. (ACL 2018)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Artikulu honetan A3 artikuluko metodoa hobetzen dugu, eta erabat gainbegiratu gabea egin. Izan ere, atzean dagoen optimizazio-helburu globala hasierako hiztegiarekiko independentea izanagatik, A3 artikuluko prozedura iteratiboa optimo lokal kaskarretan trabatuta geratzen da ausazko soluzio batetik abiatuz gero. Arazo horri aurre egiteko, artikulu honetan hasierako hiztegia eraikitzeko metodo gainbegiratu gabe bat proposatzen dugu, hizkuntzen antzekotasun estrukturalean oinarritzen dena. Horretarako, hitz bakoitzak hizkuntza bereko gainerako hitzekin duen antzekotasun-banaketari erreparatzen diogu, eta antzeko banaketa duten hizkuntza ezberdinetako hitzak lerrokatu. Horretaz gain, prozedura iteratiboa bera ere sendoago egiteko hainbat teknika proposatzen ditugu. Azkenik, A2 artikuluan hurbilpen gainbegiratuentzat landutako birpisaketa metodoa ere barneratzen dugu.[8] Proposatutako sistema aurreko metodo gainbegiratu gabeak baino sendoagoa dela erakusten dugu artikuluan. Era berean, ordura arte argitaratutako emaitzarik onenak eskuratzen ditu datu-multzo estandarretan, aurreko metodo gainbegiratuak ere gaindituz.

[8] A3 artikulua A2 artikuluaren aurretik landu genuen eta, horrenbestez, ez ditu azken horretan garatutako teknikak barneratzen.

[A5] Artetxe et al. (ICLR 2018)

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018d. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Artikulu honetan itzultzaile automatiko neuronalak corpus elebakarrak soilik erabiliz entrenatzeko metodo bat proposatzen dugu. Proposatutako sistema ohiko arretadun kodetzaile-deskodetzaile arkitekturan oinarritzen da, aldaketa txiki batzuekin: bi hizkuntzetarako kodetzaile partekatu bat erabiltzen du, eta hizkuntza bakoitzeko deskodetzaile propio bat. Kodetzaileko hitz-bektoreak A3 artikuluko lerrokatze-metodoa erabiliz hasieratzen ditugu,[9] eta entrenamenduan zehar izoztuta mantendu. Hartara, kodetzaileak hitz mailako errepresentazio elebidunak jasotzen ditu sarrera modura, eta bektore horiek konbinatuz esaldi mailako errepresentazio elebidunak lortzeaz arduratzen da, deskodetzaile bakoitzak dagokion hizkuntzako testu bihurtzen dituenak. Sistema entrenatzeko, zarata murriztea eta atzeranzko itzulpena uztartzen ditugu. Artikulu hau itzulpen automatiko neuronal gainbegiratu gabearen inguruan plazaratutako lehen lana izan zen, aldi berean argitaratutako Lample et al. (2018a) artikuluarekin batera.

[A6] Artetxe et al. (EMNLP 2018)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Artikulu honetan sintagmetan oinarritutako itzultzaile automatiko estatistikoak modu gainbegiratu gabean entrenatzeko metodo bat proposatzen dugu. Horretarako, gure sistemak hitz-bektoreak orokortzen ditu bi hizkuntzetako n-grama edo hitz-segiden bektoreak ikasteko, eta haiek lerrokatu A4 artikuluko metodoa erabiliz. Behin hori eginda, lerrokatutako bektoreak erabiltzen ditugu n-gramak itzuli eta itzulpen-taula bat eraikitzeko. Itzulpen-taula hori hizkuntza-eredu batekin konbinatuz, hasierako itzultzaile automatiko estatistiko bat lortzen dugu. Amaitzeko, hasierako sistema hori hobetzen dugu atzeranzko itzulpen iteratiboan oinarritutako doikuntza eta finketaren bidez. A5 artikuluan proposatutako printzipioak paradigma estatistikora egokitzeaz gain, artikulu honek aurreko emaitzak modu nabarmenean ere hobetzen ditu.

[9] A5 artikulua A4 artikuluaren aurretik landu genuen, eta horregatik erabili genuen A3ko lerrokatze-metodoa A4ko metodo hobetuaren ordez. Lerrokatzea ikasteko zenbaki-zerrendan oinarritutako hasieraketa erabili genuen.

[A7] Artetxe et al. (ACL 2019a)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Artikulu honek A6 artikuluko itzulpen-sistema estatistiko gainbegiratu gabea hobetzen du, eta hurbilpen neuronalarekin konbinatzeko metodo bat proposatu. Lehenengo zatiari dagokionez, hiru dira aurreko sistemarekiko proposatutako hobekuntza nagusiak: (i) izen bereziak hobeto itzultzeko helburuarekin, hasierako itzulpen-taulan Levenshtein distantzian oinarritutako bi ezaugarri berri gehitzen ditugu, (ii) doikuntza modu gainbegiratu gabean egiteko metodo sendoago bat proposatzen dugu, helburu-funtzio esplizitu bat optimizatzen duena, eta (iii) finketa iteratiboa bi noranzkoetan batera egiten dugu, itzulpen-probabilitate bakoitza estimatzeko aurkako noranzkoko atzeranzko itzulpena erabiliz. Bigarren zatiari dagokionez, noranzko bakoitzeko itzultzaile automatiko neuronal arrunt bat entrenatzen dugu atzeranzko itzulpen iteratiboaren bidez. Hasierako iterazioetan aurreko sistema estatistiko gainbegiratu gabea erabiltzen dugu corpus paralelo sintetikoa sortzeko baina, ikasketak aurrera egin ahala, aurkako noranzkoko itzultzaile neuronalarekin ordezkatzen dugu. Proposatutako sistemak aurreko metodoekiko hobekuntza nabarmenak eskuratzen ditu, WMT 2014ko sistema gainbegiratu onena gainditzera iritsiz ingelesa-alemana bikotearen kasuan.

[A8] Artetxe et al. (ACL 2019b)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Arestian aipatu bezala, hitz-bektoreen lerrokatzea hiztegi elebidunak induzitzeko erabili izan da hitz bakoitzaren beste hizkuntzako auzokide hurbilena hartuz. Artikulu honetan hitz-bektore eleaniztunetatik abiatuta hiztegi elebidunak induzitzeko beste hurbilpen bat proposatzen dugu, A6 eta A7 artikuluetan landutako itzulpen estatistiko gainbegiratu gabeko tekniketan oinarritzen dena. Zehatzagoak izanez, lerrokatutako hitz-bektoreekin itzulpen-taula bat sortzen dugu, eta hizkuntza-eredu batekin konbinatu itzultzaile automatiko estatistiko bat eraikitzeko. Behin hori eginda, corpus paralelo sintetiko bat sortzen dugu sistema hori erabiliz, eta hiztegi elebidun bat induzitu lerrokatze estatistikoko tekniken bidez. Modu horretara auzokide hurbilenean oinarritutako teknikek baino emaitza nabarmenki hobeak lortzen ditugu hitz-bektore lerrokatu berberetatik abiatuta. Artikulu hau ACL 2019ko artikulu onenaren sarirako izendatua izan zen.

[A9] Artetxe et al. (ACL 2020a)

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020d. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Accepted for publication.

Iritzi artikulu honetan ikasketa eleaniztun gainbegiratu gabearen motibazio, definizio, hurbilpen eta metodologia berrikusten ditugu, eta euretako bakoitzean zehaztasun handiagoz jokatzeko deia egin. Horrela, ikerketa-lerro hori justifikatzeko baliabide paraleloen balizko gabezia erabili izan bada ere, halako lanek planteatutako baldintzak ez direla erabat errealistak argudiatzen dugu, eta ikasketa eleaniztun gainbegiratu gabea motibatzerakoan zorroztasun handiagoz aritzeko beharra erakutsi. Era berean, ikasketa gainbegiratu gabearen aterkipean erabili izan diren ikasketa-seinale ezberdinak berrikusten ditugu, eta metodo ezberdinak aurkeztu eta alderatzerakoan erabiltzen dituzten seinaleak aintzat hartzeko deia egin. Horretaz gain, sistema eleaniztun gainbegiratu gabeak garatu eta ebaluatzeko hainbat arazo metodologiko identifikatzen ditugu, eta eurei aurre egiteko gomendioak eman. Amaitzeko, ikasketa eleaniztun gainbegiratu gabearen baitan jorratutako ikerketa-lerro ezberdinen ikuspegi bateratu bat ematen dugu, eta ebaluazio-marko komun baten alde egin. Modu horretara, besteak beste tesi hau bera tarteko arloak berriki jasan duen bilakaera azkarraren inguruan hausnartu, eta etorkizunera begira zorroztasun handiagoz jokatzeko gogoeta bat sustatu nahi izan dugu.

1.4 Tesitik kanpo utzitako artikuluak

Atal honetan tesi-garaian idatzitako gainerako artikuluak aurkeztuko ditugu. Artikulu hauek tesiaren gai nagusitik aldentzen dira edota autore nagusitzat beste norbait dute eta, hori dela eta, tesi-txosten honetatik kanpo uztea erabaki dugu. Euren artean daude, halaber, tesian zehar Facebook AI Research eta DeepMind zentroetan egindako egonaldietako lanak.

[A10] Artetxe et al. (CoNLL 2018)

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018e. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291, Brussels, Belgium. Association for Computational Linguistics.

Artikulu honetan hitz-bektoreek alderdi linguistiko kontrajarriak nola kodetzen dituzten aztertzen dugu. Behin hitz-bektoreak ikasita, semantika/sintaxia edo antzekotasuna/ahai-

detasuna moduko alderdi kontrajarrietan azaleratutako informazioa moldatzea posible dela erakusten dugu, transformazio lineal gainbegiratu gabe bat erabiliz. Orain arte uste zenaren kontra, hitz-bektoreek zuzenean ikusgai dena baino informazio gehiago kodetzen dutela erakusten dugu horrela, eta aurkikuntza horrek ebaluazio intrintseko eta estrintsekoan dituen ondorioak aztertu. Ikerketa hau A2 artikuluko deskonposaketaren harira abiatu genuen, proposatutako transformazioarekin lotura estua duena, baina tesiaren gai nagusitik urruntzen da erabat elebakarra den neurrian. Lan honek CoNLL 2018ko artikulu onenaren saria jaso zuen.

[A11] Ormazabal et al. (ACL 2019)

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Hizkuntza ezberdinetako hitz-bektoreak transformazio linealen bidez lerrokatu ahal izateko beharrezkoa da euren egitura antzekoa izatea. Hainbat autorek zalantzan jarri izan dute hori hala denik, domeinu- eta hizkuntza-ezberdintasunen arabera hitz-bektoreen egituretan alde handiak daudela erakutsiz. Artikulu honetan fenomeno hori hitz-bektoreak independenteki ikastearen ondorio eta, horrenbestez, lerrokatze-metodoen berezko muga bat ote den aztertzen dugu, ala ezberdintasun linguistikoek eragindako oztopo gaindiezin bat. Horretarako, lerrokatze-metodoak eta hainbat hizkuntzatako hitz-bektoreak batera ikasten dituzten metodoak alderatzen ditugu corpus paraleloak erabiliz. Baldintza ideal horietan lerrokatze-metodoak nabarmenki okerrago dabiltzala erakusten dugu. Modu horretara, artikulu honek tesian landutako lehen ikerketa-lerroaren mugak erakusten ditu, eta etorkizuneko ikerketa-lerro modura hizkuntza-ezberdinetako hitz-bektoreak batera ikasteko beharrezko gainbegirapena murriztea planteatu. Artikulu hau Aitor Ormazabalen gradu bukaerako lanaren baitan landu genuen.

[A12] Artetxe and Schwenk (ACL 2019)

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Artikulu honetan esaldi-bektore eleaniztunen bidez corpus paraleloak erauzteko metodo bat proposatzen dugu. Aurreko metodoak auzokide hurbilenean oinarritzen ziren, kosinu-antzekotasuna erabiliz atalase finko batekin, baina neurri horrek eskala-inkontsistentzia

arazoak dituela erakusten dugu. Horri aurre egiteko, emaniko esaldi-bikotearen eta hurbileneko hautagaien arteko aldean oinarritzen den metodo bat proposatzen dugu. Metodo horrek hobekuntza nabarmenak lortzen ditu nola ebaluazio intrintsekoan hala itzulpen automatikoan bertan. Artikulu hau Facebook AI Research-en egindako egonaldian landu genuen.

[A13] Artetxe and Schwenk (TACL 2019)

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Artikulu honetan helburu orokorreko esaldi-bektore eleaniztunak hizkuntza kopuru handietara eskala daitezkeela erakusten dugu. Horretarako bokabulario partekatuko BiLSTM kodetzaile bat erabiltzen dugu, deskodetzaile batekin parekatu eta corpus paralelo publikoak erabiliz entrenatzen duguna 93 hizkuntzatan. Hurbilpen hori hizkuntza arteko transferentzia-ikasketan, corpus paraleloen erauzketan eta hizkuntza arteko antzekotasun-bilaketan ebaluatzen dugu, emaitza sendoak eskuratuz. Artikulu hau Facebook AI Research-en egindako egonaldian landu genuen.

[A14] Gamallo et al. (CL 2019)

Pablo Gamallo, Susana Sotelo, José Ramom Pichel, and Mikel Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

Artikulu honetan aditz partikuladunak testuinguruan itzultzeko eredu distribuzional bat proposatzen dugu. Horretarako, sarrerako sintagmaren itzulpen-hautagaiak sortzen ditugu hiztegi elebidun bat eta transferentzia-erregelak baliatuz. Behin hori eginda, sarrerako sintagmaren eta itzulpen-hautagai bakoitzaren bektore-errepresentazioak eraikitzen ditugu, sintaktikoki anotatutako corpus elebakar bateko zenbaketetan oinarrituz. Azkenik, itzulpen-hautagaiak puntuatzen ditugu sarrerarekiko kosinu-antzekotasunean oinarrituz, eta puntuazio altuena duena aukeratu. Artikulu hau kanpo-kolaborazio baten emaitza da.

[A15] Artetxe et al. (SEPLN 2019)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019c. Unsupervised neural machine translation, a new paradigm solely based on monolingual text. *Procesamiento del Lenguaje Natural*, 63:151–154.

Artikulu honetan UnsupNMT proiektua aurkezten dugu, Espainiako Ekonomia, In-

dustria eta Lehiakortasun Ministerioak finantzatua *Explora* programapean. Proiektuak lotura estua du tesi honekin, biek ala biek hitz-bektore eleaniztunetan oinarrituz itzultzaile automatiko gainbegiratu gabeak ikastea baitute helburu. Hain justu ere, tesi honen hastapenetan lortutako emaitza onen harira abiatu genuen proiektua, 2 urteko iraupena duena eta egun bukatzear dena.

[A16] Artetxe et al. (ACL 2020b)

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020c. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Accepted for publication.

Tesi honetan landutako hitz-bektoreen lerrokatzeaz gain, errepresentazio eleaniztunak ikasteko beste hurbilpen gainbegiratu gabe bat agertu da berriki: hizkuntza ezberdinetako corpus elebakarrak konbinatu eta hizkuntza-eredu maskaratu sakon bat entrenatzea. Hurbilpen horrek hizkuntza ezberdinen artean lerrokatutako errepresentazioak ikasten ditu, bere helburu-funtzioan horretara bultzatzen duen termino espliziturik ez badago ere. Hainbat autorek portaera hori 3 faktoreren ondorioa dela planteatu dute: (i) partekatutako bokabulario-sarrerak izatea, heldulekutzat jokatzen dutenak, (ii) hizkuntza guztietan ikasketa batera egitea, efektu hau zabaldu eta azken buruan (iii) abstrakzio eleaniztun sakonak ematen dituena.

Artikulu honetan hipotesi hori gezurtatzen dugu, printzipio horiek guztiak urratzen dituen sistema batek antzeko emaitzak eskuratzen dituela erakutsiz. Zehatzagoak izanez, hizkuntza-eredu maskaratu elebakar bat ikasten dugu lehendabizi, eta beste hizkuntza batera transferitu hitz-bektoreen matrize berri bat ikasiz, gainerako parametroak izoztuta mantentzen ditugularik. Hurbilpen horrek ez du bokabulario partekaturik erabiltzen, hizkuntza ezberdinetako ikasketa ez du batera egiten, eta transferentzia maila lexikoan soilik egiten du. Hala eta guztiz ere, hizkuntza arteko transferentziako ataza estandarretan ohiko hurbilpenaren antzeko emaitzak lortzen dituela erakusten dugu, eredu elebakar sakonek beste hizkuntzetara orokortzeko gai diren abstrakzioak ikasten dituztela iradokitzen duena. Horretaz gain, XQuAD deituriko datu-multzo berri bat ere aurkezten dugu, SQuADeko 240 paragrafo eta 1190 galdera-erantzun biltzen dituena, haien 10 hizkuntzatako itzulpenekin batera. Artikulu hau DeepMind-en egindako egonaldian landu genuen.

[A17] Artetxe et al. (arXiv 2020a)

Mikel Artetxe, Gorka Labaka, Noe Casas, and Eneko Agirre. 2020b. Do all roads lead to Rome? Understanding the role of initialization in iterative back-translation. *arXiv preprint arXiv:2002.12867*. Under review.

A7 artikuluan proposatutako sistemak noranzko bakoitzeko itzultzaile automatiko neuronal arrunt bat entrenatzen du atzeranzko itzulpen iteratiboaren bidez, prozedura hasieratzeko itzultzaile automatiko estatistiko gainbegiratu gabe bat erabiliz. Artikulu honetan, hurbilpen horretan hasieraketak duen garrantzia aztertzen dugu. Horretarako, hainbat sistema ezberdin probatzen ditugu prozedura iteratiboa hasieratzeko: erregeletan oinarritutako itzultzaile bat, tamaina ezberdinetako corpus paraleloetan entrenatutako itzultzaile neuronal eta estatistikoak, eta A7 artikuluko itzultzaile estatistiko gainbegiratu gabe berbera. Gure esperimentuek hasierako sistemaren eragina nahiko txikia dela erakusten dute, atzeranzko itzulpen iteratiboak antzeko soluzioetara konbergitzeko joera baitauka. Hori horrela, etorkizunerako ikerketa-lerro modura prozedura iteratiboa bera hobetzea proposatzen dugu.

[A18] Artetxe et al. (arXiv 2020b)

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*. Under review.

Itzulpena sarri erabili ohi da hizkuntza arteko transferentzia-ikasketan: datu-multzo eleaniztun asko itzulpen-zerbitzu profesionalen bitartez sortuak dira, eta itzulpen automatikoaren bidez entrenamendu-multzoa nahiz test-multzoa itzultzea ohiko transferentzia-teknika bat da. Artikulu honetan halako itzulpenek artefaktuak eragin ditzaketela erakusten dugu. Hizkuntza-inferentziaren kasuan, adibidez, premisa eta hipotesia nor bere aldetik itzultzeak hain arteko teilakapen lexikoa murriztu dezake, egungo ereduen orokortze-gaitasunean eragin nabarmena duena. Fenomeno hori dela eta, orain arteko hainbat aurkikuntza berrikustea beharrezkoa dela erakusten dugu.

1.5 Oinarriak

Tesi honek hitz-bektoreak eta itzulpen automatikoa ditu oinarritzat. Izan ere, tesiaren azken helburua itzultzaile automatikoak modu gainbegiratu gabean ikastea da, eta horretarako jarraituriko hurbilpena hizkuntza ezberdinetako hitz-bektoreak lerrokatzetik abiatzen da. Hori horrela, atal honetan gai bion nondik norakoak azalduko ditugu, tesia jarraitzeko beharrezkoak diren oinarriak finkatuz.

Kontuan izan behar da, dena den, nola hitz-bektoreek hala itzulpen automatikoak ibilbide luzea dutela Hizkuntzaren Prozesamenduaren arloan, egun ere ikerketa-gai oso aktiboak izanik. Hori horrela, gure asmoa ez da arlo horietako literatura xehe-xehe aztertzea, egungo perspektibatik hurbilpen eta teknika nagusiak azaltzea baizik, gerora tesian zehar erabiliko ditugunak.

1.5.1 Hitz-bektoreak

Ikuspegi konputazionaletik, hizkuntza idatzia sinbolo diskretuen sekuentzia modura modelatu ohi da, bertako unitate atomikoak karaktere, azpihitz edo hitzak izanik, erabilitako ereduaren arabera. Halako errepresentazio sinbolikoek, baina, muga nabarmenak dituzte hizkuntza modu automatikoan prozesatzerako garaian. Adibide modura, *zerri* eta *zorri* hitzak oso antzekoak dira karaktere mailan, eta *zerri* eta *basurde* hitzak, berriz, zeharo desberdinak. Animalien artean, baina, zerriak askoz hurbilago daude basurdeetatik zorrietatik baino. Hitz, azpihitz edo karaktere bakoitzarentzat elementu ezberdin bat darabilten errepresentazio lokalek ez dute halako antzekotasun-noziorik barneratzen, eta sistema batek aurrez ikusi gabeko instantzia berrietara orokortzea zailtzen du horrek.

Arazo horri aurre egiteko, hitz-bektoreak erabili izan dira Hizkuntzaren Prozesamenduaren arloan. Hitz-bektoreak bokabularioko hitz bakoitzari espazio jarraitu bateko bektore bat esleitzen dioten errepresentazio banatuak dira. Arestian aipatutako errepresentazio lokalak ez bezala, hitz-bektoreak gai dira hitz ezberdinen arteko erlazioak modelatzeko, bektorearen osagai berberen balio ezberdinak erabiltzen baitituzte eurak errepresentatzeko. Horrela, hitz-bektoreen arteko distantziak—kosinu-antzekotasunaren bidez neurtu ohi direnak[10]—dagozkien hitzen antzekotasun semantikoaren adierazle izan ohi dira.

Halako errepresentazioak eraikitzeko informazio-iturri ezberdinak erabil daitezke. Aukera bat eskuz eraikitako ezagutza-baseak erabiltzea da, grafo-egitura izan ohi dutenak (Goikoetxea et al., 2015). Hitz-bektoreen eredu gehien-gehienak, baina, hipotesi distribuzionala delakoan oinarritzen dira. Hipotesi distribuzionalak dio antzeko testuingurutan agertu ohi diren hitzek antzeko esanahia izan ohi dutela (Harris, 1954; Firth, 1957). *"Otsalizarraren itzalpean zegoen"* eta *"otsalizarraren fruitua gorria da"* pasarteak emanda, adibidez, *otsalizarra* zuhaitz bat dela igar genezake nahiz eta hitz hori aurrez ez ezagutu, testuinguru horietan zuhaitz-izen bat agertzea bailitzateke ohikoena.

Printzipio horretan oinarrituz, corpus elebakar bateko agerkidetza-patroiak erabili ohi dira hitz-bektoreak modu automatikoan ikasteko. Horretarako hurbilpenak bi multzotan sailkatu izan dira (Baroni et al., 2014): (i) kontaketan oinarritutako ereduak, Hizkuntzaren Prozesamenduaren arloan ibilbide luzea dutenak, eta (ii) eredu prediktiboak, azken hamarkadan nagusitu direnak. Jarraian, hurbilpen horietako bakoitza xehetasun gehiagoz azalduko dugu.

Kontaketan oinarritutako ereduak

Kontaketan oinarritutako ereduen bereizgarria X agerkidetza-matrize gordin batetik abiatzea da. Matrizeko $X_{i,*}$ errenkada bokabularioko i. hitzari dagokio, eta $X_{i,j}$ posizioak

[10] u eta v bektoreen arteko kosinu-antzekotasuna $\cos(u,v) = \frac{u \cdot v}{||u||||v||}$ gisara definitzen da, $||\cdot||$ norma euklidearra izanik. Hortaz, luzera normalizatutako biderketa eskalartzat ikus daiteke.

hitz horren eta j. elementuaren arteko agerkidetza-maiztasuna—edo bertatik eratorritako neurriren bat—jasotzen du. $X_{*,j}$ zutabetzat, berriz, osagai ezberdinak erabil daitezke. Informazio-berreskurapenera zuzenduta egonik, kontaketan oinarritutako lehen ereduek dokumentuak zerabiltzaten, adibidez (Salton et al., 1975; Deerwester et al., 1990). Ohikoena, baina, errenkadatzat bezala zutabetzat ere hitzak eurak erabiltzea da (Lund and Burgess, 1996). Eredu horietan, $X_{i,j}$ posizioak i. hitza j. hitzaren testuinguruan (leiho jakin baten baitan) zenbatetan agertzen den jasotzen du. Hurbilpen horrek arestian azalduriko hipotesi distribuzionala jarraitzen du bete-betean. Izan ere, i. eta j. hitzak antzeko testuinguruetan agertzen badira, k. hitzarekin duten agerkidetza-maiztasuna antzekoa izango da ($X_{i,k} \approx X_{j,k}$), eta euren $X_{i,*}$ eta $X_{j,*}$ errenkadek ere antzeko balioak izango dituzte horrenbestez. Hori horrela, $X_{i,*}$ errenkada har dezakegu i. hitza errepresentatuko duen hitz-bektoretzat.

Oinarrizko hurbilpen horrek testuinguru guztiei garrantzia bera ematen die. Errealitatean, baina, testuinguru batzuk beste batzuk baino adierazgarriagoak suertatzen dira. Hala nola, arestiko adibidean *itzalpean* eta *fruitua* testuinguru-hitzak oso lagungarriak dira *otsalizar* hitzaren esanahia ezagutzeko. *Zegoen* eta *da* testuinguru-hitzak, ostera, ez dira bereziki esanguratsuak, beste hitz askoren testuinguruan ere agertu ohi dira eta. Arazo horri aurre egiteko, agerkidetza-maiztasun gordinen ordez euretatik eratorritako beste neurri batzuk erabili ohi dira, agerkidetza bakoitzaren esangura maila neurtzen dutenak. Ohiko aukera bat puntukako elkarrekiko informazioa (PMI, ingelesezko *pointwise mutual information*-etik) erabiltzea da, $X_{i,j}$ posizioko w_i eta w_j hitzak elkarrekin eta bakarka agertzeko probabilitate-estimazioen arteko aldea neurtzen duena (Church and Hanks, 1990):

$$\mathrm{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

Normalean balio negatiboak zeroekin ordezkatzen dituen aldaera bat erabiltzen da, PMI positibo (PPMI, ingelesezko *positive PMI*-tik) deritzona (Niwa and Nitta, 1994):

$$\mathrm{PPMI}(w_i, w_j) = \max\left(0, \mathrm{PMI}(w_i, w_j)\right)$$

Oinarrizko hurbilpenaren beste arazo bat da X matrizea oso handia izan ohi dela, koadratikoki hazten baita bokabularioaren tamainarekiko. Arazo horri aurre egiteko, dimentsionaltasun-murrizketa teknikak erabili ohi dira. Horretarako metodo ezagunena ezkutuko analisi semantiko (LSA, ingelesezko *latent semantic analysis*-etik) deritzona da (Dumais et al., 1988; Deerwester et al., 1990), informazio-berreskurapenaren arloan ezkutuko indexazio semantiko (LSI, ingelesezko *latent semantic indexing*-etik) deitzen zaiona. LSAk balio singularren deskonposaketa erabiltzen du X agerkidetza-matrizea beste hiru matrizeren biderkadura modura deskonposatzeko:

$$X = U\Sigma V^T$$

non U eta V matrize ortonormalak baitira eta Σ balio singularren matrize diagonala. Izan bedi Σ_k k balio singular handienek osaturiko matrize diagonala, eta U_k eta V_k balio singular horiei dagozkien U eta V matrizeetako zutabeek osatutako matrizeak. LSAk $\hat{X} = U_k\Sigma_k V_k$ hartzen du, eta $\hat{X}_{i,*}$ erabili bokabularioko i. sarreraren hitz-bektoretzat, k bektoreon dimentsio kopurua izanik. $\hat{X} = U_k\Sigma_k V_k$ matrize berria X jatorrizkoaren hein txikiagoko hurbilpentzat ikus daiteke, k heineko $\hat{X}$ matrize guztien artean $||\hat{X} - X||_F$ hurbilpen-errorea minimizatzen duena baita, $||\cdot||_F$ Frobeniusen norma izanik. Prozedura hori hainbat ikuspegitatik interpreta daiteke (Turney and Pantel, 2010): esanahi ezkutua azaleratzeko metodo gisara, zarata murrizteko teknika modura, ordena altuagoko agerkidetzak modelatzeko metodotzat, edota sakabanatzea murrizteko prozedura modura, jatorrizko X matrizea sakabanatua baita (nagusiki zeroz osatua) eta $\hat{X}$ matrizea, berriz, dentsoa.

Kontaketan oinarritutako ereduen inguruko literatura-azterketa xeheago baterako, ikus Turney and Pantel (2010).

Eredu prediktiboak

Agerkidetza-matrize gordin batetik abiatu beharrean, eredu prediktiboek hitz bakoitzaren testuingurua aurresaten duen neurona-sare bat entrenatzen dute **autogainbegiratzearen printzipioa** jarraituz, eta hark ikasitako errepresentazioak hartu hitz-bektoretzat. Hain justu ere, hitzen errepresentazio banatuak ikastearen ideia funtsezkoa izan da **hizkuntza-eredu neuronaletan** euren hastapenetatik (Bengio et al., 2003). Gerora, hizkuntza-eredu neuronalen bidez ikasitako hitz-bektoreak beste ataza batzuetarako ere lagungarriak izan zitezkeela erakutsi zen, eta horretara zuzenduriko eredu berriak proposatu (Collobert and Weston, 2008; Collobert et al., 2011; Turian et al., 2010; Huang et al., 2012).

Hitz-bektoreak Hizkuntzaren Prozesamenduaren erdigunean jarri zituen lana, baina, Mikolov et al. (2013a,c) izan zen. **Eredu log-linealen bidez** kalitatezko hitz-bektoreak modu eraginkorrean ikas zitezkeela erakutsi zuten bertan, eta horretarako *word2vec* tresna publiko egin. Zehatzagoak izanez, bi dira lan horretan proposatutako ereduak: skip-gram eta CBOW.

$(w_1, ..., w_N)$ hitz-sekuentziak osatutako ikasketa-corpusa emanda, **skip-gram** ereduak corpuseko hitz bakoitzeko bere testuinguruko hitzak banaka aurresaten ikasten du:

$$\mathcal{L}_{\text{SG}} = -\sum_{t} \sum_{-c \leq j \leq c, j \neq 0} \log p\left(w_t | w_{t+j}\right)$$

c testuinguru-leihoaren tamaina izanik. Bere oinarrizko formulazioan, skip-gram ereduak

softmax funtzioa erabiltzen du $p\left(w_t|w_{t+j}\right)$ probabilitatea kalkulatzeko:

$$p\left(w_t|w_{t+j}\right) = \frac{\exp\left(x_{w_{t+j}} \cdot \tilde{x}_{w_t}\right)}{\sum_{w'} \exp\left(x_{w_{t+j}} \cdot \tilde{x}_{w'}\right)}$$

x_{w_i} eta $\tilde{x}_{w_i}$ w_i hitzaren sarrerako eta irteerako bektoreak izanik, hurrenez hurren. Ikasketa gradiente jaitsiera estokastikoaren bidez egiten da eta, behin bukatuta, x_i hartzen da bokabularioko i. sarreraren hitz-bektoretzat, $\tilde{x}$ irteerako bektoreak baztertuz.

CBOW ereduak, berriz, testuinguruko hitz guztiak hartuta erdiko hitza aurresaten ikasten du:

$$\mathcal{L}_{\text{CBOW}} = -\sum_t \log p\left(w_t|w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}\right)$$

$p\left(w_t|w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}\right)$ probabilitatea kalkulatzeko softmax funtzioa erabiltzen du, testuingurua errepresentatzeko bertako hitzen bektoreen batezbestekoa hartuz:

$$p\left(w_t|w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}\right) = \frac{\exp\left(\frac{1}{2c}\sum_{-c\leq j\leq c, j\neq 0} x_{w_{t+j}} \cdot \tilde{x}_{w_t}\right)}{\sum_{w'} \exp\left(\frac{1}{2c}\sum_{-c\leq j\leq c, j\neq 0} x_{w_{t+j}} \cdot \tilde{x}_{w'}\right)}$$

Nola skip-gram hala CBOWren kasuan, baina, softmax funtzioa kalkulatzea oso garestia da, izendatzaileko batugai kopurua bokabularioaren tamainaren araberakoa baita. Mikolov et al. (2013a,c) lanaren gakoetako bat **softmax osoaren hurbilpen eraginkor** bat erabiltzea da, kalitatezko hitz-bektoreak ikasteko nahikoa dena. Horretarako bi metodo proposatu zituzten: softmax hierarkikoa lehendabiziko lanean (Mikolov et al., 2013a), eta laginketa negatiboa ondorengoan (Mikolov et al., 2013c).

Softmax hierarkikoak (Morin and Bengio, 2005; Mikolov et al., 2013a) irteerako geruza zuhaitz bitar modura errepresentatzen du hitz-maiztasunen araberako Huffman kodeketa erabiliz (Huffman, 1952). Bokabularioko sarrerak hostoei dagozkie, eta nodo bakoitzak bektore bat du, haren seme-alaba guztien probabilitate erlatiboak errepresentatzen dituena. Horri esker, hitz bakoitzaren baldintzazko probabilitateak bokabularioaren tamainarekiko denbora logaritmikoan kalkula daitezke batez bestean. Zehatzagoak izanez, $n(w_t, i)$ zuhaitzaren errotik w_t hitzera doan bidearen i. nodoa izanik, $\tilde{x}_{n(w_t,i)}$ nodo horren irteerako bektorea, $\text{ch}(n(w_t, i))$ nodo horren seme-alaba arbitrario bat, eta $L(w_t)$ bidearen luzera, softmax hierarkikoak honela definitzen du $p\left(w_t|w_{t+j}\right)$ baldintzazko probabilitatea skip-gramen kasuan:

$$p\left(w_t|w_{t+j}\right) = \prod_{i=1}^{L(w_t)-1} \sigma\left(\llbracket n(w_t, i+1) = \text{ch}(n(w_t, i))\rrbracket x_{w_{t+j}} \cdot \tilde{x}_{n(w_t,i)}\right)$$

$\sigma(x) = 1/(1+\exp(-x))$ sigmoide funtzioa izanik, eta $\llbracket a\rrbracket = 1$ izanik a egia denean eta -1 bestela. CBOWren formulazioa antzekoa da, $x_{w_{t+j}}$ ordez testuinguruko hitzen sarrerako

bektoreen batezbestekoa hartuz softmax osoarekin bezala.

Laginketa negatiboak, berriz, ikasketa-helburua bera aldatzen du: testuinguru bakoitza emanda dagokion hitza aurresan beharrean, testuinguru/hitz bikote bakoitza corpusetik ala zarata-banaketa batetik datorren aurresaten ikasten du. Horretarako sailkatzaile bitar bat erabiltzen du, bokabularioaren tamainarekiko denbora konstantea behar duena. Skip-gramen kasuan, honakoa da helburu-funtzioa:

$$\mathcal{L}_{\text{SGNS}} = -\sum_{t} \sum_{-c \leq j \leq c, j \neq 0} \left(\log \sigma \left(x_{w_{t+j}} \cdot \tilde{x}_{w_t} \right) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma \left(-x_{w_{t+j}} \cdot \tilde{x}_{w_i} \right) \right] \right)$$

k lagin negatiboen kopurua izanik. CBOWren formulazioa antzekoa da, berriro ere $x_{w_{t+j}}$ ordez testuinguruko hitzen sarrerako bektoreen batezbestekoa hartuz. $P_n(w)$ zarata-banaketa da, Mikolov et al. (2013c) lanak unigrama-banaketan oinarrituz definitzen duena:

$$P_n(w) = \frac{f(w)^{3/4}}{\sum_{w'} f(w')^{3/4}}$$

$f(w)$ w hitzaren maiztasuna izanik.

Mikolov et al. (2013c) lanean proposatutako beste hobekuntza nagusia **hitz usuen azpilaginketa** da. Teknika horrek corpuseko w_i hitz bakoitza honako probabilitatearen arabera baztertzen du:

$$P\left(w_i\right) = \max\left(0, 1 - \sqrt{\frac{t}{f(w_i)}}\right)$$

t atalasea izanik (maiztasun hori baino baxuagoko hitzak ez dira sekula baztertzen). Mikolov et al. (2013c) lanean atalasetzat 10^{-5} inguruko balioak erabiltzea gomendatzen dute. Hitz usuen azpilaginketak ikasketa azkartzeko balio du, iterazioko adibide kopurua murrizten baitu. Era berean, maiztasun handiko hitz asko funtzio-hitzak izan ohi dira, karga semantiko eskasekoak, eta hizpide dugun teknikak hitz horiei gehiegizko garrantzia ematea galarazten du.

Gerora proposatutako beste hedapen aipagarri bat sarrerako bektoreak **karaktere mailako informazioarekin** aberastea da (Bojanowski et al., 2017). Lan horrek w_i hitz bakoitzarentzat G_{w_i} multzoa definitzen du, hitz hori bera eta bere karaktereen n-gramak biltzen dituena, hitzaren hasiera eta bukaera adierazteko `<` eta `>` karaktere bereziak erabiliz. Adibide modura, *zuhaitz* hitzari ondorengo G_{zuhaitz} multzoa dagokio $n = 3$ denean:

$$G_{\text{zuhaitz}} = \{ \texttt{<zu, zuh, uha, hai, ait, itz, tz>, <zuhaitz>} \}$$

Multzoko g osagai bakoitzari z_g bektore bat esleitzen saio, eta w_i hitzaren x_{w_i} sarrerako bektorea bere multzoko osagaien bektoreak batuz eskuratzen da:

$$x_{w_i} = \sum_{g \in G_{w_i}} z_g$$

Bojanowski et al. (2017) lanean skip-gram ereduko sarrerako bektore finkoak bektoreokin ordezkatzen dituzte, hitz beraren forma ezberdinen artean (*zuhaitz*, *zuhaitza*, *zuhaitzaren*...) informazioa partekatzea ahalbidetzen duena. Era berean, horretarako *fastText* tresna publikoki eskuragarri jarri zuten, bai eta Wikipedia corpusa erabiliz tresna horrekin ikasitako 294 hizkuntzatako hitz-bektoreak ere.

Amaitzeko, **kontaketan oinarritutako ereduak eta eredu prediktiboak** familia ezberdintzat aurkeztu ohi badira ere, hainbat lanek bien artean **lotura estua** dagoela erakutsi dute. Haien artean entzutetsuenetarikoa Levy and Goldberg (2014) da, bektoreen dimentsionaltasuna behar bezain handia denean laginketa negatibodun skip-gramen soluzio optimoa $x_i \cdot \tilde{x}_j = \text{PMI}(w_i, w_j) - \log k$ dela erakusten duena. Hori horrela, kontaketan oinarritutako metodoen antzera skip-gramek ere PMI matrizea faktorizatzen duela arrazoitzen dute, ikasketan zehar matrize hori modu esplizituan agertu ez arren. Gerora, Cotterell et al. (2017) lanak faktorizazio inplizitu hori familia esponentzialeko osagai nagusien analisi (Collins et al., 2002) modura interpreta zitekeela erakutsi zuen.

1.5.2 Itzulpen automatikoa

$\mathbf{x} = x_1^n = (x_1, ..., x_n)$ jatorrizko hizkuntzako sekuentzia bat emanda, itzulpen automatikoaren helburua $\mathbf{y} = y_1^m = (y_1, ..., y_m)$ helburuko hizkuntzako sekuentzia baliokide egokiena topatzea da. Edozein $\mathbf{s}$ sekuentziako s_i osagai bakoitzari *token* deituko diogu, eta s_i^j erabiliko dugu i. tokenetik j. tokenera doan $(s_i, s_{i+1}, ..., s_{j-1}, s_j)$ azpisekuentzia izendatzeko. Oro har, sekuentziak esaldiei dagozkie eta tokenak hitzei, baina dokumentu edota azpihitz mailan aritzea ere posible da.

Weaver (1955) memorandumetik hasita, itzulpen automatikoa ibilbide luzeko arloa da. Hasierako sistemek hurbilpen sinboliko bat jarraitzen zuten, eta erregeletan oinarritzen ziren (Hutchins, 1986). Itzulpen automatiko modernoak, baina, datuak ditu abiapuntu, eta formulazio probabilistiko bat darabil (Brown et al., 1988, 1990). Horrela, $\mathbf{y}$ itzulpen-hautagai bakoitzari $p(\mathbf{y}|\mathbf{x})$ probabilitate-masa bat esleitzen zaio, helburua probabilitate altueneko $\hat{\mathbf{y}}$ itzulpena aurkitzea izanik:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$$

Sistema horiek **entrenamendu** edo ikasketa bat behar dute, non $p(\mathbf{y}|\mathbf{x})$ modelatzen duen ereduaren parametroak estimatzen baitira. Horretarako corpus paralelo bat erabili ohi da, bi hizkuntzetako sekuentzia baliokideak biltzen dituena (oro har, milioika esaldi-bikote). Behin eredua ikasita $\hat{\mathbf{y}}$ itzulpen optimoa bilatzeko prozedurari, berriz, **deskodeketa** deritzo. Praktikan, $\mathbf{y}$ itzulpen posible guztien multzoa handiegia izan ohi da soluzio zehatza aurkitzeko, eta horren ordez sorta-bilaketa moduko bilaketa-algoritmo heuristikoak erabili ohi dira.

$p(\mathbf{y}|\mathbf{x})$ modelatzeko erabilitako ereduaren arabera, corpusetan oinarritutako itzulpen

automatikoko bi hurbilpen nagusi daude: itzulpen automatiko estatistikoa, $p(\mathbf{y}|\mathbf{x})$ hainbat faktoretan deskonposatu eta horietako bakoitza estimatzeko eredu estatistiko bat darabilena, eta itzulpen automatiko neuronala, $p(\mathbf{y}|\mathbf{x})$ zuzenean estimatzeko neurona-sare bat darabilena. Jarraian, hurbilpen bakoitza xehetasun gehiagoz azalduko dugu.

Itzulpen automatiko estatistikoa

Itzulpen automatiko estatistikoa **kanal zaratatsuaren ereduan** oinarritzen da, eta Bayesen teorema darabil $p(\mathbf{y}|\mathbf{x})$ faktorizatzeko:[11]

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg\max_{\mathbf{y}} \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} = \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$$

$p(\mathbf{x})$ ez dago $\mathbf{y}$ itzulpenaren menpe, azken berdintza ematen duena. Horri esker, problema bi zatitan banatzen da: $p(\mathbf{y})$ modelatzen duen osagaiari *hizkuntza-eredu* deritzo, eta $p(\mathbf{x}|\mathbf{y})$ modelatzen duenari, berriz, *itzulpen-eredu*.

Hizkuntza-eredua $\mathbf{y}$ helburuko hizkuntzako sekuentzia bakoitzari $p(\mathbf{y})$ probabilitate-masa bat esleitzeaz arduratzen da. Katearen erregela erabiliz, probabilitate hori honela faktoriza daiteke:

$$p(\mathbf{y}) = \prod_{i=1}^{m} p\left(y_i|y_1^{i-1}\right)$$

Itzulpen automatiko estatistikoan $\mathbf{y}$ sekuentzia k. ordenako Markoven kate modura modelatu ohi da, azken k tokenak soilik erabiliz $p\left(y_i|y_1^{i-1}\right)$ hurbiltzeko:

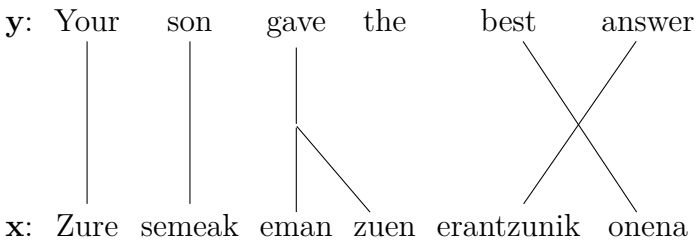
$$p\left(y_i|y_1^{i-1}\right) \approx p\left(y_i|y_{i-(k-1)}^{i-1}\right)$$

k parametroak ereduaren ordena definitzen du, eta bere balioaren arabera unigrama, bigrama, trigrama eta, orokorrean, n-grama ereduez mintzo ohi da. Baldintzazko probabilitate bakoitza estimatzeko, berriz, corpus elebakar bateko maiztasun-kontaketak erabiltzen dira:

$$p\left(y_i|y_{i-(k-1)}^{i-1}\right) = \frac{\text{count}\left(y_{i-(k-1)}^{i}\right)}{\text{count}\left(y_{i-(k-1)}^{i-1}\right)}$$

non $\text{count}(y_i^j)$ funtzioak y_i^j azpisekuentziak corpusean duen agerpen-kopurua adierazten baitu. Estimazio horiek, baina, zeroko probabilitatea esleitzen diete corpusean agertzen ez diren n-gramei. Arazo horri aurre egiteko, *back-off* eta leuntze teknikak erabili ohi dira (Kneser and Ney, 1995; Chen and Goodman, 1996).

[11] Gai honen inguruko lehen lanek frantsesetik ingelesera itzultzea zuten helburu, eta $\mathbf{f}$ eta $\mathbf{e}$ zerabiltzaten hizkuntza horietako sekuentziak izendatzeko (Brown et al., 1988, 1993). Itzulpen automatiko estatistikoko lan gehienek notazio horri eutsi diote hizkuntzak edozein izanik ere. Itzulpen automatiko neuronaleko literaturan, berriz, ohikoena $\mathbf{x}$ eta $\mathbf{y}$ erabiltzea da. Lan honetan notazio bateratu bat erabili nahi izan dugu paradigma bientzat, eta azken aukeraren alde egin.



1.2 irudia: Hitz-lerrokatzearen adibide bat, non $\mathbf{a} = (1, 2, 3, 3, 6, 5)$.

Itzulpen-eredua, berriz, $p(\mathbf{x}|\mathbf{y})$ baldintzazko probabilitatea modelatzeaz arduratzen da. Horretarako lehen ereduak token mailan aritzen ziren, eta **hitz-lerrokatzea** zuten oinarri. 1.2 irudiak erakutsi bezala, hitz-lerrokatzeak $\mathbf{x}$ sekuentziako token bakoitza gehienez ere $\mathbf{y}$ sekuentziako token batekin lotzen du. Halako lerrokatze bakoitza $\mathbf{a} = a_1^n = (a_1, ..., a_n)$ zenbaki arrunten sekuentzia modura adieraziko dugu, non $a_i = j$ izango baita baldin eta x_i tokena y_j tokenari lotuta badago. x_i tokenak inolako loturarik ez badu $a_i = 0$ adieraziko dugu, y_0 posizioan `NULL` token berezia dagoela suposatuz. Hitz-lerrokatzeko ereduek $\mathbf{y}$ sekuentzia emanda $\mathbf{x}$ sekuentzia $\mathbf{a}$ lerrokatzearen bidez itzultzeko $p(\mathbf{x}, \mathbf{a}|\mathbf{y})$ probabilitatea modelatzen dute, eta $p(\mathbf{x}|\mathbf{y})$ lerrokatze posible guztien baturatzat adierazi:

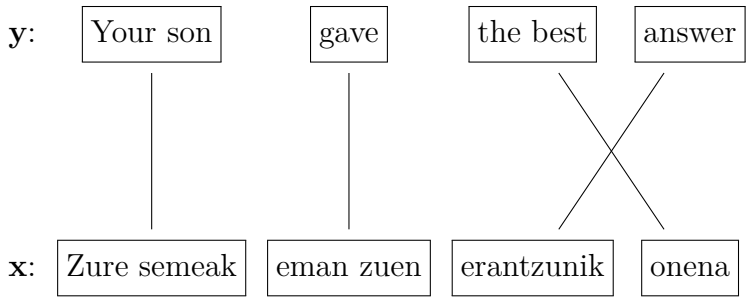
$$p(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{a}} p(\mathbf{x}, \mathbf{a}|\mathbf{y})$$

Katearen erregela erabiliz, $p(\mathbf{x}, \mathbf{a}|\mathbf{y})$ probabilitatea honela faktorizatu ohi da (Brown et al., 1993):

$$p(\mathbf{x}, \mathbf{a}|\mathbf{y}) = p(n|\mathbf{y}) \prod_{i=1}^{n} p\left(a_i|a_1^{i-1}, x_1^{i-1}, n, \mathbf{y}\right) p\left(x_i|a_1^{i}, x_1^{i-1}, n, \mathbf{y}\right)$$

non $p(n|\mathbf{y})$ terminoa x sekuentziaren luzera n izateko baldintzazko probabilitatea baita, $p\left(a_i|a_1^{i-1}, x_1^{i-1}, n, \mathbf{y}\right)$ x_i tokena y_{a_i} tokenari lerrokatuta egotekoa, eta $p\left(x_i|a_1^{i}, x_1^{i-1}, n, \mathbf{y}\right)$ $\mathbf{x}$ sekuentziako $i.$ tokena x_i izatekoa.

Faktore horiek guztiak modelatu ahal izateko hurbilpen ezberdinak erabili izan dira. Ezagunenak **IBM ereduak** deritzenak dira, geroz eta konplexutasun handiagoko 5 eredu biltzen dituztenak (Brown et al., 1993). Guztietan sinpleena den IBM 1 ereduak n luzera guztiak ekiprobableak direla suposatzen du, bai eta luzera bereko lerrokatze guztiak ere, alegia, $p(n|\mathbf{y}) = \epsilon$ eta $p\left(a_i|a_1^{i-1}, x_1^{i-1}, n, \mathbf{y}\right) = (m+1)^{-1}$, ϵ konstante txiki bat izanik. Era berean, x_i tokena sortzeko probabilitatea harekin lerrokatutako y_{a_i} tokenaren menpe soilik dagoela suposatzen du, alegia, $p\left(x_i|a_1^{i}, x_1^{i-1}, n, \mathbf{y}\right) = t\left(x_i|y_{a_i}\right)$, x_i tokenaren y_{a_i} tokenarekiko *itzulpen-probabilitatea* deritzona. Hori horrela, $p(\mathbf{x}, \mathbf{a}|\mathbf{y})$ probabilitatea



1.3 irudia: Sintagma mailako itzulpenaren adibide bat.

honela berridatz daiteke:

$$p(\mathbf{x}, \mathbf{a}|\mathbf{y}) = \frac{\epsilon}{(m+1)^n} \prod_{i=1}^{n} t\left(x_i|y_{a_i}\right)$$

Gainerako IBM ereduek aurreneko horren hainbat gabezia konpontzen dituzte: IBM 2 ereduak berrordenatze-eredu absolutu bat barneratzen du, IBM 3 ereduak y_i token bakoitza $\mathbf{x}$ sekuentziako zenbat tokenekin lerrokatu modelatzen duen emankortasun-eredu bat erabiltzen du; IBM 4 ereduak berrordenatze-eredu erlatibo bat darabil, eta IBM 5 ereduak aurreko bien defizientzia-arazoa konpontzen du, probabilitate-masaren zati bat ezinezko gertaerei esleitzea eragiten zuena. Eredu horietaz gain, IBM 2 ereduaren Vogel et al. (1996) eta Dyer et al. (2013) lanetako birparametrizazioak ere oso erabiliak izan dira. Lehena Markoven eredu ezkutuetan oinarritzen da, IBM 4 ereduarekin ere uztartu direnak (Och and Ney, 2003).

Erabilitako eredua edozein dela ere, bere parametroak **egiantz handieneko estimazioaren** bidez ikasi ohi dira corpus paralelo baten gainean. Horretarako itxaropen-maximizazio algoritmo iteratiboa (Baum, 1972; Dempster et al., 1977) erabili ohi da. Era berean, IBM 1 ez beste eredu guztiek optimo lokal bat baino gehiago dituztenez, algoritmoa hasieratzeko aurreko ereduarentzat lortutako soluzioa erabili ohi da.

Hasiera batean itzulpen-ereduak token mailan aritzen baziren ere, gerora **sintagmetan oinarritutako ereduak** nagusitu ziren (Och et al., 1999; Zens et al., 2002; Koehn et al., 2003; Och and Ney, 2004). 1.3 irudiak erakutsi bezala, eredu horietan $\mathbf{x}$ eta $\mathbf{y}$ sekuentziak I sintagmatan segmentatzen dira (token bat edo gehiagoko azpisekuentzia jarraituak),[12] eta $\mathbf{x}$ sekuentziako $\bar{x}_i$ sintagma bakoitza $\mathbf{y}$ sekuentziako $\bar{y}_i$ sintagmaren bidez itzuli. Hori horrela, sintagmetan oinarritutako itzulpen-ereduek honela modelatu ohi dute $p(\mathbf{x}|\mathbf{y})$

[12] Itzulpen automatiko estatistikoaren arloan *sintagma* terminoa edozein token-segmentu izendatzeko erabiltzen da, inolako murriztapen linguistikorik gabe. Horren ordez osagai sintaktikoak (zuhaitz sintaktiko bateko adabegiak) soilik erabiltzea kaltegarria dela erakutsi izan da enpirikoki (Koehn et al., 2003).

baldintzazko probabilitatea segmentazio jakin baterako:

$$p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{I} \phi\left(\bar{x}_i|\bar{y}_i\right) d\left(start_i, end_{i-1}\right)$$

$\phi\left(\bar{x}_i|\bar{y}_i\right)$ terminoak $\bar{y}_i$ sintagmaren itzulpena $\bar{x}_i$ izateko probabilitatea modelatzen du, eta $d\left(start_i, end_{i-1}\right)$ terminoak, berriz, itzulitako sintagmen hurrenkera. Azken horri *berrordenatze-eredu* deritzo, eta ohiko aukera bat itzulitako sintagma bakoitzak aurrekoa-rekiko duen distantzian oinarritzea da:

$$d\left(start_i, end_{i-1}\right) = \alpha^{|start_i - end_{i-1} - 1|}$$

non α konstante bat baita, $start_i$ $\mathbf{y}$ sekuentziako $i.$ sintagmari $\mathbf{x}$ sekuentzian dagokion sintagmaren hasierako posizioa, eta end_{i-1} $\mathbf{y}$ sekuentziako $(i-1).$ sintagmari $\mathbf{x}$ sekuentzian dagokion sintagmaren bukaerako posizioa. $\phi\left(\bar{x}_i|\bar{y}_i\right)$ itzulpen-probabilitateak estimatzeko, berriz, corpus paralelo bateko maiztasun-kontaketak erabiltzen dira:

$$\phi\left(\bar{x}_i|\bar{y}_i\right) = \frac{\text{count}\left(\bar{x}_i, \bar{y}_i\right)}{\text{count}\left(\bar{y}_i\right)}$$

$\text{count}\left(\bar{x}_i, \bar{y}_i\right)$ terminoak $(\bar{x}_i, \bar{y}_i)$ sintagma-bikoteak corpusean duen agerpen kopurua adierazten du. Horretarako beharrezkoa da corpusetik sintagma-bikoteak erauztea, arestiko hitz-lerrokatzeetan oinarrituz egin ohi dena. Hitz-lerrokatzeko ereduek, baina, 1-N motako lerrokatzeak soilik ematen dituzte (x_i token bakoitza gehienez ere y_{a_i} token bakarrarekin lerrokatzen dute). Arazo horri aurre egiteko, hitz-lerrokatzea bi noranzkoetan egin ohi da, 1-N eta M-1 motako lerrokatzeak lortuz, eta biak konbinatu M-N motako lerrokatzeak lortzeko.[13] Prozedura horri *simetrizazio* deritzo, eta heuristiko ezberdinak erabiliz egin ohi da. Metodo gehienak bi noranzkoetako lerrokatzeen ebakiduratik abiatzen dira, eta bien bildurako beste lerrokatze-puntu batzuk gehitu irizpide ezberdinen arabera (Och and Ney, 2003; Tillmann, 2003; Venugopal et al., 2003). Behin hori eginda, lerrokatze horrekin kontsistenteak diren sintagma-bikoteak erauzten dira. $\mathbf{x}$ eta $\mathbf{y}$ sekuentziak eta A haien lerrokatzea emanda, non $(i, j) \in A$ baldin eta x_i eta y_j lerrokatuta badaude, ondorengo $BP\left(\mathbf{x}, \mathbf{y}, A\right)$ sintagma-bikoteak erauzi ohi dira (Zens et al., 2002; Och and Ney, 2004):

$$BP\left(\mathbf{x}, \mathbf{y}, A\right) = \Big\{ \left(x_i^{i+k}, y_j^{j+l}\right) : \forall\left(i', j'\right) \in A : i \leq i' \leq i+k \iff j \leq j' \leq j+l \\ \wedge \exists\left(i', j'\right) \in A : i \leq i' \leq i+k \iff j \leq j' \leq j+l \Big\}$$

[13] Sintagma-lerrokatzeak zuzenean ikastea ere posible da (Marcu and Wong, 2002), baina praktikan hitz-lerrokatze klasikoan oinarritutako hurbilpenak nagusitu ziren.

Sintagmetan oinarritutako sistemekin batera garatutako beste hobekuntza garrantzitsu bat jatorrizko kanal zaratatsuaren eredua **eredu log-lineal** modura orokortzea izan zen (Och and Ney, 2002). Eredu log-linealek hainbat ezaugarri-funtzio konbinatzen dituzte $p(\mathbf{y}|\mathbf{x})$ modelatzeko:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y})\right)}{\sum_{\mathbf{y}'} \exp\left(\sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y}')\right)}$$

non $h_i(\mathbf{x}, \mathbf{y})$ i. ezaugarri funtzioa baita, eta λ_i hari dagokion pisua. Eskala logaritmikora pasata, eta izendatzailea $\mathbf{y}$ itzulpen-hautagaiarekiko independentea denez, $\hat{\mathbf{y}}$ itzulpen optimoa honela bila daiteke:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg\max_{\mathbf{y}} \sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y})$$

Lehen aipatu bezala, kanal zaratatsuaren jatorrizko eredua eredu log-linealen kasu berezitzat ikus daiteke, non $h_1(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{y})$, $h_2(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}|\mathbf{y})$ eta $\lambda_1 = \lambda_2 = 1$. Era berean, sintagmetan oinarritutako oinarrizko eredua $h_1(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{y})$, $h_2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{I} \log \phi(\bar{x}_i|\bar{y}_i)$ eta $h_3(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{I} |start_i - end_{i-1} - 1|$ kasu berezi gisara ere ikus daiteke, $\lambda_1 = \lambda_2 = 1$ eta $\lambda_3 = \log \alpha$ izanik. Eredu log-linealek estimazio hobeak lortzeko ezaugarri-funtzio gehiago barneratzea ahalbidetzen dute. Praktikan, honakoak dira sistema moderno gehienetan erabili ohi diren ezaugarri-funtzioak:

- $\log p(\mathbf{y}) = \sum_{i=1}^{m} \log p\left(y_i|y_1^{i-1}\right)$ **hizkuntza-eredua**, lehen azaldu duguna.
- $\sum_{i=1}^{I} \log \phi(\bar{y}_i|\bar{x}_i)$ eta $\sum_{i=1}^{I} \log \phi(\bar{x}_i|\bar{y}_i)$ aurreranzko eta atzeranzko **sintagmen itzulpen-probabilitateak**, lehen azaldu ditugunak.
- $\sum_{i=1}^{I} \log p_w(\bar{y}_i|\bar{x}_i)$ eta $\sum_{i=1}^{I} \log p_w(\bar{x}_i|\bar{y}_i)$ aurreranzko eta atzeranzko **pisu lexikoak** (Koehn et al., 2003). Pisu lexikoak arestian azaldutako hitz-lerrokatzeko $t(x_i|y_{a_i})$ itzulpen-probabilitateetan oinarritzen dira:

  $$p_w(\bar{x}|\bar{y}) = \max_a \prod_{i=1}^{n} \frac{1}{|\{j : (i,j) \in a\}|} \sum_{(i,j)\in a} t(x_i|y_j)$$

  non, notazioa sinplifikatze aldera, $\bar{x} = (x_1, ..., x_n)$ eta $\bar{y} = (y_1, ..., y_m)$ baitira, eta a sintagma-bikote horren corpus paraleloko hitz-lerrokatze bat.
- $\sum_{i=1}^{I} |start_i - end_{i-1} - 1|$ **distantzian oinarritutako berrordenatze-eredua**, lehen azaldu duguna.
- $\sum_{i=1}^{I} \log p(o_i|\bar{x}_i, \bar{y}_i)$ **berrordenatze-eredu lexikoak** (Tillmann, 2004; Koehn et al., 2005; Galley and Manning, 2008). $p(o_i|\bar{x}_i, \bar{y}_i)$ termino bakoitzak $(\bar{x}_i, \bar{y}_i)$ sintagma-bikotearen orientazio-probabilitatea modelatzen du. Sintagma-bikote baten orientazioa *monotonoa* dela esaten da $\bar{y}_i$ sintagmaren ezkerrekoa $\bar{x}_i$ sintagmaren ezkerre-

koari dagokionean, *trukatua* $\bar{y}_i$ sintagmaren ezkerrekoa $\bar{x}_i$ sintagmaren eskuinekoari dagokionean, eta *etena* bestela. Probabilitate horiek entrenamendu-corpus paraleloan estimatzen dira sintagma-bikote bakoitzarentzat.

- m ($\mathbf{y}$ itzulpenaren token kopurua), **hitz-penalizazio** deitu ohi zaiona eta itzulpen labur edo luzeagoak sortzeko joera kontrolatzen duena.

- k (sintagma kopurua), **sintagma-penalizazio** deitu ohi zaiona eta sekuentziak segmentatzerakoan sintagma labur edo luzeagoak erabiltzeko joera kontrolatzen duena.

λ_i pisuak aparteko balidazio-corpus paralelo batean optimizatu ohi dira ebaluazio-metrikaren baten arabera, gehienetan BLEU izan ohi dena (Papineni et al., 2002). Horretarako algoritmo ezagunena MERT da (Och, 2003), baina beste hainbat metodo ere proposatu izan dira (Neubig and Watanabe, 2016).

Itzulpen automatiko estatistikoaren baitan hurbilpen nagusia azaldu berri dugun sintagmetan oinarritutako eredu log-linealena bada ere, hurbilpen horren hainbat **hedapen** ere landu izan dira. Aipagarrienak eredu hierarkikoak edo sintaxian oinarritutakoak dira (Chiang, 2005), sintagmak unitate atomikotzat tratatu beharrean ez-amaierako aldagaiak barneratu eta zuhaitz-egitura bat jarraituz itzultzen dutenak, bai eta eredu faktorizatuak ere (Koehn and Hoang, 2007), token bakoitzaren formaz gain bere lema edota kategoria gramatikala moduko faktore gehigarriak ere erabiltzen dituztenak.

Deskodeketa, azkenik, sorta-bilaketa erabiliz egin ohi da. Sorta-bilaketan itzulpen partzialen hautagai-zerrenda bat mantentzen da, prozedura iteratibo baten bidez zabaltzen joaten dena. Lehenengo iterazioan hautagai bakarra itzulpen hutsa da eta, algoritmoaren urrats bakoitzean, itzulpen partzial bakoitzeko hautagai berriak sortzen dira jarraipen posible ezberdinekin. Sistema estatistikoek modu lokalean itzultzen dutenez, hautagai bakoitzeko jatorrizko sekuentziako zein token itzuli diren ere gordetzen da. Hautagaiak hedatzerakoan itzuli gabeko jatorrizko sekuentziako sintagmen baliokideak soilik erabiltzen dira, eta prozedura jatorrizko sekuentziako token guztiak itzulita daudenean bukatzen da. Behin hautagai guztiak hedatuta, iterazio bakoitzaren amaieran hedapen onenak soilik mantentzen dira, gainerako hautagaiak baztertuz. Sintagmetan oinarritutako ereduetan, baina, jatorrizko sekuentziaren segmentazio ezberdinak erabil daitezke, eta jatorrizko sekuentziako sintagmak ere edozein ordenatan itzul daitezke. Hori dela eta, hautagaiak baztertzerakoan itzulpen partzialei ereduak esleitutako probabilitateez gain etorkizuneko kostuaren estimazio bat ere erabiltzen da.

Itzulpen automatiko neuronala

Itzulpen automatiko neuronalaren ideia nagusia $p(\mathbf{y}|\mathbf{x})$ modelatzeko muturretik muturrera entrenatutako neurona-sare bat erabiltzea da. Hurbilpen hori aspaldidanik landu izan

bada ere (Allen, 1987; Chrisman, 1991), garai hartan ez zuen arrakasta handirik izan, eta itzulpen automatiko neuronal modernoa azken hamarkada honetan garatu da ikasketa sakonaren loraldiarekin bat etorriz. Hasiera batean itzulpen-eredu neuronalak sistema estatistikoetan txertatu izan ziren (Schwenk et al., 2007; Kalchbrenner and Blunsom, 2013; Cho et al., 2014), kasurik gehienetan bertako sintagmen itzulpen-probabilitateak estimatzeko. Gerora, baina, prozesu osoa neurona-sareen bidez egiten zuten sistemak agertu ziren (Sutskever et al., 2014; Bahdanau et al., 2015). Hurbilpen hori itzulpen automatikoko paradigma nagusi bihurtu da azken urteotan, sistema estatistikoekiko hobekuntza nabarmenak eskuratuz (Barrault et al., 2019).

Itzulpen automatiko neuronalean $p(\mathbf{y}|\mathbf{x})$ katearen erregela aplikatuz faktorizatzen da:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{m} p\left(y_i|y_1^{i-1}, \mathbf{x}\right)$$

Probabilitate hori modelatzeko **kodetzaile-deskodetzaile** deituriko arkitektura orokorra erabiltzen da. *Kodetzaile* deituriko osagaia $\mathbf{x}$ sarrerako sekuentziaren $\mathbf{h} = h_1^n = (h_1, ..., h_n)$ errepresentazioa sortzeaz arduratzen da, $h_i \in \mathbb{R}^d$ bakoitza d dimentsioko bektore bat izanik:

$$\mathbf{h} = \text{enc}(\mathbf{x})$$

Deskodetzaile deituriko osagaiak, berriz, $\mathbf{y}$ sekuentziako token bakoitzaren probabilitatea modelatzen du aurreko token guztietan eta $\mathbf{h}$ errepresentazioan oinarrituz:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{m} p\left(y_i|y_1^{i-1}, \mathbf{h}\right)$$

Nola kodetzailearen hala deskodetzailearen parametroak egiantz handieneko estimazioaren bidez optimizatu ohi dira corpus paralelo baten gainean gradiente jaitsiera estokastikoa erabiliz.

Arkitektura orokor horren baitan hainbat aldaera proposatu izan dira. Lehen hurbilpen modernoek sarrerako sekuentziaren **luzera finkoko errepresentazio** bat zerabilten (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014), $c \in \mathbb{R}^d$ gisara adieraziko duguna. Horretarako, halako sistema gehienek neurona-sare errepikari bat erabiltzen dute kodetzailetzat. Neurona-sare errepikariek $\mathbf{x}$ sarrera sekuentzialki prozesatzen dute, urrats bakoitzean h_i egoera ezkutua eguneratuz x_i tokena eta h_{i-1} aurreko egoera ezkutuaren arabera:

$$h_i = \text{RNN}_{\text{enc}}(x_i, h_{i-1})$$

Oinarrizko neurona-sare errepikariek $h_i = \sigma(Wv_{x_i} + Uh_{i-1} + b)$ hartzen dute (Elman, 1990), non W, U eta b ereduaren parametroak baitira eta v_{x_i} berriz x_i tokenari dagokion bektorea, gainerako parametroekin batera ikasten dena. Praktikan, baina, hainbat

geruzatako LSTM (Hochreiter and Schmidhuber, 1997) edo GRU (Cho et al., 2014) motako sareak erabili ohi dira, informazio-jarioa kontrolatzeko ateak barneratzen dituztenak. Erabiltzen den aldaera edozein izanik ere, sistema horietan h_n azken egoera ezkutua hartu ohi da sarrerako sekuentziaren c errepresentaziotzat:

$$c = h_n$$

Deskodetzailea ere beste neurona-sare errepikari bat izan ohi da, pauso bakoitzean s_i ezkutuko egoera eguneratzen duena s_{i-1} aurreko egoera ezkutuaren, y_{i-1} aurreko tokenaren eta, aukeran, c sarrerako errepresentazioaren arabera:

$$s_i = \mathrm{RNN}_{\mathrm{dec}}\left(y_{i-1}, s_{i-1}, c\right)$$

Aukera bat hasierako egoeratzat $s_0 = c$ errepresentazioa hartu eta $s_i = \mathrm{RNN}_{\mathrm{dec}}\left(y_{i-1}, s_{i-1}\right)$ neurona-sare errepikari arrunt bat erabiltzea da (Sutskever et al., 2014), baina c pausu guztietan erabiltzea ere posible da (Kalchbrenner and Blunsom, 2013). Amaitzeko, s_i egoera ezkutua, y_{i-1} aurreko tokena eta c errepresentazioa hartu, eta y_i tokenari probabilitate-masa bat esleitzen dion funtzio bat aplikatzen da:

$$p\left(y_i | y_1^{i-1}, \mathbf{h}\right) = f\left(y_i | s_i, y_{i-1}, c\right)$$

Horretarako aukerarik errazena s_i bokabularioaren tamainara eramateko transformazio lineal bat aplikatu eta softmax funtzioa aplikatzea da (Sutskever et al., 2014):

$$f\left(y_i | s_i, y_{i-1}, c\right) = \frac{\exp\left(s_i \cdot v_{y_i}\right)}{\sum_{j=1}^{|V|} \exp\left(s_i \cdot v_j\right)}$$

non v_i bokabularioko $i.$ tokenari dagokion bektore bat baita, gainerako parametroekin batera ikasten dena, eta $|V|$ bokabularioaren tamaina. Horren ordez, sarreratzat s_i, y_{i-1} edota c hartzen dituen feedforward neurona-sare bat ere erabil daiteke (Cho et al., 2014).

Oinarrizko hurbilpen horren muga garrantzitsu bat da $\mathbf{x}$ luzera aldakorreko sekuentzia bat errepresentatzeko c luzera finkoko bektore bat erabiltzen duela. Arazo horri aurre egiteko **arreta-mekanismoa** proposatu zen (Bahdanau et al., 2015), egungo sistemen funtsezko osagai bat bihurtu dena. Arreta-mekanismoak q kontsulta-bektore bat, $\mathbf{k} = (k_1, ..., k_L)$ gako-bektoreen sekuentzia bat eta haiei lotutako $\mathbf{v} = (v_1, ..., v_L)$ balio-bektoreen sekuentzia bat hartzen ditu, eta balio-bektore horien batezbesteko haztatu bat itzuli:

$$\mathrm{attn}\left(q, \mathbf{k}, \mathbf{v}\right) = \sum_{i=1}^{L} \alpha_i v_i$$

α_i koefizienteak kalkulatzeko kontsultaren eta gako bakoitzaren arteko antzekotasuna

kalkulatu eta softmax funtzioa aplikatzen da:

$$\alpha_i = \frac{\exp\left(\text{score}\left(q, k_i\right)\right)}{\sum_{j=1}^{L} \exp\left(\text{score}\left(q, k_j\right)\right)}$$

$\text{score}\left(q, k_i\right)$ bektoreen antzekotasuna neurtzeko edozein funtzio izan daiteke, ohiko aukera bat honakoa izanik (Vaswani et al., 2017):

$$\text{score}\left(q, k_i\right) = \frac{q \cdot k_i}{\sqrt{d_k}}$$

non d_k terminoak q eta k_i bektoreen dimentsio kopurua adierazten baitu (alegia, $q, k_i \in \mathbb{R}^{d_k}$). Horren ordez zenbakitzaileko biderkadura eskalarra soilik ere har daiteke (Luong et al., 2015b), bai eta feedforward neurona-sare konplexuago bat erabili ere (Bahdanau et al., 2015).

Arreta-mekanismoak irteerako y_i token bakoitza aurresateko $\mathbf{x}$ sekuentziaren zati esanguratsuenei erreparatzea ahalbidetzen du. Horrela, **arretadun itzultzaile automatiko neuronaletan** h_j bektore bakoitzak x_j tokena bere testuinguruan errepresentatzen du, eta y_i bakoitza aurresateko haien batezbesteko haztatu ezberdin bat erabiltzen da, c bektore finko bat beharrean. Arreta-mekanismoa deskodetzailearen puntu ezberdinetan txerta daiteke. Aukera bat irteerako geruzan egitea da (Luong et al., 2015b):

$$p\left(y_i | y_1^{i-1}, \mathbf{h}\right) = f\left(y_i | s_i, \text{attn}\left(s_i, \mathbf{h}, \mathbf{h}\right)\right)$$

$f\left(\cdot\right)$ feedforward neurona-sare bat izanik[14] eta $s_i = \text{RNN}_{\text{dec}}\left(y_{i-1}, s_{i-1}\right)$ ohiko neurona-sare errepikari batetik datorrelarik. Horren ordez arreta-mekanismoa deskodetzailearen sarrerako geruzan aplikatzea ere posible da, gakotzat s_{i-1} erabiliz (Bahdanau et al., 2015; Wu et al., 2016).

Gerora, $\mathbf{x}$ eta $\mathbf{y}$ arteko dependentziak ez ezik sekuentzia bion barne-dependentziak ere arreta-mekanismoaren bidez modelatzea proposatu zen (Vaswani et al., 2017). Arkitektura hori **transformer** gisara ezagutzen da, eta itzulpen automatiko neuronaleko hurbilpen nagusi bilakatu da (Barrault et al., 2019), neurona-sare errepikarietan oinarritutako aurreko sistemak atzean utziz.[15] Transformer motako kodetzaileek $\mathbf{h}$ bektore-sekuentzia bat hartu eta luzera bereko beste sekuentzia bat itzultzen dute arreta-mekanismoa $\mathbf{h}$ sekuentziari berari aplikatuz:

$$\text{T}_{\text{enc}}\left(\mathbf{h}\right) = \left(f\left(\text{attn}\left(W_q h_i, W_k h_1^n, W_v h_1^n\right)\right)\right)_{i=1}^{n}$$

[14] Zehazki, Luong et al. (2015b) lanean $f\left(y_i | s_i, a_i\right) = \text{softmax}\left(W_1 \tanh\left(W_2 s_i + W_3 a_i\right)\right)$ hartzen dute.

[15] Neurona-sare errepikari eta transformerrez gain, neurona-sare konboluzionaletan oinarritutako sistemak ere badira (Kalchbrenner et al., 2016; Gehring et al., 2017a,b; Wu et al., 2019a), baina lerro hori ez da hainbeste landu orain arte.

$(a_i)_{i=1}^n = (a_1, ..., a_n)$ izanik eta $Wh_1^n = (Wh_1, ..., Wh_n)$. W_q, W_k eta W_v matrizeak sistemaren parametroak dira, eta $f(\cdot)$ feedforward neurona-sare bat.[16] Transformer sistemetan halako hainbat geruza erabili ohi dira, bakoitzaren irteera hurrengoaren sarreratzat erabiliz. Lehen geruzako sarreratzat, berriz, $\mathbf{x}$ sekuentziako token bakoitza eta haien posizioak errepresentatzen dituen $(v_{x_1} + p_1, ..., v_{x_n} + p_n)$ bektore-sekuentzia erabiltzen da.[17] Deskodetzaileak halako beste transformazio bat aplikatzen dio $\mathbf{s}$ bektore-sekuentziari, bi ezberdintasun nagusirekin: (i) arreta-mekanismoko gako- eta balio-bektoretzat s_i kontsulta-bektorearen aurreko s_1^i bektoreak soilik erabiltzen dira, modelatu nahi den $p\left(y_i|y_1^{i-1}, \mathbf{h}\right)$ probabilitatea uneko y_i tokenaren aurreko y_1^{i-1} tokenek soilik baldintzatzen baitute, eta (ii) $\mathbf{x}$ sarrerako sekuentzia aintzat hartzeko arreta-mekanismoa birritan aplikatzen da, $\mathbf{s}$ beraren gainean lehendabizi eta kodetzailetik datorren $\mathbf{h}$ gainean ondoren:

$$\mathrm{T}_{\mathrm{dec}}(\mathbf{s}, \mathbf{h}) = \left(f\left(\mathrm{attn}\left(W_q\,\mathrm{attn}\left(\tilde{W}_q s_i, \tilde{W}_k s_1^i, \tilde{W}_v s_1^i\right), W_k h_1^n, W_v h_1^n\right)\right)\right)_{i=1}^m$$

non W_q, W_k, W_v, $\tilde{W}_q$, $\tilde{W}_k$ eta $\tilde{W}_v$ matrizeak ereduaren parametroak baitira. Kodetzailearen antzera halako hainbat geruza erabiltzen dira, eta $p\left(y_i|y_1^{i-1}, \mathbf{h}\right)$ probabilitateak lortzeko azken geruzaren irteerari transformazio lineal bat eta softmax funtzioa aplikatzen zaizkio. Azkenik, aipatzekoa da Vaswani et al. (2017) lanean azalduriko sistemarekiko 3 hedapen barneratzen dituztela: (i) hainbat arreta-buru konbinatzea, (ii) feedforward nahiz arreta-modulu bakoitzaren ostean geruza-normalizazioa aplikatzea (Ba et al., 2016), eta (iii) modulu horietako bakoitza eta hurrengoaren artean hondar-konexio bat erabiltzea.

Itzulpen automatiko estatistikoan ez bezala, itzulpen automatiko neuronalean ez da bideragarria **bokabulario handiekin** lan egitea. Izan ere, sistema neuronalek irteerako y token posible bakoitza v_y bektore baten bidez errepresentatzen dute, eta eurak gordetzeko memoria ez ezik softmax funtzioa aplikatzen duen azken geruzaren kostua ere bokabularioaren tamainaren arabera hazten da. Hori dela eta, hasierako sistemek maiztasun handieneko 30.000-80.000 hitzetara mugatzen zuten bokabularioa, eta ez ziren gai gainerako hitzak itzultzeko (Sutskever et al., 2014; Bahdanau et al., 2015). Arazo horri aurre egiteko hainbat metodo proposatu izan dira: hitz ezezagunak kopiatzeko mekanismo bat erabiltzea (Luong et al., 2015c), bokabularioz kanpoko hitzak karaktere mailako moduluen bidez itzultzea (Luong and Manning, 2016) edo garrantziaren araberako laginketaren bidez bokabularioaren azpimultzoak erabiltzea (Jean et al., 2015), adibidez. Denborarekin nagusitu den soluzioa, baina, azpihitz mailako bokabularioak erabiltzea da, maiztasun txikiko hitzak hainbat tokenetan segmentatuz. Segmentazio hori egiteko eredu

[16] Zehazki, Vaswani et al. (2017) lanean $f(u) = W_1 \max(0, W_2 u + b_2) + b_1$ hartzen dute.

[17] Arreta-mekanismoa gako-balio bikoteen permutazioekiko inbariantea da eta, ondorioz sekuentziako osagaien posizioa esplizituki errepresentatu behar da haien ordena modelatu ahal izateko. Posizio-bektoreak ikasi egin daitezke, edo kodeketa finkoren bat erabili (Vaswani et al., 2017; Shaw et al., 2018).

ezberdinak daude, erabilienak BPE (Sennrich et al., 2016b), WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016) eta SentencePiece tresnako unigrama-eredua (Kudo, 2018; Kudo and Richardson, 2018) izanik.

Lehen aipatu bezala, bere jatorrizko formulazioan, itzulpen automatiko neuronalak corpus paraleloak soilik erabiltzen ditu $p(y|x)$ zuzenean modelatzen ikasteko. Sistema horiek entrenatzeko **corpus elebakarrak** ere aprobetxatu ahal izateko hainbat proposamen egin izan dira. Aukera bat itzulpen automatiko estatistikoaren antzera hizkuntza-eredu bat barneratzea da, izan zuzenean (Gulcehre et al., 2015) edo kanal zaratatsuaren ereduaren bidez (Yu et al., 2017; Yee et al., 2019; Ng et al., 2019; Yu et al., 2019). Ikasketa dualak, berriz, ziklo-kontsistentzia eta hizkuntza-ereduak uztartzen ditu (He et al., 2016). Teknikarik erabiliena, baina, **atzeranzko itzulpena** delakoa da (Sennrich et al., 2016a; Edunov et al., 2018). Metodo horrek aurkako noranzkoan itzultzen duen itzultzaile neuronal arrunt bat entrenatzen du lehendabizi, eta helburuko hizkuntzako corpus elebakar bat jatorrizko hizkuntzara itzultzeko erabili. Behin hori eginda, jatorrizko corpus paraleloa eta atzeranzko sistemaren bidez sortutako corpus paralelo sintetikoa konbinatzen dira, eta aurreranzko sistema entrenatzeko erabili.

Sistema estatistikoen antzera, sistema neuronaletan ere **deskodeketa** sorta-bilaketaren bidez egin ohi da. Itzulpen automatiko neuronalaren kasuan, baina, itzulpena ez da modu lokalean egiten, y_i token bakoitza sortzeko $p\left(y_i|y_1^{i-1}, \mathbf{h}\right)$ probabilitatea jatorrizko sekuentzia osoak baldintzatzen baitu bere $\mathbf{h}$ errepresentazioaren bidez. Horri esker, ez dago jatorrizko sekuentziako zein zati itzuli diren kontrolatu beharrik, ez eta, horrenbestez, etorkizuneko kostuaren estimaziorik egin beharrik ere. Sorkuntza noiz bukatu erabakitzeko sekuentziaren amaiera adierazten duen token berezi bat erabiltzen da, entrenamendu-corpuseko helburuko hizkuntzako sekuentzia bakoitzaren bukaeran eransten dena.

1.6 Erlazionatutako lana

Atal honetan tesi honekin erlazionatutako lanen literatura-azterketa bat aurkeztuko dugu. 1.6.1 atalean hitz-bektoreen hizkuntza arteko lerrokatzea izango dugu hizpide, eta 1.6.2 atalean itzulpen automatiko gainbegiratu gabea. Kontuan izan behar da ikerketa-arlo horiek oso aktiboak izan direla azken urteotan, eta tesi honetan eginiko ekarpenek ere eragin dute euren bilakaeran. Hori horrela, gai horien bilakaera eta egungo egoera azaltzeaz gain, tesi hau osatzen duten artikuluak euren testuinguruan jartzeko ere balio du atal honek.

1.6.1 Hitz-bektoreen hizkuntza arteko lerrokatzea

Hitz-bektoreen hizkuntza arteko lerrokatzea azken urteotan garatu den ikerketa-gai bat bada ere, lehenago ere baziren hizkuntza ezberdinetako hitzen errepresentazio banatuak

ikasten zituzten metodoak. **Kontaketan oinarritutako hitz-bektoreen garaian**, horretarako hurbilpen nagusia agerkidetza-matrizeko testuingurutzat bi hizkuntzek partekatutako elementuak erabiltzean zetzan. Hala nola, hizkuntza arteko informazio-berreskurapenerako LSI metodoa itzulitako dokumentu-bikoteen gainean aplikatzea proposatu zen (Dumais et al., 1997). Hitz baten ordain jakin batek hitz hori agertzen den dokumentuen itzulpenetan agertzeko joera izango duenez, hitz bientzat antzeko errepresentazioak ikasten ditu metodo horrek. Horri esker, posible da kontsultako hitzak hizkuntza batean egonik beste hizkuntzako dokumentuak bilatzea. Testuingurutzat hitzak darabiltzaten ereduen kasuan, berriz, hizkuntza bien arteko lotura hiztegi elebidunen bidez egin izan da. Horretarako agerkidetza-maiztasunak hizkuntza bakoitzarentzat bere aldetik kalkulatu ohi dira, eta hiztegiaren arabera baliokideak diren testuinguru-hitzak zutabe berean jarri. Teknika hori oso erabilia izan da hiztegi elebidunen indukzioan, hasierako hiztegi batetik hasi eta modu horretara lortutako bektoreak itzulpen-bikote berriak erauzteko erabiliz (Fung and McKeown, 1997; Rapp, 1999; Garera et al., 2009; Laroche and Langlais, 2010). Hitz-bektore prediktiboen helduerarekin lan horiek atzean geratu baziren ere, egungo metodoen oinarrizko hainbat ideiek badituzte aurrekariak garai hartan. Hala nola, autore batzuk hizkuntza ezberdinetako bektore-errepresentazioak azpiespazio komun batera proiektatzeko korrelazio kanonikoaren analisia erabiltzea proposatu zuten (Gaussier et al., 2004; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011), eredu prediktiboen hastapenetan proposatu zen bezala (Faruqui and Dyer, 2014). Era berean, garai hartako *bootstrapping* teknikak (Peirsman and Padó, 2008; Peirsman and Padó, 2010; Vulić and Moens, 2013) egungo metodo gainbegiratu gabeek darabilten autoikasketa iteratiboaren aurrekaritzat jo daitezke, tesi honetako Artetxe et al. (2017) lanean proposatu genuena.

Dena dela, **eredu prediktiboen helduerarekin** lan horiek atzean geratu ziren, eta helburu orokorreko hitz-bektore eleaniztunak ikasteko metodoen belaunaldi berri bat garatu zen. Hasiera batean gehien landu zen hurbilpena eredu elebakarrak hainbat hizkuntzatako hitz-bektoreak batera ikasteko hedatzea izan zen. Horretarako, ohiko termino elebakarraz gain, termino eleaniztun bat erabili ohi da helburu-funtzioan, hizkuntza ezberdinetako hitz baliokideak elkarrengandik hurbil egotera bultzatzen dituena baliabide paraleloren baten arabera. Hurbilpen ezagun bat skip-gram eredua corpus paraleloak erabiltzeko hedatzea da, hitz bakoitzaren hizkuntza bereko testuinguru-hitzez gain harekin lerrokatutako esaldikoak ere ikasketan barneratuz (Gouws et al., 2015; Luong et al., 2015a; Coulmance et al., 2015). Bestelako ereduetan oinarritutako metodoak ere proposatu izan dira (Klementiev et al., 2012b; Kočiský et al., 2014; Chandar A P et al., 2014), bai eta hiztegi elebidunak (Gouws and Søgaard, 2015; Duong et al., 2016) edo dokumentu konparagarriak (Vulić and Moens, 2015, 2016) darabiltzatenak ere. Familia horren inguruko literatura-azterketa sakonago baterako, ikus Ruder et al. (2019) lana.

Denborarekin nagusitu den hurbilpena, baina, hizpide dugun **hitz-bektoreen hizkuntza arteko lerrokatzearena** izan da. Familia horretako metodoek hizkuntza bakoitzeko

hitz-bektoreak modu independentean entrenatzen dituzte corpus elebakarrak erabiliz. Behin hori eginda, transformazio lineal bat ikasten dute hitz-bektore horiek espazio komun batera proiektatzeko. Ohikoena jatorrizko hizkuntzako hitz-bektoreak helburuko hizkuntzara proiektatzea da, baina hizkuntza bakoitzarentzat transformazio lineal bat ikasten duten metodoak ere badira, bai eta bi hizkuntza baino gehiagorekin lan egiten dutenak ere. Ikasketa, berriz, hiztegi elebidun bat erabiliz egin ohi da, modu batera ala bestera bertako hitz-bikoteen arteko antzekotasuna maximizatuz.

Ikerketa-lerro honen inguruko **aurreneko lanek** ataza **erregresio linealeko** problema gisara formulatu zuten, proiektatutako hitz-bektoreen eta haien ordainen arteko distantzia euklidearren karratuen batura minimizatuz. Guztietan lehena Mikolov et al. (2013b) izan zen, jatorrizko hizkuntzako hitz bakoitzaren helburuko hizkuntzako auzokide hurbilena hartuz itzulpen berriak induzi zitezkeela ere erakutsi zuena. Gerora ere, hiztegi elebidunen indukzioa izan da hitz-bektoreen hizkuntza arteko lerrokatzea ebaluatzeko erabili izan den ataza nagusia. Dinu et al. (2015) lanean L2 erregularizazioa barneratu zuten. Shigeto et al. (2015) lanak, berriz, jatorrizko hizkuntzako bektoreak helburuko hizkuntzara proiektatu beharrean, helburuko hizkuntzakoak jatorrizko hizkuntzara lerrokatzea egokiagoa dela erakutsi zuen.

Erregresioaz gain, hitz-bektoreen lerrokatzearen hastapenetan proposatutako beste hurbilpen bat **korrelazio kanonikoaren analisia** izan zen (Faruqui and Dyer, 2014). Metodo horrek hizkuntza biak espazio komun batera proiektatzen ditu, dimentsio bakoitzeko hizkuntza bien arteko korrelazioa maximizatuz, beti ere dimentsiook aurrekoekiko korrelazio gabeak izatearen murriztapenarekin. Korrelazioa aldagai bakoitzaren bariantzarekiko inbariantea denez, dimentsio guztiek bariantza bera izateko murriztapen gehigarria erabiltzen da.

Handik gutxira proposatutako beste aldaera bat, gerora funtsezkoa bihurtu dena, **ortogonaltasunarena** izan zen. Transformazio ortogonalak transformazio linealen azpimultzo bat dira, jatorrizko espazioko biderkadura eskalarra aldatzen ez dutenak. Halako transformazioei dagozkien Q matrize ortogonalek $Q^TQ = QQ^T = I$ propietatea betetzen dute, I identitate-matrizea izanik eta, horrenbestez, $Q^{-1} = Q^T$ (alegia, matrize ortogonal baten alderantzizkoa bere iraulia da). Bektore-bikote batzuen arteko distantzia euklidearren karratuen batura minimizatzen duen transformazio ortogonal optimoa bilatzeari *Procrustesen problema ortogonal* deritzo, eta balio singularren deskonposaketa bidezko soluzio itxi bat du (Schönemann, 1966). Ortogonaltasunaren murriztapena hainbat ikuspuntutatik motibatu izan da hitz-bektoreen lerrokatzean. Xing et al. (2015) lanean hitz-bektoreak ikasteko erabiltzen den helburu-funtzioa (biderkadura eskalarren arabera definitua), lerrokatzea ikasteko erabiltzen dena (distantzia euklidearren karratuak) eta ebaluazio garaian erabiltzen den antzekotasun-neurria (kosinua) ez direla kontsistenteak argudiatu zuten. Arazo horri aurre egiteko hitz-bektoreen eta lerrokatzearen ikasketan ere kosinu-antzekotasuna erabiltzea proposatu zuten, hitz-bektoreak unitate bateko luzera izatera behartuz. Lerrokatzearen ikasketan luzera-normalizazioa mantendu eta kosinu-

antzekotasuna maximizatzeko, ortogonalitatearen murriztapena erabiltzea proposatu zuten. Gure Artetxe et al. (2016) lanean, berriz, ortogonaltasuna inbariantza elebakarra bermatzeko murriztapentzat aurkeztu genuen, murriztapenik gabeko transformazio linealak ataza elebakarretan kaltegarriak izan daitezkeela erakutsiz. Antzeko motibazio bat jarraituz, Zhang et al. (2016) lanean ortogonaltasuna datu gutxiko transferentzia-ikasketako ataza baterako aplikatu zuten. Smith et al. (2017) lanean, azkenik, lerrokatzea bere buruarekiko kontsistentea izan dadin—alegia, proiektatzen diren hitz-bektoreak jatorrizko hizkuntzakoak ala helburuko hizkuntzakoak izan euren arteko antzekotasunak berdinak izan daitezen—transformazio lineala ortogonala behar dela izan erakutsi zuten.

Hasiera batean erregresio lineala, korrelazio kanonikoaren analisia eta metodo ortogonalak hurbilpen ezberdintzat proposatu izan baziren ere, tesi honetan aurkezten ditugun Artetxe et al. (2016) eta, batez ere, Artetxe et al. (2018a) lanetan familia horiek guztiak **orokortzen dituen marko bat** proposatu genuen. Marko horretako transformazio nagusia—hizkuntzen arteko lerrokatzeaz arduratzen dena—ortogonala da, eta aurreko metodoen arteko ezberdintasunak hautazko beste urrats batzuei dagozkie: aurreprozesua, zuritzea, birpisaketa, deszuritzea eta dimentsionaltasun-murrizketa. Aurreko metodoak orokortzeaz gain, urrats bakoitzaren eragina enpirikoki aztertu genuen, eta modu horretara ikasitakoari esker aldaera hobe bat proposatu.

Orain arte aipaturiko lan guztiek hiztegi elebidun bat erabiltzen dute ikasketarako, datu-multzo estandar gehienetan 5.000 sarrera inguru izan ohi dituena. Hori horrela, arlo honetako ikerketa-lerro garrantzitsu bat lerrokatzea ikasteko beharrezkoa den **gainbegirapena murriztea** izan da. Kategoria gramatikalen etiketatzearen hizkuntza arteko transferentziaren inguruko Zhang et al. (2016) lanean 10 hitz-bikote soilik erabili zituzten bi hizkuntzaren arteko transformazio ortogonala ikasteko. Artetxe et al. (2017) lanean erakutsi genuenez, baina, hurbilpen horrek emaitza kaskarrak ematen ditu lerrokatze finagoak eskatzen dituzten atazetan. Vulić and Korhonen (2016) lanean, berriz, ikasketa-hiztegiaren jatorriak, tamainak eta fidagarritasunak lerrokatzean daukan eragina aztertu zuten. Analisi horretan oinarrituz, entrenamendu-hiztegia automatikoki sortzeko dokumentu mailako lerrokatzeak soilik behar dituen Vulić and Moens (2016) laneko metodoa erabiltzea proposatu zuten. Horren ordez bi hizkuntzetan berdin idazten diren hitzak hartzea ere aztertu zuten, baina hurbilpen horrek emaitza okerragoak ematen dituela erakutsi zuten. Aurrekoari lotuta, Yehezkel Lubin et al. (2019) lanean entrenamendu-hiztegian sarrera okerrak egotearen eragina aztertu zuten, eta zarata mota horrekiko sendoagoa den metodo bat diseinatu.

Tesi honetan aurkezten dugun Artetxe et al. (2017) lanean, berriz, entrenamendu-hiztegiarekiko menpekotasuna arintzeko **autoikasketan** oinarritutako metodo bat proposatu genuen. Zehatzagoak izanez, gure metodoak hitz-bektoreen lerrokatzea eta hiztegi elebidunen indukzioa txandakatzen ditu modu iteratiboan: hasierako hiztegia erabiliz bi hizkuntzetako hitz-bektoreak lerrokatzen ditu lehendabizi transformazio ortogonal bat erabiliz, lerrokatutako hitz-bektoreak hiztegi berri bat induzitzeko erabili, eta hiztegi

berri horren bidez hitz-bektoreak berriro lerrokatu, urrats horiek prozesuak konbergitu arte errepikatuz. Lan horretan erakutsi genuenez, metodo horrek jatorrizko hizkuntzako hitz bakoitzaren eta harengandik hurbilen dagoen helburuko hizkuntzakoaren arteko batez besteko distantzia euklidearraren karratua minimizatzen du modu inplizituan. Optimizazio-helburu hori abiapuntuko entrenamendu-hiztegiarekiko independentea da, eta proposatutako metodoak haren optimo lokal batera konbergitzen du. Enpirikoki, aurreko metodoek 5.000 sarrerako hiztegiekin lortzen zituzten pareko emaitzak lortu genituen 25 hitz-bikote edo zenbakien zerrenda batetik soilik abiatuta. Hasierako hiztegiak gutxieneko kalitate hori izatea, baina, beharrezkoa da, osterantzean metodoa optimo lokal kaskarretan trabatuta geratzen baita.

Aldi berean argitaratutako Hauer et al. (2017) lanean ere gurearen antzeko *bootstrapping* metodo bat proposatu zuten. Kasu horretan, baina, entrenamendu-hiztegia modu inkrementalean zabaltzen dute, iterazio bakoitzean itzulpen fidagarrienak gehituz (baldin eta hitz horiek dagoeneko hiztegian ez bazeuden). Hori dela eta, metodo horrek ez du gurearen berme teorikorik, eta abiapuntuko hiztegiarekiko menpekotasun handiagoa izatea espero daiteke. Horretaz gain, lan horretan edizio-distantzia eta hitzen maiztasuna uztartzen dituen heuristiko bat erabiltzen dute abiapuntuko hiztegia modu automatikoan sortzeko. Era berean, ez dute ortogonaltasunaren murriztapena erabiltzen, eta noranzko bakoitzeko lerrokatze ezberdin bat ikasten dute, hiztegia induzitzerakoan bi noranzkoetako antzekotasunak konbinatuz.

Gerora argitaratutako Ruder et al. (2018) lanean, berriz, gure Artetxe et al. (2017) lana aldagai ezkutuko eredu probabilistiko gisara berrinterpretatu zuten, guk proposatutako autoikasketa iteratiboak hura optimizatzeko itxaropen-maximizazio algoritmotzat jokatuko lukeelarik. Zehatzagoak izanez, hiztegiak grafo gisara errepresentatzen dituzte, nodoak hizkuntza bietako hitzak eta ertzak hiztegiko sarrerak izanik. Euren ereduko aldagai ezkutua grafo horretako ertz-multzoa da, haren gaineko a priori banaketa ezberdinak erabiltzea ahalbidetzen duena. Guk proposatutako ereduak IBM 1 ereduaren baliokidea den lerrokatze-banaketa bat erabiltzen duela erakutsi zuten, 1-N motako loturak onartzen dituena. Horren ordez, grafo bipartigarrien gaineko banaketa uniforme bat erabiltzea proposatu zuten, 1-1 motako loturak soilik onartzen dituena.

Gainbegirapena murrizteko ikerketa-lerroa muturrera eramanez, hainbat autore hitz-bektoreen hizkuntza arteko lerrokatzea modu **erabat gainbegiratu gabean** ikasten ahalegindu ziren, inolako entrenamendu-hiztegirik erabili gabe. Guztietan lehena Miceli Barone (2016) izan zen. Lan horretan hizkuntza ezberdinen egitura semantikoa corpus elebakarretako hitz-banaketak aztertuz errepresentazio komunak ikasteko behar beste antzekoa izan beharko litzatekeela planteatu zuten. Hipotesi hori egiaztatzeko, ikasketa antagonikoan oinarritutako sistema bat proposatu zuten, bi osagai dituena: (i) *sortzailea*, jatorrizko hizkuntzako hitz-bektore bat hartu eta helburuko hizkuntzako espaziora eramateaz arduratzen dena, eta (ii) *diskriminatzailea*, hitz-bektore jakin bat helburuko hizkuntzakoa den ala kodetzaileak transformatutako jatorrizko hizkuntzakoa

den bereizteaz arduratzen dena. Diskriminatzailea helburuko hizkuntzako hitz-bektoreei ahalik eta probabilitate-masa handiena esleitzeko entrenatzen da, eta sortzailea, berriz, diskriminatzaileak haren irteerari ahalik eta probabilitate-masa handiena eslei diezaion. Bi azpisareen arteko lehia horrek aurrera egin ahala, diskriminatzaileak hitz-bektoreen jatorria hobeto bereizten ikasten du, eta sortzailea transformazio hobeak ikastera bultzatzen du horrek, diskriminatzaileak bere irteera helburuko hizkuntzako hitz-bektoreekin nahas dezan. Ondo bidean, prozesuaren bukaeran kodetzaileak transformatutako jatorrizko hizkuntzako hitz-bektoreak eta helburuko hizkuntzakoak bereizezinak izango dira, bi hizkuntzetako hitz-bektoreen lerrokatze egoki bat litzatekeena. Oinarrizko hurbilpen horretan, baina, sortzaileak kolapsatu eta jatorrizko hizkuntzako hitz-bektore guztiak helburuko hizkuntzako puntu gutxi batzuetara eramateko joera duela ikusi zuten. Horri aurre egiteko, autokodetzaile antagonikoak erabiltzea proposatu zuten (Makhzani et al., 2016), *deskodetzaile* deituriko hirugarren osagai bat erabiltzen dutena. Deskodetzailea sortzaileak transformatutako hitz-bektore bat emanik jatorrizko hitz-bektorea berreskuratzeaz arduratzen da. Osagai horrekin batera sortzailea ere entrenatzen da, transformatutako hitz-bektoreak jatorrizkoen informazioa mantentzera bultzatuz. Hizpide dugun artikuluak sistema hori hizkuntza bietako nolabaiteko informazio semantikoren bat lerrokatzeko gai zela erakutsi zuen, baina autorearen beraren hitzetan proposaturiko metodoa ez zen lehiakorra beste errepresentazio eleaniztun batzuen ondoan. Ikasketa antagonikoaren ohiko ezegonkortasun arazoak medio, autoreak galdera irekitzat utzi zuen ea bide hori emankorragoa izan zitekeen, ala paradigma horren funtsezko muga batera iritsi ote zen. Ondorengo lanek erantzun zuzena aurrenekoa zela erakutsiko zuten, oinarrizko hurbilpen hori finduz emaitza positiboak eskuratuko baitzituzten.

Zhang et al. (2017a) lanean diskriminatzailea erregularizatzeko hainbat teknika probatu zituzten. Era berean, entrenamendua ez zela konbergentea ikusi zuten, bukaerako ereduak emaitza kaskarrak ematen baitzituen, eta eredurik onena automatikoki aukeratzeko irizpidetzat sortzailearen galera erabiltzea proposatu zuten. Horretaz gain, sistemaren hiperparametro ezberdinak ere findu zituzten, bai eta oinarrizko arkitekturaren aldaera ezberdinak probatu ere, emaitzarik onenak autokodetzaile antagonikoarekin eskuratu bazituzten ere. Lan hori emaitza positiboak erakusten lehena izan zen, baina nahiko baldintza berezietan: ez zituzten datu-multzo estandarrak erabili, esperimentuak eskala txikian egin zituzten, dimentsionaltasun eta bokabulario-tamaina txikiko hitz-bektoreak erabiliz, eta erreferentziatzat erabili zuten sistema gainbegiratua oso ahula zen, 50 edo 100 hitz-bikoterekin soilik entrenatua normalean erabili ohi diren 5.000 bikoteren ordez.

Eskala handian eta baldintza estandarretan emaitza sendoak erakusten lehenak, berriz, Lample et al. (2018b) izan ziren. Lan horretan ez zuten deskodetzailerik erabili, eta horren ordez entrenamenduan zehar sortzaileko transformazio-matrizea matrize ortogonal batetik hurbil manten zedin eguneratzea proposatu zuten, transformazio ortogonalak

alderanzgarriak baitira definizioz.[18] Horretaz gain, gainbegiratu gabeko balidazio-irizpide modura induzitutako hiztegiko batez besteko kosinu-antzekotasuna hartzea proposatu zuten. Lan horren beste gakoetako bat ikasketa antagonikoa eta guk proposatutako autoikasketa iteratiboa uztartzea izan zen. Izan ere, behin ikasketa antagonikoa bukatuta, lerrokatutako hitz-bektoreekin hiztegi elebidun bat induzitu eta autoikasketa iteratiboa hasieratzeko erabiltzea proposatu zuten. Hori horrela, bukaerako lerrokatzea autoikasketaren bidez lortzen da, ikasketa antagonikoaren funtzioa abiapuntuko hiztegia modu gainbegiratu gabean lortzea izanik.

Tesi honetan aurkezten dugun Artetxe et al. (2018b) lanean autoikasketa metodoa hobetu eta, ikasketa antagonikoa erabili beharrean, hasierako hiztegia modu gainbegiratu gabean sortzeko hurbilpen sinpleago bat proposatu genuen. Gure metodoa hitz bakoitzak hizkuntza bereko gainerako hitzekiko duen antzekotasun-banaketan oinarritzen da, eta antzeko banaketa duten hitz-bikoteak identifikatzen ditu hasierako hiztegia osatzeko. Horretaz gain, autoikasketa metodoa bera ere hobetu genuen, hiztegia modu estokastikoan induzituz, bi noranzkoetako itzulpenak konbinatuz, eta maiztasun gutxiko hitzak baztertuz, besteak beste. Era berean, aurrez aipaturiko Zhang et al. (2017a) eta Lample et al. (2018b) sistemek baldintza zailagoetan huts egiten zutela erakutsi genuen, metodo gainbegiratu gabeen sendotasun eta egonkortasunaren arazoa erabat ebatzi gabe zegoela erakutsiz. Gure sistemak, berriz, probatutako datu-multzo guztietan emaitza positiboak eskuratu zituen, aurreko metodoek baino emaitza hobeak lortuz.

Metodo gainbegiratu gabeen arrakastak oihartzun handia izan zuen eta, denbora gutxian, gai horren inguruko **lan mordoa** argitaratu ziren. Yang et al. (2018a) lanean guk proposaturiko hasieraketa gainbegiratu gabea erabili zuten, bi hizkuntzetako hitz-bektoreak lerrokatzeko *batezbestekoen alde maximoa* deituriko neurri bat minimizatu, eta soluzio hori are gehiago fintzeko autoikasketa iteratiboa aplikatu. Hoshen and Wolf (2018) lanak ere gure autoikasketaren antzeko algoritmo iteratibo bat proposatu zuen, hasieraketarako osagai nagusien analisia erabiltzen duena. Ezegonkortasunaren arazoari aurre egiteko, 500 berrekite egin zituzten, eta irizpide gainbegiratu gabe baten arabera exekuziorik onena aukeratu. Zhang et al. (2017b) lanean proiektatutako jatorrizko hizkuntzako hitz-bektoreen eta helburuko hizkuntzakoen arteko Wasserstein distantzia minimizatzeko bi metodo proposatu zituzten: (i) Wasserstein sare antagoniko sortzaileak erabiltzea, zeinetan diskriminatzaileak Wasserstein distantzia estimatzen baitu, eta (ii) autoikasketa iteratiboaren antzera proiekzioaren ikasketa eta hitzen esleipena txandakatzea. Esleipen optimoa kalkulatzeko garraio optimoko problema bat ebaztea eskatzen du Wasserstein distantziak, eta Sinkhorn distantzian oinarritutako hurbilpen bat erabili zuten horretarako (Cuturi, 2013). Grave et al. (2019) lanean ere antzeko formulazio bat jarraitu zuten, baina proiekzioa eta hitzen esleipena txandaka optimizatu beharrean

[18] Aipatzeko da Zhang et al. (2017a) lanean ere aldaera hori probatu zutela, kasu horretan murriztapena bermatzeko matrize ortogonalen parametrizazio bat erabiliz. Dena dela, euren kasuan autokodetzaile antagonikoekin emaitza hobeak eskuratu zituzten.

prozedura estokastiko bat proposatu zuten, hasieraketarako problemaren erlaxazio konbexu bat erabiliz. Xu et al. (2018) lanean, berriz, hizkuntza bietako hitz-bektoreak proiektatzeko metodo bat proposatu zuten, optimizazio-helburutzat Sinkhorn distantzia eta ziklo-kontsistentzia uztartzen zituena. Hasieraketarako, Wasserstein sare antagoniko sortzaile bat erabili zuten. Mukherjee et al. (2018) lanean, bestalde, lerrokatutako hitz-bektoreen arteko dependentzia estatistikoa maximizatzea proposatu zuten, hura neurtzeko elkarrekiko informazioaren aldaera koadratiko bat erabiliz. Aurreko hurbilpen askok bezala, euren ikasketa-algoritmoak ere transformazioaren ikasketa eta hiztegi elebidunaren indukzioa txandakatzen ditu modu iteratiboan, eta ziklo-kontsistentzia ere barneratzen du erregularizatzailetzat. Dou et al. (2018) lanean autokodetzaile bariazionaletan inspiratutako metodo bat proposatu zuten. Euren ereduak jatorrizko eta helburuko hitz-bektoreak ezkutuko aldagai berberak sortzen dituela suposatzen du. Hizkuntza bietako hitz-bektoreak ezkutuko espazio horretara eramateko kodetzaile bana ikasten dute, bai eta bertatik jatorrizko hitz-bektoreak berreskuratzeko deskodetzaile bana ere. Autokodetzaile bariazionalen antzera aldagai ezkutua banaketa aurrezarri bat izatera behartu beharrean, hizkuntza bietako aldagai ezkutuek banaketa bera izatea soilik eskatzen dute. Horretarako diskriminatzaile bat erabili zuten, ikasketa antagonikoa erabiliz entrenatu zutena. Mohiuddin and Joty (2019) lanean aurrekoaren antzeko arkitektura bat proposatu zuten, baina hizkuntza biak aldagai ezkutu bakarrera proiektatu beharrean, hizkuntza bietako aldagai ezkutuen arteko transformazio lineal bana ikastea proposatu zuten. Aldarmaki et al. (2018) lanean hitz-bektoreak proiektatu gabe hiztegi elebidunak induzitzeko metodo bat proposatu zuten. Horretarako, jatorrizko hizkuntzako edozein bi hitzen arteko distantzia eta haien helburuko hizkuntzako ordainen artekoa ahalik eta antzekoenak izatea bilatzen dute, distantziok jatorrizko hitz-bektoreen gainean kalkulatzen dituztelarik. Irizpide horren araberako hiztegi optimoa bilatzeko algoritmo iteratibo bat proposatu zuten eta, behin ikasketa amaituta, induzitutako hiztegia hitz-bektoreak lerrokatzeko erabili, metodo gainbegiratu arrunt bat erabiliz. Alvarez-Melis and Jaakkola (2018) lanean antzeko ideia bat landu zuten, hiztegi elebidunen indukzioa garraio optimoko problema gisara formulatuz. Horretarako Gromov-Wasserstein distantzian oinarritu ziren, espazio ezberdinetako laginak alderatu beharrean espazio metrikoak eurak zuzenean alderatzen dituena. Horrela, euren metodoak hizkuntza bakoitzeko antzekotasun-matrizearen gainean egiten du lan, eta Sinkhorn distantzian oinarrituriko hurbilpena erabili esleipen-problema ebazteko. Bestalde, Hartmann et al. (2019) lanean aztertu berri ditugun hainbat hasieraketa-metodo alderatu zituzten. Emaitzarik onenak sare antagoniko sortzaile arruntekin lortu zituzten, ezegonkortasunaren arazoari aurre egiteko irizpide gainbegiratu gabe bat baliatuz eta hasierako soluzio hori fintzeko guk proposaturiko autoikasketa iteratibo estokastikoa erabiliz.

Ikusi dugunez, hiztegi elebidunen indukzioak lotura estua du hitz-bektoreen hizkuntza arteko lerrokatzearekin. Izan ere, lerrokatze-metodo askoren funtsezko osagai bat izateaz gain, eurak ebaluatzeko erabili izan den ataza nagusia—lan gehien-gehienetan, bakarra—

ere bada. Horretarako oinarrizko hurbilpena kosinu-antzekotasunaren arabera jatorrizko hizkuntzako hitz bakoitzaren helburuko hizkuntzako auzokide hurbilena hartzea da. Dinu et al. (2015) lanean, baina, hurbilpen horrek ***hubness*** delako arazoa daukala erakutsi zuten. Fenomeno horrek puntu gutxi batzuk—*hub* deritzenak—puntu askoren auzokide hurbilenak izatea eragiten du, eta dimentsionaltasun handiko bektore-espazioen berezko propietate bat da (Radovanović et al., 2010a,b). Horren ondorioz, helburuko hizkuntzako hitz gutxi batzuk jatorrizko hizkuntzako hitz askoren itzulpentzat agertzeko joera dute induzitutako hiztegian, nahiz eta desegokiak izan. Arazo horri aurre egiteko hainbat teknika proposatu izan dira. Dinu et al. (2015) lanean bertan jatorrizko hizkuntzako hitz jakin bat itzultzeko helburuko hizkuntzako hitz bakoitzaren auzokide hurbilenen zerrendan daukan posizioari erreparatzea proposatu zuten, jatorrizko hitz hori posizio altuenean daukan hautagaia hartuz eta, berdinketen kasuan, kosinu-antzekotasun altuena duena. Smith et al. (2017) lanean, berriz, hautagaiak puntuatzeko softmax funtzioa aurkako noranzkoan aplikatzea proposatu zuten, tenperatura kontrolatzeko hiperparametro bat erabiliz. Huang et al. (2019) lanean teknika horren orokortze bat proposatu zuten, emaitza hobeak lortzen dituena. Lample et al. (2018b) lanean CSLS neurria proposatu zuten, bi bektoreen arteko kosinu-antzekotasunaren eta bektore horiek beste hizkuntzako k auzokide hurbilenekin duten batezbesteko kosinu-antzekotasunaren arteko aldea hartzen duena. Sinplea izan arren, teknika horrek emaitza onak eskuratzen ditu, eta gerora argitaratutako lan gehienek erabili izan dute. Joulin et al. (2018) lanean, berriz, hitz-bektoreak lerrokatzeko transformazio lineala ikasterakoan CSLS neurria bera optimizatzea proposatu zuten. Horren aurretik ere, hiztegi-indukzioan beharrean transformazio linealaren ikasketan bertan hubness-aren arazoa arintzeko metodoak landu izan ziren. Horrela, Shigeto et al. (2015) lanean erregresio lineala aurkako noranzkoan egitea lagungarria dela erakutsi zuten, eta Lazaridou et al. (2015) lanean, berriz, marjina maximoko optimizazio-helburu bat erabiltzea proposatu zuten.

Orain arte aipaturiko lanek bi hizkuntzatako hitz-bektoreak soilik lerrokatzen bazituzten ere, **hainbat hizkuntza lerrokatzeko** metodoak ere proposatu izan dira. Oinarrizko hurbilpen bat hizkuntza bat—normalean ingelesa—pibotetzat hartu eta gainerako hizkuntzak bertara proiektatzea da. Ammar et al. (2016) lanean lehen aipatu dugun korrelazio kanonikoaren analisian oinarritutako metodo gainbegiratuaren aldaera bat erabili zuten horretarako. Anastasopoulos and Neubig (2019) lanean pibote-hizkuntzaren eragina aztertu zuten, eta pibote optimoa hizkuntza-bikotearen araberakoa dela ondorioztatu. Alaux et al. (2019) lanean pibotearen hurbilenak zeharkako itzulpena (pibotea ez diren hizkuntzen artekoa) kaltetzen duela erakutsi zuten, eta gainerako hizkuntza-bikoteak ere aintzat hartzen dituen hurbilpen gainbegiratu gabe bat proposatu. Euren metodoak lehen ikusi ditugun hainbat teknika uztartzen ditu: oinarrizko algoritmoa Grave et al. (2019) lanean oinarritzen da, baina Alvarez-Melis and Jaakkola (2018) lanaren antzera Gromov-Wasserstein probleman oinarritutako hasieraketa bat erabiltzen dute, eta azken iterazioetan Joulin et al. (2018) laneko helburu-funtzio bera optimizatu.

Chen and Cardie (2018) lanean ere hizkuntza-konbinazio guztiak aintzat hartzen dituen hurbilpen gainbegiratu gabe bat proposatu zuten, baina euren kasuan Lample et al. (2018b) laneko metodoa izan zen hedatu zutena. Heyman et al. (2019) lanean, berriz, bi hizkuntzarekin hasi eta hizkuntza berriak inkrementalki gehitzea proposatu zuten. Horretarako, gure Artetxe et al. (2018b) laneko metodo gainbegiratu gabearen hedapen bat erabili zuten. Kementchedjhieva et al. (2018) lanean Procrustesen analisi orokortua (Gower, 1975) erabiltzea proposatu zuten, arestian aipaturiko metodo ortogonalak hainbat bektore-espazio lerrokatzeko hedatzen dituena. Nakashole and Flauger (2017) lanean entrenamendu-hiztegi txikiko hizkuntza-bikoteak lerrokatzeko metodo gainbegiratu bat proposatu zuten, bien arteko zubi-lanak egiteko pibote-hizkuntzak erabiltzen dituena. Jawanpuria et al. (2019) lanean, bestalde, jatorrizko eta helburuko hizkuntzako hitz-bektoreak espazio komun batera eramateko transformazio ortogonal bana ez ezik, espazio horretako antzekotasun-neurria—Mahalanabis metrika gisara definitua—ere modu gainbegiratuan ikastea proposatu zuten, eta hurbilpen hori hainbat hizkuntzarekin lan egiteko orokortu.

Arestian aipatu bezala, lan gehien-gehienek hiztegi elebidunen indukzioa erabiltzen dute ebaluazio-ataza bakartzat. Hainbat autorek **ebaluazio-protokolo horren zenbait arazo** identifikatu dituzte. Alde batetik, ebaluazio-hiztegi gehienak automatikoki sortuak izan dira, eta haien kalitatea zalantzan jarri izan da. Kementchedjhieva et al. (2019) lanean, zehazki, MUSE datu-multzo ezaguneko hiztegi batzuk aztertu zituzten (Lample et al., 2018b), eta bi arazo identifikatu: (i) kategoria gramatikalen banaketa ez da adierazgarria, izen berezien proportzioa oso altua izanik, eta (ii) erreferentziazko itzulpenetan hutsuneak daude. Arazo biek sistema ezberdinen ebaluazioan eragin nabarmena izan dezaketela erakutsi zuten. Braune et al. (2018) lanean, berriz, ebaluazio-hiztegi gehienak domeinu orokorrekoak eta maiztasun altuko hitzez osatuak direla argudiatu zuten,[19] eta hitz arraro nahiz domeinuz kanpokoak itzultzerakoan emaitza nabarmenki kaskarragoak lortzen zirela erakutsi. Baldintza zail horietan emaitza hobeak lortzeko karaktere mailako informazioa eta edizio-distantzia erabiltzea ere proposatu zuten. Riley and Gildea (2018) lanean ere, informazio ortografikoa erabiltzeko antzeko proposamen bat egin zuten. Horretaz gain, datu-multzo gehienak morfologia kontrolatu gabe sortu ziren, eta horri lotutako bi arazo identifikatu zituzten Czarnowska et al. (2019) lanean: (i) lexema arruntenen kasuan ere forma flexionatu gutxi batzuk soilik jasotzen dituzte hiztegiek, eta (ii) ebaluazio-hiztegiko lema batzuk entrenamendu-hiztegian ere agertzen dira forma flexionatu ezberdinekin. Hori ikusirik, lexema bakoitzaren forma flexionatu gehienak biltzen dituzten hiztegiak sortu zituzten, entrenamendu eta ebaluaziorako azpimultzoek

[19] Azken puntuari buruz, aipatzekoa da MUSE datu-multzoak (Lample et al., 2018b) maiztasun altueneko 5.000 hitzak hartzen dituela entrenamendurako eta hurrengo 1.500 hitzak ebaluaziorako. Dinu et al. (2015) lanean proposatu eta guk (Artetxe et al., 2017, 2018a) hizkuntza gehiagotara zabalduriko ebaluazio-hiztegiek, berriz, uniformeki banatutako bost maiztasun-tarte hartzen dituzte beregain, maiztasunaren faktorea behar bezala kontrolatuz. Bi kasuetan, bokabularioa 200.000 hitzetakoa da.

ez zutela lemarik partekatzen bermatuz. Datu-multzo hori egungo ereduek morfologiaren ikuspuntutik duten orokortze-gaitasuna aztertzeko erabili zuten. Bestalde, Bakarov et al. (2018) lanean ebaluazio-ataza ezberdinen arteko korrelazio falta aztertu zuten. Fujinuma et al. (2019) lanean grafoen modularitatean oinarritutako ebaluazio-neurri intrintseko gainbegiratu gabe bat proposatu zuten, eta hiru atazarekin korrelazio sendoa zeukala erakutsi. Glavaš et al. (2019) lanean analisi zabalago bat egin zuten eta, neurri handi batean, eredu ezberdinen kalitatea atazaren araberakoa dela ondorioztatu. Horrekin batera, hitz-bektoreen hizkuntza arteko lerrokatze-metodoak hiztegi elebidunen indukziorako neurrira diseinatzea beste atazetarako kaltegarria izan daitekeela erakutsi zuten. Tesi honetan aurkezten dugun Artetxe et al. (2019a) lanean aurreko hori osatu genuen, eta itzulpen automatiko gainbegiratu gabeko teknikak erabilita hiztegi elebidunen indukzioan emaitza hobeak lor daitezkeela erakutsi. Hortaz, Glavaš et al. (2019) lanean hiztegi-indukzioan soilik pentsatzea beste atazetarako desegokia izan daitekeela erakutsi bazuten, lerrokatze-metodoetan soilik pentsatzea hiztegi-indukziorako kaltegarria izan daitekeela erakutsi genuen guk, arlo honetako lan gehienen planteamendu orokorra berrikustea eskatzen duena.

Bestalde, paradigma honen **muga posibleek** ere zeresan ugari eman dute. Alde batetik, zalantzan jarri izan da transformazio linealak hizkuntza ezberdinetako hitz-bektoreak zehaztasunez lerrokatzeko behar beste adierazkorrak direnik. Nakashole and Flauger (2018) lanean bektore-espazioko azpieremu ezberdinetarako transformazio lineal lokalak ikasi zituzten, eta azpieremuak elkarrengandik geroz eta urrunago egon haientzako ikasitako transformazioak orduan eta ezberdinagoak zirela ikusi zuten. Hori horrela, maila globalean bi hizkuntzen arteko egiazko erlazioa ez dela lineala ondorioztatu zuten, maila lokalean transformazio linealen bidez hurbil badaiteke ere. Ideia horretan sakonduz, azpieremuen nozioa barneratzen duen lerrokatze-metodo bat proposatu zuten Nakashole (2018) lanean. Doval et al. (2018) lanean, berriz, lerrokatzea fintzeko bigarren transformazio bat ikastea proposatu zuten, entrenamendu-hiztegiko hitz bakoitza dagokion bektorearen eta bere ordainaren arteko batezbestekora hurbiltzen duena. Horretaz gain, transformazio ez-linealak ikasteko korrelazio kanonikoaren analisiaren aldaera ezberdinak erabiltzea ere proposatu izan da: Lu et al. (2015) lanean korrelazio kanonikoaren analisi sakona (Andrew et al., 2013) erabili zuten, eta Zhao and Gilman (2020) lanean, berriz, kernel bidezko korrelazio kanonikoaren analisia (Lai and Fyfe, 2000).

Linealtasunari estuki lotuta, hizkuntza ezberdinetako hitz-bektoreak modu gainbegiratu gabean lerrokatu ahal izateko haien egiturak antzekoak izatea ezinbestekoa dela esan izan da, **isomorfismo edo isometriaren** kontzeptua erabiliz formalizatu izan dena, eta premisa horren muga posibleak ere aztertu izan dira. Søgaard et al. (2018) lanean hizkuntza ezberdinetako hitz-bektoreen isomorfismo-maila neurtzeko neurri bat proposatu zuten, haien auzokide hurbilenen grafoen laplacetarren balio propioen antzekotasunean oinarritzen dena. Neurri horrek Lample et al. (2018b) laneko metodo gainbegiratu gabeak hizkuntza-bikote ezberdinentzat lorturiko emaitzekin korrelazio estua zeukala

erakutsi zuten. Horretaz gain, metodo horrek urruneko hizkuntza-bikote batzuekin, domeinu ezberdineko corpusekin eta algoritmo ezberdinekin ikasitako hitz-bektoreekin erabat huts egiten zuela erakutsi zuten. Autoikasketa algoritmo iteratiboa hizkuntza bietan berdin idatzitako hitzekin hasieratuta, berriz, emaitza sendoagoak lortu zituzten. Hartmann et al. (2018) lanean Lample et al. (2018b) laneko metodo gainbegiratu gabea algoritmo ezberdinekin ikasitako ingelesezko hitz-bektoreak elkarren artean lerrokatzeko ere ez zela gai erakutsi zuten. Patra et al. (2019) lanean beste isomorfismo-neurri bat proposatu zuten, Gromov-Hausdorff distantziaren hurbilpen batean oinarritzen dena. Horretaz gain, isomorfismoaren beharra arintzeko metodo erdigainbegiratu bat ere proposatu zuten,[20] ohiko ikasketa antagoniko gainbegiratu gabea, hiztegi elebidun bateko sarreren lerrokatzea eta ortogonalitatearen murriztapen erlaxatu bat uztartzen dituena. Vulić et al. (2019) lanean gure Artetxe et al. (2018b) laneko sistemaren analisi enpiriko sakon bat egin zuten, euren aurretiko esperimentuen arabera metodo gainbegiratu gabeen artean sendoena zena. Hala eta guztiz ere, 15 hizkuntzaren 210 konbinazioetatik 87tan huts egiten zuela erakutsi zuten, metodo gainbegiratu gabeen ezegonkortasuna oraindik ere ebatzi gabeko problema bat dela erakutsiz. Metodo bera hasieratzeko hiztegi elebidun txiki bat erabilita, berriz, emaitza hobeak lortu zituzten. Hori horrela, ikuspegi praktiko batetik gainbegiratze ahuleko hurbilpenak egokiagoak izan daitezkeela defendatu zuten. Horren harira, tesi honetan aurkezten dugun Artetxe et al. (2020d) iritzi-artikuluan metodo erabat gainbegiratu gabeen motibazioaren inguruan hausnartu genuen. Doval et al. (2020) lanean Lample et al. (2018b) eta gure Artetxe et al. (2018b) lanetako metodoen beste analisi enpiriko bat aurkeztu zuten, euren kasuan ere baldintza zailenetan gainbegirapena lagungarria dela ondorioztatuz. Wang et al. (2019) lanean, berriz, gainbegirapen ahultzat Wikipediako artikulu lerrokatuak erabiltzeko ikasketa antagonikoaren hedapen bat proposatu zuten. Wada et al. (2019) lanean metodo gainbegiratu gabeek corpus elebakar txikiekin ere huts egiten zutela erakutsi zuten, eta baldintza horietarako hizkuntza-eredu eleaniztunetan oinarritutako metodo bat proposatu. Gure Ormazabal et al. (2019) lanean hizkuntza ezberdinetako hitz-bektoreen egitura hain ezberdina izatearen arrazoia haien ikasketa modu independentean egitea dela erakutsi genuen. Izan ere, bi hizkuntzetako hitz-bektoreak batera ikasita lerrokatzearen bidez baino hitz-bektore isomorfikoagoak lortzen direla ikusi genuen, bai eta hiztegien indukzioan emaitza hobeak eskuratzen direla eta hubness-aren arazoa arinagoa dela ere. Alderaketa hori egiteko, baina, corpus paraleloak erabili behar izan genituen, arestian ikusi bezala lehen familiako metodoek gainbegirapen sendoa behar izaten baitute. Hori horrela, etorkizunera begira hizkuntza ezberdinetako hitz-bektoreak batera ikasten dituzten metodo gainbegiratu gabeak garatzea ikerketa-lerro interesgarria izan litekeela iradoki genuen.

[20] Aipatzekoa da autoikasketa iteratiboaren inguruko gure jatorrizko lana (Artetxe et al., 2017) eta antzerakoak ere metodo erdigainbegiratutzat ikus daitezkeela. Kasu horietan, baina, gainbegirapena ahula da eta hasieraketarako soilik erabiltzen da.

Azken horri lotuta, aipatzekoa da, urte batzuk atzera eginez, Cao et al. (2016) lanean proposatu zutela CBOW ereduaren hedapen elebidun gainbegiratu gabe bat, bi hizkuntzetako hitz-bektoreen batezbestekoa eta bariantza antzekoak izan zitezen termino gehigarri bat erabiltzen zuena. Euren esperimentuak, baina, nahiko baldintza berezietan egin zituzten, eta ikerketa-lerro horrek ez zuen jarraipenik izan une hartan. Berrikiago, bi hizkuntzatako hitz-bektoreak batera entrenatzeko itzultzaile automatiko gainbegiratu gabe baten bidez sortutako corpus paralelo sintetiko bat erabiltzea proposatu zen (Marie and Fujita, 2019). Hurbilpen horrek ez du inolako gainbegirapenik eskatzen, corpus elebakarrak soilik erabiltzen baititu, baina corpus paraleloak darabiltzan edozein eredu gainbegiratu aplikatzea ahalbidetzen du. Arestian aipaturiko gure Artetxe et al. (2019a) lana ere ideia horretan bertan oinarritzen da, baina sortutako corpus paralelo sintetikoa hiztegi elebidun bat erauzteko erabiltzen du zuzenean, hitz-bektore elebidunak entrenatu beharrean. Itzulpen automatiko gainbegiratu gabearen testuinguruan, bi hizkuntzatako corpus elebakarrak elkartu eta hitz-bektoreen eredu elebakar arrunt bat entrenatzea proposatu zuten Lample et al. (2018c) lanean. Ideia horretan oinarrituz, Wang et al. (2020) lanean hitz-bektoreak modu horretara ikasi, bokabularioa banatu, eta bi hizkuntzak lerrokatzea proposatu zuten. Modu horretara, hizkuntza bietako hitz-bektoreak ez dira modu erabat independentean entrenatzen, bi hizkuntzek partekatutako hitzek zubi-lanak egiten baitituzte, eta gainerako hitzak modu finagoan lerrokatzen dira bigarren urrats batean. Zhou et al. (2019) lanean, azkenik, hitz-bektoreak puntu finkotzat tratatu beharrean Gauss-en nahaste-eredu batek definitutako probabilitate-dentsitateak erabili zituzten, atazaren berezko ziurgabetasuna modelatzeko baliagarria dena, eta bi hizkuntzatako dentsitateak lerrokatu.

1.6.2 Itzupen automatiko gainbegiratu gabea

Itzulpen automatiko estatistikoa bere lehen urratsak ematen ari zelarik, Rapp (1995) lanak corpus elebakarrak erabiliz ere hitz mailako itzulpenak induzitzea posible izan zitekeela erakutsi zuen simulazio baten bidez. Gerora, asko eta asko izan dira **hiztegi elebidunen indukzioan** lan egin dutenak. Aurreko atalean ikusi bezala, metodo horietako gehienek hitz-bektoreak dituzte oinarritzat, izan prediktiboak ala kontaketan oinarrituak, baina ataza konkretu horretara zuzenduriko teknikak ere proposatu izan dira (Koehn and Knight, 2000, 2002; Tamura et al., 2012; Irvine and Callison-Burch, 2013b; Wijaya et al., 2017). Metodo gehienek, baina, hasierako hiztegi elebidun bat eskatzen dute eta, ondorioz, ez dira erabat gainbegiratu gabeak. Horretaz gain, sistema horiek hitz mailan itzultzen dute testu mailan beharrean eta, horrenbestez, ezin daitezke itzultzaile automatikotzat hartu.

Nolanahi ere, **hiztegi elebidunen indukzioa itzulpen automatiko estatistikoko sistemetan txertatzeko** proposamenak ere egin izan dira. Metodo gehienak erdigainbegiratuak dira, eta hiztegi-indukzioa corpus paraleloekin entrenatutako itzultzaile arrunt

baten estaldura hobetzeko erabiltzen dute, haren itzulpen-taula zabalduz (Daumé III and Jagarlamudi, 2011; Irvine and Callison-Burch, 2013a, 2014). Horretaz gain, corpus paraleloen beharra erabat ezabatzeko aukera ere aztertu izan da. Klementiev et al. (2012a) lanak corpus elebakarrak erabiliz sintagmetan oinarritutako eredu log-lineal baten osagaiak ikastea bideragarria izan zitekeela erakutsi zuen. Euren esperimentuetan, baina, itzulpen-taulako sarrerak (sintagma mailako itzulpen-hautagaiak, baina ez haien itzulpen-probabilitateak) aurrez emanak direla suposatzen dute eta, horiek lortzeko, corpus paralelo bat erabili zuten. Horretaz gain, 49.795 sarrerako hiztegi elebidun bat ere erabili zuten. Irvine and Callison-Burch (2016) lanean, berriz, itzulpen-taulako sarrerak corpus elebidunik gabe induzitu zituzten, baina euren metodoak ere hiztegi elebidun bat eskatzen du. Euren esperimentu gehienetan, gainera, corpus elebidun txiki bat erabili zuten, eta corpus elebidunik erabili gabe erakutsi zuten emaitza bakarra bi paragraforen itzulpenaren adibidea izan zen.

Hori horrela, testu elebakarrak soilik erabiliz itzulpen automatikoko sistema osoak entrenatzeko lehen saiakerak **deszifratzearen** eskutik etorri ziren. Ikerketa-lerro horren aitzindaritzat Knight et al. (2006) lana har dezakegu.[21] Itzulpen automatiko estatistikoaren antzera, $\mathbf{x}$ testu zifratua emanda $\mathbf{y}$ testu laua berreskuratzearen problema kanal zaratatsuaren eredua jarraituz formulatu zuten bertan, $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y})$ bilatuz. Deszifratzearen kasuan, baina, $p(\mathbf{x}|\mathbf{y})$ eredua corpus elebakarrak soilik erabiliz ikasi behar da. Horretarako, lehenik eta behin $p(\mathbf{y})$ modelatzen duen bigrama bidezko hizkuntza-eredu bat ikastea proposatu zuten. Behin hori eginda, $p(\mathbf{x}|\mathbf{y})$ ereduaren parametroak ikasten dituzte, $\mathbf{x}$ behatutako testu zifratuaren $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y})$ probabilitatea maximizatuz. $p(\mathbf{x}|\mathbf{y})$ modelatzeko $\mathbf{y}$ sekuentziako osagaiak $t(x_i|y_i)$ probabilitatearen arabera banaka ordezkatzen direla suposatzen dute, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{n} t(x_i|y_i)$ hartuz. Ikasketa itxaropen-maximizazio algoritmoaren bidez egiten dute, programazio dinamikoa erabiliz $O(nv^2)$ denbora behar duena iterazioko, v bokabularioaren tamaina izanik eta n $\mathbf{x}$ sekuentziaren luzera. Lan horretan bertan marko orokor horren hainbat aplikazio posible ere aztertu zituzten: ordezkapen-kodeen deszifratzea, idazkera ezezagunen deszifratzea, eta hizkuntzen deszifratzea edo itzulpen automatikoa. Jarraian ikusiko dugun bezala, ataza horiek euren ibilbide propioa jarraitu dute gerora.

Hasiera batean gehien landu zen aplikazioa **ordezkapen-kodeen deszifratzearena** izan zen. Knight et al. (2006) lanean bertan euren itxaropen-maximizazio bidezko

[21] Kriptoanalisia historia luzeko jakintzagaia izan arren, aipatutako lana aitzindari izan zen ataza horri Hizkuntzaren Prozesamenduaren arlotik begiratu eta bere aplikazio posibleak aztertzen, gerora beste hainbat lanetan garatu zirenak. Aipatzekoa da, era berean, deszifratzea itzulpen automatikoaren beraren jaiotzarako inspirazio-iturri nagusietako bat izan zela. Izan ere, Weaver (1955) memoranduma turkiera jakin gabe turkierazko testu bat deszifratzea lortu zuteneko bigarren mundu-gerrako pasadizo batekin hasten da, eta honela motibatzen du itzulpena konputagailuen bidez automatizatzeko ideia: *Errusierako artikulu bat ikusten dudanean, honakoa pentsatzen dut: "Hau errealitatean ingelesez idatzita dago, baina sinbolo arraro batzuk erabili dituzte. Orain deskodetzen ahaleginduko naiz."* Kriptografia modernoak, bestalde, bere ibilbide propioa jarraitu du—besteak beste, gako publikoko metodoen agerpena tarteko—, eta ez da esanguratsua hizpide dugun gai honekiko.

metodoa ordezkapen-kode sinpleak[22] hausteko gai dela erakutsi zuten, eta japonierazko silaba mailako ordezkapen-kode baterako egokitu gerora (Ravi and Knight, 2009). Ravi and Knight (2011a) lanean inferentzia bayestarrean oinarritutako metodo bat proposatu zuten, karaktere mailako hizkuntza-ereduez gain hitz-zerrendak ere erabiltzen dituena. Lan hori Zodiac-408 kode homofoniko[23] ezaguna automatikoki hausten lehena izan zen. Berg-Kirkpatrick and Klein (2013) lanean, berriz, oinarrizko itxaropen-maximizazio algoritmoan ausazko berrekiteek duten garrantzia erakutsi zuten, eta algoritmo hori Zodiac-408 nahiz prozedura berberaz sorturiko testu zifratu zailagoak hausteko gai dela erakutsi. Beste hurbilpen bat deszifratzea optimizazio-problema konbinatorio gisara formulatzea da, non hizkuntza-ereduak gidatuta baliokidetza-taula determinista optimoa bilatu nahi den. Horretarako osokoen programazioa erabiltzea proposatu izan da (Ravi and Knight, 2008), bai eta baliokidetza-taula inkrementalki eraiki eta soluzio optimoa A* algoritmoaren bidez bilatzea ere (Corlett and Penn, 2010). Nuhn et al. (2013) lanean A* beharrean sorta-bilaketan oinarritutako metodo bat proposatu zuten, hitz-zerrendarik gabe Zodiac-408 deszifratzeko gai dena. Gerora metodo hori hobetu eta Beale kodea modu automatikoan lehen aldiz hausteko aplikatu zuten (Nuhn et al., 2014). Hauer et al. (2014) lanean sorta-bilaketaren ordez Monte Carlo zuhaitz-bilaketa erabiltzea proposatu zuten, eta karaktere mailako eta hitz mailako hizkuntza-ereduak uztartu. Berrikiago, deszifratzearen problemari aurre egiteko neurona-sareetan oinarritutako metodoak erabiltzea ere proposatu izan da. Batetik, n-grametan oinarritutako hizkuntza-ereduen ordez hizkuntza-eredu neuronalak erabiltzea lagungarria dela erakutsi zuten (Kambhatla et al., 2018), eta sare antagoniko sortzaileetan oinarritutako metodo bat ere proposatu zuten (Gomez et al., 2018). Deszifratzearen erronka nagusietako bat bere kostu konputazional altua da. Hain zuzen ere, Nuhn and Ney (2013) lanean ordezkapen-kode sinpleak ere esleipen-problema koadratikoaren baliokideak direla erakutsi zuten eta, horrenbestez, NP-zailak. Corlett and Penn (2013) lanean, berriz, euren aurreko laneko A* algoritmoaren (Corlett and Penn, 2010) exekuzio-denbora teoriko eta praktikoaren arteko alde nabarmenaren zergatia aztertu zuten. Azkenik, aipatu ditugun lanek aspaldidanik eskuz hautsitako edota esperimentuetarako neurrira sortutako testu zifratuak erabili bazituzten ere, aipatzekoa da deszifratze automatikoa giltzarri izan zela Copiale kodea lehen aldiz hausteko (Knight et al., 2011).

Ordezkapen-kodeek idazkera ezaguneko testu lau bat berariaz eraldatzen dute bere edukia benetako hartzaileak soilik uler dezan. **Idazkera ezezagunen deszifratzean**, berriz, sarrera testu laua da, eta idazkera-sistema bera da ezagutzen ez dena. Problema

[22] Ordezkapen-kode sinpleetan testu lauko karaktere ezberdin bakoitza testu zifratuko karaktere ezberdin bati dagokio, eta ordezkapena baliokidetza-taula finko baten arabera egiten da, testuingurua kontuan izan gabe.

[23] Kode homofonikoetan testu lauko karaktere bakoitza kodetzeko hainbat karaktere ezberdin erabil daitezke. Zodiac-408 kodeak, gainera, ez du zuriunerik erabiltzen hitzak bereizteko, ataza are zailagoa egiten duena.

horren aplikazio nagusia aitzinako idazkera galduak deszifratzea da,[24] adibide ezagun bat 1.1 atalean aipatutako egiptoar hieroglifikoena izanik. Knight et al. (2006) lan aitzindariaren aurretik, Knight and Yamada (1999) artikuluan landu zuten dagoeneko aplikazio konkretu hori. Bertan, arestian aipaturiko itxaropen-maximizazio metodoa erabili zuten transkripzio fonetikoen corpus batetik abiatuta idazkera ezezaguneko testu baten transkripzio probableena bilatzeko, eta idazkera ezberdineko hiru hizkuntzari aplikatu: gaztelania, japoniera eta txinera. Snyder et al. (2010) lanean, berriz, murriztapen linguistikoak barneratzen dituen eredu bayestar bat proposatu zuten. Euren metodoa aitzina galdutako ugaritera modu automatikoan deszifratzeko erabili zuten, bertako karaktereak hizkuntza horri lotutako hebraitarrekoetara mapatu eta bi hizkuntza horien arteko hitz sustraikideak identifikatuz. Berg-Kirkpatrick and Klein (2011) lanean eredu sinpleago bat erabili zuten eta inferentzian zentratu. Euren metodoa ugaritera deszifratzeko aplikatu bazuten ere, ez zuten baldintza erabat errealistetan egin, ugaritera eta hebraitarreko hitz sustraikideen zerrenda banatik abiatu baitziren. Berrikiago, idazkera ezezagunak deszifratzeko karaktere mailako eredu neuronal bat proposatu zen (Luo et al., 2019), atazaren berezko murriztapenak barneratzeko ikasketa kostu minimoko fluxu-problema gisara formulatuz. Metodo hori ugaritera ez ezik Lineal B ere modu automatikoan deszifratzen lehena izan zen.

Orain arte ikusiriko deszifratze-atazak itzulpen automatikoaren antzera formula badaitezke ere, **hizkuntza batetik beste batera itzultzea problema konplexuagoa** da. Izan ere, hizkuntzen arteko baliokidetzak hitz mailan modelatu ohi dira karaktere mailan beharrean eta, ondorioz, bokabularioaren tamaina askoz handiagoa izan ohi da. Hori gutxi balitz, baliokidetza horiek ez dira deterministak izaten (hitz bakoitzak hainbat itzulpen ezberdin izan ditzake), ez eta banakakoak ere (hitz jakin baten itzulpena hainbat hitz izan daitezke, eta hitzen hurrenkera ere alda daiteke). Bestalde, aurreko atazetan testu zifratua laburra izan ohi bada ere, kasu honetan hizkuntza bietako corpus elebakar handiak izan ohi dira. Datu gehiago izatea eredu hobeak ikasteko lagungarria bada ere, ezaugarri horrek, bokabulario-tamaina handiekin batera, kostu konputazionalarena are eta erronka handiagoa bihurtzen du.

Hizkuntzen deszifratzearen helburua testu elebakarrak soilik erabiliz itzulpen automatiko estatistikoko sistemak entrenatzea da, deszifratzearen hurbilpena aipatu berri ditugun berezitasunetara egokituz. Horretan lehena Ravi and Knight (2011b) lana izan zen, bi hurbilpen ezberdin proposatu zituena: (i) itxaropen-maximizazio metodoaren egokitzapen bat, hitzen ordezkapena, txertaketa, ezabaketa eta berrordenatze lokala modelatzen dituen eredu baten parametroak ikasteko erabil daitekeena, eta (ii) IBM 3 ereduaren parametroak testu elebakarretatik estimatzeko metodo bayestar bat, inferentzia egiteko Gibbs laginketa darabilena. Gerora, lehen metodoaren eskalagarritasuna hobetzeko hainbat hurbilpen proposatu izan dira: Nuhn et al. (2012) lanean itzulpen-

[24] Horretaz gain Knight et al. (2006) lanean antzeko hurbilpen bat jarraituz karaktere-kodeketa ezezagunak deszifratzea ere proposatu zuten.

hautagaiak murrizteko testuinguru-bektoreak erabiltzea proposatu zuten, Nuhn and Ney (2014) lanean sorta-bilaketa erabiltzea eta hizkuntza-ereduaren nahiz uneko baliokidetzen araberako hautagai aurrezarri batzuk soilik hedatzea, eta Kim et al. (2017) lanean, berriz, hasieraketarako hitz-klaseak baliatzea eta baliokidetza-matrize sakabanatu bat erabiltzea. Horretaz gain, laginketa-prozedura hobetuz metodo bayestarra eskalagarriagoa egiteko proposamenak ere egin izan dira (Dou and Knight, 2012; Ravi, 2013). Dou and Knight (2013) lanean, berriz, mendekotasun-erlazioetan oinarritutako bigramak erabiltzea lagungarria dela erakutsi zuten. Era berean, deszifratze-ereduetan hitz-bektoreak barneratzea ere proposatu zen, ikasketan zehar hizkuntza bietako hitz-bektoreak lerrokatzeko transformazio lineal bat ikasiz (Dou et al., 2015), aurreko azpiatalean ikusi ditugun metodoen antzera. Pourdamghani and Knight (2017) lanean, bestalde, hurbileko hizkuntzen artean itzultzeko karaktere mailako eredu bat proposatu zuten.

Itzultzaile automatikoak entrenatzeko corpus elebakarrak soilik erabiltzen aitzindariak izan baziren ere, aipatu berri ditugun lanek **muga nabarmenak** zituzten, eta nahiko baldintza berezietan ebaluatu izan ziren. Zenbait kasutan ebaluazioa hitz mailan egin zuten (Dou and Knight, 2013; Dou et al., 2015; Kim et al., 2017), hiztegi indukzioaren antzera, eta baldintza horietan hitz-bektoreen lerrokatze-metodoek askoz emaitza hobeak lortzen dituztela erakutsi zen (Zhang et al., 2017a). BLEU edo antzerako metrikak erabiliz esaldi mailan ebaluatu izan direnean, berriz, domeinu berezietako datu-multzo oso txikiak erabili izan dira. Adibidez, gehien erabili izan den OPUS datu-multzoak 20.000 esaldi baino gutxiago ditu hizkuntza bakoitzean, eta bokabularioaren tamaina ez da 1.000 hitzetara heltzen (Ravi and Knight, 2011b; Ravi, 2013; Nuhn et al., 2012; Nuhn and Ney, 2014). Ravi (2013) lanak datu-multzo handiago bat ere erabili zuen, baina 5,3 BLEU puntu bakarrik lortu zituen bertan, sarrerako testua kopiatze hutsak 3,0 BLEU ematen zituelarik. Beste kasu batzuetan deszifratzeko teknikak corpus paraleloekin batera erabili izan dira, paradigma erdigainbegiratuari dagokiona (Dou and Knight, 2012, 2013; Dou et al., 2014).

Hori horrela, eskala eta baldintza estandarretan emaitza sendoak lortzen lehenak tesi honetan aurkezten dugun Artetxe et al. (2018d) eta harekin batera argitaratutako Lample et al. (2018a) lanak izan ziren. Deszifratzearen bidea alde batera utzi eta ataza honetarako hurbilpen berri bat proposatu zuten biek: **itzulpen automatiko neuronal gainbegiratu gabea**. Horretarako neurona-sare errepikarietan oinarritutako arretadun kodetzaile-deskodetzaile arkitektura bat erabili zuten. Bi metodoek kodetzaile partekatu bat darabilte bi hizkuntzetarako, eta lerrokatutako hitz-bektoreak erabiltzen dituzte bere sarrerako geruza hasieratzeko. Gure kasuan Artetxe et al. (2017) laneko metodoa erabili genuen horretarako, eta hasieratutako bektore horiek izoztuta mantendu genituen ikasketan zehar. Lample et al. (2018a) lanaren kasuan, berriz, Lample et al. (2018b) laneko metodoa erabili zuten lerrokatzea ikasteko, eta ez zuten izozterik aplikatu. Hasieraketa horri esker kodetzaileak hitz mailako errepresentazio elebidunak jasotzen ditu sarreratzat, eta bere lana esaldi mailako errepresentazio elebidunak sortzea da.

Deskodetzailea, berriz, errepresentazio horiek hartu eta helburuko hizkuntzako sekuentzia bakoitzari probabilitate-masa bat esleitzeaz arduratzen da. Gure lanean hizkuntza bakoitzarentzat deskodetzaile bereizi bat erabili genuen, eta Lample et al. (2018a) lanean, berriz, bi hizkuntzentzako deskodetzaile partekatu bat. Corpus elebakarrak erabiliz ereduaren parametroak ikasteko bi mekanismo nagusi erabiltzen dituzten bi metodoek. Alde batetik, autokodetzea, sekuentzia bat kodetu eta hizkuntza berera deskodetzean jatorrizko sekuentzia berreskuratzeko probabilitatea maximizatzen duena. Bere oinarrizko bertsioan kopiatze-ataza tribial bati dagokio hori eta, soluzio endekatuak ekiditeko, sarrerako sekuentziari zarata bat aplikatzen zaio, hitz batzuk ausaz ezabatu edota euren ordena aldatuz. Bigarren mekanismoa atzeranzko itzulpena da, corpus elebakarreko sekuentzia bat hartu, sistema bera erabiliz beste hizkuntzara itzuli, eta itzulpen-bikote hori eredua entrenatzeko erabiltzen duena, beti ere alde sintetikoa sarreratzat hartuz eta jatorrizko sekuentzia irteeratzat. Bi horietaz gain, Lample et al. (2018a) lanean ikasketa antagonikoa ere erabili zuten, kodetzailearen irteera hizkuntzarekiko independentea izan dadin hizkuntza hori aurresaten duen diskriminatzaile bat erabiliz.

Gerora, oinarrizko hurbilpen horren hainbat hobekuntza proposatu izan dira. Yang et al. (2018b) lanean hizkuntza bakoitzarentzat kodetzaile ezberdin bat erabiltzea proposatu zuten, euren artean parametroen azpimultzo bat soilik partekatuz. Horretaz gain, kodetzaileen irteera hizkuntzarekiko independentea izan dadin ez ezik, deskodetzaile bakoitzaren irteera dagokion hizkuntzakoa izan dadin ere ikasketa antagonikoa erabiltzea proposatu zuten. Era berean, neurona-sare errepikarien ordez transformer arkitektura erabiltzen lehenak izan ziren, gerora nagusitu dena. Sun et al. (2019) lanean ikasketan zehar hitz-bektoreak lerrokatuta mantentzeko erregularizazio-metodo bat proposatu zuten. Lample et al. (2018c) lanean, berriz, hitz-bektore horiek hasieratzeko beste modu bat proposatu zuten, hurbileko hizkuntza-bikoteei zuzendua: independenteki ikasiriko hitz-bektoreak lerrokatu beharrean, bi hizkuntzetako corpus elebakarrak nahastu, azpihitz-bokabulario partekatu bat ikasi, eta hitz-bektoreen eredu elebakar arrunt bat entrenatzea, partekatutako azpihitzek zubi-lanak egiten dituztelarik. Wei et al. (2019) lanean autokodetzearen hedapen bariazional bat eta errefortzu-ikasketan oinarritutako atzeranzko itzulpenaren aldaera bat proposatu zituzten. Leng et al. (2019) lanean urruneko hizkuntzak zuzenean itzuli beharrean, zubi-lanak egiteko tarteko hizkuntzak erabili eta hainbat urratsetan itzultzea proposatu zuten. Adibide modura, danieratik galegora zuzenean itzuli beharrean, lehendabizi danieratik ingelesera, ondoren ingelesetik gaztelaniara eta, azkenik, gaztelaniatik galegora itzulita emaitza hobeak eskura daitezkeela erakutsi zuten. Bikote guztien artean itzultzeko sistema gainbegiratu gabeak erabili zituzten, hizkuntza horietako corpus elebakarrak soilik erabiliz. Bi hizkuntzen artean itzultzeko tarteko hizkuntza egokienak aukeratzeko metodo automatiko bat ere proposatu zuten, konbinazio posible guztiak probatzea konputazionalki garestiegia baita. Sen et al. (2019) lanean, ostera, hainbat hizkuntzaren artean itzultzeko eredu eleaniztun bat proposatu zuten, eta ohiko sistema elebidunekin baino emaitza hobeak lortzen zituela

erakutsi. Guk proposatutako eredu elebidunaren antzera, hizkuntza bakoitzarentzat deskodetzaile ezberdin bat erabili zuten, eta kodetzailea, berriz, hizkuntza guztiek partekatzen dute. Atzeranzko itzulpena egiteko konbinazio posibleak hizkuntza kopuruaren arabera koadratikoki hazten direnez, jatorrizko edo helburuko hizkuntzatzat ingelesa duten konbinazioak soilik erabili zituzten.

Bestalde, tesi honetan aurkezten dugun Artetxe et al. (2018c) lanak eta harekin batera argitaratutako Lample et al. (2018c) lanak paradigma berri bat proposatu zuten: **itzulpen automatiko estatistiko gainbegiratu gabea**. Biek ala biek corpus elebakarrak erabiltzen dituzte sintagmetan oinarritutako itzulpen estatistikoko sistemak eraikitzeko baina, deszifratzearen hurbilpena jarraitu beharrean, itzulpen automatiko neuronal gainbegiratu gaberako garatutako printzipioetan oinarritzen dira. Lehenik eta behin, hasierako itzulpen-taula bat sortzen dute hitz-bektoreen hizkuntza arteko lerrokatzearen bidez. Horretarako hitzen bektoreak ez ezik n-grama luzeagoen bektoreak ere ikasteko skip-gram ereduaren hedapen ezberdinak erabiltzen dituzte, bi hizkuntzetako bektoreak modu gainbegiratu gabean lerrokatu, jatorrizko hizkuntzako n-grama bakoitzaren helburuko hizkuntzako k auzokide hurbilpenak hartu, eta haien itzulpen-probabilitateak estimatu softmax funtzioa aplikatuz. Behin hori eginda, itzulpen-taula hori hizkuntza-eredu batekin konbinatzen dute hasierako itzulpen-sistema bat lortzeko. Azkenik, atzeranzko itzulpen iteratiboa erabiltzen dute soluzio hori fintzeko. Teknika horrek unean uneko sistema erabiltzen du corpus elebakar bat itzultzeko, eta horrela lortutako corpus paralelo sintetikoa aurkako noranzkoan sistema gainbegiratu arrunt bat entrenatzeko erabili. Prozedura hori behin eta berriz errepikatzen da, urrats bakoitzean noranzkoa txandakatuz. Eredu log-linealaren pisuak doitzeko atzeranzko itzulpenean oinarritutako hurbilpen bat proposatu genuen guk. Lample et al. (2018c) lanean, berriz, pisu aurrezarriak erabili zituzten, inolako doikuntzarik gabe. Artetxe et al. (2019b) lanean azaldu berri dugun hurbilpenaren hiru hobekuntza proposatu genituen: (i) hasierako itzulpen-taulan edizio-distantzian oinarritutako beste bi ezaugarri gehitzea, (ii) eredu log-linearen pisuak doitzeko ziklo-kontsistentzia eta hizkuntza-ereduak uztartzen dituen metodo gainbegiratu gabe bat erabiltzea, eta (iii) noranzko bakoitzeko itzulpen-probabilitateak estimatzeko aurkako noranzkoko atzeranzko itzulpena erabiltzea prozedura iteratiboan.

Horretaz gain, hainbat autore hurbilpen **neuronala eta estatistikoa konbinatzen** saiatu izan dira. Horretan lehena Lample et al. (2018c) lana izan zen. Arestian azalduriko euren sistema neuronal gainbegiratu gabea entrenatzeko eredu horrek berak sortutako atzeranzko itzulpenak ez ezik, aurretiaz entrenatutako itzultzaile automatiko estatistiko gainbegiratu gabe batenak ere erabiltzea proposatu zuten bertan. Marie and Fujita (2018) lanean, berriz, arkitektura gainbegiratu gabe berezi bat erabili beharrean, sistema estatistiko gainbegiratu gabe baten bidez sortutako corpus paralelo sintetikoa itzultzaile neuronal arrunt bat hutsetik entrenatzeko erabiltzea proposatu zuten. Prozedura hori bi noranzkoetan egiten dute, eta horrela ikasitako sistema neuronal biak atzeranzko itzulpenaren bidez corpus paralelo sintetiko gehiago sortzeko erabili. Behin hori eginda,

corpus paralelo sintetiko hedatua beste itzultzaile neuronal bana hutsetik entrenatzeko erabiltzen dute, prozedura modu iteratiboan errepikatuz. Marie and Fujita (2020) lanean hurbilpen hori findu eta prozedura iteratiboan sistema neuronal eta estatistikoen ikasketa txandakatzea proposatu zuten. Tesi honetan aurkezten dugun Artetxe et al. (2019b) lanean, berriz, aldiro-aldiro sistema neuronal bat hutsetik ikasi beharrean, aurkako noranzkoko bi eredu neuronal elkarrekin ikastea proposatu genuen, atzeranzko itzulpenaren bidez elkar elikatzen direnak. Hasierako iterazioetarako aurretiaz entrenaturiko sistema estatistiko gainbegiratu gabe bat erabili genuen. Ren et al. (2019b) lanean, azkenik, noranzko bakoitzerako eredu neuronal eta estatistiko bana ikasteko itxaropen-maximizazioan oinarritutako metodo bat proposatu zuten.

Orain arte ikusi ditugunez gain, Kim et al. (2018) lanean itzulpen automatiko gainbegiratu gaberako beste hurbilpen bat proposatu zuten, bi urrats konbinatzen dituena: **hitzez hitzeko itzulpena eta itzulpen horren zuzenketa**. Lehen urratsean hitz-bektoreen hizkuntza arteko lerrokatzea erabiltzen dute sarrerako testua hitzez hitz itzultzeko. Hitz bakoitza modu isolatuan itzuli beharrean, hautagai bakoitzaren kosinu-antzekotasuna eta hizkuntza-eredu bat konbinatzen dituzte, eta sekuentzia osoa sorta-bilaketaren bidez itzuli. Horrek hitz bakoitza itzultzeko bere testuingurua aintzat hartzea ahalbidetzen badu ere, 1-1 motako baliokidetzak soilik onartzen ditu, inolako berrordenatzerik gabe. Hori dela eta, bigarren urratsean arretadun kodetzaile-deskodetzaile neuronal bat erabiltzen dute, hitzez hitzeko itzulpenaren akatsak zuzentzeaz arduratzen dena. Eredu hori entrenatzeko helburuko hizkuntzako corpus elebakarreko sekuentziak hartzen dituzte, eta hitzez hitzeko itzulpena simulatzen duen zarata artifiziala aplikatu. Horretarako, hitzen ordena ausaz aldatzen dute, ausazko hitz batzuk sartu, eta beste hitz batzuk ausaz ezabatu. Zaratadun sekuentzia hori hartuta, jatorrizko testua berreskuratzeko entrenatzen dute kodetzaile-deskodetzailea. Pourdamghani et al. (2019) lanean, berriz, kodetzaile-deskodetzaile hori entrenatzeko beste hurbilpen bat proposatu zuten: helburuko hizkuntza eta beste hirugarren hizkuntza baten arteko corpus paralelo bat erabiltzea. Hirugarren hizkuntza horretako esaldiak hitzez hitz itzultzen dituzte lehenengo urratseko metodo berbera erabiliz, eta kodetzaile-deskodetzailea itzulpen hori hartu eta corpus paraleloko erreferentziazko itzulpena aurresateko entrenatzen dute. Yang et al. (2019) lanean hizpide dugun arkitekturan oinarrituz hitz-bektoreen lerrokatze hobe bat ikasteko metodo bat proposatu zuten, errefortzu-ikasketa darabilena.

Orain arte ikusiriko hurbilpen gehienek atzeranzko itzulpena badarabilte ere, Wu et al. (2019b) lanean itzultzaile neuronal gainbegiratu gabeak entrenatzeko beste teknika bat proposatu zuten, ***erauzi-editatu*** deitu ziotena. Hurbilpen horrek kodetzaile partekatua erabiltzen du jatorrizko eta helburuko hizkuntzetako sekuentzien bektore-errepresentazioak lortzeko, eta jatorrizko hizkuntzako sekuentzia bakoitzaren helburuko hizkuntzako corpuseko k auzokide hurbilenak hartu. Ondoren, helburuko hizkuntzako sekuentzia horietako bakoitzaren errepresentazioa eta jatorrizko hizkuntzakoarena konbinatzen dituzte. Errepresentazio horietako bakoitza deskodetzaileari pasatu eta helburuko

hizkuntzako sekuentzia berriak sortzen dituzte gero, jatorrizko hizkuntzako hasierako sekuentziaren antza izango dutenak. Antzekotasun hori neurtzeko beste azpisare bat erabiltzen dute, eta jatorrizko sekuentzia emanda sistemaren irteera modu horretara erauzitako sekuentzien ahalik eta antzekoena izan dadin entrenatzen dute eredua.[25]

Orain arte azaldu ditugun hurbilpenek hitz-bektore eleaniztunak erabiltzen dituzte hasieraketarako. Azkenaldian indarra hartzen ari den beste aukera bat **hizkuntza-eredu maskaratuen** bidez neurona-sare osoa hasieratzea da. Hurbilpen hori proposatzen lehena Conneau and Lample (2019) lana izan zen. Horretarako BERTen aldaera eleaniztunaren (Devlin et al., 2019) antzeko metodo bat erabiltzea proposatu zuten, hainbat hizkuntzatako corpus elebakarrak nahastu, azpihitz bokabulario komun bat ikasi, corpus konbinatuko token batzuk ausaz maskaratu, eta transformer kodetzaile bat maskaratutako token horiek berreskuratzeko entrenatzen duena. Eredu hori itzultzaile automatiko neuronal gainbegiratu gabe baten kodetzailea eta deskodetzailea hasieratzeko erabili zuten, autokodetzea eta atzeranzko itzulpenaren bidez entrenatu zutena. Ren et al. (2019a) lanean, berriz, hizkuntza-eredu maskaratuaren beste aldaera bat proposatu zuten hasieraketarako, hitz-bektore eleaniztunak barneratzen dituena. Itzulpen automatiko estatistiko gainbegiratu gabearen antzera, euren metodoak n-gramen itzulpenak induzitzen ditu lehendabizi hitz-bektoreen lerrokatzearen bidez. Behin hori eginda, corpus konbinatuko n-grama batzuk ausaz maskaratu eta, jatorrizko tokenak berreskuratu beharrean, haien itzulpena aurresaten ikasten dute. Azaldu berri ditugun metodoek hizkuntza-eredu maskaraturako transformer kodetzaile bat erabiltzen dute, eta haren pisuak itzulpen automatiko gainbegiratu gabeko ereduaren kodetzailea eta deskodetzailea hasieratzeko erabiltzen dituzte, bakoitza bere aldetik. Song et al. (2019) lanean, berriz, hizkuntza-eredu maskaraturako transformer kodetzaile-deskodetzaile bat erabili eta eredu osoa hasieratzeko baliatzea proposatu zuten. Horretarako sarrerako n-grama batzuk ausaz aukeratu, bertako tokenak maskaratu, eta kodetzaile-deskodetzaile osoa maskaratutako tokenak aurresateko entrenatzen dute. Deskodetzailea autoerregresiboa da, baina kodetzailearen sarreran maskaratutako tokenak soilik aurresaten ditu, eta gainerako tokenak maskaratu egiten dira bere sarreran. Liu et al. (2020) lanean, berriz, hurbilpen sinpleago bat proposatu zuten. Euren metodoak hainbat esaldi hartu, ausaz berrordenatu, n-grama batzuk ausaz aukeratu, n-grama bakoitza maskara-token bakar batekin ordezkatu, eta eredu osoa jatorrizko sekuentzia berreskuratzeko entrenatzen du, deskodetzailean inolako maskaratzerik aplikatu gabe.

Itzulpen automatiko gainbegiratu gabeak izan duen oihartzunaren erakusgarri, aipatze-

[25] Itzulpen automatiko gainbegiratu gabeak bi hizkuntzetako corpus elebakarrak independenteak direla suposatzen du. Hori horrela, lehen pausoan erauzitako helburuko hizkuntzako sekuentziak jatorrizkoaren antzekoak izan litezke, baina ez, printzipioz behintzat, paraleloak. Hori dela eta, hizpide dugun metodoa jatorrizko eta helburuko sekuentzien antzekotasuna estimatzean oinarritzen da. Beste lan batzuk, berriz, Wikipedia bezalako corpus konparagarriak erabiltzen dituzte, eta corpus horietan sekuentzia paralelo batzuk daudela suposatzen dute. Hurbilpen horiek sekuentzia paraleloak automatikoki erauzi eta gainbegirapentzat erabiltzen dituzte (Ruiter et al., 2019; Wu et al., 2019c).

koa da **WMT ataza partekatuan**—itzulpen automatikoaren inguruko garrantzitsuena dena—gai horren inguruko azpiataza bat antolatu dutela 2018ko ediziotik aurrera. Lehen urte hartan (Bojar et al., 2018) 3 parte-hartzaile izan ziren: Graça et al. (2018), Stojanovski et al. (2018) eta Del et al. (2018). Lehen biek hitz-bektore eleaniztunen bidez hasieratutako hurbilpen neuronala jarraitu zuten, eta besteak, berriz, hurbilpen estatistikoa. 2019ko edizioan (Barrault et al., 2019), berriz, 5 parte-hartzaile izan ziren: Marie et al. (2019), Li et al. (2019), Stojanovski et al. (2019), Kvapilíková et al. (2019) eta Liu et al. (2019). Parte-hartzaile guztiek hurbilpen neuronala eta estatistikoa konbinatu zituzten, eta lehen postuetan geratu zirenek hizkuntza-eredu maskaratuen bidezko hasieraketa erabili zuten eredu neuronalarentzat.

Bukatzeko, alemana-ingelesa eta frantsesa-ingelesa moduko hizkuntza-bikoteetan emaitza ikusgarriak lortu izan badira ere, zenbait autorek emaitza negatiboen berri eman dute corpus txikiak, zaratatsuak, domeinu ezberdinetakoak edota urruneko hizkuntzetakoak erabiltzean (Neubig and Hu, 2018; Guzmán et al., 2019). Halako faktoreen eragina hobeto ulertzeko asmoz, Marchisio et al. (2020) eta Kim et al. (2020) lanek egungo sistema gainbegiratu gabeen **analisi enpiriko** bana aurkeztu zuten. Lehenak gure Artetxe et al. (2019b) lanean aurkeztutako sistema aztertu zuen, eta bigarrenak, berriz, Conneau and Lample (2019) lanekoa. Biek ere bi hizkuntzetako corpusak antzeko domeinuetakoak izatea bereziki garrantzitsua dela ondorioztatu zuten, domeinuak ezberdinak direnean emaitzak okertu eta ezegonkortasun arazoak agertzen baitira. Horretaz gain, hizkuntzen arteko distantzia linguistikoak ere eragin nabarmena duela erakutsi zuten.

2 Conclusions

In this thesis, we have proved that it is possible to align independently trained word representations in different languages based solely on their structural similarity. In addition, we have shown that it is possible to learn machine translation systems from monolingual corpora alone by leveraging such aligned representations, obtaining competitive results in standard benchmarks. Ultimately, the contributions made at this thesis, along with other contemporaneous developments, have played a central role in the recent emergence and popularization of unsupervised cross-lingual learning in general and unsupervised machine translation in particular. This new paradigm contrasts with the heavy dependency on parallel corpora that has long characterized this field, and opens new research avenues for the future. More concretely, the main **contributions** made at this thesis, as well as the corresponding conclusions we draw, are as follows:

- We have developed a new mathematical framework that generalizes several cross-lingual word embedding alignment methods (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Shigeto et al., 2015; Xing et al., 2015; Zhang et al., 2016; Smith et al., 2017). The mapping into a common space is performed by an orthogonal transformation, and differences in previous methods are explained in terms of additional normalization, whitening, re-weighting, de-whitening and dimensionality reduction steps. Such a decomposition into several interpretable transformations allowed us to gain new insights into the behavior of existing methods. In particular, we found that re-weighting is greatly beneficial, but must be done in the target language for the length normalization performed by cosine similarity to be effective in nearest neighbor retrieval, which explains the effectiveness of inverse regression (Shigeto et al., 2015). In addition, our results showed that whitening can bring small improvements, but only if de-whitened appropriately, which previous methods failed to do. Furthermore, we observed that the improvements brought by dimensionality reduction greatly overlap with those of re-weighting, with the latter being more effective. Finally, we showed that applying length normalization followed by mean

centering as a preprocessing step is highly beneficial. Based on these observations, we have also designed a new variant that outperforms previous methods in bilingual lexicon induction.

- We have proposed an iterative self-learning approach to learn cross-lingual word embedding mappings. In its basic form, our approach works by implicitly optimizing an unsupervised objective function following an alternating optimization procedure, yet it critically relies on a good initialization to avoid poor local optima. We have also proposed an unsupervised initialization method accordingly, which is based on the intra-lingual similarity distribution of monolingual embeddings, and developed techniques to make the self-learning procedure more robust. Together, our method is able to align word embeddings in different languages in a completely unsupervised manner, obtaining results that are competitive with those of previous supervised methods. Interestingly, the only training signals used throughout the process are the co-occurrence counts coming from monolingual corpora. From this, we can conclude that the co-occurrence patterns of equivalent words in different languages tend to be similar, which suffices to learn high-quality cross-lingual representations.

- We have designed both a neural and a phrase-based statistical machine translation system that can be trained using monolingual corpora only. In both cases, our approach makes use of our unsupervised cross-lingual embedding alignment method for initialization, and heavily relies on back-translation to further train the model. Interestingly, our pure statistical approach obtains better results than our pure neural approach, which contrasts with the superiority of neural machine translation in the supervised scenario. In relation to that, we believe that the modular architecture of phrase-based statistical machine translation is particularly suitable for the unsupervised setting, as it decomposes the translation process into several meaningful components that can be designed separately. In particular, it relies on a language model as a central component, which can be naturally learned from monolingual corpora. Nevertheless, it is well known that statistical machine translation has severe limitations (e.g., the locality or the sparsity problem), which also apply in the unsupervised scenario. In fact, our best unsupervised results are obtained by training two conventional neural machine translation systems in opposite directions through iterative back-translation, using our unsupervised statistical machine translation system for warmup purposes only. This approach is already competitive with the state-of-the-art in supervised statistical machine translation as demonstrated by our WMT 2014 results,[1] which suggests that purely statistical unsupervised approaches are unlikely to bring further improvements. As

[1] Machine translation research has almost exclusively focused on neural approaches in recent years, and the best performing WMT 2014 participants remain representative of the state-of-the-art for this approach.

such, we believe that the neural paradigm will play a central role in future research on unsupervised machine translation, and phrase-based machine translation will either be used in conjunction with it or eliminated altogether. In relation to that, several authors have since then reported strong results using a pure neural system with large-scale pretraining (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020), which seems to confirm this trend. More broadly, our work shows that it is possible to train high-quality machine translation systems in an unsupervised manner by using cross-lingual representations for initialization along with iterative or on-the-fly variants of back-translation. Our neural, statistical and hybrid approaches, as well as most other works on the topic, are based on this general formula despite their fundamentally different nature, which we conclude is an effective way to train machine translation systems using monolingual corpora.

- We have proposed a new method to induce bilingual dictionaries from cross-lingual word embeddings that is based on our unsupervised statistical machine translation system. Our approach builds a phrase-table based on the cross-lingual embeddings, combines it with a language model, and uses the resulting machine translation system to generate a synthetic parallel corpus, from which we induce a bilingual dictionary through statistical word alignment. We have shown that this approach outperforms direct retrieval methods like nearest neighbor and CSLS by a substantial margin. Even if, at a high level, unsupervised machine translation works by generalizing from word level to text level translation, this suggests that the techniques employed to that end can also improve the word level translation quality. At the same time, existing work on cross-lingual word embedding mappings has almost exclusively been evaluated on bilingual lexicon induction through direct retrieval, but our work shows that more elaborated approaches can obtain substantially better results in this task. This complements the study by Glavaš et al. (2019), who observe that bilingual dictionary induction results do not always correlate well with downstream performance, showing that some methods designed specifically for this task perform poorly in others. This prompts to reconsider common evaluation practices in this area, in that future work in bilingual dictionary induction should not focus exclusively in cross-lingual word embedding mappings with direct retrieval, nor should cross-lingual word embedding mappings be evaluated in bilingual lexicon induction alone.

- We have critically examined the motivations, definition, methodology and approaches for unsupervised cross-lingual learning. We have argued that, contrary to the common narrative, the strict unsupervised scenario—involving enough monolingual data *and* no parallel data—is not entirely realistic, and call for a more rigorous motivation of this research area. In addition, we have identified different monolingual and cross-lingual signals—stemming from common assumptions and

varying amounts of linguistic knowledge—that can be exploited when learning cross-lingual models from monolingual data, and advocate for future work being more aware and explicit about their use. Finally, we have described several methodological issues in the validation and evaluation of unsupervised cross-lingual models, and advocated for a more holistic view of this research area, including cross-lingual word embeddings, deep multilingual pretraining and unsupervised machine translation. All in all, given the fast pace in which the field is moving, we believe that establishing a rigorous basis and best practices is essential to keep making meaningful progress in this topic.

In terms of **publications**, this thesis comprises 9 papers published in top-tier conferences (5 at ACL, 2 at EMNLP, 1 at AAAI, and 1 at ICLR), including a best paper award nomination at ACL 2019. In addition, we published 7 other peer reviewed papers during this PhD (3 at ACL, 1 at CoNLL, 1 at TACL, 1 at CL, and 1 at SEPLN), including the CoNLL 2018 best paper award, as well as 2 other papers that are currently under review.

The software used to conduct our research has been released as several **open source** projects: VecMap[2] comprises all of our cross-lingual word embedding alignment code, UNdreaMT[3] implements our unsupervised neural machine translation system, Monoses[4] implements our unsupervised statistical machine translation system, and phrase2vec[5] is our extension of word2vec to learn n-gram embeddings, which is used by Monoses. Several authors have directly built on our code to implement their methods (Riley and Gildea, 2018; Ruder et al., 2018; Yehezkel Lubin et al., 2019; Sen et al., 2019; Kvapilíková et al., 2019), and many others have used it as part of their experiments.

At the same time, the results presented in this thesis have been **corroborated by several independent studies**. In the case of cross-lingual word embedding mappings, most recent research has focused on unsupervised methods, and our system VecMap has been reported to outperform other approaches in this setting. In particular, Glavaš et al. (2019) compared several popular methods (Lample et al., 2018b; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018) in a variety of tasks, concluding that *"The results highlight VecMap as the most robust choice among unsupervised models: besides being the only model to produce successful runs for all language pairs, it also significantly outperforms other unsupervised models—both when considering all language pairs and only the subset where other models produce successful runs."* Vulić et al. (2019) reported that their preliminary experiments further verified these findings, and Doval et al. (2020) also found VecMap to be more robust than MUSE in their comparative study. Additionally, our work on unsupervised cross-lingual embedding alignment was included in the LREC 2020 reproducibility challenge, and its participants were able to

[2] https://github.com/artetxem/vecmap
[3] https://github.com/artetxem/undreamt
[4] https://github.com/artetxem/monoses
[5] https://github.com/artetxem/phrase2vec

successfully replicate our results (Garneau et al., 2020; Pluciński et al., 2020). In the case of unsupervised machine translation, this type of studies have been more scarce and have generally focused in a single method—presumably due to its more recent emergence and higher computational cost—, but our results have also been reported to be reproducible (Marchisio et al., 2020).

It is remarkable the fast pace in which this research area has advanced since its recent emergence. For instance, the first unsupervised neural machine translation system proposed at this thesis, as well as the concurrent work by Lample et al. (2018a), obtained 14-15 BLEU points in English-French WMT 2014, whereas our last system, presented barely a year later, obtained 36 BLEU points in the exact same benchmark. We expect that unsupervised machine translation and, more generally, unsupervised cross-lingual learning, will become a consolidated research area and keep making significant progress in the upcoming years. More concretely, the main research lines that we would like to explore in the **future** are as follows:

- Several recent studies have concluded that existing unsupervised machine translation and cross-lingual embedding alignment methods are highly sensitive to the linguistic distance and domain similarity of the training data, showing that they often break completely in more challenging scenarios (Vulić et al., 2019; Guzmán et al., 2019; Marchisio et al., 2020; Kim et al., 2020). It should be noted that the presence of poor local optima and the difficulty of the underlying optimization problem have been one of the main challenges in unsupervised cross-lingual embedding alignment since its early days—starting with our initial work on self-learning, which required weak supervision for initialization—, and significant progress has been made at mitigating these issues. As such, we believe that, rather than revealing an inherent limitation of unsupervised cross-lingual learning, these recent negative results are indicative that such problems have not been solved completely. This way, we think that further investigating the effect of linguistic distance, typology, and the size, quality and domain of the training data in unsupervised cross-lingual learning and developing models that are more robust in these axes is an important research direction.

- As discussed above, all unsupervised machine translation systems proposed at this thesis, as well as most other contemporaneous work, follow the same general principle, which uses iterative back-translation to train a bidirectional machine translation system initialized through cross-lingual representations. Despite its contrasted effectiveness and generality, this approach is heuristic in nature, as it does not formalize the objective to be optimized. For that reason, we would like to develop a more principled formulation of unsupervised machine translation, which we believe could be helpful to better understand the shortcomings of current approaches

and design stronger variants. We think that the variational interpretation of back-translation (Cotterell and Kreutzer, 2018), the dual learning formulation of machine translation (He et al., 2016), as well as the prior work on statistical decipherment (Knight et al., 2006; Ravi and Knight, 2011b) can be helpful to that end.

- Deep multilingual pretraining (e.g., multilingual BERT) has recently emerged as an alternative to embedding alignment for learning cross-lingual representations from monolingual corpora. While both approaches have been developed independently and treated as different research topics by the community, our recent work shows that a monolingual BERT model can be transferred to new languages by learning a new set of embeddings, obtaining results that are competitive with multilingual BERT (Artetxe et al., 2020c). This suggests that, akin to cross-lingual word embeddings, deep multilingual pretraining might mostly be learning a lexical level alignment. In the future, we would like to further study this possible connection. If confirmed, we believe that ideas from the cross-lingual embedding alignment literature could potentially be used to efficiently transfer existing deep models to new languages. This has a great practical interest, as state-of-the-art deep multilingual pretraining methods are computationally very expensive, and research in this area has consequently been dominated by a few industry actors. For instance, Liu et al. (2020) used 256 GPUs for more than 2 weeks to pretrain their model, whereas our embedding alignment method runs in less than 15 minutes in a single GPU (Artetxe et al., 2018b).

- In addition to being a relevant task on its own, translation is intimately connected to cross-lingual learning. As such, we believe that unsupervised machine translation—or specific techniques developed for it—can also be helpful in other multilingual tasks like cross-lingual transfer learning. In particular, while most work in this area has focused on natural language understanding problems, we think that machine translation is likely to play a more central role in cross-lingual generation problems. More broadly, unsupervised machine translation provides a straightforward way for generating synthetic parallel data, which can be used to train conventional supervised systems. Our work on bilingual lexicon induction through unsupervised machine translation already shows the potential of this idea, which Marie and Fujita (2019) further applied to learn better cross-lingual word embeddings. In the future, we would like to delve deeper into this direction, and explore ways to exploit unsupervised machine translation in other cross-lingual tasks.

Glossary

a priori banaketa *prior distribution*

aditz partikuladun *phrasal verb*

agerkidetza *co-occurrence*

aldagai ezkutu *latent variable*

alderantzizko matrize *inverse matrix*

arreta-mekanismo *attention mechanism*

atalase *threshold*

ataza partekatu *shared task*

ate *gate*

atzeranzko itzulpen *back-translation*

ausazko berrekite *random restart*

autogainbegiratze *self-supervision*

autoikasketa *self-learning*

autokodetzaile antagoniko *adversarial autoencoder*

autokodetzaile bariazional *variational autoencoder*

autokodetze *autoencoding*

auzokide hurbilen *nearest neighbor*

azpihitz *subword*

azpilaginketa *subsampling*

baldintzazko probabilitate *conditional probability*

balidazio *validation*

balio *value*

balio propio *eigenvalue*

balio singularren deskonposaketa *singular value decomposition*

banaketa *distribution*

batezbesteko haztatu *weighted average*

batezbestekoen alde maximo *maximum mean discrepancy*

Bayesen teorema *Bayes' theorem, Bayes' rule*

behe-borne *lower bound*

berrordenatze-eredu *reordering model, distorsion model*

biderkadura eskalar *dot product, scalar product*

birpisaketa *re-weighting*

bokabulario *vocabulary*

defizientzia *deficiency*

deskodeketa *decoding*

deszuritzea *de-whitening*

dimentsionaltasun-murrizketa *dimensionality reduction*

doikuntza *tuning*

ebaluazio estrintseko *extrinsic evaluation*

ebaluazio intrintseko *intrinsic evaluation*

egiantz handieneko estimazio *maximum likelihood estimation*

egoera ezkutu *hidden state*

emankortasun *fertility*

entrenamendu *training*

erdigainbegiratu *semisupervised*

eredu faktorizatu *factorized model*

eredu log-lineal *log-linear model*

eredu prediktibo *predictive model*

errefortzu-ikasketa *reinforcement learning*

errepresentazio banatu *distributed representation*

esleipen-problema koadratiko *quadratic assignment problem*

euklidear *Euclidean*

ez-amaierako *non-terminal*

ezaugarri-funtzio *feature function*

ezkutuko analisi semantiko *latent semantic analysis*

ezkutuko indexazio semantiko *latent semantic indexing*

faktorizatu *factorize*

familia esponentzialeko osagai nagusien analisi *exponential family principal component analysis*

feedforward neurona-sare *feedforward neural network*

finketa *refinement*

flexio *inflection*

gainbegiratu *supervised*

gainbegiratu gabe *unsupervised*

gainbegiratze *supervision*

gako *key*

garraio optimoko problema *optimal transport problem*

garrantziaren araberako laginketa *importance sampling*

Gauss-en nahaste-eredu *Gaussian mixture model*

geruza-normalizazio *layer normalization*

goi-borne *upper bound*

gradiente jaitsiera estokastiko *stochastic gradient descent*

hein *rank*

hipotesi distribuzional *distributional hypothesis*

hitz-bektore *word embedding*

hizkuntza arteko *cross-lingual*

hizkuntza-eredu *language model*

hizkuntza-eredu maskaratu *masked language model*

hizkuntza-inferentzia *natural language inference*

hiztegi *dictionary*

hiztegi elebidunen indukzio *bilingual lexicon induction*

hondar-konexio *residual connection*

ikasketa antagoniko *adversarial learning*

ikasketa dual *dual learning*

ikasketa sakon *deep learning*

inferentzia bayestar *Bayesian inference*

informazio-berreskurapen *information retrieval*

itxaropen-maximizazio *expectation-maximization*

itzulpen automatiko estatistiko *statistical machine translation*

itzulpen automatiko neuronal *neural machine translation*

itzulpen-eredu *translation model*

itzulpen-taula *phrase table*

izendatzaile *denominator*

kanal zaratatsuaren eredu *noisy channel model*

katearen erregela *chain rule*

kategoria gramatikal *part-of-speech*

kode homofoniko *homophonic cipher*

kodetzaile-deskodetzaile *encoder-decoder*

kontaketan oinarritutako eredu *count-based model*

kontsulta *query*

korrelazio kanonikoaren analisi *canonical correlation analysis*

kosinu-antzekotasun *cosine similarity*

kostu minimoko fluxu-problema *minimum-cost flow problem*

lagin-efizientzia *sample efficiency*

laginketa negatibo *negative sampling*

laplacetar *Laplacian*

lema *lemma*

lerrokatze *alignment*

leuntze *smoothing*

marjina maximo *max-margin*

matrize irauli *transpose matrix*

matrize sakabanatu *sparse matrix*

mendekotasun-erlazio *dependency relation*

metaikasketa *meta-learning*

modularitate *modularity*

Monte Carlo zuhaitz-bilaketa *Monte Carlo Tree Search*

neurona-sare *neural network*

neurona-sare errepikari *recurrent neural network*

neurona-sare konboluzional *convolutional neural network*

norma *norm*

NP-zail *NP-hard*

optimizazio-problema konbinatorio *combinatorial optimization problem*

ordezkapen-kode *substitution cipher*

ordezkapen-kode sinple *simple substitution cipher*

orientazio eten *discontinuous orientation*

orientazio monotono *monotone orientation*

orientazio trukatu *swap orientation*

osagai nagusien analisi *principal component analysis*

osokoen programazio *integer programming*

Procrustesen analisi orokortu *generalized Procrustes analysis*

Procrustesen problema ortogonal *orthogonal Procrustes problem*

puntukako elkarrekiko informazio *pointwise mutual information*

sailkatzaile *classifier*

sakabanatze *sparsity*

sare antagoniko sortzaile *generative adversarial network*

sekuentzia *sequence*

singtagma *phrase*

sintagmetan oinarritutako itzulpen automatiko estatistiko *phrase-based statistical machine translation*

soluzio endekatu *degenerated solution*

sorta-bilaketa *beam search*

sustraikide *cognate*

teilakatze lexiko *lexical overlap*

testu lau *plain text*

transferentzia-ikasketa *transfer learning*

zarata murrizte *denoising*

zarata-banaketa *noise distribution*

zenbakitzaile *numerator*

ziklo-kontsistentzia *cycle consistency*

zuritzea *whitening*

Bibliography

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyper-alignment for multilingual word embeddings. In *Proceedings of the Seventh International Conference on Learning Representations*.

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics*, 6:185–196.

Robert B Allen. 1987. Several studies on natural language and back-propagation. In *Proceedings of the IEEE First International Conference on Neural Networks*, volume 2, pages 335–341. IEEE Piscataway, NJ.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Antonios Anastasopoulos and Graham Neubig. 2019. Should all cross-lingual embeddings speak English? *arXiv preprint arXiv:1911.03058*.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA. PMLR.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019c. Unsupervised neural machine translation, a new paradigm solely based on monolingual text. *Procesamiento del Lenguaje Natural*, 63:151–154.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*. Under review.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018d. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Mikel Artetxe, Gorka Labaka, Noe Casas, and Eneko Agirre. 2020b. Do all roads lead to Rome? Understanding the role of initialization in iterative back-translation. *arXiv preprint arXiv:2002.12867*. Under review.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018e. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020c. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Accepted for publication.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020d. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Accepted for publication.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Third International Conference on Learning Representations*.

Amir Bakarov, Roman Suvorov, and Ilya Sochenkov. 2018. The limitations of cross-language word embeddings evaluation. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 94–100, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Leonard E Baum. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3(1):1–8.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Taylor Berg-Kirkpatrick and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 313–321, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2013. Decipherment with a million random restarts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 874–878, Seattle, Washington, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193, New Orleans, Louisiana. Association for Computational Linguistics.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827, Osaka, Japan. The COLING 2016 Organizing Committee.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

Yu Chen and Andreas Eisele. 2012. MultiUN v2: UN documents with multilingual alignments. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2500–2504, Istanbul, Turkey. European Language Resources Association (ELRA).

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations

using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Lonnie Chrisman. 1991. Learning recursive distributed representations for holistic computation. *Connection Science*, 3(4):345–366.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Michael Collins, S. Dasgupta, and Robert E Schapire. 2002. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems 14*, pages 617–624. MIT Press.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7057–7067.

Eric Corlett and Gerald Penn. 2010. An exact A* method for deciphering letter-substitution ciphers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1040–1047, Uppsala, Sweden. Association for Computational Linguistics.

Eric Corlett and Gerald Penn. 2013. Why letter substitution puzzles are not hard to solve: A case study in entropy and probabilistic search-complexity. In *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, pages 83–92, Sofia, Bulgaria. Association for Computational Linguistics.

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*.

Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. 2017. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 175–181, Valencia, Spain. Association for Computational Linguistics.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA. Association for Computational Linguistics.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Maksym Del, Andre Tättar, and Mark Fishel. 2018. Phrase-based unsupervised machine translation with compositional phrase embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 361–367, Belgium, Brussels. Association for Computational Linguistics.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Third International Conference on Learning Representations (Workshop Track)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 557–565, Doha, Qatar. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying Bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised bilingual lexicon induction via latent variable models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 621–626, Brussels, Belgium. Association for Computational Linguistics.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304, Brussels, Belgium. Association for Computational Linguistics.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2020. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4013–4023, Marseille, France. European Language Resources Association.

Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. 2019a. *Ethnologue: Languages of Africa and Europe, Twenty-Second Edition*. SIL International.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. 2019b. *Ethnologue: Languages of Asia, Twenty-Second Edition*. SIL International.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. 2019c. *Ethnologue: Languages of the Americas and the Pacific, Twenty-Second Edition*. SIL International.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference*

on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4952–4962, Florence, Italy. Association for Computational Linguistics.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Fifth Workshop on Very Large Corpora*.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu, Hawaii. Association for Computational Linguistics.

Pablo Gamallo, Susana Sotelo, José Ramom Pichel, and Mikel Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 129–137, Boulder, Colorado. Association for Computational Linguistics.

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand, and Luc Lamontagne. 2020. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings: Making the method robustly reproducible as well. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5546–5554, Marseille, France. European Language Resources Association.

Eric Gaussier, J.M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 526–533, Barcelona, Spain.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017a. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1434–1439, Denver, Colorado. Association for Computational Linguistics.

Aidan N. Gomez, Sicong Huang, Ivan Zhang, Bryan M. Li, Muhammad Osama, and Lukasz Kaiser. 2018. Unsupervised cipher cracking using discrete GANs. In *Proceedings of the Sixth International Conference on Learning Representations*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

John C Gower. 1975. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51.

Miguel Graça, Yunsu Kim, Julian Schamper, Jiahui Geng, and Hermann Ney. 2018. The RWTH aachen university English-German and German-English unsupervised neural machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 377–385, Belgium, Brussels. Association for Computational Linguistics.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2018. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 582–586, Brussels, Belgium. Association for Computational Linguistics.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2019. Comparing unsupervised word translation methods step by step. In *Advances in Neural Information Processing Systems 32*, pages 6033–6043.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Bradley Hauer, Ryan Hayward, and Grzegorz Kondrak. 2014. Solving substitution ciphers with combined language models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2314–2325, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 619–624, Valencia, Spain. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*, pages 820–828.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Jiaji Huang, Qiang Qiu, and Kenneth Church. 2019. Hubless nearest neighbor search for bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4072–4080, Florence, Italy. Association for Computational Linguistics.

David A Huffman. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101.

William John Hutchins. 1986. *Machine translation: past, present, future*. Ellis Horwood Chichester.

Ann Irvine and Chris Callison-Burch. 2013a. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013b. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–523, Atlanta, Georgia. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170, Ann Arbor, Michigan. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2016. End-to-end statistical machine translation with zero or small parallel texts. *Natural Language Engineering*, 22(4):517–548.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Nishant Kambhatla, Anahita Mansouri Bigvand, and Anoop Sarkar. 2018. Decipherment of substitution ciphers with neural language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 869–874, Brussels, Belgium. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium. Association for Computational Linguistics.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868, Brussels, Belgium. Association for Computational Linguistics.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Yunsu Kim, Julian Schamper, and Hermann Ney. 2017. Unsupervised training for large vocabulary translation using sparse lexicon and word classes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 650–656, Valencia, Spain. Association for Computational Linguistics.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012a. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140, Avignon, France. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012b. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2011. The Copiale cipher. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 2–9, Portland, Oregon. Association for Computational Linguistics.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 499–506, Sydney, Australia. Association for Computational Linguistics.

Kevin Knight and Kenji Yamada. 1999. A computational approach to deciphering unknown scripts. In *Unsupervised Learning in Natural Language Processing*.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229, Baltimore, Maryland. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*, volume 5, pages 79–86.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, page 711–715. AAAI Press.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ivana Kvapilíková, Dominik Macháček, and Ondřej Bojar. 2019. CUNI systems for the unsupervised news translation task in WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 241–248, Florence, Italy. Association for Computational Linguistics.

Pei Ling Lai and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *Proceedings of the Sixth International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China. Coling 2010 Organizing Committee.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, pages 175–183, Florence, Italy. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019. The NiuTrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282, Florence, Italy. Association for Computational Linguistics.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, Denver, Colorado. Association for Computational Linguistics.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155, Florence, Italy. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015c. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial autoencoders. In *Proceedings of the Fourth International Conference on Learning Representations*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Daniel Marcu and Daniel Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Benjamin Marie and Atsushi Fujita. 2019. Unsupervised joint training of bilingual word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3224–3230, Florence, Italy. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2020. Iterative training of unsupervised neural and statistical machine translation systems. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(5).

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy. Association for Computational Linguistics.

Yuxian Meng, Xiangyuan Ren, Zijun Sun, Xiaoya Li, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Large-scale pretraining for neural machine translation with tens of billions of sentence pairs. *arXiv preprint arXiv:1909.11861*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3857–3867, Minneapolis, Minnesota. Association for Computational Linguistics.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *AISTATS'05*, pages 246–252.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium. Association for Computational Linguistics.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 221–227, Melbourne, Australia. Association for Computational Linguistics.

Ndapandula Nakashole and Raphael Flauger. 2017. Knowledge distillation for bilingual dictionary induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2497–2506, Copenhagen, Denmark. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Graham Neubig and Taro Watanabe. 2016. Optimization for statistical machine translation: A survey. *Computational Linguistics*, 42(1):1–54.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 156–164, Jeju Island, Korea. Association for Computational Linguistics.

Malte Nuhn and Hermann Ney. 2013. Decipherment complexity in 1:1 substitution ciphers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 615–621, Sofia, Bulgaria. Association for Computational Linguistics.

Malte Nuhn and Hermann Ney. 2014. EM decipherment for large vocabularies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 759–764, Baltimore, Maryland. Association for Computational Linguistics.

Malte Nuhn, Julian Schamper, and Hermann Ney. 2013. Beam search for solving substitution ciphers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1576, Sofia, Bulgaria. Association for Computational Linguistics.

Malte Nuhn, Julian Schamper, and Hermann Ney. 2014. Improved decipherment of homophonic ciphers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1764–1768, Doha, Qatar. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Yves Peirsman and Sebastian Padó. 2008. Semantic relations in bilingual lexicons. *ACM Trans. Speech Lang. Process.*, 8(2).

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California. Association for Computational Linguistics.

Kamil Pluciński, Mateusz Lango, and Michał Zimniewicz. 2020. A closer look on unsupervised cross-lingual word embeddings mapping. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5555–5562, Marseille, France. European Language Resources Association.

Maurice Pope. 1999. *The story of decipherment: From Egyptian hieroglyphs to Maya script*, revised edition. Thames and Hudson.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.

Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193.

Alexandre Rafalovitch, Robert Dale, et al. 2009. United Nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, Maryland, USA. Association for Computational Linguistics.

Sujith Ravi. 2013. Scalable decipherment for machine translation via hash sampling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 362–371, Sofia, Bulgaria. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2008. Attacking decipherment problems optimally with low-order N-gram models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 812–819, Honolulu, Hawaii. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2009. Probabilistic methods for a Japanese syllable cipher. In *International Conference on Computer Processing of Oriental Languages*, pages 270–281. Springer.

Sujith Ravi and Kevin Knight. 2011a. Bayesian inference for Zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 239–247, Portland, Oregon, USA. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011b. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019a. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, Hong Kong, China. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019b. Unsupervised neural machine translation with SMT as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248.

Parker Riley and Daniel Gildea. 2018. Orthographic features for bilingual lexicon induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–394, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Peter H Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk, Marta R. Costa-jussà, and Jose A. R. Fonollosa. 2007. Smooth bilingual *n*-gram translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 430–438, Prague, Czech Republic. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. CCMatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the Fifth International Conference on Learning Representations*.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057, Uppsala, Sweden. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5926–5936, Long Beach, California, USA. PMLR.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2018. The LMU Munich unsupervised machine translation systems. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 513–521, Belgium, Brussels. Association for Computational Linguistics.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2019. The LMU Munich unsupervised machine translation system for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 393–399, Florence, Italy. Association for Computational Linguistics.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36, Jeju Island, Korea. Association for Computational Linguistics.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA. Association for Computational Linguistics.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ashish Venugopal, Stephan Vogel, and Alex Waibel. 2003. Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Sapporo, Japan. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Beijing, China. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124, Florence, Italy. Association for Computational Linguistics.

Haozhou Wang, James Henderson, and Paola Merlo. 2019. Weakly-supervised concept-based adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4419–4430, Hong Kong, China. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *Proceedings of the Eigth International Conference on Learning Representations*.

Warren Weaver. 1955. Translation. *Machine translation of languages*, 14:15–23.

Xiangpeng Wei, Yue Hu, Luxi Xing, and Li Gao. 2019. Unsupervised neural machine translation with future rewarding. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 281–290, Hong Kong, China. Association for Computational Linguistics.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463, Copenhagen, Denmark. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019a. Pay less attention with lightweight and dynamic convolutions. In *Proceedings of the Seventh International Conference on Learning Representations*.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019b. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1173–1183, Minneapolis, Minnesota. Association for Computational Linguistics.

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019c. Machine translation with weakly paired documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4375–4384, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Peng Chen, Tianyu Liu, and Xu Sun. 2019. MAAM: A morphology-aware alignment model for unsupervised bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3190–3196, Florence, Italy. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Shuangzhi Wu, Jingjing Xu, Dongdong Zhang, and Xu Sun. 2018a. Learning unsupervised word mapping by maximizing mean discrepancy. *arXiv preprint arXiv:1811.00275*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018b. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia. Association for Computational Linguistics.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5696–5701, Hong Kong, China. Association for Computational Linguistics.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning vector-spaces with noisy supervised lexicon. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 460–465, Minneapolis, Minnesota. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pages 3320–3328.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. 2017. The neural noisy channel. In *Proceedings of the Fifth International Conference on Learning Representations*.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2019. Putting machine translation in context with the noisy channel model. *arXiv preprint arXiv:1910.00553*.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence*, pages 18–32, Berlin, Heidelberg. Springer Berlin Heidelberg.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.

Jiawei Zhao and Andrew Gilman. 2020. Non-linearity in mapping based cross-lingual word embeddings. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3583–3589, Marseille, France. European Language Resources Association.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1588–1598, Minneapolis, Minnesota. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A

Appendix

This appendix includes a copy of the publications related to this thesis in the recommended reading order.

Learning principled bilingual mappings of word embeddings while preserving monolingual invariance

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group, University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

Mapping word embeddings of different languages into a single space has multiple applications. In order to map from a source space into a target space, a common approach is to learn a linear mapping that minimizes the distances between equivalences listed in a bilingual dictionary. In this paper, we propose a framework that generalizes previous work, provides an efficient exact method to learn the optimal linear transformation and yields the best bilingual results in translation induction while preserving monolingual performance in an analogy task.

1 Introduction

Bilingual word embeddings have attracted a lot of attention in recent times (Zou et al., 2013; Kočiský et al., 2014; Chandar A P et al., 2014; Gouws et al., 2014; Gouws and Søgaard, 2015; Luong et al., 2015; Wick et al., 2016). A common approach to obtain them is to train the embeddings in both languages independently and then learn a mapping that minimizes the distances between equivalences listed in a bilingual dictionary. The learned transformation can also be applied to words missing in the dictionary, which can be used to induce new translations with a direct application in machine translation (Mikolov et al., 2013b; Zhao et al., 2015).

The first method to learn bilingual word embedding mappings was proposed by Mikolov et al. (2013b), who learn the linear transformation that minimizes the sum of squared Euclidean distances for the dictionary entries. Subsequent work has proposed alternative optimization objectives to learn better mappings. Xing et al. (2015) incorporate length normalization in the training of word embeddings and try to maximize the cosine similarity instead, introducing an orthogonality constraint to preserve the length normalization after the projection. Faruqui and Dyer (2014) use canonical correlation analysis to project the embeddings in both languages to a shared vector space.

Beyond linear mappings, Lu et al. (2015) apply deep canonical correlation analysis to learn a non-linear transformation for each language. Finally, additional techniques have been used to address the hubness problem in Mikolov et al. (2013b), both through the neighbor retrieval method (Dinu et al., 2015) and the training itself (Lazaridou et al., 2015). We leave the study of non-linear transformation and other additions for further work.

In this paper, we propose a general framework to learn bilingual word embeddings. We start with a basic optimization objective (Mikolov et al., 2013b) and introduce several meaningful and intuitive constraints that are equivalent or closely related to previously proposed methods (Faruqui and Dyer, 2014; Xing et al., 2015). Our framework provides a more general view of bilingual word embedding mappings, showing the underlying connection between the existing methods, revealing some flaws in their theoretical justification and providing an alternative theoretical interpretation for them. Our experiments on an existing English-Italian word translation induction and an English word analogy task give strong empirical evidence in favor of our theoretical reasoning, while showing that one of our models clearly outperforms previous alternatives.

2 Learning bilingual mappings

Let X and Z denote the word embedding matrices in two languages for a given bilingual dictionary so that their ith row X_{i*} and Z_{i*} are the word embeddings of the ith entry in the dictionary. Our goal is to find a linear transformation matrix W so that XW best approximates Z, which we formalize minimizing the sum of squared Euclidean distances following Mikolov et al. (2013b):

$$\arg\min_W \sum_i \|X_{i*}W - Z_{i*}\|^2$$

Alternatively, this is equivalent to minimizing the (squared) Frobenius norm of the residual matrix:

$$\arg\min_W \|XW - Z\|_F^2$$

Consequently, W will be the so called least-squares solution of the linear matrix equation $XW = Z$. This is a well-known problem in linear algebra and can be solved by taking the Moore-Penrose pseudoinverse $X^+ = \left(X^T X\right)^{-1} X^T$ as $W = X^+ Z$, which can be computed using SVD.

2.1 Orthogonality for monolingual invariance

Monolingual invariance is needed to preserve the dot products after mapping, avoiding performance degradation in monolingual tasks (e.g. analogy). This can be obtained requiring W to be an orthogonal matrix ($W^T W = I$). The exact solution under such orthogonality constraint is given by $W = VU^T$, where $Z^T X = U\Sigma V^T$ is the SVD factorization of $Z^T X$ (cf. Appendix A). Thanks to this, the optimal transformation can be efficiently computed in linear time with respect to the vocabulary size. Note that orthogonality enforces an intuitive property, and as such it could be useful to avoid degenerated solutions and learn better bilingual mappings, as we empirically show in Section 3.

2.2 Length normalization for maximum cosine

Normalizing word embeddings in both languages to be unit vectors guarantees that all training instances contribute equally to the optimization goal. As long as W is orthogonal, this is equivalent to maximizing the sum of cosine similarities for the dictionary entries, which is commonly used for similarity computations:

$$\begin{aligned}&\arg\min_W \sum_i \left\| \frac{X_{i*}}{\|X_{i*}\|} W - \frac{Z_{i*}}{\|Z_{i*}\|} \right\|^2 \\ &= \arg\max_W \sum_i \cos\left(X_{i*}W, Z_{i*}\right)\end{aligned}$$

This last optimization objective coincides with Xing et al. (2015), but their work was motivated by an hypothetical inconsistency in Mikolov et al. (2013b), where the optimization objective to learn word embeddings uses dot product, the objective to learn mappings uses Euclidean distance and the similarity computations use cosine. However, the fact is that, as long as W is orthogonal, optimizing the squared Euclidean distance of length-normalized embeddings is equivalent to optimizing the cosine, and therefore, the mapping objective proposed by Xing et al. (2015) is equivalent to that used by Mikolov et al. (2013b) with orthogonality constraint and unit vectors. In fact, our experiments show that orthogonality is more relevant than length normalization, in contrast to Xing et al. (2015), who introduce orthogonality only to ensure that unit length is preserved after mapping.

2.3 Mean centering for maximum covariance

Dimension-wise mean centering captures the intuition that two randomly taken words would not be expected to be semantically similar, ensuring that the expected product of two random embeddings in any dimension and, consequently, their cosine similarity, is zero. As long as W is orthogonal, this is equivalent to maximizing the sum of dimension-wise covariance for the dictionary entries:

$$\begin{aligned}&\arg\min_W \|C_m X W - C_m Z\|_F^2 \\ &= \arg\max_W \sum_i \operatorname{cov}\left(XW_{*i}, Z_{*i}\right)\end{aligned}$$

where C_m denotes the centering matrix

This equivalence reveals that the method proposed by Faruqui and Dyer (2014) is closely related to our framework. More concretely, Faruqui and Dyer (2014) use Canonical Correlation Analysis (CCA) to project the word embeddings in both languages to a shared vector space. CCA maximizes

the dimension-wise covariance of both projections (which is equivalent to maximizing the covariance of a single projection if the transformations are constrained to be orthogonal, as in our case) but adds an implicit restriction to the two mappings, making different dimensions have the same variance and be uncorrelated among themselves[1]:

$$\arg\max_{A,B} \sum_i \text{cov}\left(XA_{*i}, ZB_{*i}\right)$$
$$\text{s.t.} \quad A^T X^T C_m X A = B^T Z^T C_m Z B = I$$

Therefore, the only fundamental difference between both methods is that, while our model enforces monolingual invariance, Faruqui and Dyer (2014) do change the monolingual embeddings to meet this restriction. In this regard, we think that the restriction they add could have a negative impact on the learning of the bilingual mapping, and it could also degrade the quality of the monolingual embeddings. Our experiments (cf. Section 3) show empirical evidence supporting this idea.

3 Experiments

In this section, we experimentally test the proposed framework and all its variants in comparison with related methods. For that purpose, we use the translation induction task introduced by Mikolov et al. (2013b), which learns a bilingual mapping on a small dictionary and measures its accuracy on predicting the translation of new words. Unfortunately, the dataset they use is not public. For that reason, we use the English-Italian dataset on the same task provided by Dinu et al. (2015)[2]. The dataset contains monolingual word embeddings trained with the word2vec toolkit using the CBOW method with negative sampling (Mikolov et al., 2013a)[3]. The English embeddings were trained on a 2.8 billion word corpus (ukWaC + Wikipedia + BNC), while the 1.6 billion word corpus itWaC was used to train the Italian embeddings. The dataset also contains a bilingual dictionary learned from Europarl, split into a training set of 5,000 word pairs and a test set of 1,500 word pairs, both of them uniformly distributed in frequency bins. Accuracy is the evaluation measure.

Apart from the performance of the projected embeddings in bilingual terms, we are also interested in the monolingual quality of the source language embeddings after the projection. For that purpose, we use the word analogy task proposed by Mikolov et al. (2013a), which measures the accuracy on answering questions like "what is the word that is similar to *small* in the same sense as *biggest* is similar to *big*?" using simple word vector arithmetic. The dataset they use consists of 8,869 semantic and 10,675 syntactic questions of this type, and is publicly available[4]. In order to speed up the experiments, we follow the authors and perform an approximate evaluation by reducing the vocabulary size according to a frequency threshold of 30,000 (Mikolov et al., 2013a). Since the original embeddings are the same in all the cases and it is only the transformation that is applied to them that changes, this affects all the methods in the exact same way, so the results are perfectly comparable among themselves. With these settings, we obtain a coverage of 64.98%.

We implemented the proposed method in Python using NumPy, and make it available as an open source project[5]. The code for Mikolov et al. (2013b) and Xing et al. (2015) is not publicly available, so we implemented and tested them as part of the proposed framework, which only differs from the original systems in the optimization method (exact solution instead of gradient descent) and the length normalization approach in the case of Xing et al. (2015) (postprocessing instead of constrained training). As for the method by Faruqui and Dyer (2014), we used their original implementation in Python and MATLAB[6], which we extended to cover cases where the dictionary contains more than one entry for the same word.

[1] While CCA is typically defined in terms of correlation (thus its name), correlation is invariant to the scaling of variables, so it is possible to constrain the canonical variables to have a fixed variance, as we do, in which case correlation and covariance become equivalent

[2] http://clic.cimec.unitn.it/~georgiana.dinu/down/

[3] The context window was set to 5 words, the dimension of the embeddings to 300, the sub-sampling to 1e-05 and the number of negative samples to 10

[4] https://code.google.com/archive/p/word2vec/

[5] https://github.com/artetxem/vecmap

[6] https://github.com/mfaruqui/crosslingual-cca

	EN-IT	EN AN.
Original embeddings	-	76.66%
Unconstrained mapping	34.93%	73.80%
+ length normalization	33.80%	73.61%
+ mean centering	38.47%	73.71%
Orthogonal mapping	36.73%	76.66%
+ length normalization	36.87%	76.66%
+ mean centering	39.27%	76.59%

Table 1: Our results in bilingual and monolingual tasks.

	EN-IT	EN AN.
Original embeddings	-	76.66%
Mikolov et al. (2013b)	34.93%	73.80%
Xing et al. (2015)	36.87%	76.66%
Faruqui and Dyer (2014)	37.80%	69.64%
Our method	39.27%	76.59%

Table 2: Comparison of our method to other work.

3.1 Results of our framework

The rows in Table 1 show, respectively, the results for the original embeddings, the basic mapping proposed by Mikolov et al. (2013b) (cf. Section 2) and the addition of orthogonality constraint (cf. Section 2.1), with and without length normalization and, incrementally, mean centering. In all the cases, length normalization and mean centering were applied to all embeddings, even if missing from the dictionary.

The results show that the orthogonality constraint is key to preserve monolingual performance, and it also improves bilingual performance by enforcing a relevant property (monolingual invariance) that the transformation to learn should intuitively have. The contribution of length normalization alone is marginal, but when followed by mean centering we obtain further improvements in bilingual performance without hurting monolingual performance.

3.2 Comparison to other work

Table 2 shows the results for our best performing configuration in comparison to previous work. As discussed before, (Mikolov et al., 2013b) and (Xing et al., 2015) were implemented as part of our framework, so they correspond to our uncostrained mapping with no preprocessing and orthogonal mapping with length normalization, respectively.

As it can be seen, the method by Xing et al. (2015) performs better than that of Mikolov et al. (2013b) in the translation induction task, which is in line with what they report in their paper. Moreover, thanks to the orthogonality constraint their monolingual performance in the word analogy task does not degrade, whereas the accuracy of Mikolov et al. (2013b) drops by 2.86% in absolute terms with respect to the original embeddings.

Since Faruqui and Dyer (2014) take advantage of CCA to perform dimensionality reduction, we tested several values for it and report the best (180 dimensions). This beats the method by Xing et al. (2015) in the bilingual task, although it comes at the price of a considerable degradation in monolingual quality.

In any case, it is our proposed method with the orthogonality constraint and a global preprocessing with length normalization followed by dimension-wise mean centering that achieves the best accuracy in the word translation induction task. Moreover, it does not suffer from any considerable degradation in monolingual quality, with an anecdotal drop of only 0.07% in contrast with 2.86% for Mikolov et al. (2013b) and 7.02% for Faruqui and Dyer (2014).

When compared to Xing et al. (2015), our results in Table 1 reinforce our theoretical interpretation for their method (cf. Section 2.2), as it empirically shows that its improvement with respect to Mikolov et al. (2013b) comes solely from the orthogonality constraint, and not from solving any inconsistency.

It should be noted that the implementation by Faruqui and Dyer (2014) also length-normalizes the word embeddings in a preprocessing step. Following the discussion in Section 2.3, this means that our best performing configuration is conceptually very close to the method by Faruqui and Dyer (2014), as they both coincide on maximizing the average dimension-wise covariance and length-normalize the embeddings in both languages first, the only difference being that our model enforces monolingual invariance after the normalization while theirs does change the monolingual embeddings to make different dimensions have the same variance and be uncorrelated among themselves. However, our model performs considerably better than any configuration from Faruqui and Dyer (2014) in both the monolingual and the bilingual task, supporting our hypothesis that these two constraints that are implicit in their method are not only conceptually confusing,

but also have a negative impact.

4 Conclusions

This paper develops a new framework to learn bilingual word embedding mappings, generalizing previous work and providing an efficient exact method to learn the optimal transformation. Our experiments show the effectiveness of the proposed model and give strong empirical evidence in favor of our reinterpretation of Xing et al. (2015) and Faruqui and Dyer (2014). It is the proposed method with the orthogonality constraint and a global preprocessing with length normalization and dimension-wise mean centering that achieves the best overall results both in monolingual and bilingual terms, surpassing those previous methods. In the future, we would like to study non-linear mappings (Lu et al., 2015) and the additional techniques in (Lazaridou et al., 2015).

Acknowledgments

This research was partially supported by the European Commision (QTLeap FP7-ICT-2013-10-610516), a Google Faculty Award, and the Spanish Ministry of Economy and Competitiveness (TADEEP TIN2015-70214-P). Mikel Artetxe enjoys a doctoral grant from the Spanish Ministry of Education, Culture and Sports.

A Proof of solution under orthogonality

Constraining W to be orthogonal ($W^TW = I$), the original minimization problem can be reformulated as follows (cf. Section 2.1):

$$\begin{aligned}
&\arg\min_W \sum_i \|X_{i*}W - Z_{i*}\|^2 \\
&= \arg\min_W \sum_i \left(\|X_{i*}W\|^2 + \|Z_{i*}\|^2 - 2X_{i*}WZ_{i*}^T\right) \\
&= \arg\max_W \sum_i X_{i*}WZ_{i*}^T \\
&= \arg\max_W \mathrm{Tr}\left(XWZ^T\right) \\
&= \arg\max_W \mathrm{Tr}\left(Z^TXW\right)
\end{aligned}$$

In the above expression, $\mathrm{Tr}(\cdot)$ denotes the trace operator (the sum of all the elements in the main diagonal), and the last equality is given by its cyclic property. At this point, we can take the SVD of Z^TX as $Z^TX = U\Sigma V^T$, so $\mathrm{Tr}\left(Z^TXW\right) = \mathrm{Tr}\left(U\Sigma V^TW\right) = \mathrm{Tr}\left(\Sigma V^TWU\right)$. Since V^T, W and U are orthogonal matrices, their product V^TWU will also be an orthogonal matrix. In addition to that, given that Σ is a diagonal matrix, its trace after an orthogonal transformation will be maximal when the values in its main diagonal are preserved after the mapping, that is, when the orthogonal transformation matrix is the identity matrix. This will happen when $V^TWU = I$ in our case, so the optimal solution will be $W = VU^T$.

References

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015), workshop track*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 224–229.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 270–280.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Min-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*, pages 151–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *Thirtieth AAAI conference on Artificial Intelligence (AAAI)*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations

Mikel Artetxe and **Gorka Labaka** and **Eneko Agirre**
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

Using a dictionary to map independently trained word embeddings to a shared space has shown to be an effective approach to learn bilingual word embeddings. In this work, we propose a multi-step framework of linear transformations that generalizes a substantial body of previous work. The core step of the framework is an orthogonal transformation, and existing methods can be explained in terms of the additional normalization, whitening, re-weighting, de-whitening and dimensionality reduction steps. This allows us to gain new insights into the behavior of existing methods, including the effectiveness of inverse regression, and design a novel variant that obtains the best published results in zero-shot bilingual lexicon extraction. The corresponding software is released as an open source project.

1 Introduction

Bilingual word embeddings have attracted a lot of attention in recent times. Most methods to learn them use some sort of bilingual signal at the document level, either in the form of document-aligned or label-aligned comparable corpora (Søgaard et al. 2015; Vulić and Moens 2016; Mogadala and Rettinger 2016) or, more commonly, in the form of parallel corpora (Gouws, Bengio, and Corrado 2015; Luong, Pham, and Manning 2015).

An alternative approach that we address in this paper is to independently train the embeddings for each language on monolingual corpora, and then map them to a shared space based on a bilingual dictionary (Mikolov, Le, and Sutskever 2013; Lazaridou, Dinu, and Baroni 2015). This requires minimal bilingual supervision compared to other approaches, while allowing to leverage large amounts of monolingual corpora with competitive results (Vulić and Korhonen 2016; Artetxe, Labaka, and Agirre 2017). Moreover, the learned mappings can also be applied to words that were missing in the training dictionary, and thus induce their translations, with improvements in machine translation (Zhao, Hassan, and Auli 2015).

Authors have proposed different methods to learn such word embedding mappings, but their approach and motivations are often divergent, making it difficult to get a general understanding of the topic. In this work, we tackle this issue and propose a multi-step framework that generalizes previous work. The core step of the framework, which maps both languages to a shared space using an orthogonal transformation, is shared by all variants, and the differences between previous methods are exclusively explained in terms of their normalization, whitening, re-weighting and dimensionality reduction behavior. We analyze the effect of each of these steps with experimental support, which allows us to gain new insights into the behavior of existing methods. Based on these insights, we design a novel variant that improves the state-of-the-art in bilingual lexicon extraction.

Our framework is highly related to the zero-shot learning paradigm, where a multi-class classifier trained over a subset of the labels learns to predict unseen labels by exploiting a common representation for them (Palatucci et al. 2009). In our scenario, these labels correspond to the target language words and their common representation is provided by their corresponding embeddings. This is a prototypical zero-shot learning problem, and similar mapping techniques have also been used in other zero-shot tasks like image labeling (Shigeto et al. 2015; Lazaridou, Dinu, and Baroni 2015) and drug discovery (Larochelle, Erhan, and Bengio 2008).

The remaining of this paper is organized as follows. Section 2 discusses related work. Section 3 explains the proposed multi-step framework and shows the equivalence with previous methods. Section 4 then presents the experimental settings, while Section 5 discusses the obtained results. Section 6 concludes the paper.

2 Related work

For the sake of space, we will focus on related work directly relevant to embedding mappings and bilingual lexicon extraction. Bilingual embedding mapping methods work by independently training the word embeddings in two languages, and then mapping them to a shared space based on a bilingual dictionary. Even if the literature in the topic is quite broad, existing methods can be classified in the following four groups:

1. **Regression methods** map the embeddings in one language to maximize their similarity with the other language. For that purpose, methods in this group use a least-squares objective function that learns the linear transformation minimizing the sum of squared Euclidean distances for the dictionary entries. This approach was first

proposed by Mikolov, Le, and Sutskever (2013), and later adopted by many other authors that incorporated L2 regularization (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015; Vulić and Korhonen 2016). Even if the linear transformation is usually learned from the source language into the target language, Shigeto et al. (2015) argue that it is better to map the target language into the source language as a way to address the hubness problem[1].

2. **Canonical methods** map the embeddings in both languages to a shared space where their similarity is maximized. This is usually done through Canonical Correlation Analysis (CCA) as first proposed by Faruqui and Dyer (2014), who motivate their method as a way to improve the quality of monolingual embeddings using bilingual data. With a similar motivation, Lu et al. (2015) extend this work and use Deep Canonical Correlation Analysis to learn non-linear mappings. CCA was also extended to the multilingual scenario by Ammar et al. (2016) taking English as the pivot language.

3. **Orthogonal methods** map the embeddings in one or both languages to maximize their similarity, but constrain the transformation to be orthogonal. This constraint has been introduced with different motivations. Xing et al. (2015) allege inconsistencies in previous approaches, and orthogonality serves to preserve the length normalization performed by their method to address them. Artetxe, Labaka, and Agirre (2016) motivate orthogonality as a way to preserve monolingual invariance, preventing the degradation in monolingual tasks observed for other techniques. Zhang et al. (2016) focus on a transfer-learning scenario with only ten translation pairs for training, and incorporate orthogonality as a hard regularizer. Finally, Smith et al. (2017) point out that the mapping should be orthogonal in order to be self-consistent.

4. **Margin methods** map the embeddings in one language to maximize the margin between the correct translations and the rest of the candidates. This approach was proposed by Lazaridou, Dinu, and Baroni (2015) as a way to address the hubness problem, with the addition of intruder negative sampling to generate more informative training examples.

As it can be seen, the previous work on embedding mappings is quite diverse, with many authors working under different scenarios and motivations. In an attempt to provide a more general view, Artetxe, Labaka, and Agirre (2016) show the equivalence of different objective functions under orthogonality and different normalization procedures, and clarify that regression, canonical and orthogonal methods essentially differ on the constraints imposed on the mapping. In contrast, our framework decomposes these differences into several interpretable steps, which allows us to gain additional insights into the behavior of previous methods and design new variants addressing their deficiencies. We also cover additional methods, including most references in this section (see Table 1).

A practical application of embedding mappings, as well as the main evaluation task, is bilingual lexicon extraction, that is, the zero-shot translation of words that were missing in the training dictionary. This is usually done through **nearest neighbor retrieval**, taking the closest embedding in the target language according to some similarity metric (usually cosine). However, Dinu, Lazaridou, and Baroni (2015) argue that this approach suffers from the hubness problem, and propose using **inverted nearest neighbor retrieval**[2] instead, which takes the target embedding that has the source embedding ranked highest in its nearest neighbor list. Ties are solved by taking the candidate with the highest cosine similarity. Finally, **inverted softmax retrieval** (Smith et al. 2017) also works by reversing the direction of the query, but instead of using the cosine in the similarity computations, it uses a softmax function with a hyperparameter to control the temperature, which is tuned in the training dictionary. In this paper we revisit these techniques, and show that the alternatives to nearest neighbor mitigated deficiencies in previous mapping methods, while our method learns better mappings.

3 Proposed framework

Let X and Z be the word embedding matrices in two languages for a given bilingual dictionary so that their ith row X_{i*} and Z_{i*} are the embeddings of the ith entry. We aim to learn the transformation matrices W_X and W_Z so the mapped embeddings XW_X and ZW_Z are close to each other.

We next propose a multi-step framework to learn such mappings that allows to generalize previous work. The ith step of the framework applies a linear transformation to the output embeddings of the previous step in each language. This way, if $X_{(i)}$ denotes the output embeddings in the source language at step i and $W_{X(i)}$ the linear transformation at step i, we will have $X_{(i)} = X_{(i-1)}W_{X(i)}$ and $W_X = \prod_i W_{X(i)}$, and analogously for the target language. As it is clear from this last expression, the composition of several linear transformations is another linear transformation, so the purpose of our framework is not to improve the expressive power of linear mappings, but rather to decompose them into several meaningful steps. More concretely, our framework consists of the following steps:

- **Step 0: Normalization (optional)**: In this optional preprocessing step, the word embeddings in each language are independently normalized. This can involve length normalization (making all embeddings have a unit Euclidean norm), and mean centering (making each component have a zero mean). Note that this is done as a preprocessing step, obtaining the initial embedding matrices $X_{(0)}$ and $Z_{(0)}$ that will be mapped by the following ones.

[1] Hubness (Radovanović, Nanopoulos, and Ivanović 2010a; 2010b) refers to the phenomenon of some points (known as *hubs*) being the nearest neighbors of many other points in high-dimensional spaces, and has been reported to severely affect bilingual embedding mappings (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015; Shigeto et al. 2015; Smith et al. 2017).

[2] Note that the original paper refers to this method as *globally-corrected* retrieval.

		S0 (l)	S0 (m)	S1	S2	S3	S4 (src)	S4 (trg)	S5
OLS	Mikolov, Le, and Sutskever (2013)			x	x	src	trg	trg	
	Shigeto et al. (2015)			x	x	trg	src	src	
CCA	Faruqui and Dyer (2014)	x	x	x	x				x
Orth.	Xing et al. (2015)	x			x				
	Zhang et al. (2016)				x				
	Artetxe, Labaka, and Agirre (2016)	x	x		x				
	Smith et al. (2017)	x			x				x
	Proposed (Section 5)	x	x	x	x	trg	src	trg	x

Table 1: Equivalence of the proposed framework with previous methods. (l) and (m) denote length normalization and mean centering, respectively.

- **Step 1: Whitening (optional)**. This optional step applies a whitening or sphering transformation to the embeddings in each language, which makes their different components have a unit variance and be uncorrelated among themselves, turning their covariance matrices into the identity matrix[3]. For that purpose, we adopt the Mahalanobis or ZCA whitening, taking $W_{X(1)} = (X^T X)^{-\frac{1}{2}}$ and $W_{Z(1)} = (Z^T Z)^{-\frac{1}{2}}$.
- **Step 2: Orthogonal mapping**. This step maps the embeddings in both languages to a shared space. Both transformations are constrained to be orthogonal, preserving the dot product for each of the languages on their own. More concretely, we take $W_{X(2)} = U$ and $W_{Z(2)} = V$, where $USV^T = X_{(1)}^T Z_{(1)}$ is the SVD factorization of $X_{(1)}^T Z_{(1)}$. This maximizes the summative cross-covariance of the mapped embeddings $\mathrm{Tr}(X_{(1)} W_{X(2)} W_{Z(2)}^T Z_{(1)}^T)$. Moreover, the ith component of the mapped embeddings corresponds to the direction of maximum cross-covariance being orthogonal to the previous ones, and S_{ii} is its corresponding cross-covariance value. Note that when whitening is applied at step 1, the variance in all directions is 1, so the cross-covariance is equivalent to the cross-correlation.
- **Step 3: Re-weighting (optional)**: This optional step re-weights each component according to its cross-correlation, increasing the relevance of those that best match across languages. So as to simplify the formalization, we will only consider this step if step 1 was applied before, in which case the cross-correlations correspond to the singular values in S (step 2). The re-weighting can be applied to the source language embeddings ($W_{X(3)} = S$ and $W_{Z(3)} = I$), or to the target language embeddings ($W_{X(3)} = I$ and $W_{Z(3)} = S$).
- **Step 4: De-whitening (optional)**: This optional step restores the original variance in every direction, and it is thus only meaningful if step 1 was applied before. The embeddings in a given language can be de-whitened with respect to the original variance in that same language, but also with respect to the original variance in the other language, as both languages are in the same space after step 2. In either case, de-whitening language A with respect to B requires $W_{A(4)} = W_{B(2)}^T W_{B(1)}^{-1} W_{B(2)}$.
- **Step 5: Dimensionality reduction (optional)**: This optional step keeps the first n components of the resulting embeddings and drops the rest, which is obtained by $W_{X(5)} = W_{Z(5)} = (I_n \ 0)^T$. This can be seen as an extreme form of re-weighting, where the first n components are re-weighted by one and the remaining ones by zero.

An interesting aspect of this framework is that the mapping of both languages to a common space is reduced to a single step that is shared by all variants (step 2). Moreover, this mapping is orthogonal and, therefore, preserves monolingual invariance. Therefore, different variants, including existing methods, will only differ on their treatment of normalization, whitening/de-whitening, re-weighting and dimensionality reduction, which are easier to interpret. More concretely, the equivalence of this framework with existing methods, detailed in Table 1, is as follows:

- **Regression methods** correspond to the case where both languages are whitened, re-weighting is applied to the source language, and both languages are de-whitened with respect to the target language (or inversely if the regression is applied from the target language into the source language). This equivalence is directly given by the close-form solution of the unregularized variant, known as Ordinary Least Squares (OLS)[4], and we leave the analysis of $L2$ regularization for future work.

[3]Note that our use of the variance and covariance concepts at this step and the following ones assumes that the embeddings are already mean centered (i.e. we take $X^T X$ as (proportional to) the covariance matrix of X, $Z^T Z$ as the covariance matrix of Z, and $X^T Z$ as the cross-covariance matrix of X and Z).

[4]The optimal solution of OLS is given by $W_{OLS} = X^+ Z$, where $X^+ = \left(X^T X\right)^{-1} X^T$ is the Moore-Penrose pseudoinverse of X. At the same time, by simple algebraic development of our claimed equivalence, $W_X = \left(X^T X\right)^{-1} X^T Z V$ and $W_Z = V$, where V is an orthogonal matrix given by the SVD factorization at step 2. Therefore, both solutions are equivalent up to the orthogonal transformation V of the resulting space, which is invariant with respect to the dot product.

- **Canonical methods** (CCA) correspond to the case where both languages are whitened, none is de-whitened, re-weighting is not used, and dimensionality reduction is applied. The equivalence is given by the SVD solution of CCA (see for instance Lu and Foster (2014)).
- **Orthogonal methods** correspond to the simplest case without any whitening, re-weighting and de-whitening. The equivalence is directly given by the transformation learned at step 2, which is equivalent to the solutions of Artetxe, Labaka, and Agirre (2016) and Smith et al. (2017).

As it can be seen, our framework covers all mapping families with the exception of margin based ones, which were only explored by Lazaridou, Dinu, and Baroni (2015) and surpassed by subsequent work.

4 Experimental settings

For easier comparison with related work, we performed our experiments in the bilingual lexicon extraction scenario proposed by Dinu, Lazaridou, and Baroni (2015) and used by subsequent authors. Their public English-Italian dataset[5] includes monolingual word embeddings in both languages together with a bilingual dictionary split in a training set and a test set. Artetxe, Labaka, and Agirre (2017) extended this dataset to English-German and English-Finnish, which we also use in our experiments. In all cases, the embeddings were trained with the word2vec toolkit with CBOW and negative sampling (Mikolov et al. 2013)[6]. The training and test sets were derived from dictionaries built from Europarl word alignments and available at OPUS (Tiedemann 2012), taking 1,500 random entries uniformly distributed in 5 frequency bins as the test set and the 5,000 most frequent pairs of the remaining word pairs as the training set. The corpora used consisted of 2.8 billion words for English (ukWaC + Wikipedia + BNC), 1.6 billion words for Italian (itWaC), 0.9 billion words for German (SdeWaC), and 2.8 billion words for Finnish (Common Crawl from WMT 2016).

In addition to these languages, we further extended the dataset to English-Spanish using the exact same settings described above. For that purpose, we used the WMT News Crawl 2007-2012 corpus[7] for Spanish, which consists of 386 million words. Tokenization was performed using standard Moses tools. Note that the resulting Spanish corpus has a different domain to the previous ones (news vs web crawling), and it is also smaller, which explains the lower accuracy numbers in the next section.

The goal of our experiments is twofold. On the one hand, we want to analyze the effect of each of the steps of our framework on their own, and interpret the results in relation to the behavior of previous methods. On the other hand, we want to identify the best variant of our framework, and compare it with existing methods proposed in the literature. Given that the effect of normalization was already analyzed in detail by Artetxe, Labaka, and Agirre (2016), we leave this factor aside in our experiments and use their recommended configuration, which performs length normalization followed by mean centering. Moreover, we use cosine similarity with standard nearest neighbor as our retrieval method unless otherwise specified, which allows us to better evaluate the quality of the mapping itself. The remaining factors are analyzed independently, and their best combination is then compared to the state-of-the-art. The code and resources to reproduce our experiments are available at `https://github.com/artetxem/vecmap`.

[5] `http://clic.cimec.unitn.it/~georgiana.dinu/down/`

[6] The context window was set to 5 words, the dimension of the embeddings to 300, the sub-sampling to 1e-05 and the number of negative samples to 10, and the vocabulary was restricted to the 200,000 most frequent words.

[7] `http://www.statmt.org/wmt13/translation-task.html`

5 Results and discussion

From Section 5.1 to 5.4, we respectively analyze the effect of whitening/de-whitening, re-weighting, dimensionality reduction and the retrieval method. Section 5.5 then compares the proposed system with other methods in the literature.

5.1 Whitening and de-whitening (steps 1 and 4)

As discussed before, existing methods have a very different behavior with respect to whitening. While orthogonal methods do not perform any whitening, both CCA and OLS whiten both languages, and the latter also de-whitens them with respect to one of the languages, depending on the direction of regression (see Table 1).

Table 2 shows our results for different whitening/de-whitening strategies. In addition to the said variants implicitly used by existing methods, it also includes our proposed variant: the more intuitive choice of de-whitening each language with respect to the original variance in that same language.

As it can be seen, the results show that, for most language pairs, whitening and de-whitening each language with respect to itself brings a small improvement over not whitening at all. The only exception is English-Finnish, whose accuracy drops almost one point with respect to not applying any whitening or de-whitening. A possible explanation of why proper whitening and de-whitening helps is a hypothetical bias that would otherwise push directions with high variance together.

But, more importantly, the results show that the whitening/de-whitening behavior of both CCA and OLS is not only counterintuitive, but also harmful. In the case of the former, simply whitening both languages causes a huge accuracy drop of 7-9 points, suggesting that the variances of the original embeddings are relevant and should not be ignored by any means. In the case of the latter, de-whitening with respect to either language causes an accuracy drop of 2-4 points, showing that this de-whitening strategy is better than not de-whitening at all, but worse than the natural choice of de-whitening with respect to the language in question.

5.2 Re-weighting (step 3)

As seen in the Section 3, neither orthogonal methods nor CCA use re-weighting, while OLS re-weights either the

Motivation	S1	S4 (src)	S4 (trg)	EN-IT	EN-DE	EN-FI	EN-ES
Orth.				39.27%	41.87%	**30.62%**	31.40%
CCA	x			32.27%	33.00%	22.05%	23.73%
OLS	x	src	src	37.33%	38.47%	25.35%	28.87%
	x	trg	trg	38.00%	36.60%	26.33%	28.80%
New	x	src	trg	**39.47%**	**41.93%**	29.71%	**31.67%**

Table 2: Accuracy for different whitening (S1) and de-whitening (S4) configurations. All settings use length normalization and mean centering, and do not re-weight nor apply dimensionality reduction.

Mot.	S3	EN-IT	EN-DE	EN-FI	EN-ES
Orth. / CCA		39.47%	41.93%	29.71%	31.67%
OLS	src	38.53%	41.73%	28.65%	30.47%
	trg	**43.80%**	**44.27%**	**32.79%**	**36.47%**

Table 3: Accuracy for different re-weighting (S3) configurations. All settings use length normalization, mean centering, and whitening/de-whitening with respect to the original language.

S3	S5	EN-IT	EN-DE	EN-FI	EN-ES
		39.47%	41.93%	29.71%	31.67%
	x	42.53%	**44.53%**	32.09%	33.80%
trg		43.80%	44.27%	32.79%	36.47%
	x	**44.00%**	44.27%	**32.94%**	**36.53%**

Table 4: Accuracy for different dimensionality reduction (S5) and re-weighting (S3) configurations. All settings use length normalization, mean centering, whitening, and de-whitening with respect to the original language.

source or the target language depending on the direction of regression. Table 3 shows the results obtained for all these different re-weighting strategies.

As it can be seen, re-weighting the target language is highly beneficial, bringing an improvement of 3-5 points in all cases, while re-weighting the source language is always harmful. Interestingly, which side to re-weight should not be a relevant factor when using the dot product, so this difference must be explained by the length normalization performed by cosine similarity. Note that, when re-weighting the source language, this length normalization is applied to each source language word on its own, but its nearest neighbor list is not affected in any way, as its similarity with respect to all target language words is only scaled by a constant normalization factor. As a consequence, for the length normalization of cosine similarity to be effective in nearest neighbor retrieval, the re-weighting must be applied in the target language, which can explain why we obtain better results for it.

This behavior is also consistent with the findings of Shigeto et al. (2015) regarding the direction of regression. Recall that these authors claim that mapping the target language into the source language is better than mapping the source language into the target language, which respectively correspond to re-weighting the target language and the source language according to our framework (see Table 1). While Shigeto et al. (2015) explain the relevance of the regression direction in terms of the emergence of hubs in the subsequent nearest neighbor retrieval, our work identifies that the origin of this problem is in the implicit re-weighting direction and its relation with the length normalization performed by cosine similarity.

5.3 Dimensionality reduction (step 5)

As discussed before, CCA is always used with dimensionality reduction, while OLS never is. Dimensionality reduction is typically not applied in orthogonal methods either, although Smith et al. (2017) recently introduced it for the first time.

Table 4 shows our results with and without dimensionality reduction. When performing dimensionality reduction, we always chose the number of dimensions that yield the highest accuracy in the training dictionary, and then evaluate in the test set. As discussed in Section 3, dimensionality reduction can be seen as an extreme form of re-weighting, so we performed these experiments with and without re-weighting the target language so as to better understand how these two steps interact.

As it can be seen, dimensionality reduction has a positive effect in all cases. However, its impact is very small when using target language re-weighting (an improvement of 0.20 points in the best case), and much bigger when not using any re-weighting (improvements of 2-3 points). This suggests that re-weighting and dimensionality reduction have an overlapping effect, which reinforces our interpretation that dimensionality reduction is just an extreme form of re-weighting that removes the components with smallest cross-correlation. In relation to that, it is remarkable that re-weighting gives considerably better results than dimensionality reduction alone, which can be attributed to its smooth rescaling of embedding components in contrast to the binary discarding performed by dimensionality reduction. The only exception in this regard is English-German, for which dimensionality reduction alone gives slightly better results.

All in all, we can conclude that, in spite of being con-

Retrieval method	EN-IT	EN-DE	EN-FI	EN-ES
Nearest neighbor	44.00%	**44.27%**	**32.94%**	36.53%
Inverted nearest neighbor	43.07%	42.20%	31.18%	32.53%
Inverted softmax	**45.27%**	44.13%	**32.94%**	**36.60%**

Table 5: Accuracy for different retrieval methods. All settings use length normalization, mean centering, whitening, target language re-weighting, de-whitening with respect to the original language, and dimensionality reduction tuned in training.

	EN-IT	EN-DE	EN-FI	EN-ES
Mikolov, Le, and Sutskever (2013)	34.93% (**)	35.00% (**)	25.91% (**)	27.73% (**)
Faruqui and Dyer (2014)	38.40% (*)	37.13% (*)	27.60% (*)	26.80% (*)
Shigeto et al. (2015)	41.53% (**)	43.07% (**)	31.04% (**)	33.73% (**)
Dinu, Lazaridou, and Baroni (2015)	37.7% / 38.53% (*)	38.93% (*)	29.14% (*)	30.40% (*)
Lazaridou, Dinu, and Baroni (2015)	40.2%	-	-	-
Xing et al. (2015)	36.87% (**)	41.27% (**)	28.23% (**)	31.20% (**)
Zhang et al. (2016)	36.73% (**)	40.80% (**)	28.16% (**)	31.07% (**)
Artetxe, Labaka, and Agirre (2016)	39.27%	41.87% (*)	30.62% (*)	31.40% (*)
Smith et al. (2017)	43.1% / 44.53% (**)	43.33% (**)	29.42% (**)	35.13% (**)
Proposed (nearest neighbor)	44.00%	**44.27%**	**32.94%**	36.53%
Proposed (inverted softmax)	**45.27%**	44.13%	**32.94%**	**36.60%**

Table 6: Accuracy of our method in comparison with previous work. (*) means that the results were obtained using the original implementation from the authors, while (**) means that the results were obtained using our custom implementation as part of our proposed framework. The rest of the results were reported in the original papers. For methods that were not originally proposed for bilingual lexicon extraction, we used nearest neighbor retrieval.

nected, re-weighting tends to work considerably better than dimensionality reduction thanks to its smooth nature. Moreover, combining them has a small but positive impact, and should be the preferred configuration to use.

5.4 Retrieval method

Most previous work uses standard nearest neighbor for bilingual lexicon extraction (see Section 2), but alternative retrieval methods have been proposed to address the hubness problem attributed to it (Dinu, Lazaridou, and Baroni 2015; Smith et al. 2017). Table 5 reports the results for each of these methods. In the case of inverted softmax, we tune the inverse temperature to optimize the accuracy in the training set, which we find to work better than maximizing the log-likelihood as originally proposed by Smith et al. (2017). To speed up the computations, we take a random sample of 1,500 words to estimate the partition function of the softmax during tuning, but use the entire source vocabulary in the test set. Similarly, we use the entire source vocabulary as pivots when using inverted nearest neighbor.

As it can be seen, inverted softmax performs at par with standard nearest neighbor retrieval for all language pairs except for English-Italian, where it brings an improvement of 1.27 points. Note that this number is considerably smaller than the nearly 5 points reported by Smith et al. (2017) for the same dataset. At the same time, inverted nearest neighbor performs worse than standard nearest neighbor in our experiments. This suggests that alternative retrieval methods are not fully complementary with the improvements brought by our framework. We hypothesize that this is connected to our previous discussion on re-weighting. Recall that our work explains that, for the length normalization performed by cosine similarity to be effective in nearest neighbor retrieval, the re-weighting should be performed in the opposite side. Nevertheless, most previous work was not applying re-weighting properly, and alternative retrieval methods would mitigate the problem by reversing the direction of nearest neighbor. Note, thus, that the alternative methods were alleviating an inherent flaw of the mapping methods during retrieval, while our framework learns better mappings.

5.5 Comparison with the state-of-the-art

Having analyzed the different steps of the proposed framework on their own, we next analyze how it performs in comparison to other methods proposed in the literature. For that purpose, we choose the recommended variant of our framework as discussed throughout the section, which is the one using whitening, re-weighting the target language, de-whitening with respect the original language, and applying dimensionality reduction (see Table 1). The obtained results are given in Table 6. Note that we only tried limited combinations of well-motivated steps and, given that we tested in several pairs of languages, we think that our conclusions are well supported. Moreover, note that our implementation of inverted softmax optimizes accuracy and uses the entire source vocabulary for computing the partition function at test time as described in Section 5.4, which performs better than the variant reported in Smith et al. (2017) as shown

by its corresponding line in the table (43.1 vs. 44.53 for EN-IT).

As it can be seen, our system obtains the best published results in all the four language pairs. Moreover, it also surpasses the previous state-of-the-art even when using standard nearest neighbor retrieval, which shows the superiority of the mapping method itself.

6 Conclusions and future work

In this work, we propose a new framework to learn bilingual embedding mappings that generalizes a substantial body of previous work (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Shigeto et al. 2015; Xing et al. 2015; Zhang et al. 2016; Artetxe, Labaka, and Agirre 2016; Smith et al. 2017). A key aspect of our framework is that the mapping to a common space is reduced to a single orthogonal transformation that is shared by all variants, and their differences are exclusively explained in terms of their normalization, whitening, re-weighting, de-whitening and dimensionality reduction behavior. This allows us to gain new insights into existing mapping methods, as follows:

- Whitening can bring small improvements, but only if de-whitened appropriately. Our work shows that the implicit de-whitening behavior of both OLS methods (Mikolov, Le, and Sutskever 2013) and CCA methods (Faruqui and Dyer 2014) is flawed.
- Re-weighting is very helpful, but, contrary to most previous work, it should be performed in the target language for the length normalization performed by cosine similarity to be effective in nearest neighbor retrieval. This explains why mapping the target language into the source language performs better than mapping the source language into the target language for regression methods (Shigeto et al. 2015).
- Dimensionality reduction is an extreme form of re-weighting. Even if it was shown to be beneficial with CCA methods (Faruqui and Dyer 2014) and orthogonal methods (Smith et al. 2017), smooth re-weighting gives even better results. Using both of them together is not harmful, bringing further improvements in some cases, and should be the default configuration to try.

Moreover, we also shed light on the relation between mapping methods and retrieval methods when inducing bilingual lexicons:

- The use of alternative retrieval methods to nearest neighbor (Dinu, Lazaridou, and Baroni 2015; Smith et al. 2017) mitigated deficiencies in the implicit re-weighting behavior of previous mapping methods. When re-weighting is properly applied in the target language, inverted softmax (Smith et al. 2017) performs at par with standard nearest neighbor in most cases, while inverted nearest neighbor gives considerably worse results.

Based on these insights, we propose a new variant that obtains the best published results in bilingual lexicon extraction for all the four language pairs tested. We release our implementation as an open source project, which allows to replicate several previous methods as well as our improved variant (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Dinu, Lazaridou, and Baroni 2015; Shigeto et al. 2015; Xing et al. 2015; Zhang et al. 2016; Smith et al. 2017; Artetxe, Labaka, and Agirre 2016). In the future, we would like to incorporate L2 regularization in our framework and extend our analysis to max-margin methods (Lazaridou, Dinu, and Baroni 2015) and non-linear mappings (Lu et al. 2015). Moreover, we would like to introduce hyperparameters to control the intensity of whitening/de-whitening and re-weighting, which we believe could bring further improvements with proper tuning. Finally, we would like to adapt and evaluate our framework in other zero-shot learning scenarios.

Acknowledgments

This research was partially supported by a Google Faculty Award, the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the Basque Government (MODELA KK-2016/00082) and the UPV/EHU (excellence research group). Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

Ammar, W.; Mulcaire, G.; Tsvetkov, Y.; Lample, G.; Dyer, C.; and Smith, N. A. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Artetxe, M.; Labaka, G.; and Agirre, E. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2289–2294. Austin, Texas: Association for Computational Linguistics.

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 451–462. Vancouver, Canada: Association for Computational Linguistics.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 462–471. Gothenburg, Sweden: Association for Computational Linguistics.

Gouws, S.; Bengio, Y.; and Corrado, G. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, 748–756.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*.

Lazaridou, A.; Dinu, G.; and Baroni, M. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 270–280. Beijing, China: Association for Computational Linguistics.

Lu, Y., and Foster, D. P. 2014. Large scale canonical correlation analysis with iterative least squares. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 91–99.

Lu, A.; Wang, W.; Bansal, M.; Gimpel, K.; and Livescu, K. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 250–256. Denver, Colorado: Association for Computational Linguistics.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 151–159. Denver, Colorado: Association for Computational Linguistics.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mogadala, A., and Rettinger, A. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 692–702. San Diego, California: Association for Computational Linguistics.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc. 1410–1418.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11(Sep):2487–2531.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 186–193. ACM.

Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; and Matsumoto, Y. 2015. *Ridge Regression, Hubness, and Zero-Shot Learning*. Cham: Springer International Publishing. 135–151.

Smith, S. L.; Turban, D. H.; Hamblin, S.; and Hammerla, N. Y. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR 2017)*.

Søgaard, A.; Agić, v.; Martínez Alonso, H.; Plank, B.; Bohnet, B.; and Johannsen, A. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1713–1722. Beijing, China: Association for Computational Linguistics.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Vulić, I., and Korhonen, A. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 247–257. Berlin, Germany: Association for Computational Linguistics.

Vulić, I., and Moens, M.-F. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research* 55(1):953–994.

Xing, C.; Wang, D.; Liu, C.; and Lin, Y. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1006–1011. Denver, Colorado: Association for Computational Linguistics.

Zhang, Y.; Gaddy, D.; Barzilay, R.; and Jaakkola, T. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1307–1317. San Diego, California: Association for Computational Linguistics.

Zhao, K.; Hassan, H.; and Auli, M. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1527–1536. Denver, Colorado: Association for Computational Linguistics.

Learning bilingual word embeddings with (almost) no bilingual data

Mikel Artetxe **Gorka Labaka** **Eneko Agirre**
IXA NLP group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

Most methods to learn bilingual word embeddings rely on large parallel corpora, which is difficult to obtain for most language pairs. This has motivated an active research line to relax this requirement, with methods that use document-aligned corpora or bilingual dictionaries of a few thousand words instead. In this work, we further reduce the need of bilingual resources using a very simple self-learning approach that can be combined with any dictionary-based mapping technique. Our method exploits the structural similarity of embedding spaces, and works with as little bilingual evidence as a 25 word dictionary or even an automatically generated list of numerals, obtaining results comparable to those of systems that use richer resources.

1 Introduction

Multilingual word embeddings have attracted a lot of attention in recent times. In addition to having a direct application in inherently crosslingual tasks like machine translation (Zou et al., 2013) and crosslingual entity linking (Tsai and Roth, 2016), they provide an excellent mechanism for transfer learning, where a model trained in a resource-rich language is transferred to a less-resourced one, as shown with part-of-speech tagging (Zhang et al., 2016), parsing (Xiao and Guo, 2014) and document classification (Klementiev et al., 2012).

Most methods to learn these multilingual word embeddings make use of large parallel corpora (Gouws et al., 2015; Luong et al., 2015), but there have been several proposals to relax this requirement, given its scarcity in most language pairs. A possible relaxation is to use document-aligned or label-aligned comparable corpora (Søgaard et al., 2015; Vulić and Moens, 2016; Mogadala and Rettinger, 2016), but large amounts of such corpora are not always available for some language pairs.

An alternative approach that we follow here is to independently train the embeddings for each language on monolingual corpora, and then learn a linear transformation to map the embeddings from one space into the other by minimizing the distances in a bilingual dictionary, usually in the range of a few thousand entries (Mikolov et al., 2013a; Artetxe et al., 2016). However, dictionaries of that size are not readily available for many language pairs, specially those involving less-resourced languages.

In this work, we reduce the need of large bilingual dictionaries to much smaller seed dictionaries. Our method can work with as little as 25 word pairs, which are straightforward to obtain assuming some basic knowledge of the languages involved. The method can also work with trivially generated seed dictionaries of numerals (i.e. 1-1, 2-2, 3-3, 4-4...) making it possible to learn bilingual word embeddings without any real bilingual data. In either case, we obtain very competitive results, comparable to other state-of-the-art methods that make use of much richer bilingual resources.

The proposed method is an extension of existing mapping techniques, where the dictionary is used to learn the embedding mapping and the embedding mapping is used to induce a new dictionary iteratively in a self-learning fashion (see Figure 1). In spite of its simplicity, our analysis of the implicit optimization objective reveals that the method is exploiting the structural similarity of independently trained embeddings.

We analyze previous work in Section 2. Section 3 describes the self-learning framework, while Section 4 presents the experiments. Section 5 analyzes the underlying optimization objective, and Section 6 presents an error analysis.

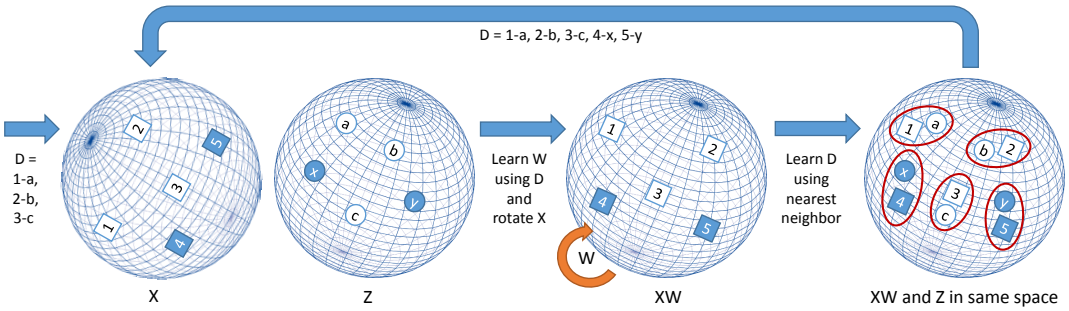


Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D, which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

2 Related work

We will first focus on bilingual embedding mappings, which are the basis of our proposals, and then on other unsupervised and weakly supervised methods to learn bilingual word embeddings.

2.1 Bilingual embedding mappings

Methods to induce bilingual mappings work by independently learning the embeddings in each language using monolingual corpora, and then learning a transformation from one embedding space into the other based on a bilingual dictionary.

The first of such methods is due to Mikolov et al. (2013a), who learn the linear transformation that minimizes the sum of squared Euclidean distances for the dictionary entries. The same optimization objective is used by Zhang et al. (2016), who constrain the transformation matrix to be orthogonal. Xing et al. (2015) incorporate length normalization in the training of word embeddings and maximize the cosine similarity instead, enforcing the orthogonality constraint to preserve the length normalization after the mapping. Finally, Lazaridou et al. (2015) use max-margin optimization with intruder negative sampling.

Instead of learning a single linear transformation from the source language into the target language, Faruqui and Dyer (2014) use canonical correlation analysis to map both languages to a shared vector space. Lu et al. (2015) extend this work and apply deep canonical correlation analysis to learn non-linear transformations.

Artetxe et al. (2016) propose a general framework that clarifies the relation between Mikolov et al. (2013a), Xing et al. (2015), Faruqui and Dyer (2014) and Zhang et al. (2016) as variants of the same core optimization objective, and show that a new variant is able to surpass them all. While most of the previous methods use gradient descent, Artetxe et al. (2016) propose an efficient analytical implementation for those same methods, recently extended by Smith et al. (2017) to incorporate dimensionality reduction.

A prominent application of bilingual embedding mappings, with a direct application in machine translation (Zhao et al., 2015), is bilingual lexicon extraction, which is also the main evaluation method. More specifically, the learned mapping is used to induce the translation of source language words that were missing in the original dictionary, usually by taking their nearest neighbor word in the target language according to cosine similarity, although Dinu et al. (2015) and Smith et al. (2017) propose alternative retrieval methods to address the hubness problem.

2.2 Unsupervised and weakly supervised bilingual embeddings

As mentioned before, our method works with as little as 25 word pairs, while the methods discussed previously use thousands of pairs. The only exception in this regard is the work by Zhang et al. (2016), who only use 10 word pairs with good results on transfer learning for part-of-speech tagging. Our experiments will show that, although their method captures coarse-grained relations, it fails on finer-grained tasks like bilingual lexicon induction.

Bootstrapping methods similar to ours have been previously proposed for traditional count-based vector space models (Peirsman and Padó, 2010; Vulić and Moens, 2013). However, while previous techniques incrementally build a high-

Algorithm 1 Traditional framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)
1: $W \leftarrow$ LEARN_MAPPING(X, Z, D)
2: $D \leftarrow$ LEARN_DICTIONARY(X, Z, W)
3: EVALUATE_DICTIONARY(D)

dimensional model where each axis encodes the co-occurrences with a specific word and its equivalent in the other language, our method works with low-dimensional pre-trained word embeddings, which are more widely used nowadays.

A practical aspect for reducing the need of bilingual supervision is on the design of the seed dictionary. This is analyzed in depth by Vulić and Korhonen (2016), who propose using document-aligned corpora to extract the training dictionary. A more common approach is to rely on shared words and cognates (Peirsman and Padó, 2010; Smith et al., 2017), eliminating the need of bilingual data in practice. Our use of shared numerals exploits the same underlying idea, but relies on even less bilingual evidence and should thus generalize better to distant language pairs.

Miceli Barone (2016) and Cao et al. (2016) go one step further and attempt to learn bilingual embeddings without any bilingual evidence. The former uses adversarial autoencoders (Makhzani et al., 2016), combining an encoder that maps the source language embeddings into the target language, a decoder that reconstructs the original embeddings, and a discriminator that distinguishes mapped embeddings from real target language embeddings, whereas the latter adds a regularization term to the training of word embeddings that pushes the mean and variance of each dimension in different languages close to each other. Although promising, the reported performance in both cases is poor in comparison to other methods.

Finally, the induction of bilingual knowledge from monolingual corpora is closely related to the decipherment scenario, for which models that incorporate word embeddings have also been proposed (Dou et al., 2015). However, decipherment is only concerned with translating text from one language to another and relies on complex statistical models that are designed specifically for that purpose, while our approach is more general and learns task-independent multilingual embeddings.

Algorithm 2 Proposed self-learning framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)
1: **repeat**
2: $W \leftarrow$ LEARN_MAPPING(X, Z, D)
3: $D \leftarrow$ LEARN_DICTIONARY(X, Z, W)
4: **until** convergence criterion
5: EVALUATE_DICTIONARY(D)

3 Proposed self-learning framework

As discussed in Section 2.1, a common evaluation task (and practical application) of bilingual embedding mappings is to induce bilingual lexicons, that is, to obtain the translation of source words that were missing in the training dictionary, which are then compared to a gold standard test dictionary for evaluation. This way, one can say that the seed (train) dictionary is used to learn a mapping, which is then used to induce a better dictionary (at least in the sense that it is larger). Algorithm 1 summarizes this framework.

Following this observation, we propose to use the output dictionary in Algorithm 1 as the input of the same system in a self-learning fashion which, assuming that the output dictionary was indeed better than the original one, should serve to learn a better mapping and, consequently, an even better dictionary the second time. The process can then be repeated iteratively to obtain a hopefully better mapping and dictionary each time until some convergence criterion is met. Algorithm 2 summarizes this alternative framework that we propose.

Our method can be combined with any embedding mapping and dictionary induction technique (see Section 2.1). However, efficiency turns out to be critical for a variety of reasons. First of all, by enclosing the learning logic in a loop, the total training time is increased by the number of iterations. Even more importantly, our framework requires to explicitly build the entire dictionary at each iteration, whereas previous work tends to induce the translation of individual words on-demand later at runtime. Moreover, from the second iteration onwards, it is this induced, full dictionary that has to be used to learn the embedding mapping, and not the considerably smaller seed dictionary as it is typically done. In the following two subsections, we respectively describe the embedding mapping method and the dictionary in-

duction method that we adopt in our work with these efficiency requirements in mind.

3.1 Embedding mapping

As discussed in Section 2.1, most previous methods to learn embedding mappings use variants of gradient descent. Among the more efficient exact alternatives, we decide to adopt the one by Artetxe et al. (2016) for its simplicity and good results as reported in their paper. We next present their method, adapting the formalization to explicitly incorporate the dictionary as required by our self-learning algorithm.

Let X and Z denote the word embedding matrices in two languages so that X_{i*} corresponds to the ith source language word embedding and Z_{j*} corresponds to the jth target language embedding. While Artetxe et al. (2016) assume these two matrices are aligned according to the dictionary, we drop this assumption and represent the dictionary explicitly as a binary matrix D, so that $D_{ij} = 1$ if the ith source language word is aligned with the jth target language word. The goal is then to find the optimal mapping matrix W^* so that the sum of squared Euclidean distances between the mapped source embeddings $X_{i*}W$ and target embeddings Z_{j*} for the dictionary entries D_{ij} is minimized:

$$W^* = \arg\min_W \sum_i \sum_j D_{ij} ||X_{i*}W - Z_{j*}||^2$$

Following Artetxe et al. (2016), we length normalize and mean center the embedding matrices X and Z in a preprocessing step, and constrain W to be an orthogonal matrix (i.e. $WW^T = W^TW = I$), which serves to enforce monolingual invariance, preventing a degradation in monolingual performance while yielding to better bilingual mappings. Under such orthogonality constraint, minimizing the squared Euclidean distance becomes equivalent to maximizing the dot product, so the above optimization objective can be reformulated as follows:

$$W^* = \arg\max_W \operatorname{Tr}\left(XWZ^TD^T\right)$$

where $\operatorname{Tr}(\cdot)$ denotes the trace operator (the sum of all the elements in the main diagonal). The optimal orthogonal solution for this problem is given by $W^* = UV^T$, where $X^TDZ = U\Sigma V^T$ is the singular value decomposition of X^TDZ. Since the dictionary matrix D is sparse, this can be efficiently computed in linear time with respect to the number of dictionary entries.

3.2 Dictionary induction

As discussed in Section 2.1, practically all previous work uses nearest neighbor retrieval for word translation induction based on embedding mappings. In nearest neighbor retrieval, each source language word is assigned the closest word in the target language. In our work, we use the dot product between the mapped source language embeddings and the target language embeddings as the similarity measure, which is roughly equivalent to cosine similarity given that we apply length normalization followed by mean centering as a preprocessing step (see Section 3.1). This way, following the notation in Section 3.1, we set $D_{ij} = 1$ if $j = \operatorname{argmax}_k (X_{i*}W) \cdot Z_{k*}$ and $D_{ij} = 0$ otherwise[1].

While we find that independently computing the similarity measure between all word pairs is prohibitively slow, the computation of the entire similarity matrix XWZ^T can be easily vectorized using popular linear algebra libraries, obtaining big performance gains. However, the resulting similarity matrix is often too large to fit in memory when using large vocabularies. For that reason, instead of computing the entire similarity matrix XWZ^T in a single step, we iteratively compute submatrices of it using vectorized matrix multiplication, find their corresponding maxima each time, and then combine the results.

4 Experiments and results

In this section, we experimentally test the proposed method in bilingual lexicon induction and crosslingual word similarity. Subsection 4.1 describes the experimental settings, while Subsections 4.2 and 4.3 present the results obtained in each of the tasks. The code and resources necessary to reproduce our experiments are available at `https://github.com/artetxem/vecmap`.

4.1 Experimental settings

For easier comparison with related work, we evaluated our mappings on **bilingual lexicon induction** using the public **English-Italian** dataset by Dinu et al. (2015), which includes monolingual word embeddings in both languages together with a bilingual dictionary split in a training set and a

[1] Note that we induce the dictionary entries starting from the source language words. We experimented with other alternatives in development, with minor differences.

test set[2]. The embeddings were trained with the word2vec toolkit with CBOW and negative sampling (Mikolov et al., 2013b)[3], using a 2.8 billion word corpus for English (ukWaC + Wikipedia + BNC) and a 1.6 billion word corpus for Italian (itWaC). The training and test sets were derived from a dictionary built form Europarl word alignments and available at OPUS (Tiedemann, 2012), taking 1,500 random entries uniformly distributed in 5 frequency bins as the test set and the 5,000 most frequent of the remaining word pairs as the training set.

In addition to English-Italian, we selected two other languages from different language families with publicly available resources. We thus created analogous datasets for **English-German** and **English-Finnish**. In the case of German, the embeddings were trained on the 0.9 billion word corpus SdeWaC, which is part of the WaCky collection (Baroni et al., 2009) that was also used for English and Italian. Given that Finnish is not included in this collection, we used the 2.8 billion word Common Crawl corpus provided at WMT 2016[4] instead, which we tokenized using the Stanford Tokenizer (Manning et al., 2014). In addition to that, we created training and test sets for both pairs from their respective Europarl dictionaries from OPUS following the exact same procedure used for English-Italian, and the word embeddings were also trained using the same configuration as Dinu et al. (2015).

Given that the main focus of our work is on **small seed dictionaries**, we created random subsets of 2,500, 1,000, 500, 250, 100, 75, 50 and 25 entries from the original training dictionaries of 5,000 entries. This was done by shuffling once the training dictionaries and taking their first k entries, so it is guaranteed that each dictionary is a strict subset of the bigger dictionaries.

In addition to that, we explored using automatically generated dictionaries as a shortcut to practical unsupervised learning. For that purpose, we created **numeral dictionaries**, consisting of words matching the `[0-9]+` regular expression in both vocabularies (e.g. 1-1, 2-2, 3-3, 1992-1992 etc.). The resulting dictionary had 2772 entries for English-Italian, 2148 for English-German, and 2345 for English-Finnish. While more sophisticated approaches are possible (e.g. involving the edit distance of all words), we believe that this method is general enough that should work with practically any language pair, as Arabic numerals are often used even in languages with a different writing system (e.g. Chinese and Russian).

While bilingual lexicon induction is a standard evaluation task for seed dictionary based methods like ours, it is unsuitable for bilingual corpus based methods, as statistical word alignment already provides a reliable way to derive dictionaries from bilingual corpora and, in fact, this is how the test dictionary itself is built in our case. For that reason, we carried out some experiments in **crosslingual word similarity** as a way to test our method in a different task and allowing to compare it to systems that use richer bilingual data. There are no many crosslingual word similarity datasets, and we used the RG-65 and WordSim-353 crosslingual datasets for English-German and the WordSim-353 crosslingual dataset for English-Italian as published by Camacho-Collados et al. (2015) [5].

As for the **convergence criterion**, we decide to stop training when the improvement on the average dot product for the induced dictionary falls below a given threshold from one iteration to the next. After length normalization, the dot product ranges from -1 to 1, so we decide to set this threshold at 1e-6, which we find to be a very conservative value yet enough that training takes a reasonable amount of time. The curves in the next section confirm that this was a reasonable choice.

This convergence criterion is usually met in less than 100 iterations, each of them taking 5 minutes on a modest desktop computer (Intel Core i5-4670 CPU with 8GiB of RAM), including the induction of a dictionary of 200,000 words at each iteration.

4.2 Bilingual lexicon induction

For the experiments on bilingual lexicon induction, we compared our method with those proposed by Mikolov et al. (2013a), Xing et al. (2015), Zhang et al. (2016) and Artetxe et al. (2016), all of them implemented as part of the framework proposed by the latter. The results ob-

[2] http://clic.cimec.unitn.it/~georgiana.dinu/down/

[3] The context window was set to 5 words, the dimension of the embeddings to 300, the sub-sampling to 1e-05 and the number of negative samples to 10, and the vocabulary was restricted to the 200,000 most frequent words

[4] http://www.statmt.org/wmt16/translation-task.html

[5] http://lcl.uniroma1.it/similarity-datasets/

	English-Italian			English-German			English-Finnish		
	5,000	**25**	**num.**	**5,000**	**25**	**num.**	**5,000**	**25**	**num.**
Mikolov et al. (2013a)	34.93	0.00	0.00	35.00	0.00	0.07	25.91	0.00	0.00
Xing et al. (2015)	36.87	0.00	0.13	41.27	0.07	0.53	28.23	0.07	0.56
Zhang et al. (2016)	36.73	0.07	0.27	40.80	0.13	0.87	28.16	0.14	0.42
Artetxe et al. (2016)	39.27	0.07	0.40	**41.87**	0.13	0.73	**30.62**	0.21	0.77
Our method	**39.67**	**37.27**	**39.40**	40.87	**39.60**	**40.27**	28.72	**28.16**	**26.47**

Table 1: Accuracy (%) on bilingual lexicon induction for different seed dictionaries

tained with the 5,000 entry, 25 entry and the numerals dictionaries for all the 3 language pairs are given in Table 1.

The results for the 5,000 entry dictionaries show that our method is comparable or even better than the other systems. As another reference, the best published results using nearest-neighbor retrieval are due to Lazaridou et al. (2015), who report an accuracy of 40.20% for the full English-Italian dictionary, almost at pair with our system (39.67%).

In any case, the main focus of our work is on smaller dictionaries, and it is under this setting that our method really stands out. The 25 entry and numerals columns in Table 1 show the results for this setting, where all previous methods drop dramatically, falling below 1% accuracy in all cases. The method by Zhang et al. (2016) also obtains poor results with small dictionaries, which reinforces our hypothesis in Section 2.2 that their method can only capture coarse-grain bilingual relations for small dictionaries. In contrast, our proposed method obtains very competitive results for all dictionaries, with a difference of only 1-2 points between the full dictionary and both the 25 entry dictionary and the numerals dictionary in all three languages. Figure 2 shows the curve of the English-Italian accuracy for different seed dictionary sizes, confirming this trend.

Finally, it is worth mentioning that, even if all the three language pairs show the same general behavior, there are clear differences in their absolute accuracy numbers, which can be attributed to the linguistic proximity of the languages involved. In particular, the results for English-Finnish are about 10 points below the rest, which is explained by the fact that Finnish is a non-indoeuropean agglutinative language, making the task considerably more difficult for this language pair. In this regard, we believe that the good results with small dictionaries are a strong indication of the robustness of our method, showing that it is able to learn good bilingual mappings from very little bilingual evidence even for distant language pairs where the structural similarity of the embedding spaces is presumably weaker.

4.3 Crosslingual word similarity

In addition to the baseline systems in Section 4.2, in the crosslingual similarity experiments we also tested the method by Luong et al. (2015), which is the state-of-the-art for bilingual word embeddings based on parallel corpora (Upadhyay et al., 2016)[6]. As this method is an extension of word2vec, we used the same hyperparameters as for the monolingual embeddings when possible (see Section 4.1), and leave the default ones otherwise. We used Europarl as our parallel corpus to train this method as done by the authors, which consists of nearly 2 million parallel sentences.

As shown in the results in Table 2, our method obtains the best results in all cases, surpassing the rest of the dictionary-based methods by 1-3 points depending on the dataset. But, most importantly, it does not suffer from any significant degradation for using smaller dictionaries and, in fact, our method gets better results using the 25 entry dictionary or the numeral list as the only bilingual evidence than any of the baseline systems using much richer resources.

The relatively poor results of Luong et al. (2015) can be attributed to the fact that the dictionary based methods make use of much bigger monolingual corpora, while methods based on parallel corpora are restricted to smaller corpora. However, it is not clear how to introduce monolingual corpora on those methods. We did run some experiments with BilBOWA (Gouws et al., 2015), which supports training in monolingual corpora in addition to bilingual corpora, but obtained very poor results[7]. All in all, our experiments show

[6] We also tested English-German pre-trained embeddings from Klementiev et al. (2012) and Chandar A P et al. (2014). They both had coverage problems that made the results hard to compare, and, when considering the correlations for the word pairs in their vocabulary, their performance was poor.

[7] Upadhyay et al. (2016) report similar problems using

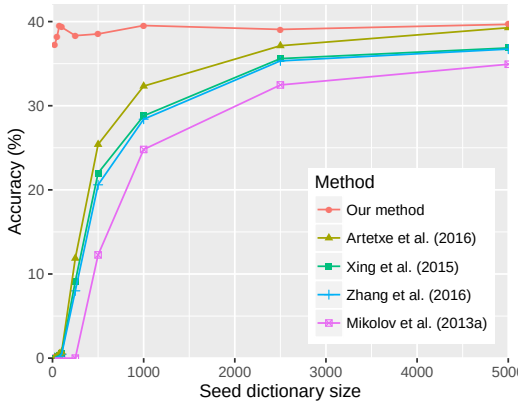


Figure 2: Accuracy on English-Italian bilingual lexicon induction for different seed dictionaries

that it is better to use large monolingual corpora in combination with very little bilingual data rather than a bilingual corpus of a standard size alone.

5 Global optimization objective

It might seem somehow surprising at first that, as seen in the previous section, our simple self-learning approach is able to learn high quality bilingual embeddings from small seed dictionaries instead of falling in degenerated solutions. In this section, we try to shed light on our approach, and give empirical evidence supporting our claim.

More concretely, we argue that, for the embedding mapping and dictionary induction methods described in Section 3, the proposed self-learning framework is implicitly solving the following global optimization problem[8]:

$$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$$
$$\text{s.t.} \quad WW^T = W^TW = I$$

Contrary to the optimization objective for W in Section 3.1, the global optimization objective does not refer to any dictionary, and maximizes the similarity between each source language word and its closest target language word. Intuitively, a random solution would map source language embeddings to seemingly random locations in the target language space, and it would thus be unlikely that they have any target language word nearby, making the optimization value small. In contrast, a good solution would map source language words close to their translation equivalents in the target language space, and they would thus have their corresponding embeddings nearby, making the optimization value large. While it is certainly possible to build degenerated solutions that take high optimization values for small subsets of the vocabulary, we think that the structural similarity between independently trained embedding spaces in different languages is strong enough that optimizing this function yields to meaningful bilingual mappings when the size of the vocabulary is much larger than the dimensionality of the embeddings.

The reasoning for how the self-learning framework is optimizing this objective is as follows. At the end of each iteration, the dictionary D is updated to assign, for the current mapping W, each source language word to its closest target language word. This way, when we update W to maximize the average similarity of these dictionary entries at the beginning of the next iteration, it is guaranteed that the value of the optimization objective will improve (or at least remain the same). The reason is that the average similarity between each word and what were previously the closest words will be improved if possible, as this is what the updated W directly optimizes (see Section 3.1). In addition to that, it is also possible that, for some source words, some other target words get closer after the update. Thanks to this, our self-learning algorithm is guaranteed to converge to a local optimum of the above global objective, behaving like an alternating optimization algorithm for it.

It is interesting to note that the above reasoning is valid no matter what the the initial solution is, and, in fact, the global optimization objective does not depend on the seed dictionary nor any other

		IT	DE	
	Bi. data	**WS**	**RG**	**WS**
Luong et al. (2015)	Europarl	.331	.335	.356
Mikolov et al. (2013a)	5k dict	.627	.643	.528
Xing et al. (2015)	5k dict	.614	.700	.595
Zhang et al. (2016)	5k dict	.616	.704	.596
Artetxe et al. (2016)	5k dict	.617	.716	.597
Our method	5k dict	.624	.742	**.616**
	25 dict	.626	**.749**	.612
	num.	**.628**	.739	.604

Table 2: Spearman correlations on English-Italian and English-German crosslingual word similarity

BilBOWA.

[8]While we restrict our formal analysis to the embedding mapping and dictionary induction method that we use, the general reasoning should be valid for other choices as well.

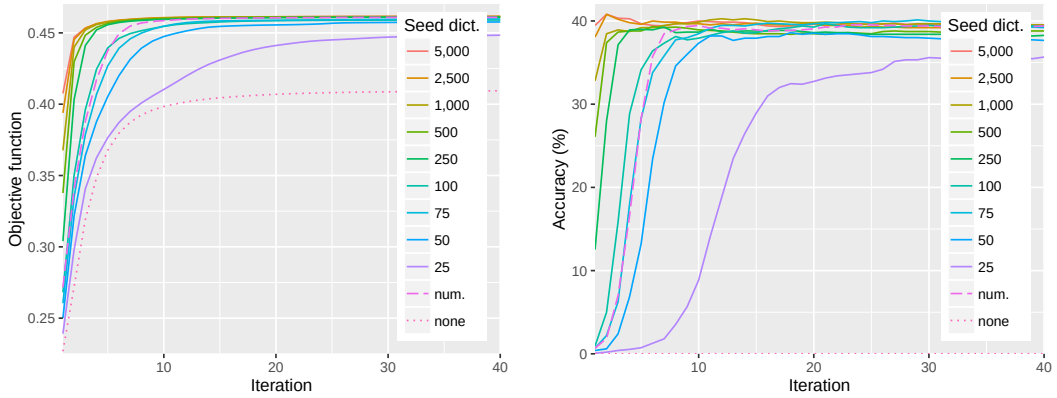


Figure 3: Learning curve on English-Italian according to the global objective function (left) and the accuracy on bilingual lexicon induction (right)

bilingual resource. For that reason, it should be possible to use a random initialization instead of a small seed dictionary. However, we empirically observe that this works poorly in practice, as our algorithm tends to get stuck in poor local optima when the initial solution is not good enough.

The general behavior of our method is reflected in Figure 3, which shows the learning curve for different seed dictionaries according to both the objective function and the accuracy on bilingual lexicon induction. As it can be seen, the objective function is improved from iteration to iteration and converges to a local optimum just as expected. At the same time, the learning curves show a strong correlation between the optimization objective and the accuracy, as it can be clearly observed that improving the former leads to an improvement of the latter, confirming our explanations. Regarding random initialization, the figure shows that the algorithm gets stuck in a poor local optimum of the objective function, which is the reason of the bad performance (0% accuracy) on bilingual lexicon induction, but the proposed optimization objective itself seems to be adequate.

Finally, we empirically observe that our algorithm learns similar mappings no matter what the seed dictionary was. We first repeated our experiments on English-Italian bilingual lexicon induction for 5 different dictionaries of 25 entries, obtaining an average accuracy of 38.15% and a standard deviation of only 0.75%. In addition to that, we observe that the overlap between the predictions made when starting with the full dictionary and the numerals dictionary is 76.00% (60.00% for the 25 entry dictionary). At the same time, 37.00% of the test cases are correctly solved by both instances, and it is only 5.07% of the test cases that one of them gets right and the other wrong (34.00% and 8.94% for the 25 entry dictionary). This suggests that our algorithm tends to converge to similar solutions even for disjoint seed dictionaries, which is in line with our view that we are implicitly optimizing an objective that is independent from the seed dictionary, yet a seed dictionary is necessary to build a good enough initial solution to avoid getting stuck in poor local optima. For that reason, it is likely that better methods to tackle this optimization problem would allow learning bilingual word embeddings without any bilingual evidence at all and, in this regard, we believe that our work opens exciting opportunities for future research.

6 Error analysis

So as to better understand the behavior of our system, we performed an error analysis of its output in English-Italian bilingual lexicon induction when starting with the 5,000 entry, the 25 entry and the numeral dictionaries in comparison with the baseline method of Artetxe et al. (2016) with the 5,000 entry dictionary. For that purpose, we took 100 random examples from the test set in the [1-5K] frequency bin, another 100 from the [5K-20K] frequency bin and 30 from the [100K-200K] frequency bin, and manually analyzed each of the errors made by all the 4 different variants.

Our analysis first reveals that, in all the cases, about a third of the translations taken as erroneous according to the gold standard are not so in real-

ity. This corresponds to both different morphological variants of the gold standard translations (e.g. *dichiarato/dichiarò*) and other valid translations that were missing in the gold standard (e.g. *climb* → *salita* instead of the gold standard *scalato*). This phenomenon is considerably more pronounced in the first frequency bins, which already have a much higher accuracy according to the gold standard.

As for the actual errors, we observe that nearly a third of them correspond to named entities for all the different variants. Interestingly, the vast majority of the proposed translations in these cases are also named entities (e.g. *Ryan* → *Jason*, *John* → *Paolo*), which are often highly related to the original ones (e.g. *Volvo* → *BMW*, *Olympus* → *Nikon*). While these are clear errors, it is understandable that these methods are unable to discriminate between named entities to this degree based solely on the distributional hypothesis, in particular when it comes to common proper names (e.g. *John*, *Andy*), and one could design alternative strategies to address this issue like taking the edit distance as an additional signal.

For the remaining errors, all systems tend to propose translations that have some degree of relationship with the correct ones, including near-synonyms (e.g. *guidelines* → *raccomandazioni*), antonyms (e.g. *sender* → *destinatario*) and words in the same semantic field (e.g. *nominalism* → *intuizionismo* / *innatismo*, which are all philosophical doctrines). However, there are also a few instances where the relationship is weak or unclear (e.g. *loch* → *giardini*, *sweep* → *serrare*). We also observe a few errors that are related to multiwords or collocations (e.g. *carrier* → *aereo*, presumably related to the multiword *air carrier* / *linea aerea*), as well as some rare word that is repeated across many translations (*Ferruzzi*), which could be attributed to the hubness problem (Dinu et al., 2015; Lazaridou et al., 2015).

All in all, our error analysis reveals that the baseline method of Artetxe et al. (2016) and the proposed algorithm tend to make the same kind of errors regardless of the seed dictionary used by the latter, which reinforces our interpretation in the previous section regarding an underlying optimization objective that is independent from any training dictionary. Moreover, it shows that the quality of the learned mappings is much better than what the raw accuracy numbers might suggest, encouraging the incorporation of these techniques in other applications.

7 Conclusions and future work

In this work, we propose a simple self-learning framework to learn bilingual word embedding mappings in combination with any embedding mapping and dictionary induction technique. Our experiments on bilingual lexicon induction and crosslingual word similarity show that our method is able to learn high quality bilingual embeddings from as little bilingual evidence as a 25 word dictionary or an automatically generated list of numerals, obtaining results that are competitive with state-of-the-art systems using much richer bilingual resources like larger dictionaries or parallel corpora. In spite of its simplicity, a more detailed analysis shows that our method is implicitly optimizing a meaningful objective function that is independent from any bilingual data which, with a better optimization method, might allow to learn bilingual word embeddings in a completely unsupervised manner.

In the future, we would like to delve deeper into this direction and fine-tune our method so it can reliably learn high quality bilingual word embeddings without any bilingual evidence at all. In addition to that, we would like to explore non-linear transformations (Lu et al., 2015) and alternative dictionary induction methods (Dinu et al., 2015; Smith et al., 2017). Finally, we would like to apply our model in the decipherment scenario (Dou et al., 2015).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and Flavio Merenda for his help with the error analysis.

This research was partially supported by a Google Faculty Award, the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the Basque Government (MODELA KK-2016/00082) and the UPV/EHU (excellence research group). Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2289–2294. https://aclweb.org/anthology/D16-1250.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation* 43(3):209–226.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 1–7. http://www.aclweb.org/anthology/P15-2001.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1818–1827. http://aclweb.org/anthology/C16-1171.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 1853–1861. http://papers.nips.cc/paper/5270-an-autoencoder-approach-to-learning-bilingual-word-representations.pdf.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 836–845. http://www.aclweb.org/anthology/P15-1081.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 462–471. http://www.aclweb.org/anthology/E14-1049.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*. pages 748–756.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 1459–1474. http://www.aclweb.org/anthology/C12-1089.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 270–280. http://www.aclweb.org/anthology/P15-1027.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 250–256. http://www.aclweb.org/anthology/N15-1028.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics, Denver, Colorado, pages 151–159. http://www.aclweb.org/anthology/W15-1521.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial autoencoders. In *Proceedings of the 4rd International Conference on Learning Representations (ICLR 2016), workshop track*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. http://www.aclweb.org/anthology/P14-5010.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, pages 121–126. http://anthology.aclweb.org/W16-1614.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 692–702. http://www.aclweb.org/anthology/N16-1083.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 921–929. http://www.aclweb.org/anthology/N10-1135.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), conference track*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1713–1722. http://www.aclweb.org/anthology/P15-1165.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 589–598. http://www.aclweb.org/anthology/N16-1072.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1661–1670. http://www.aclweb.org/anthology/P16-1157.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 247–257. http://www.aclweb.org/anthology/P16-1024.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1613–1624. http://www.aclweb.org/anthology/D13-1168.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research* 55(1):953–994.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 119–129. http://www.aclweb.org/anthology/W14-1613.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1006–1011. http://www.aclweb.org/anthology/N15-1104.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies. Association for Computational Linguistics, San Diego, California, pages 1307–1317. http://www.aclweb.org/anthology/N16-1156.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1527–1536. http://www.aclweb.org/anthology/N15-1176.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1393–1398. http://www.aclweb.org/anthology/D13-1141.

A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings

Mikel Artetxe and **Gorka Labaka** and **Eneko Agirre**
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

Recent work has managed to learn cross-lingual word embeddings without parallel data by mapping monolingual embeddings to a shared space through adversarial training. However, their evaluation has focused on favorable conditions, using comparable corpora or closely-related languages, and we show that they often fail in more realistic scenarios. This work proposes an alternative approach based on a fully unsupervised initialization that explicitly exploits the structural similarity of the embeddings, and a robust self-learning algorithm that iteratively improves this solution. Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems. Our implementation is released as an open source project at https://github.com/artetxem/vecmap.

1 Introduction

Cross-lingual embedding mappings have shown to be an effective way to learn bilingual word embeddings (Mikolov et al., 2013; Lazaridou et al., 2015). The underlying idea is to independently train the embeddings in different languages using monolingual corpora, and then map them to a shared space through a linear transformation. This allows to learn high-quality cross-lingual representations without expensive supervision, opening new research avenues like unsupervised neural machine translation (Artetxe et al., 2018b; Lample et al., 2018).

While most embedding mapping methods rely on a small seed dictionary, adversarial training has recently produced exciting results in fully unsupervised settings (Zhang et al., 2017a,b; Conneau et al., 2018). However, their evaluation has focused on particularly favorable conditions, limited to closely-related languages or comparable Wikipedia corpora. When tested on more realistic scenarios, we find that they often fail to produce meaningful results. For instance, none of the existing methods works in the standard English-Finnish dataset from Artetxe et al. (2017), obtaining translation accuracies below 2% in all cases (see Section 5).

On another strand of work, Artetxe et al. (2017) showed that an iterative self-learning method is able to bootstrap a high quality mapping from very small seed dictionaries (as little as 25 pairs of words). However, their analysis reveals that the self-learning method gets stuck in poor local optima when the initial solution is not good enough, thus failing for smaller training dictionaries.

In this paper, we follow this second approach and propose a new unsupervised method to build an initial solution without the need of a seed dictionary, based on the observation that, given the similarity matrix of all words in the vocabulary, each word has a different distribution of similarity values. Two equivalent words in different languages should have a similar distribution, and we can use this fact to induce the initial set of word pairings (see Figure 1). We combine this initialization with a more robust self-learning method, which is able to start from the weak initial solution and iteratively improve the mapping. Coupled together, we provide a fully unsupervised cross-lingual mapping method that is effective in realistic settings, converges to a good solution in all cases tested, and sets a new state-of-the-art in bilingual lexicon extraction, even surpassing previous supervised methods.

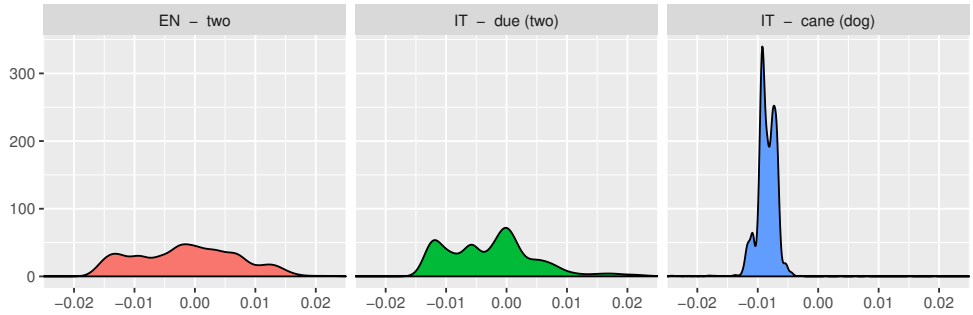


Figure 1: Motivating example for our unsupervised initialization method, showing the similarity distributions of three words (corresponding to the smoothed density estimates from the normalized square root of the similarity matrices as defined in Section 3.2). Equivalent translations (two and *due*) have more similar distributions than non-related words (two and *cane* - meaning dog). This observation is used to build an initial solution that is later improved through self-learning.

2 Related work

Cross-lingual embedding mapping methods work by independently training word embeddings in two languages, and then mapping them to a shared space using a linear transformation.

Most of these methods are **supervised**, and use a bilingual dictionary of a few thousand entries to learn the mapping. Existing approaches can be classified into regression methods, which map the embeddings in one language using a least-squares objective (Mikolov et al., 2013; Shigeto et al., 2015; Dinu et al., 2015), canonical methods, which map the embeddings in both languages to a shared space using canonical correlation analysis and extensions of it (Faruqui and Dyer, 2014; Lu et al., 2015), orthogonal methods, which map the embeddings in one or both languages under the constraint of the transformation being orthogonal (Xing et al., 2015; Artetxe et al., 2016; Zhang et al., 2016; Smith et al., 2017), and margin methods, which map the embeddings in one language to maximize the margin between the correct translations and the rest of the candidates (Lazaridou et al., 2015). Artetxe et al. (2018a) showed that many of them could be generalized as part of a multi-step framework of linear transformations.

A related research line is to adapt these methods to the **semi-supervised** scenario, where the training dictionary is much smaller and used as part of a bootstrapping process. While similar ideas where already explored for traditional count-based vector space models (Peirsman and Padó, 2010; Vulić and Moens, 2013), Artetxe et al. (2017) brought this approach to pre-trained low-dimensional word embeddings, which are more widely used nowadays. More concretely, they proposed a self-learning approach that alternates the mapping and dictionary induction steps iteratively, obtaining results that are comparable to those of supervised methods when starting with only 25 word pairs.

A practical approach for reducing the need of bilingual supervision is to design **heuristics to build the seed dictionary**. The role of the seed lexicon in learning cross-lingual embedding mappings is analyzed in depth by Vulić and Korhonen (2016), who propose using document-aligned corpora to extract the training dictionary. A more common approach is to rely on shared words and cognates (Peirsman and Padó, 2010; Smith et al., 2017), while Artetxe et al. (2017) go further and restrict themselves to shared numerals. However, while these approaches are meant to eliminate the need of bilingual data in practice, they also make strong assumptions on the writing systems of languages (e.g. that they all use a common alphabet or Arabic numerals). Closer to our work, a recent line of **fully unsupervised** approaches drops these assumptions completely, and attempts to learn cross-lingual embedding mappings based on distributional information alone. For that purpose, existing methods rely on adversarial training. This was first proposed by Miceli Barone (2016), who combine an encoder that maps source language embeddings into the target language, a decoder that reconstructs the source language embeddings from the mapped embeddings, and a discriminator that discriminates between the mapped embeddings and the true target language embed-

dings. Despite promising, they conclude that their model "is not competitive with other cross-lingual representation approaches". Zhang et al. (2017a) use a very similar architecture, but incorporate additional techniques like noise injection to aid training and report competitive results on bilingual lexicon extraction. Conneau et al. (2018) drop the reconstruction component, regularize the mapping to be orthogonal, and incorporate an iterative refinement process akin to self-learning, reporting very strong results on a large bilingual lexicon extraction dataset. Finally, Zhang et al. (2017b) adopt the earth mover's distance for training, optimized through a Wasserstein generative adversarial network followed by an alternating optimization procedure. However, all this previous work used comparable Wikipedia corpora in most experiments and, as shown in Section 5, face difficulties in more challenging settings.

3 Proposed method

Let X and Z be the word embedding matrices in two languages, so that their ith row X_{i*} and Z_{i*} denote the embeddings of the ith word in their respective vocabularies. Our goal is to learn the linear transformation matrices W_X and W_Z so the mapped embeddings XW_X and ZW_Z are in the same cross-lingual space. At the same time, we aim to build a dictionary between both languages, encoded as a sparse matrix D where $D_{ij} = 1$ if the jth word in the target language is a translation of the ith word in the source language.

Our proposed method consists of four sequential steps: a pre-processing that normalizes the embeddings (§3.1), a fully unsupervised initialization scheme that creates an initial solution (§3.2), a robust self-learning procedure that iteratively improves this solution (§3.3), and a final refinement step that further improves the resulting mapping through symmetric re-weighting (§3.4).

3.1 Embedding normalization

Our method starts with a pre-processing that length normalizes the embeddings, then mean centers each dimension, and then length normalizes them again. The first two steps have been shown to be beneficial in previous work (Artetxe et al., 2016), while the second length normalization guarantees the final embeddings to have a unit length. As a result, the dot product of any two embeddings is equivalent to their cosine similarity and directly related to their Euclidean distance[1], and can be taken as a measure of their similarity.

3.2 Fully unsupervised initialization

The underlying difficulty of the mapping problem in its unsupervised variant is that the word embedding matrices X and Z are unaligned across both axes: neither the ith vocabulary item X_{i*} and Z_{i*} nor the jth dimension of the embeddings X_{*j} and Z_{*j} are aligned, so there is no direct correspondence between both languages. In order to overcome this challenge and build an initial solution, we propose to first construct two alternative representations X' and Z' that are aligned across their jth dimension X'_{*j} and Z'_{*j}, which can later be used to build an initial dictionary that aligns their respective vocabularies.

Our approach is based on a simple idea: while the axes of the original embeddings X and Z are different in nature, both axes of their corresponding similarity matrices $M_X = XX^T$ and $M_Z = ZZ^T$ correspond to words, which can be exploited to reduce the mismatch to a single axis. More concretely, assuming that the embedding spaces are perfectly isometric, the similarity matrices M_X and M_Z would be equivalent up to a permutation of their rows and columns, where the permutation in question defines the dictionary across both languages. In practice, the isometry requirement will not hold exactly, but it can be assumed to hold approximately, as the very same problem of mapping two embedding spaces without supervision would otherwise be hopeless. Based on that, one could try every possible permutation of row and column indices to find the best match between M_X and M_Z, but the resulting combinatorial explosion makes this approach intractable.

In order to overcome this problem, we propose to first sort the values in each row of M_X and M_Z, resulting in matrices $\mathrm{sorted}(M_X)$ and $\mathrm{sorted}(M_Z)$[2]. Under the strict isometry condition, equivalent words would get the exact same vector across languages, and thus, given a word and its row in $\mathrm{sorted}(M_X)$, one could apply nearest neighbor retrieval over the rows of $\mathrm{sorted}(M_Z)$ to find its corresponding translation.

On a final note, given the singular value decomposition $X = USV^T$, the similarity matrix

[1] Given two length normalized vectors u and v, $u \cdot v = \cos(u, v) = 1 - ||u - v||^2/2$.

[2] Note that the values in each row are sorted independently from other rows.

is $M_X = US^2U^T$. As such, its square root $\sqrt{M_X} = USU^T$ is closer in nature to the original embeddings, and we also find it to work better in practice. We thus compute $\text{sorted}(\sqrt{M_X})$ and $\text{sorted}(\sqrt{M_Z})$ and normalize them as described in Section 3.1, yielding the two matrices X' and Z' that are later used to build the initial solution for self-learning (see Section 3.3).

In practice, the isometry assumption is strong enough so the above procedure captures some cross-lingual signal. In our English-Italian experiments, the average cosine similarity across the gold standard translation pairs is 0.009 for a random solution, 0.582 for the optimal supervised solution, and 0.112 for the mapping resulting from this initialization. While the latter is far from being useful on its own (the accuracy of the resulting dictionary is only 0.52%), it is substantially better than chance, and it works well as an initial solution for the self-learning method described next.

3.3 Robust self-learning

Previous work has shown that self-learning can learn high-quality bilingual embedding mappings starting with as little as 25 word pairs (Artetxe et al., 2017). In this method, training iterates through the following two steps until convergence:

1. Compute the optimal orthogonal mapping maximizing the similarities for the current dictionary D:

$$\arg\max_{W_X, W_Z} \sum_i \sum_j D_{ij}((X_{i*}W_X) \cdot (Z_{j*}W_Z))$$

   An optimal solution is given by $W_X = U$ and $W_Z = V$, where $USV^T = X^TDZ$ is the singular value decomposition of X^TDZ.

2. Compute the optimal dictionary over the similarity matrix of the mapped embeddings $XW_XW_Z^TZ^T$. This typically uses nearest neighbor retrieval from the source language into the target language, so $D_{ij} = 1$ if $j = \text{argmax}_k\,(X_{i*}W_X) \cdot (Z_{k*}W_Z)$ and $D_{ij} = 0$ otherwise.

The underlying optimization objective is independent from the initial dictionary, and the algorithm is guaranteed to converge to a local optimum of it. However, the method does not work if starting from a completely random solution, as it tends to get stuck in poor local optima in that case. For that reason, we use the unsupervised initialization procedure at Section 3.2 to build an initial solution. However, simply plugging in both methods did not work in our preliminary experiments, as the quality of this initial method is not good enough to avoid poor local optima. For that reason, we next propose some key improvements in the dictionary induction step to make self-learning more robust and learn better mappings:

- **Stochastic dictionary induction**. In order to encourage a wider exploration of the search space, we make the dictionary induction stochastic by randomly keeping some elements in the similarity matrix with probability p and setting the remaining ones to 0. As a consequence, the smaller the value of p is, the more the induced dictionary will vary from iteration to iteration, thus enabling to escape poor local optima. So as to find a fine-grained solution once the algorithm gets into a good region, we increase this value during training akin to simulated annealing, starting with $p = 0.1$ and doubling this value every time the objective function at step 1 above does not improve more than $\epsilon = 10^{-6}$ for 50 iterations.

- **Frequency-based vocabulary cutoff**. The size of the similarity matrix grows quadratically with respect to that of the vocabularies. This does not only increase the cost of computing it, but it also makes the number of possible solutions grow exponentially[3], presumably making the optimization problem harder. Given that less frequent words can be expected to be noisier, we propose to restrict the dictionary induction process to the k most frequent words in each language, where we find $k = 20,000$ to work well in practice.

- **CSLS retrieval**. Dinu et al. (2015) showed that nearest neighbor suffers from the hubness problem. This phenomenon is known to occur as an effect of the curse of dimensionality, and causes a few points (known as *hubs*) to be nearest neighbors of many other points (Radovanović et al., 2010a,b). Among the existing solutions to penalize the similarity score of hubs, we adopt the Cross-domain

[3]There are m^n possible combinations that go from a source vocabulary of n entries to a target vocabulary of m entries.

Similarity Local Scaling (CSLS) from Conneau et al. (2018). Given two mapped embeddings x and y, the idea of CSLS is to compute $\mathrm{r_T}(x)$ and $\mathrm{r_S}(y)$, the average cosine similarity of x and y for their k nearest neighbors in the other language, respectively. Having done that, the corrected score $\mathrm{CSLS}(x, y) = 2\cos(x, y) - \mathrm{r_T}(x) - \mathrm{r_S}(y)$. Following the authors, we set $k = 10$.

- **Bidirectional dictionary induction**. When the dictionary is induced from the source into the target language, not all target language words will be present in it, and some will occur multiple times. We argue that this might accentuate the problem of local optima, as repeated words might act as strong attractors from which it is difficult to escape. In order to mitigate this issue and encourage diversity, we propose inducing the dictionary in both directions and taking their corresponding concatenation, so $D = D_{X \to Z} + D_{Z \to X}$.

In order to build the **initial dictionary**, we compute X' and Z' as detailed in Section 3.2 and apply the above procedure over them. As the only difference, this first solution does not use the stochastic zeroing in the similarity matrix, as there is no need to encourage diversity (X' and Z' are only used once), and the threshold for vocabulary cutoff is set to $k = 4,000$, so X' and Z' can fit in memory. Having computed the initial dictionary, X' and Z' are discarded, and the remaining iterations are performed over the original embeddings X and Z.

3.4 Symmetric re-weighting

As part of their multi-step framework, Artetxe et al. (2018a) showed that re-weighting the target language embeddings according to the cross-correlation in each component greatly improved the quality of the induced dictionary. Given the singular value decomposition $USV^T = X^T DZ$, this is equivalent to taking $W_X = U$ and $W_Z = VS$, where X and Z are previously whitened applying the linear transformations $(X^T X)^{-\frac{1}{2}}$ and $(Z^T Z)^{-\frac{1}{2}}$, and later de-whitened applying $U^T (X^T X)^{\frac{1}{2}} U$ and $V^T (Z^T Z)^{\frac{1}{2}} V$.

However, re-weighting also accentuates the problem of local optima when incorporated into self-learning as, by increasing the relevance of dimensions that best match for the current solution, it discourages to explore other regions of the search space. For that reason, we propose using it as a final step once self-learning has converged to a good solution. Unlike Artetxe et al. (2018a), we apply re-weighting symmetrically in both languages, taking $W_X = US^{\frac{1}{2}}$ and $W_Z = VS^{\frac{1}{2}}$. This approach is neutral in the direction of the mapping, and gives good results as shown in our experiments.

4 Experimental settings

Following common practice, we evaluate our method on **bilingual lexicon extraction**, which measures the accuracy of the induced dictionary in comparison to a gold standard.

As discussed before, **previous evaluation** has focused on favorable conditions. In particular, existing unsupervised methods have almost exclusively been tested on Wikipedia corpora, which is comparable rather than monolingual, exposing a strong cross-lingual signal that is not available in strictly unsupervised settings. In addition to that, some datasets comprise unusually small embeddings, with only 50 dimensions and around 5,000-10,000 vocabulary items (Zhang et al., 2017a,b). As the only exception, Conneau et al. (2018) report positive results on the English-Italian dataset of Dinu et al. (2015) in addition to their main experiments, which are carried out in Wikipedia. While this dataset does use strictly monolingual corpora, it still corresponds to a pair of two relatively close indo-european languages.

In order to get a wider picture of how our method compares to previous work in different conditions, including more challenging settings, we carry out our experiments in the widely used **dataset** of Dinu et al. (2015) and the subsequent extensions of Artetxe et al. (2017, 2018a), which together comprise English-Italian, English-German, English-Finnish and English-Spanish. More concretely, the dataset consists of 300-dimensional CBOW embeddings trained on WacKy crawling corpora (English, Italian, German), Common Crawl (Finnish) and WMT News Crawl (Spanish). The gold standards were derived from dictionaries built from Europarl word alignments and available at OPUS (Tiedemann, 2012), split in a test set of 1,500 entries and a training set of 5,000 that we do not use in our experiments. The datasets are freely available. As a non-european agglutinative language, the English-Finnish pair is particularly challeng-

	ES-EN				IT-EN				TR-EN			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Zhang et al. (2017a), $\lambda = 1$	71.43	68.18	**10**	13.2	60.38	56.45	**10**	12.3	0.00	0.00	0	13.0
Zhang et al. (2017a), $\lambda = 10$	70.24	66.37	**10**	13.0	57.64	52.60	**10**	12.6	21.07	17.95	**10**	13.2
Conneau et al. (2018), code	76.18	75.82	**10**	25.1	**67.32**	**67.00**	**10**	25.9	32.64	14.34	5	25.3
Conneau et al. (2018), paper	76.15	75.81	**10**	25.1	67.21	60.22	9	25.5	29.79	16.48	7	25.5
Proposed method	**76.43**	**76.28**	**10**	**0.6**	66.96	66.92	**10**	**0.9**	**36.10**	**35.93**	**10**	**1.7**

Table 1: Results of unsupervised methods on the dataset of Zhang et al. (2017a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

	EN-IT				EN-DE				EN-FI				EN-ES			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Zhang et al. (2017a), $\lambda = 1$	0.00	0.00	0	47.0	0.00	0.00	0	47.0	0.00	0.00	0	45.4	0.00	0.00	0	44.3
Zhang et al. (2017a), $\lambda = 10$	0.00	0.00	0	46.6	0.00	0.00	0	46.0	0.07	0.01	0	44.9	0.07	0.01	0	43.0
Conneau et al. (2018), code	45.40	13.55	3	46.1	47.27	42.15	9	45.4	1.62	0.38	0	44.4	36.20	21.23	6	45.3
Conneau et al. (2018), paper	45.27	9.10	2	45.4	0.07	0.01	0	45.0	0.07	0.01	0	44.7	35.47	7.09	2	44.9
Proposed method	**48.53**	**48.13**	**10**	**8.9**	**48.47**	**48.19**	**10**	**7.3**	**33.50**	**32.63**	**10**	**12.9**	**37.60**	**37.33**	**10**	**9.1**

Table 2: Results of unsupervised methods on the dataset of Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

ing due to the linguistic distance between them. For completeness, we also test our method in the Spanish-English, Italian-English and Turkish-English datasets of Zhang et al. (2017a), which consist of 50-dimensional CBOW embeddings trained on Wikipedia, as well as gold standard dictionaries[4] from Open Multilingual WordNet (Spanish-English and Italian-English) and Google Translate (Turkish-English). The lower dimensionality and comparable corpora make an easier scenario, although it also contains a challenging pair of distant languages (Turkish-English).

Our method is implemented in Python using NumPy and CuPy. Together with it, we also test the **methods** of Zhang et al. (2017a) and Conneau et al. (2018) using the publicly available implementations from the authors[5]. Given that Zhang et al. (2017a) report using a different value of their hyperparameter λ for different language pairs ($\lambda = 10$ for English-Turkish and $\lambda = 1$ for the rest), we test both values in all our experiments to better understand its effect. In the case of Conneau et al. (2018), we test both the default hyperparameters in the source code as well as those reported in the paper, with iterative refinement activated in both cases. Given the instability of these methods, we perform 10 runs for each, and report the best and average accuracies, the number of successful runs (those with >5% accuracy) and the average runtime. All the experiments were run in a single Nvidia Titan Xp.

5 Results and discussion

We first present the main results (§5.1), then the comparison to the state-of-the-art (§5.2), and finally ablation tests to measure the contribution of each component (§5.3).

5.1 Main results

We report the results in the dataset of Zhang et al. (2017a) at Table 1. As it can be seen, the proposed method performs at par with that of Conneau et al. (2018) both in Spanish-English and Italian-English, but gets substantially better results in the more challenging Turkish-English pair. While we are able to reproduce the results reported by Zhang et al. (2017a), their method gets the worst results of all by a large margin. Another disadvantage of that model is that different

[4]The test dictionaries were obtained through personal communication with the authors. The rest of the language pairs were left out due to licensing issues.

[5]Despite our efforts, Zhang et al. (2017b) was left out because: 1) it does not create a one-to-one dictionary, thus difficulting direct comparison, 2) it depends on expensive proprietary software 3) its computational cost is orders of magnitude higher (running the experiments would have taken several months).

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Faruqui and Dyer (2014)	38.40*	37.13*	27.60*	26.80*
	Shigeto et al. (2015)	41.53†	43.07†	31.04†	33.73†
	Dinu et al. (2015)	37.7	38.93*	29.14*	30.40*
	Lazaridou et al. (2015)	40.2	-	-	-
	Xing et al. (2015)	36.87†	41.27†	28.23†	31.20†
	Zhang et al. (2016)	36.73†	40.80†	28.16†	31.07†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Artetxe et al. (2017)	39.67	40.87	28.72	-
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Artetxe et al. (2018a)	45.27	44.13	**32.94**	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017a), $\lambda = 1$	0.00*	0.00*	0.00*	0.00*
	Zhang et al. (2017a), $\lambda = 10$	0.00*	0.00*	0.01*	0.01*
	Conneau et al. (2018), code‡	45.15*	46.83*	0.38*	35.38*
	Conneau et al. (2018), paper‡	45.1	0.01*	0.01*	35.44*
	Proposed method	**48.13**	**48.19**	32.63	**37.33**

Table 3: Accuracy (%) of the proposed method in comparison with previous work. *Results obtained with the official implementation from the authors. †Results obtained with the framework from Artetxe et al. (2018a). The remaining results were reported in the original papers. For methods that do not require supervision, we report the average accuracy across 10 runs. ‡For meaningful comparison, runs with <5% accuracy are excluded when computing the average, but note that, unlike ours, their method often gives a degenerated solution (see Table 2).

language pairs require different hyperparameters: $\lambda = 1$ works substantially better for Spanish-English and Italian-English, but only $\lambda = 10$ works for Turkish-English.

The results for the more challenging dataset from Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a) are given in Table 2. In this case, our proposed method obtains the best results in all metrics for all the four language pairs tested. The method of Zhang et al. (2017a) does not work at all in this more challenging scenario, which is in line with the negative results reported by the authors themselves for similar conditions (only %2.53 accuracy in their large Gigaword dataset). The method of Conneau et al. (2018) also fails for English-Finnish (only 1.62% in the best run), although it is able to get positive results in some runs for the rest of language pairs. Between the two configurations tested, the default hyperparameters in the code show a more stable behavior.

These results confirm the robustness of the proposed method. While the other systems succeed in some runs and fail in others, our method converges to a good solution in all runs without exception and, in fact, it is the only one getting positive results for English-Finnish. In addition to being more robust, our method also obtains substantially better accuracies, surpassing previous methods by at least 1-3 points in all but the easiest pairs. Moreover, our method is not sensitive to hyperparameters that are difficult to tune without a development set, which is critical in realistic unsupervised conditions.

At the same time, our method is significantly faster than the rest. In relation to that, it is interesting that, while previous methods perform a fixed number of iterations and take practically the same time for all the different language pairs, the runtime of our method adapts to the difficulty of the task thanks to the dynamic convergence criterion of our stochastic approach. This way, our method tends to take longer for more challenging language pairs (1.7 vs 0.6 minutes for es-en and tr-en in one dataset, and 12.9 vs 7.3 minutes for en-fi and en-de in the other) and, in fact, our (relative) execution times correlate surprisingly well with the linguistic distance with English (closest/fastest is German, followed by Italian/Spanish, followed by Turkish/Finnish).

	EN-IT				EN-DE				EN-FI				EN-ES			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Full system	48.53	48.13	10	8.9	48.47	48.19	10	7.3	33.50	32.63	10	12.9	37.60	37.33	10	9.1
- Unsup. init.	0.07	0.02	0	16.5	0.00	0.00	0	17.3	0.07	0.01	0	13.8	0.13	0.02	0	15.9
- Stochastic	48.20	48.20	10	2.7	48.13	48.13	10	2.5	0.28	0.28	0	4.3	37.80	37.80	10	2.6
- Cutoff (k=100k)	46.87	46.46	10	114.5	48.27	48.12	10	105.3	31.95	30.78	10	162.5	35.47	34.88	10	185.2
- CSLS	0.00	0.00	0	15.0	0.00	0.00	0	13.8	0.00	0.00	0	13.1	0.00	0.00	0	14.1
- Bidirectional	46.00	45.37	10	5.6	48.27	48.03	10	5.5	31.39	24.86	8	7.8	36.20	35.77	10	7.3
- Re-weighting	46.07	45.61	10	8.4	48.13	47.41	10	7.0	32.94	31.77	10	11.2	36.00	35.45	10	9.1

Table 4: Ablation test on the dataset of Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

5.2 Comparison with the state-of-the-art

Table 3 shows the results of the proposed method in comparison to previous systems, including those with different degrees of supervision. We focus on the widely used English-Italian dataset of Dinu et al. (2015) and its extensions. Despite being fully unsupervised, our method achieves the best results in all language pairs but one, even surpassing previous supervised approaches. The only exception is English-Finnish, where Artetxe et al. (2018a) gets marginally better results with a difference of 0.3 points, yet ours is the only unsupervised system that works for this pair. At the same time, it is remarkable that the proposed system gets substantially better results than Artetxe et al. (2017), the only other system based on self-learning, with the additional advantage of being fully unsupervised.

5.3 Ablation test

In order to better understand the role of different aspects in the proposed system, we perform an ablation test, where we separately analyze the effect of initialization, the different components of our robust self-learning algorithm, and the final symmetric re-weighting. The obtained results are reported in Table 4.

In concordance with previous work, our results show that self-learning does not work with random initialization. However, the proposed unsupervised initialization is able to overcome this issue without the need of any additional information, performing at par with other character-level heuristics that we tested (e.g. shared numerals).

As for the different self-learning components, we observe that the stochastic dictionary induction is necessary to overcome the problem of poor local optima for English-Finnish, although it does not make any difference for the rest of easier language pairs. The frequency-based vocabulary cutoff also has a positive effect, yielding to slightly better accuracies and much faster runtimes. At the same time, CSLS plays a critical role in the system, as hubness severely accentuates the problem of local optima in its absence. The bidirectional dictionary induction is also beneficial, contributing to the robustness of the system as shown by English-Finnish and yielding to better accuracies in all cases.

Finally, these results also show that symmetric re-weighting contributes positively, bringing an improvement of around 1-2 points without any cost in the execution time.

6 Conclusions

In this paper, we show that previous unsupervised mapping methods (Zhang et al., 2017a; Conneau et al., 2018) often fail on realistic scenarios involving non-comparable corpora and/or distant languages. In contrast to adversarial methods, we propose to use an initial weak mapping that exploits the structure of the embedding spaces in combination with a robust self-learning approach. The results show that our method succeeds in all cases, providing the best results with respect to all previous work on unsupervised and supervised mappings.

The ablation analysis shows that our initial solution is instrumental for making self-learning work without supervision. In order to make self-learning robust, we also added stochasticity to dictionary induction, used CSLS instead of nearest neighbor, and produced bidirectional dictionaries. Results also improved using smaller in-

termediate vocabularies and re-weighting the final solution. Our implementation is available as an open source project at https://github.com/artetxem/vecmap.

In the future, we would like to extend the method from the bilingual to the multilingual scenario, and go beyond the word level by incorporating embeddings of longer phrases.

Acknowledgments

This research was partially supported by the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, Denver, Colorado. Association for Computational Linguistics.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Proceedings, Part I*, pages 135–151.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.

UNSUPERVISED NEURAL MACHINE TRANSLATION

Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

ABSTRACT

In spite of the recent success of neural machine translation (NMT) in standard benchmarks, the lack of large parallel corpora poses a major practical problem for many language pairs. There have been several proposals to alleviate this issue with, for instance, triangulation and semi-supervised learning techniques, but they still require a strong cross-lingual signal. In this work, we completely remove the need of parallel data and propose a novel method to train an NMT system in a completely unsupervised manner, relying on nothing but monolingual corpora. Our model builds upon the recent work on unsupervised embedding mappings, and consists of a slightly modified attentional encoder-decoder model that can be trained on monolingual corpora alone using a combination of denoising and backtranslation. Despite the simplicity of the approach, our system obtains 15.56 and 10.21 BLEU points in WMT 2014 French $\rightarrow$ English and German $\rightarrow$ English translation. The model can also profit from small parallel corpora, and attains 21.81 and 15.24 points when combined with 100,000 parallel sentences, respectively. Our implementation is released as an open source project[1].

1 INTRODUCTION

Neural machine translation (NMT) has recently become the dominant paradigm to machine translation (Bahdanau et al., 2014; Sutskever et al., 2014). As opposed to the traditional statistical machine translation (SMT), NMT systems are trained end-to-end, take advantage of continuous representations that greatly alleviate the sparsity problem, and make use of much larger contexts, thus mitigating the locality problem. Thanks to this, NMT has been reported to significantly improve over SMT both in automatic metrics and human evaluation (Wu et al., 2016).

Nevertheless, for the same reasons described above, NMT requires a large parallel corpus to be effective, and is known to fail when the training data is not big enough (Koehn & Knowles, 2017). Unfortunately, the lack of large parallel corpora is a practical problem for the vast majority of language pairs, including low-resource languages (e.g. Basque) as well as many combinations of major languages (e.g. German-Russian). Several authors have recently tried to address this problem using pivoting or triangulation techniques (Chen et al., 2017) as well as semi-supervised approaches (He et al., 2016), but these methods still require a strong cross-lingual signal.

In this work, we eliminate the need of cross-lingual information and propose a novel method to train NMT systems in a completely unsupervised manner, relying solely on monolingual corpora. Our approach builds upon the recent work on unsupervised cross-lingual embeddings (Artetxe et al., 2017; Zhang et al., 2017). Thanks to a shared encoder for both translation directions that uses these fixed cross-lingual embeddings, the entire system can be trained, with monolingual data, to reconstruct its input. In order to learn useful structural information, noise in the form of random token swaps is introduced in this input. In addition to denoising, we also incorporate backtranslation

[1] https://github.com/artetxem/undreamt

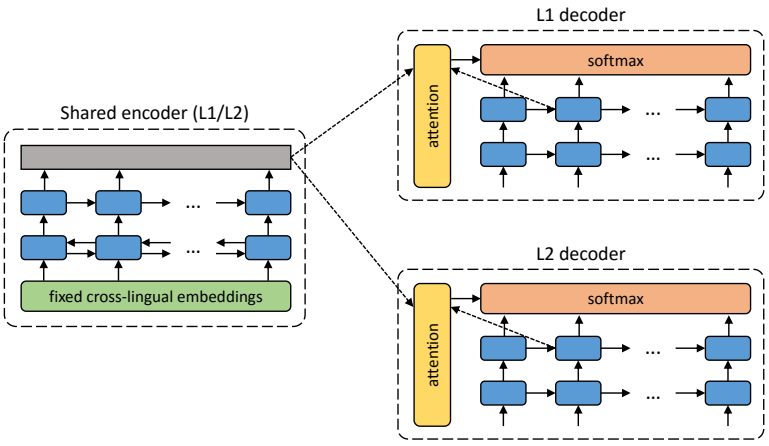


Figure 1: Architecture of the proposed system. For each sentence in language L1, the system is trained alternating two steps: *denoising*, which optimizes the probability of encoding a noised version of the sentence with the shared encoder and reconstructing it with the L1 decoder, and *on-the-fly backtranslation*, which translates the sentence in inference mode (encoding it with the shared encoder and decoding it with the L2 decoder) and then optimizes the probability of encoding this translated sentence with the shared encoder and recovering the original sentence with the L1 decoder. Training alternates between sentences in L1 and L2, with analogous steps for the latter.

(Sennrich et al., 2016a) into the training procedure to further improve results. Figure 1 summarizes this general schema of the proposed system.

In spite of the simplicity of the approach, our experiments show that the proposed system can reach up to 15.56 BLEU points for French → English and 10.21 BLEU points for German → English in the standard WMT 2014 translation task using nothing but monolingual training data. Moreover, we show that combining this method with a small parallel corpus can further improve the results, obtaining 21.81 and 15.24 BLEU points with 100,000 parallel sentences, respectively. Our manual analysis confirms the effectiveness of the proposed approach, revealing that the system is learning non-trivial translation relations that go beyond a word-by-word substitution.

The remaining of this paper is organized as follows. Section 2 analyzes the related work. Section 3 then describes the proposed method. The experimental settings are discussed in Section 4, while Section 5 presents and discusses the obtained results. Section 6 concludes the paper.

2 Related work

We will first discuss unsupervised cross-lingual embeddings, which are the basis of our proposal, in Section 2.1. Section 2.2 then addresses statistical decipherment, an SMT-inspired approach to build a machine translation system in an unsupervised manner. Finally, Section 2.3 presents previous work on training NMT systems in different low-resource scenarios.

2.1 Unsupervised cross-lingual embeddings

Most methods for learning cross-lingual word embeddings rely on some bilingual signal at the document level, typically in the form of parallel corpora (Gouws et al., 2015; Luong et al., 2015a). Closer to our scenario, embedding mapping methods independently train the embeddings in different languages using monolingual corpora, and then learn a linear transformation that maps them to a shared space based on a bilingual dictionary (Mikolov et al., 2013a; Lazaridou et al., 2015; Artetxe et al., 2016; Smith et al., 2017). While the dictionary used in these earlier work typically contains a few thousands entries, Artetxe et al. (2017) propose a simple self-learning extension that gives comparable results with an automatically generated list of numerals, which is used as a shortcut for

practical unsupervised learning. Alternatively, adversarial training has also been proposed to learn such mappings in an unsupervised manner (Miceli Barone, 2016; Zhang et al., 2017).

2.2 Statistical decipherment for machine translation

There is a considerable body of work in statistical decipherment techniques to induce a machine translation model from monolingual data, which follows the same noisy-channel model used by SMT (Ravi & Knight, 2011; Dou & Knight, 2012). More concretely, they treat the source language as ciphertext, and model the process by which this ciphertext is generated as a two-stage process involving the generation of the original English sequence and the probabilistic replacement of the words in it. The English generative process is modeled using a standard n-gram language model, and the channel model parameters are estimated using either expectation maximization or Bayesian inference. This approach was shown to benefit from the incorporation of syntactic knowledge of the languages involved (Dou & Knight, 2013; Dou et al., 2015). More in line with our proposal, the use of word embeddings has also been shown to bring significant improvements in statistical decipherment for machine translation (Dou et al., 2015).

2.3 Low-resource neural machine translation

There have been several proposals to exploit resources other than direct parallel corpora to train NMT systems. The scenario that is most often considered is one where two languages have little or no parallel data between them but are well connected through a third language (e.g. there might be little direct resources for German-Russian but plenty for German-English and English-Russian). The most basic approach in this scenario is to independently translate from the source language to the pivot language and from the pivot language to the target language. It has however been shown that the use of more advanced models like a teacher-student framework can bring considerable improvements over this basic baseline (Firat et al., 2016b; Chen et al., 2017). In the same line, Johnson et al. (2017) show that a multilingual extension of a standard NMT architecture performs reasonably well even for language pairs for which no direct data was given during training.

In addition to that, there have been several attempts to exploit monolingual corpora for NMT in combination with the more scarce parallel corpora. A simple yet effective approach is to create a synthetic parallel corpus by backtranslating a monolingual corpus in the target language (Sennrich et al., 2016a). At the same time, Currey et al. (2017) showed that training an NMT system to directly copy target language text is also helpful and complementary with backtranslation. Finally, Ramachandran et al. (2017) pre-train the encoder and the decoder in language modeling.

To the best of our knowledge, the more ambitious scenario where an NMT model is trained from monolingual corpora alone has never been explored to date, but He et al. (2016) made an important contribution in this direction. More concretely, their method trains two agents to translate in opposite directions (e.g. French $\rightarrow$ English and English $\rightarrow$ French), and make them teach each other through a reinforcement learning process. While promising, this approach still requires a parallel corpus of a considerable size for a warm start (1.2 million sentences in the reported experiments), whereas our work does not use any parallel data at all.

3 Proposed method

This section describes the proposed unsupervised NMT approach. Section 3.1 first presents the architecture of the proposed system, and Section 3.2 then describes the method to train it in an unsupervised manner.

3.1 System architecture

As shown in Figure 1, the proposed system follows a fairly standard encoder-decoder architecture with an attention mechanism (Bahdanau et al., 2014). More concretely, we use a two-layer bidirectional RNN in the encoder, and another two-layer RNN in the decoder. All RNNs use GRU cells with 600 hidden units (Cho et al., 2014), and the dimensionality of the embeddings is set to 300. As for the attention mechanism, we use the global attention method proposed by Luong et al. (2015b) with the general alignment function. There are, however, three important aspects in which

our system differs from the standard NMT, and these are critical so the system can be trained in an unsupervised manner as described next in Section 3.2:

1. **Dual structure**. While NMT systems are typically built for a specific translation direction (e.g. either French $\rightarrow$ English or English $\rightarrow$ French), we exploit the dual nature of machine translation (He et al., 2016; Firat et al., 2016a) and handle both directions together (e.g. French $\leftrightarrow$ English).
2. **Shared encoder**. Our system makes use of one and only one encoder that is shared by both languages involved, similarly to Ha et al. (2016), Lee et al. (2017) and Johnson et al. (2017). For instance, the exact same encoder would be used for both French and English. This universal encoder is aimed to produce a language independent representation of the input text, which each decoder should then transform into its corresponding language.
3. **Fixed embeddings in the encoder**. While most NMT systems randomly initialize their embeddings and update them during training, we use pre-trained cross-lingual embeddings in the encoder that are kept fixed during training. This way, the encoder is given language independent word-level representations, and it only needs to learn how to compose them to build representations of larger phrases. As discussed in Section 2.1, there are several unsupervised methods to train these cross-lingual embeddings from monolingual corpora, so this is perfectly feasible in our scenario. Note that, even if the embeddings are cross-lingual, we use separate vocabularies for each language. This way, the word *chair*, which exists both in French and English (meaning "flesh" in the former), would get a different vector in each language, although they would both be in a common space.

3.2 Unsupervised training

As NMT systems are typically trained to predict the translations in a parallel corpus, such supervised training procedure is infeasible in our scenario, where we only have access to monolingual corpora. However, thanks to the architectural modifications proposed above, we are able to train the entire system in an unsupervised manner using the following two strategies:

1. **Denoising**. Thanks to the use of a shared encoder, and exploiting the dual structure of machine translation, the proposed system can be directly trained to reconstruct its own input. More concretely, the whole system can be optimized to take an input sentence in a given language, encode it using the shared encoder, and reconstruct the original sentence using the decoder of that language. Given that we use pre-trained cross-lingual embeddings in the shared encoder, this encoder should learn to compose the embeddings of both languages in a language-independent fashion, and each decoder should learn to decompose this representation into their corresponding language. At inference time, we simply replace the decoder with that of the target language, so it generates the translation of the input text from the language-independent representation given by the encoder.

   Nevertheless, this ideal behavior is severely compromised by the fact that the resulting training procedure is essentially a trivial copying task. As such, the optimal solution for this task would not need to capture any real knowledge of the languages involved, as there would be many degenerated solutions that blindly copy all the elements in the input sequence. If this were the case, the system would at best make very literal word-by-word substitutions when used to translate from one language to another at inference time.

   In order to avoid such degenerated solutions and make the encoder truly learn the compositionality of its input words in a language independent manner, we propose to introduce random noise in the input sentences. The idea is to exploit the same underlying principle of denoising autoencoders (Vincent et al., 2010), where the system is trained to reconstruct the original version of a corrupted input sentence (Dai & Le, 2015; Hill et al., 2016). For that purpose, we alter the word order of the input sentence by making random swaps between contiguous words. More concretely, for a sequence of N elements, we make $N/2$ random swaps of this kind. This way, the system needs to learn about the internal structure of the languages involved to be able to recover the correct word order. At the same time, by discouraging the system to rely too much on the word order of the input sequence, we can better account for the actual word order divergences across languages. This training procedure can be seen as an instance of contrastive estimation (Smith & Eisner, 2005), where the

neighborhood is defined by local swaps in our case, although other functions would also be possible.

2. **On-the-fly backtranslation**. In spite of the denoising strategy, the training procedure above is still a copying task with some synthetic alterations that, most importantly, involves a single language at each time, without considering our final goal of translating between two languages. In order to train our system in a true translation setting without violating the constraint of using nothing but monolingual corpora, we propose to adapt the back-translation approach proposed by Sennrich et al. (2016a) to our scenario. More concretely, given an input sentence in one language, we use the system in inference mode with greedy decoding to translate it to the other language (i.e. apply the shared encoder and the decoder of the other language). This way, we obtain a pseudo-parallel sentence pair, and train the system to predict the original sentence from this synthetic translation.

   Note that, contrary to standard backtranslation, which uses an independent model to back-translate the entire corpus at one time, we take advantage of the dual structure of the proposed architecture to backtranslate each mini-batch on-the-fly using the model that is being trained itself. This way, as training progresses and the model improves, it will produce better synthetic sentence pairs through backtranslation, which will serve to further improve the model in the following iterations.

During training, we alternate these different training objectives from mini-batch to mini-batch. This way, given two languages L1 and L2, each iteration would perform one mini-batch of denoising for L1, another one for L2, one mini-batch of on-the-fly backtranslation from L1 to L2, and another one from L2 to L1. Moreover, by further assuming that we have access to a small parallel corpus, the system can also be trained in a semi-supervised fashion by combining these steps with directly predicting the translations in this parallel corpus just as in standard NMT.

4 Experimental settings

We make our experiments comparable with previous work by using the French-English and German-English **datasets** from the WMT 2014 shared task[2]. Following common practice, the systems are evaluated on newstest2014 using tokenized BLEU scores as computed by the `multi-bleu.perl` script[3]. As for the training data, we test the proposed system under three different settings:

- **Unsupervised**: This is the main scenario under consideration in our work, where the system has access to nothing but monolingual corpora. For that purpose, we used the News Crawl corpus with articles from 2007 to 2013.
- **Semi-supervised**: We assume that, in addition to monolingual corpora, we also have access to a small in-domain parallel corpus. This scenario has a great practical interest, as we might often have some parallel data from which we could potentially benefit, but it is insufficient to train a full traditional NMT system. For that purpose, we used the same monolingual data from the unsupervised settings together with either 10,000 or 100,000 random sentence pairs from the News Commentary parallel corpus.
- **Supervised**: This is the traditional scenario in NMT where we have access to a large parallel corpus. While not the focus of our work, this setting should provide an approximate upper-bound for the proposed system. For that purpose, we used the combination of all parallel corpora provided at WMT 2014, which comprise Europarl, Common Crawl and News Commentary for both language pairs plus the UN and the Gigaword corpus for French-English. For direct comparison with the semi-supervised scenario, we also ran separate experiments using the same subsets of News Commentary alone.

Note that, to be faithful to our target scenario, we did not make use of any parallel data in these language pairs for development or tuning purposes. Instead, we used Spanish-English WMT data for our preliminary experiments, where we also decided all the hyperparameters without any rigorous exploration.

[2] `http://www.statmt.org/wmt14/translation-task.html`

[3] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl`

As for the **corpus preprocessing**, we perform tokenization and truecasing using standard Moses tools.[4] We then apply byte pair encoding (BPE) as proposed by Sennrich et al. (2016b) using the implementation provided by the authors[5]. Learning was done on the monolingual corpus of each language independently, using 50,000 operations. While BPE is known to be an effective way to overcome the rare word problem in standard NMT, it is less clear how it would perform in our more challenging unsupervised scenario, as it might be difficult to learn the translation relations between subword units. For that reason, we also run experiments at the word level in this unsupervised scenario, limiting the vocabulary to the most frequent 50,000 tokens and replacing the rest with a special token <UNK>. We accelerate training by discarding all sentences with more than 50 elements (either BPE units or actual tokens).

Given that the proposed system uses pre-trained **cross-lingual embeddings** in the encoder as described in Section 3.1, we use the monolingual corpora described above to independently train the embeddings for each language using word2vec (Mikolov et al., 2013b). More concretely, we use the skip-gram model with ten negative samples, a context window of ten words, 300 dimensions, a sub-sampling of 10^{-5}, and ten training iterations. We then use the public implementation[6] of the method proposed by Artetxe et al. (2017) to map these embeddings to a shared space, using the recommended configuration with numeral-based initialization. In addition to being a component of the proposed system, the resulting embeddings are also used to build a simple **baseline system** that translates a sentence word-by-word, replacing each word by their nearest neighbor in the other language and leaving out-of-vocabularies unchanged.

The **training** of the proposed system itself is done using the procedure described in Section 3.2 with the cross-entropy loss function and a batch size of 50 sentences. For the unsupervised systems, we try using denoising alone as well as the combination of both denoising and backtranslation, in order to better analyze the contribution of the latter. We use Adam as our optimizer with a learning rate of $\alpha = 0.0002$ (Kingma & Ba, 2015). During training, we use dropout regularization with a drop probability $p = 0.3$. Given that we restrict ourselves not to use any parallel data for development purposes, we perform a fixed number of iterations (300,000) to train each variant. Using our PyTorch implementation, training each system took about 4-5 days on a single Titan X GPU for the full unsupervised variant. Although we observed that the system had not fully converged after this number of iterations in our preliminary experiments, we decide to stop training at this point in order to accelerate experimentation due to hardware constraints.

As described in Section 3.2, we use greedy **decoding** at training time for backtranslation, but actual inference at test time was done using beam-search with a beam size of 12 following common practice (Sutskever et al., 2014; Sennrich et al., 2016a;b; He et al., 2016). We do not use any length or coverage penalty, which might further improve the reported results.

5 Results and discussion

We discuss the quantitative results in Section 5.1, and present a qualitative analysis in Section 5.2.

5.1 Quantitative analysis

The BLEU scores obtained by all the tested variants are reported in Table 1.

As it can be seen, the proposed **unsupervised system** obtains very strong results considering that it was trained on nothing but monolingual corpora, reaching 14-15 BLEU points in French-English and 6-10 BLEU points in German-English depending on the variant and direction (rows 3 and 4). This is much stronger than the baseline system of word-by-word substitution (row 1), with improvements of at least 40% in all cases, and up to 140% in some (e.g. from 6.25 to 15.13 BLEU points in English $\rightarrow$ French). This shows that the proposed system is able to go beyond very literal translations, effectively learning to use context information and account for the internal structure of the languages.

The results also show that **backtranslation** is essential for the proposed system to work properly. In fact, the denoising technique alone is below the baseline (row 1 vs 2), while big improvements are

[4] https://github.com/moses-smt/mosesdecoder
[5] https://github.com/rsennrich/subword-nmt
[6] https://github.com/artetxem/vecmap

Table 1: BLEU scores in newstest2014. Unsupervised systems are trained in the News Crawl monolingual corpus, semi-supervised systems are trained in the News Crawl monolingual corpus and a subset of the News Commentary parallel corpus, and supervised systems (provided for comparison) are trained in either these same subsets or the full parallel corpus, all from WMT 2014. For GNMT, we report the best single model scores from Wu et al. (2016).

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	6. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	7. Comparable NMT (10k parallel)	1.88	1.66	1.33	0.82
	8. Comparable NMT (100k parallel)	10.40	9.19	8.11	5.29
	9. Comparable NMT (full parallel)	20.48	19.89	15.04	11.05
	10. GNMT (Wu et al., 2016)	-	38.95	-	24.61

seen when introducing backtranslation (row 2 vs 3). Test perplexities also confirm this: for instance, the proposed system with denoising alone obtains a per-word perplexity of 634.79 for French $\rightarrow$ English, whereas the one with backtranslation achieves a much lower perplexity of 44.74. We emphasize, however, that the proposed training procedure would not work using backtranslation alone without denoising, as the initial translations would be meaningless sentences produced by a random NMT model, encouraging the system to completely ignore the input sentence and simply learn a language model of the target language. We thus conclude that both denoising and backtranslation play an essential role during training: denoising forces the system to capture broad word-level equivalences, while backtranslation encourages it to learn more subtle relations in an increasingly natural setting.

As for the role of **subword** translation, we observe that **BPE** is slightly beneficial when German is the target language, detrimental when French is the target language, and practically equivalent when English is the target language (row 3 vs 4). This might be a bit surprising considering that the word-level system does not handle out-of-vocabularies in any way, so it always fails to translate rare words. Having a closer look, however, we observe that, while BPE manages to correctly translate some rare words, it also introduces some new errors. In particular, it sometimes happens that a subword unit from a rare word gets prefixed to a properly translated word, yielding to translations like *SevAgency* (split as *S- ev- Agency*). Moreover, we observe that BPE is of little help when translating infrequent named entities. For instance, we observed that our system translated *Tymoshenko* as *Ebferchenko* (split as *Eb- fer- chenko*). While standard NMT would easily learn to copy this kind of named entities using BPE, such relations are much more challenging to model under our unsupervised learning procedure. This way, we believe that a better handling of rare words and, in particular, named entities and numerals, could further improve the results in the future.

In addition to that, the results of the **semi-supervised system** (rows 5 and 6) show that the proposed model can greatly benefit from a small parallel corpus. Note that these semi-supervised systems differ from the full unsupervised system (row 4) in the use of either 10,000 or 100,000 parallel sentences from News Crawl, so that their training alternates between denoising, backtranslation and, additionally, maximizing the translation probability of these parallel sentences as described in Section 3.2. As it can be seen, 10,000 parallel sentences alone bring an improvement of 1-3 BLEU points, while 100,000 sentences bring an improvement of 4-7 points. These results are much better than those of a comparable NMT system trained in the same parallel data (rows 7 and 8), showing the potential interest of our approach beyond the strictly unsupervised scenario. In fact, the semi-supervised system trained in 100,000 parallel sentences (row 6) even surpasses the comparable NMT system trained in the full parallel corpus (row 9) in all cases but one, presumably because the domain of both the monolingual and the parallel corpora that it uses matches that of the test set.

Table 2: Sample French→English translations from newstest2014 by the full proposed system with BPE. See text for comments.

Source	Reference	Proposed system (full)
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	There was a shooting in Los Angeles International Airport.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beaucoup de spéculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-opération ciblée.	Such growing controversy surrounding the agency prompted early speculation that tonight's incident was the result of a targeted cyber operation.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.
Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.	The total number of deaths in October is the highest since April 2008, when 1,073 people were killed.	The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.
À l'exception de l'opéra, la province reste le parent pauvre de la culture en France.	With the exception of opera, the provinces remain the poor relative of culture in France.	At an exception, opera remains of the state remains the poorest parent culture.

As for the **supervised system**, it is remarkable that the comparable NMT model (rows 7-9), which uses the proposed architecture but trains it to predict the translations in the corresponding parallel corpus, obtains poor results compared to the state of the art in NMT (e.g. GNMT in row 10). Note that the comparable NMT system is equivalent to the semi-supervised system (rows 5 and 6), except that it does not use any monolingual corpora nor, consequently, denoising and backtranslation. As such, the comparable NMT differs from standard NMT in the use of a shared encoder with fixed embeddings (Section 3.1) and input corruption (Section 3.2).

The relatively poor results of the comparable NMT model suggest that these additional constraints in our system, which were introduced to enable unsupervised learning, may also be a factor limiting its potential performance, so we believe that the system **could be further improved in the future** by progressively relaxing these constraints during training. For instance, using fixed cross-lingual embeddings in the encoder is necessary in the early stages of training, as it forces the encoder to use a common word representation for both languages, but it might also limit what it can ultimately learn in the process. For that reason, one could start to progressively update the weights of the encoder embeddings as training progresses. Similarly, one could also decouple the shared encoder into two independent encoders at some point during training, or progressively reduce the noise level. At the same time, note that we did not perform any rigorous hyperparameter exploration, and favored efficiency over performance in the experimental design due to hardware constraints. As such, we think that there is a considerable margin to improve these results by using larger models, longer training times, and incorporating several well-known NMT techniques (e.g. ensembling and length/coverage penalty).

5.2 Qualitative Analysis

In order to better understand the behavior of the proposed system, we manually analyzed some translations for French → English, and present some illustrative examples in Table 2.

Our analysis shows that the proposed system is able to produce high-quality translations, adequately modeling non-trivial translation relations. For instance, in the first example it translates the expression *a eu lieu* (literally "has had place") as *occurred*, going beyond a literal word-by-word substitution. At the same time, it correctly translates *l'aéroport international de Los Angeles* as *Los Angeles International Airport*, properly modeling structural differences between the languages. As shown by the second example, the system is also capable of producing high-quality translations for considerably longer and more complex sentences.

Nevertheless, our analysis also points that the proposed system has limitations and, perhaps not surprisingly, its translation quality often lags behind that of a standard supervised NMT system. In particular, we observe that the proposed model has difficulties to preserve some concrete details from source sentences. For instance, in the third example *April* and *2008* are properly translated, but *octobre* ("October") is mistranslated as *May* and *1 073* as *1 064*. While these clearly point to some adequacy issues, they are also understandable given the unsupervised nature of the system, and it is remarkable that the system managed to at least replace a month by another month and a number by another close number. We believe that incorporating character level information might help to mitigate some of these issues, as it could for instance favor *October* as the translation of *octobre* instead of the selected *May*.

Finally, there are also some cases where there are both fluency and adequacy problems that severely hinders understanding the original message from the proposed translation. For instance, in the last example our system preserves most keywords in the original sentence, but it would be difficult to correctly guess its meaning just by looking at its translation. In concordance with our quantitative analysis, this suggests that there is still room for improvement, opening new research avenues for the future.

6 Conclusions and future work

In this work, we propose a novel method to train an NMT system in a completely unsupervised manner. We build upon existing work on unsupervised cross-lingual embeddings (Artetxe et al., 2017; Zhang et al., 2017), and incorporate them in a modified attentional encoder-decoder model. By using a shared encoder with these fixed cross-lingual embeddings, we are able to train the system from monolingual corpora alone, combining denoising and backtranslation.

The experiments show the effectiveness of our proposal, obtaining significant improvements in the BLEU score over a baseline system that performs word-by-word substitution in the standard WMT 2014 French-English and German-English benchmarks. Our manual analysis confirms the quality of the proposed system, showing that it is able to model complex cross-lingual relations and produce high-quality translations. Moreover, we show that combining our method with a small parallel corpus can bring further improvements, showing its potential interest beyond the strictly unsupervised scenario.

Our work opens exciting opportunities for future research, as our analysis reveals that, in spite of the solid results, there is still a considerable room for improvement. In particular, we observe that the performance of a comparable supervised NMT system is considerably below the state of the art, which suggests that the architectural modifications introduced by our proposal (Section 3.1) are also limiting its potential performance. For that reason, we would like to explore progressively relaxing these constraints during training as discussed in Section 5.1. Additionally, we would like to incorporate character level information into the model, which we believe that could be very helpful to address some of the adequacy issues observed in our manual analysis (Section 5.2). Finally, we would like to explore other neighborhood functions for denoising, and analyze their effect in relation to the typological divergences of different language pairs.

Acknowledgments

This research was partially supported by a Google Faculty Award, the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the Basque Government (MODELA KK-2016/00082), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe enjoys a doctoral grant from the Spanish MECD. Kyunghyun Cho thanks support by eBay, TenCent, Facebook, Google, NVIDIA and CIFAR, and was partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, Austin, Texas,

November 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1250.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL http://aclweb.org/anthology/P17-1042.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2014 International Conference on Learning Representations*, 2014.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1925–1935, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL http://aclweb.org/anthology/P17-1176.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1179.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pp. 148–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4715.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28*, pp. 3079–3087. 2015. URL http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf.

Qing Dou and Kevin Knight. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 266–275, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D12-1025.

Qing Dou and Kevin Knight. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1668–1676, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1173.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 836–845, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-1081.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 866–875, San Diego, California, June 2016a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1101.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 268–277, Austin, Texas, November 2016b. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1026.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 748–756, 2015.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*, 2016.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*, pp. 820–828. 2016. URL http://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1367–1377, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1162.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a ViÃ©gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1081.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, 2015.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-3204.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 270–280, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-1027.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1051.

Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, Denver, Colorado, June 2015a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-1521.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015b. Association for Computational Linguistics. URL http://aclweb.org/anthology/D15-1166.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 121–126, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://anthology.aclweb.org/W16-1614.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. 2013b.

Prajit Ramachandran, Peter Liu, and Quoc Le. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 383–391, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1039.

Sujith Ravi and Kevin Knight. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 12–21, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1002.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1162.

Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 354–362, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219884. URL http://www.aclweb.org/anthology/P05-1044.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR 2017)*, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. URL http://arxiv.org/abs/1609.08144.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL http://aclweb.org/anthology/P17-1179.

Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

While modern machine translation has relied on large parallel corpora, a recent line of work has managed to train Neural Machine Translation (NMT) systems from monolingual corpora only (Artetxe et al., 2018c; Lample et al., 2018). Despite the potential of this approach for low-resource settings, existing systems are far behind their supervised counterparts, limiting their practical interest. In this paper, we propose an alternative approach based on phrase-based Statistical Machine Translation (SMT) that significantly closes the gap with supervised systems. Our method profits from the modular architecture of SMT: we first induce a phrase table from monolingual corpora through cross-lingual embedding mappings, combine it with an n-gram language model, and fine-tune hyperparameters through an unsupervised MERT variant. In addition, iterative backtranslation improves results further, yielding, for instance, 14.08 and 26.22 BLEU points in WMT 2014 English-German and English-French, respectively, an improvement of more than 7-10 BLEU points over previous unsupervised systems, and closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 BLEU points. Our implementation is available at https://github.com/artetxem/monoses.

1 Introduction

Neural Machine Translation (NMT) has recently become the dominant paradigm in machine translation (Vaswani et al., 2017). In contrast to more rigid Statistical Machine Translation (SMT) architectures (Koehn et al., 2003), NMT models are trained end-to-end, exploit continuous representations that mitigate the sparsity problem, and overcome the locality problem by making use of unconstrained contexts. Thanks to this additional flexibility, NMT can more effectively exploit large parallel corpora, although SMT is still superior when the training corpus is not big enough (Koehn and Knowles, 2017).

Somewhat paradoxically, while most machine translation research has focused on resource-rich settings where NMT has indeed superseded SMT, a recent line of work has managed to train an NMT system without any supervision, relying on monolingual corpora alone (Artetxe et al., 2018c; Lample et al., 2018). Given the scarcity of parallel corpora for most language pairs, including less-resourced languages but also many combinations of major languages, this research line opens exciting opportunities to bring effective machine translation to many more scenarios. Nevertheless, existing solutions are still far behind their supervised counterparts, greatly limiting their practical usability. For instance, existing unsupervised NMT systems obtain between 15-16 BLEU points in WMT 2014 English-French translation, whereas a state-of-the-art NMT system obtains around 41 (Artetxe et al., 2018c; Lample et al., 2018; Yang et al., 2018).

In this paper, we explore whether the rigid and modular nature of SMT is more suitable for these unsupervised settings, and propose a novel unsupervised SMT system that can be trained on monolingual corpora alone. For that purpose, we present a natural extension of the skip-gram model (Mikolov et al., 2013b) that simultaneously learns word and phrase embeddings, which are then mapped to a cross-lingual space through self-learning (Artetxe et al., 2018b). We use the resulting cross-lingual phrase embeddings to induce a phrase table, and combine it with a language model and a distance-based distortion model to build a standard phrase-based SMT system. The weights of this model are tuned in an unsupervised manner through an iterative Minimum Error Rate Training (MERT) variant, and the entire system

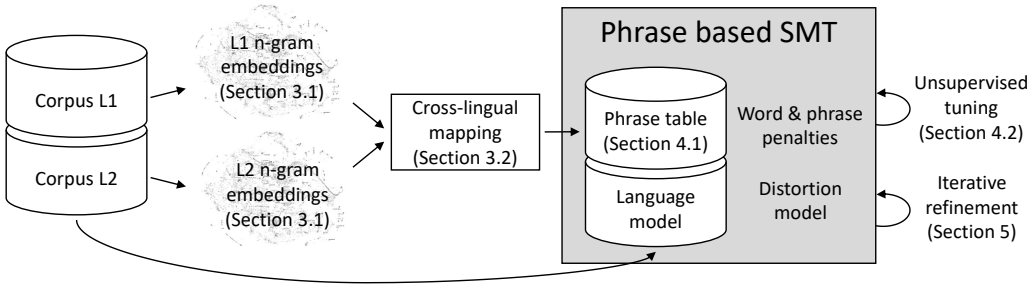


Figure 1: Architecture of our system, with references to sections.

is further improved through iterative backtranslation. The architecture of the system is sketched in Figure 1. Our experiments on WMT German-English and French-English datasets show the effectiveness of our proposal, where we obtain improvements above 7-10 BLEU points over previous unsupervised NMT-based approaches, closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 points.

The remaining of this paper is structured as follows. Section 2 introduces phrase-based SMT. Section 3 presents our unsupervised approach to learn cross-lingual n-gram embeddings, which are the basis of our proposal. Section 4 describes the proposed unsupervised SMT system itself, while Section 5 discusses its iterative refinement through backtranslation. Section 6 describes the experiments run and the results obtained. Section 7 discusses the related work on the topic, and Section 8 concludes the paper.

2 Background: phrase-based SMT

While originally motivated as a noisy channel model (Brown et al., 1990), phrase-based SMT is now formulated as a log-linear combination of several statistical models that score translation candidates (Koehn et al., 2003). The parameters of these scoring functions are estimated independently based on frequency counts, and their weights are then tuned in a separate validation set. At inference time, a decoder tries to find the translation candidate with the highest score according to the resulting combined model. The specific scoring models found in a standard SMT system are as follows:

- **Phrase table**. The phrase table is a collection of source language n-grams and a list of their possible translations in the target language along with different scores for each of them. So as to translate longer sequences, the decoder combines these partial n-gram translations, and ranks the resulting candidates according to their corresponding scores and the rest of scoring functions. In order to build the phrase table, SMT computes word alignments in both directions from a parallel corpus, symmetrizes these alignments using different heuristics (Och and Ney, 2003), extracts the set of consistent phrase pairs, and scores them based on frequency counts. For that purpose, standard SMT uses 4 scores for each phrase table entry: the direct and inverse lexical weightings, which are derived from word level alignments, and the direct and inverse phrase translation probabilities, which are computed at the phrase level.
- **Language model**. The language model assigns a probability to a word sequence in the target language. Traditional SMT uses n-gram language models for that, which use simple frequency counts over a large monolingual corpus with back-off and smoothing.
- **Reordering model**. The reordering model accounts for different word orders across languages, scoring translation candidates according to the position of each translated phrase in the target language. Standard SMT combines two such models: a distance based distortion model that penalizes deviation from a monotonic order, and a lexical reordering model that incorporates phrase orientation frequencies from a parallel corpus.
- **Word and phrase penalties**. The word and phrase penalties assign a fixed score to every generated word and phrase, and are useful to control the length of the output text and the preference for shorter or longer phrases.

Having trained all these different models, a tuning process is applied to optimize their weights in the resulting log-linear model, which typically maximizes some evaluation metric in a separate validation parallel corpus. A common choice is to optimize the BLEU score through Minimum Error Rate Training (MERT) (Och, 2003).

3 Cross-lingual n-gram embeddings

Section 3.1 presents our proposed extension of skip-gram to learn n-gram embeddings, while Section 3.2 describes how we map them to a shared space to obtain cross-lingual n-gram embeddings.

3.1 Learning n-gram embeddings

Negative sampling skip-gram takes word-context pairs (w, c), and uses logistic regression to predict whether the pair comes from the true distribution as sampled from the training corpus, or it is one of the k draws from a noise distribution (Mikolov et al., 2013b):

$$\log \sigma\left(w \cdot c\right) + \sum_{i=1}^{k} \mathbb{E}_{c_N \sim P_D}\left[\log \sigma\left(-w \cdot c_N\right)\right]$$

In its basic formulation, both w and c correspond to words that co-occur within a certain window in the training corpus. So as to learn embeddings for non-compositional phrases like *New York Times* or *Toronto Maple Leafs*, Mikolov et al. (2013b) propose to merge them into a single token in a pre-processing step. For that purpose, they use a scoring function based on their co-occurence frequency in the training corpus, with a discounting coefficient δ that penalizes rare words, and iteratively merge those above a threshold:

$$\text{score}(w_i, w_j) = \frac{\text{count}\left(w_i, w_j\right) - \delta}{\text{count}\left(w_i\right) \times \text{count}\left(w_j\right)}$$

However, we also need to learn representations for compositional n-grams in our scenario, as there is not always a 1:1 correspondence for n-grams across languages even for compositional phrases. For instance, the phrase *he will come* would typically be translated as *vendrá* into Spanish, so one would need to represent the entire phrase as a single unit to properly model this relation.

One option would be to merge all n-grams regardless of their score, but this is not straightforward given their overlapping nature, which is further accentuated when considering n-grams of different lengths. While we tried to randomly generate multiple consistent segmentations for each sentence and train the embeddings over the resulting corpus, this worked poorly in our preliminary experiments. We attribute this to the complex interactions arising from the stochastic segmentation (e.g. the co-occurrence distribution changes radically, even for unigrams), severely accentuating the sparsity problem, among other issues.

As an alternative approach, we propose a generalization of skip-gram that learns n-gram embeddings on-the-fly, and has the desirable property of unigram invariance: our proposed model learns the exact same embeddings as the original skip-gram for unigrams, while simultaneously learning additional embeddings for longer n-grams. This way, for each word-context pair (w, c) at distance d within the given window, we update their corresponding embeddings w and c with the usual negative sampling loss. In addition to that, we look at all n-grams p of different lengths that are at the same distance d, and for each pair (p, c), we update the embedding p through negative sampling. In order to enforce unigram invariance, the context c and negative samples c_N, which always correspond to unigrams, are not updated for (p, c). This allows to naturally learn n-gram embeddings according to their co-occurrence patterns as modeled by skip-gram, without introducing subtle interactions that affect its fundamental behavior.

We implemented the above procedure as an extension of *word2vec*, and use it to train monolingual n-gram embeddings with a window size of 5, 300 dimensions, 10 negative samples, 5 iterations and subsampling disabled. So as to keep the model size within a reasonable limit, we restrict the vocabulary to the most frequent 200,000 unigrams, 400,000 bigrams and 400,000 trigrams.

3.2 Cross-lingual mapping

Cross-lingual mapping methods take independently trained word embeddings in two languages, and learn a linear transformation to map them to a shared cross-lingual space (Mikolov et al., 2013a; Artetxe et al., 2018a). Most mapping methods are supervised, and rely on a bilingual dictionary, typically in the range of a few thousand entries, although a recent line of work has managed to achieve comparable results in a fully unsupervised manner based on either self-learning (Artetxe et al., 2017, 2018b) or adversarial training (Zhang et al., 2017a,b; Conneau et al., 2018).

In our case, we use the method of Artetxe et al.

(2018b) to map the n-gram embeddings to a shared cross-lingual space using their open source implementation VecMap[1]. Originally designed for word embeddings, this method builds an initial mapping by connecting the intra-lingual similarity distribution of embeddings in different languages, and iteratively improves this solution through self-learning. The method applies a frequency-based vocabulary cut-off, learning the mapping over the 20,000 most frequent words in each language. We kept this cut-off to learn the mapping over the most frequent 20,000 unigrams, and then apply the resulting mapping to the entire embedding space, including longer n-grams.

4 Unsupervised SMT

As discussed in Section 2, phrase-based SMT follows a modular architecture that combines several scoring functions through a log-linear model. Among the scoring functions found in standard SMT systems, the distortion model and word/phrase penalties are parameterless, while the language model is trained on monolingual corpora, so they can all be directly integrated into our unsupervised system. From the remaining models, typically trained on parallel corpora, we decide to leave the lexical reordering out, as the distortion model already accounts for word reordering. As for the phrase table, we learn cross-lingual n-gram embeddings as discussed in Section 3, and use them to induce and score phrase translation pairs as described next (Section 4.1). Finally, we tune the weights of the resulting log-linear model using an unsupervised procedure based on back-translation (Section 4.2).

Unless otherwise specified, we use Moses[2] with default hyperparameters to implement these different components of our system. We use KenML (Heafield et al., 2013), bundled in Moses by default, to estimate our 5-gram language model with modified Kneser-Ney smoothing, pruning n-grams longer than 3 with a single occurrence.

4.1 Phrase table induction

Given the lack of a parallel corpus from which to **extract phrase translation pairs**, every n-gram in the target language could be taken as a potential translation candidate for each n-gram in the source language. So as to keep the size of the phrase table within a reasonable limit, we train cross-lingual phrase embeddings as described in Section 3, and limit the translation candidates for each source phrase to its 100 nearest neighbors in the target language.

In order to estimate their corresponding **phrase translation probabilities**, we apply the softmax function over the cosine similarities of their respective embeddings. More concretely, given the source language phrase $\bar{e}$ and the translation candidate $\bar{f}$, their direct phrase translation probability is computed as follows[3]:

$$\phi(\bar{f}|\bar{e}) = \frac{\cos(\bar{e},\bar{f})/\tau}{\sum_{\bar{f}'}\cos(\bar{e},\bar{f}')/\tau}$$

Note that, in the above formula, $\bar{f}'$ iterates across all target language embeddings, and τ is a constant temperature parameter that controls the confidence of the predictions. In order to tune it, we induce a dictionary over the cross-lingual embeddings themselves with nearest neighbor retrieval, and use maximum likelihood estimation over it. However, inducing the dictionary in the same direction as the probability predictions leads to a degenerated solution (softmax approximates the hard maximum underlying nearest neighbor as τ approaches 0), so we induce the dictionary in the opposite direction and apply maximum likelihood estimation over it:

$$\min_{\tau} \sum_{\bar{f}} \log \phi(\bar{f}|\,\mathrm{NN}_{\bar{e}}(\bar{f})) + \sum_{\bar{e}} \log \phi(\bar{e}|\,\mathrm{NN}_{\bar{f}}(\bar{e}))$$

So as to optimize τ, we use Adam with a learning rate of 0.0003 and a batch size of 200, implemented in PyTorch.

In order to compute the **lexical weightings**, we align each word in the target phrase with the one in the source phrase most likely generating it, and take the product of their respective translation probabilities:

$$\mathrm{lex}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j w(\bar{f}_i|\bar{e}_j)\right)$$

The constant ϵ guarantees that each target language word will get a minimum probability mass, which is useful to model NULL alignments. In our experiments, we set $\epsilon = 0.001$, which we find to

[1] https://github.com/artetxem/vecmap
[2] http://www.statmt.org/moses/

[3] The inverse phrase translation probability $\phi(\bar{e}|\bar{f})$ is defined analogously.

Algorithm 1 Unsupervised tuning

Input: $m_{s\to t}$ (source-to-target models)
Input: $m_{t\to s}$ (target-to-source models)
Input: c_s (source validation corpus)
Input: c_t (target validation corpus)
Output: $w_{s\to t}$ (source-to-target weights)
Output: $w_{t\to s}$ (target-to-source weights)

$w_{t\to s} \leftarrow$ DEFAULT_WEIGHTS
repeat
  $bt_s \leftarrow$ TRANSLATE($m_{t\to s}, w_{t\to s}, c_t$)
  $w_{s\to t} \leftarrow$ MERT($m_{s\to t}, bt_s, c_t$)
  $bt_t \leftarrow$ TRANSLATE($m_{s\to t}, w_{s\to t}, c_s$)
  $w_{t\to s} \leftarrow$ MERT($m_{t\to s}, bt_t, c_s$)
until convergence

work well in practice. Finally, the word translation probabilities $w(\bar{f}_i|\bar{e}_j)$ are computed using the same formula defined for phrase translation probabilities (see above), with the difference that the partition function goes over unigrams only.

4.2 Unsupervised tuning

As discussed in Section 2, standard SMT uses MERT over a small parallel corpus to tune the weights of the different scoring functions combined through its log-linear model. Given that we only have access to monolingual corpora in our scenario, we propose to generate a synthetic parallel corpus through backtranslation (Sennrich et al., 2016) and apply MERT tuning over it, iteratively repeating the process in both directions (see Algorithm 1). For that purpose, we reserve a random subset of 10,000 sentences from each monolingual corpora, and run the proposed algorithm over them for 10 iterations, which we find to be enough for convergence.

5 Iterative refinement

The procedure described in Section 4 suffices to train an SMT system from monolingual corpora which, as shown by our experiments in Section 6, already outperforms previous unsupervised systems. Nevertheless, our proposed method still makes important simplifications that could compromise its potential performance: it does not use any lexical reordering model, its phrase table is limited by the underlying embedding vocabulary (e.g. it does not include phrases longer than trigrams, see Section 3.1), and the phrase translation probabilities and lexical weightings are estimated based on cross-lingual embeddings.

Algorithm 2 Iterative refinement

Input: c_s (source language corpus)
Input: c_t (target language corpus)
Input/Output: $m_{t\to s}$ (target-to-source models)
Input/Output: $w_{t\to s}$ (target-to-source weights)
Output: $m_{s\to t}$ (source-to-target models)
Output: $w_{s\to t}$ (source-to-target weights)

$train_s, val_s \leftarrow$ SPLIT(c_s)
$train_t, val_t \leftarrow$ SPLIT(c_t)
repeat
  $btt_s \leftarrow$ TRANSLATE($m_{t\to s}, w_{t\to s}, train_t$)
  $btv_s \leftarrow$ TRANSLATE($m_{t\to s}, w_{t\to s}, val_t$)
  $m_{s\to t} \leftarrow$ TRAIN($btt_s, train_t$)
  $w_{s\to t} \leftarrow$ MERT($m_{s\to t}, btv_s, val_t$)
  $btt_t \leftarrow$ TRANSLATE($m_{s\to t}, w_{s\to t}, train_s$)
  $btv_t \leftarrow$ TRANSLATE($m_{s\to t}, w_{s\to t}, val_s$)
  $m_{t\to s} \leftarrow$ TRAIN($btt_t, train_s$)
  $w_{t\to s} \leftarrow$ MERT($m_{t\to s}, btv_t, val_s$)
until convergence

In order to overcome these limitations, we propose an iterative refinement procedure based on backtranslation (Sennrich et al., 2016). More concretely, we generate a synthetic parallel corpus by translating the monolingual corpus in one of the languages with the initial system, and train and tune a standard SMT system over it in the opposite direction. Note that this new system does not have any of the initial restrictions: the phrase table is built and scored using standard word alignment with an unconstrained vocabulary, and a lexical reordering model is also learned. Having done that, we use the resulting system to translate the monolingual corpus in the other language, and train another SMT system over it in the other direction. As detailed in Algorithm 2, this process is repeated iteratively until some convergence criterion is met.

While this procedure would be expected to produce a more accurate model at each iteration, it also happens to be very expensive computationally. In order to accelerate our experiments, we use a random subset of 2 million sentences from each monolingual corpus for training[4], in addition to the 10,000 separate sentences that are held out as a validation set for MERT tuning, and perform a fixed number of 3 iterations of the above algorithm. Moreover, we use FastAlign (Dyer et al., 2013) instead of GIZA++ to make word alignment faster. Other than that, training over the synthetic

[4]Note that we reuse the original language model, which is trained in the full corpus.

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Artetxe et al. (2018c)	15.56	15.13	10.21	6.55	-	-
Lample et al. (2018)	14.31	15.05	-	-	13.33	9.64
Yang et al. (2018)	15.58	16.97	-	-	14.62	10.86
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**	**23.05**	**18.23**

Table 1: Results of the proposed method in comparison to existing unsupervised NMT systems (BLEU).

parallel corpus is done through standard Moses tools with default settings.

6 Experiments and results

In order to make our experiments comparable to previous work, we use the French-English and German-English datasets from the WMT 2014 shared task. As discussed throughout the paper, our system is trained on monolingual corpora alone, so we take the concatenation of all News Crawl monolingual corpora from 2007 to 2013 as our training data, which we tokenize and truecase using standard Moses tools. The resulting corpus has 749 million tokens in French, 1,606 million tokens in German, and 2,109 million tokens in English. Following common practice, the systems are evaluated in newstest2014 using tokenized BLEU scores as computed by the `multi-bleu.perl` script included in Moses. In addition to that, we also report results in German-English newstest2016 (from WMT 2016), as this was used by some previous work in unsupervised NMT (Lample et al., 2018; Yang et al., 2018)[5]. So as to be faithful to our target scenario, we did not use any parallel data in these language pairs, not even for development purposes. Instead, we ran all our preliminary experiments on WMT Spanish-English data, where we made all development decisions.

We present the results of our final system in comparison to other previous work in Section 6.1. Section 6.2 then presents an ablation study of our proposed method, where we analyze the contribution of its different components. Section 6.3 compares the obtained results to those of different supervised systems, analyzing the effect of some of the inherent limitations of our method in a standard phrase-based SMT system. Finally, Section 6.4 presents some translation examples from our system.

[5]Note that we use the same model trained in WMT 2014 for these experiments, so it is likely that our results could be further improved by using the more extensive data from WMT 2016.

6.1 Main results

We report the results obtained by our proposed system in Table 1. As it can be seen, our system obtains the best published results by a large margin, surpassing previous unsupervised NMT systems by around 10 BLEU points in French-English (both directions), and more than 7 BLEU points in German-English (both directions and datasets).

This way, while previous progress in the task has been rather incremental (Yang et al., 2018), our work represents an important step towards high-quality unsupervised machine translation, with improvements over 50% in all cases. This suggests that, in contrast to previous NMT-based approaches, phrase-based SMT may provide a more suitable framework for unsupervised machine translation, which is in line with previous results in low-resource settings (Koehn and Knowles, 2017).

6.2 Ablation analysis

We present ablation results of our proposed system in Table 2. The first row corresponds to the initial system with our induced phrase table (Section 4.1) and default weights as used by Moses, whereas the second row uses our unsupervised MERT procedure to tune these weights (Section 4.2). The remaining rows represent different iterations of our refinement procedure (Section 5), which uses backtranslation to iteratively train a standard SMT system from a synthetic parallel corpus.

The results show that the initial system with default weights (first row) is already better than previous unsupervised NMT systems (Table 1) by a substantial margin (2-6 BLEU points). Our unsupervised tuning procedure further improves results, bringing an improvement of over 1 BLEU

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Unsupervised SMT	21.16	20.13	13.86	10.59	18.01	13.22
+ unsupervised tuning	22.17	22.22	14.73	10.64	18.21	13.12
+ iterative refinement (it1)	24.81	26.53	16.01	13.45	20.76	16.94
+ iterative refinement (it2)	**26.13**	**26.57**	17.30	13.95	22.80	18.18
+ iterative refinement (it3)	25.87	26.22	**17.43**	**14.08**	**23.05**	**18.23**

Table 2: Ablation results (BLEU). The last row corresponds to our full system. Refer to the text for more details.

		WMT-14				WMT-16	
		FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Supervised	NMT (transformer)	-	41.8	-	28.4	-	-
	WMT best	35.0	35.8	29.0	20.6	40.2	34.2
	SMT (europarl)	30.61	30.82	20.83	16.60	26.38	22.12
	+ w/o lexical reord.	30.54	30.33	20.37	16.34	25.99	22.20
	+ constrained vocab.	30.04	30.10	19.91	16.32	25.66	21.53
	+ unsup. tuning	29.32	29.46	17.75	15.45	23.35	19.86
Unsup.	Proposed system	25.87	26.22	17.43	14.08	23.05	18.23

Table 3: Results of the proposed method in comparison to supervised systems (BLEU). Transformer results reported by Vaswani et al. (2017). SMT variants are incremental (e.g. 2nd includes 1st). Refer to the text for more details.

point in both French-English directions, although its contribution is somewhat weaker for German-to-English (almost 1 BLEU point in WMT 2014 but only 0.2 in WMT 2016), and does not make any difference for English-to-German.

The proposed iterative refinement method has a much stronger positive effect, with improvements over 2.5 BLEU points in all cases, and up to 5 BLEU points in some. Most gains come in the first iteration, while the second iteration brings weaker improvements and the algorithm seems to converge in the third iteration, with marginal improvements for German-English and a small drop in performance for French-English.

6.3 Comparison with supervised systems

So as to put our results into perspective, Table 3 comprises the results of different supervised methods in the same test sets. More concretely, we report the results of the Transformer (Vaswani et al., 2017), an NMT system based on self-attention that is the current state-of-the-art in machine translation, along with the scores obtained by the best performing system in each WMT shared task at the time, and those of a standard phrase-based SMT system trained on Europarl and tuned on newstest2013 using Moses. We also report the effect of removing lexical reordering from the latter as we do in our initial system (Section 4), restricting the vocabulary to the most frequent unigram, bigram and trigrams as we do when training our embeddings (Section 3), and using our unsupervised tuning procedure over a subset of the monolingual corpus (Section 4.2) instead of using standard MERT tuning over newstest2013.

Quite surprisingly, our proposed system, trained exclusively on monolingual corpora, is relatively close to a comparable phrase-based SMT system trained on Europarl, with differences below 5 BLEU points in all cases and as little as 2.5 in some. Note that both systems use the exact same language model trained on News Crawl, making them fully comparable in terms of the monolingual corpora they have access to. While more of a baseline than the state-of-the-art, note that Moses+Europarl is widely used as a reference system in machine translation. As such, we think that our results are very encouraging, as they show

Source	Reference	Proposed system
D'autres révélations ont fait état de documents divulgués par Snowden selon lesquels la NSA avait intercepté des données et des communications émanant du téléphone portable de la chancelière allemande Angela Merkel et de ceux de 34 autres chefs d'État.	Other revelations cited documents leaked by Snowden that the NSA monitored German Chancellor Angela Merkel's cellphone and those of up to 34 other world leaders.	Other disclosures have reported documents disclosed by Snowden suggested the NSA had intercepted communications and data from the mobile phone of German Chancellor Angela Merkel and those of 32 other heads of state.
La NHTSA n'a pas pu examiner la lettre d'information aux propriétaires en raison de l'arrêt de 16 jours des activités gouvernementales, ce qui a ralenti la croissance des ventes de véhicules en octobre.	NHTSA could not review the owner notification letter due to the 16-day government shutdown, which tempered auto sales growth in October.	The NHTSA could not consider the letter of information to owners because of halting 16-day government activities, which slowed the growth in vehicle sales in October.
Le M23 est né d'une mutinerie, en avril 2012, d'anciens rebelles, essentiellement tutsi, intégrés dans l'armée en 2009 après un accord de paix.	The M23 was born of an April 2012 mutiny by former rebels, principally Tutsis who were integrated into the army in 2009 following a peace agreement.	M23 began as a mutiny in April 2012, former rebels, mainly Tutsi integrated into the national army in 2009 after a peace deal.
Tunks a déclaré au Sunday Telegraph de Sydney que toute la famille était «extrêmement préoccupée» du bien-être de sa fille et voulait qu'elle rentre en Australie.	Tunks told Sydney's Sunday Telegraph the whole family was "extremely concerned" about his daughter's welfare and wanted her back in Australia.	Tunks told The Times of London from Sydney that the whole family was "extremely concerned" of the welfare of her daughter and wanted it to go in Australia.

Table 4: Randomly chosen translation examples from French→English newstest2014.

that our fully unsupervised system is already quite close to this competitive baseline.

In addition to that, the results for the constrained variants of this SMT system justify some of the simplifications required by our approach. In particular, removing lexical reordering and constraining the phrase table to the most frequent n-grams, as we do for our initial system, has a relatively small effect, with a drop of less than 1 BLEU point in all cases, and as little as 0.28 in some. Replacing standard MERT tuning with our unsupervised variant does cause a considerable drop in performance, although it is below 2.5 BLEU points even in the worst case, and our unsupervised tuning method is still better than using default weights as reported in Table 2. This shows the importance of tuning in SMT, suggesting that these results could be further improved if one had access to a small parallel corpus for tuning.

6.4 Qualitative results

Table 4 shows some of the translations produced by the proposed system for French→English. Note that these examples where randomly taken from the test set, so they should be representative of the general behavior of our approach.

While the examples reveal certain adequacy issues (e.g. *The Times of London from Sidney* instead of *Sydney's Sunday Telegraph*), and the produced output is not perfectly grammatical (e.g. *go in Australia*), our translations are overall quite accurate and fluent, and one could get a reasonable understanding of the original text from them. This suggests that unsupervised machine translation can indeed be a usable alternative in low resource settings.

7 Related work

Similar to our approach, statistical decipherment also attempts to build machine translation systems from monolingual corpora. For that purpose, existing methods treat the source language as ciphertext, and model its generation through a noisy channel model involving two steps: the generation of the original English sentence and the probabilistic replacement of the words in it (Ravi and Knight, 2011; Dou and Knight, 2012). The English generative process is modeled using an n-gram language model, and the channel model parameters are estimated using either expectation maximization or Bayesian inference. Subsequent work has attempted to enrich these models with additional information like syntactic knowledge (Dou and Knight, 2013) and word embeddings (Dou et al., 2015). Nevertheless, these systems work in a word-by-word basis and have

only been shown to work in limited settings, being often evaluated in word-level translation. In contrast, our method builds a fully featured phrase-based SMT system, and achieves competitive performance in a standard machine translation task.

More recently, Artetxe et al. (2018c) and Lample et al. (2018) have managed to train a standard attentional encoder-decoder NMT system from monolingual corpora alone. For that purpose, they use a shared encoder for both languages with pre-trained cross-lingual embeddings, and train the entire system using a combination of denoising, backtranslation and, in the case of Lample et al. (2018), adversarial training. This method was further improved by Yang et al. (2018), who use a separate encoder for each language, sharing only a subset of their parameters, and incorporate two generative adversarial networks. However, our results in Section 6.1 show that our SMT-based approach obtains substantially better results.

Our method is also connected to some previous approaches to improve machine translation using monolingual corpora. In particular, the generation of a synthetic parallel corpus through backtranslation (Sennrich et al., 2016), which is a key component of our unsupervised tuning and iterative refinement procedures, has been previously used to improve NMT. In addition, there have been several proposals to extend the phrase table of SMT systems by inducing translation candidates and/or scores from monolingual corpora, using either statistical decipherment methods (Dou and Knight, 2012, 2013) or cross-lingual embeddings (Zhao et al., 2015; Wang et al., 2016). While all these methods exploit monolingual corpora to enhance an existing machine translation system previously trained on parallel corpora, our approach learns a fully featured phrase-based SMT system from monolingual corpora alone.

8 Conclusions and future work

In this paper, we propose a novel unsupervised SMT system that can be trained on monolingual corpora alone. For that purpose, we extend the skip-gram model (Mikolov et al., 2013b) to simultaneously learn word and phrase embeddings, and map them to a cross-lingual space adapting previous unsupervised techniques (Artetxe et al., 2018b). The resulting cross-lingual phrase embeddings are used to induce a phrase table, which coupled with an n-gram language model and distance-based distortion yields an unsupervised phrase-based SMT system. We further improve results tuning the weights with our unsupervised MERT variant, and obtain additional improvements retraining the entire system through iterative backtranslation. Our implementation is available as an open source project at `https://github.com/artetxem/monoses`.

Our experiments on standard WMT French-English and German-English datasets confirm the effectiveness of our proposal, where we obtain improvements above 10 and 7 BLEU points over previous NMT-based approaches, respectively, closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 points.

In the future, we would like to extend our approach to semi-supervised scenarios with small parallel corpora, which we expect to be particularly helpful for tuning purposes. Moreover, we would like to try a hybrid approach with NMT, using our unsupervised SMT system to generate a synthetic parallel corpus and training an NMT system over it through iterative backtranslation.

Acknowledgments

This research was partially supported by the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A bilingual graph-based semantic model for statistical machine translation. In *IJCAI*, pages 2950–2956.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536, Denver, Colorado. Association for Computational Linguistics.

An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

While machine translation has traditionally relied on large amounts of parallel corpora, a recent research line has managed to train both Neural Machine Translation (NMT) and Statistical Machine Translation (SMT) systems using monolingual corpora only. In this paper, we identify and address several deficiencies of existing unsupervised SMT approaches by exploiting subword information, developing a theoretically well founded unsupervised tuning method, and incorporating a joint refinement procedure. Moreover, we use our improved SMT system to initialize a dual NMT model, which is further fine-tuned through on-the-fly back-translation. Together, we obtain large improvements over the previous state-of-the-art in unsupervised machine translation. For instance, we get 22.5 BLEU points in English-to-German WMT 2014, 5.5 points more than the previous best unsupervised system, and 0.5 points more than the (supervised) shared task winner back in 2014.

1 Introduction

The recent advent of neural sequence-to-sequence modeling has resulted in significant progress in the field of machine translation, with large improvements in standard benchmarks (Vaswani et al., 2017; Edunov et al., 2018) and the first solid claims of human parity in certain settings (Hassan et al., 2018). Unfortunately, these systems rely on large amounts of parallel corpora, which are only available for a few combinations of major languages like English, German and French.

Aiming to remove this dependency on parallel data, a recent research line has managed to train unsupervised machine translation systems using monolingual corpora only. The first such systems were based on Neural Machine Translation (NMT), and combined denoising autoencoding and back-translation to train a dual model initialized with cross-lingual embeddings (Artetxe et al., 2018c; Lample et al., 2018a). Nevertheless, these early systems were later superseded by Statistical Machine Translation (SMT) based approaches, which induced an initial phrase-table through cross-lingual embedding mappings, combined it with an n-gram language model, and further improved the system through iterative back-translation (Lample et al., 2018b; Artetxe et al., 2018b).

In this paper, we develop a more principled approach to unsupervised SMT, addressing several deficiencies of previous systems by incorporating subword information, applying a theoretically well founded unsupervised tuning method, and developing a joint refinement procedure. In addition to that, we use our improved SMT approach to initialize an unsupervised NMT system, which is further improved through on-the-fly back-translation.

Our experiments on WMT 2014/2016 French-English and German-English show the effectiveness of our approach, as our proposed system outperforms the previous state-of-the-art in unsupervised machine translation by 5-7 BLEU points in all these datasets and translation directions. Our system also outperforms the supervised WMT 2014 shared task winner in English-to-German, and is around 2 BLEU points behind it in the rest of translation directions, suggesting that unsupervised machine translation can be a usable alternative in practical settings.

The remaining of this paper is organized as follows. Section 2 first discusses the related work in the topic. Section 3 then describes our principled unsupervised SMT method, while Section 4 discusses our hybridization method with NMT. We then present the experiments done and the results obtained in Section 5, and Section 6 concludes the paper.

2 Related work

Early attempts to build machine translation systems with monolingual corpora go back to statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012). These methods see the source language as ciphertext produced by a noisy channel model that first generates the original English text and then probabilistically replaces the words in it. The English generative process is modeled using an n-gram language model, and the channel model parameters are estimated using either expectation maximization or Bayesian inference. This basic approach was later improved by incorporating syntactic knowledge (Dou and Knight, 2013) and word embeddings (Dou et al., 2015). Nevertheless, these methods were only shown to work in limited settings, being most often evaluated in word-level translation.

More recently, the task got a renewed interest after the concurrent work of Artetxe et al. (2018c) and Lample et al. (2018a) on unsupervised NMT which, for the first time, obtained promising results in standard machine translation benchmarks using monolingual corpora only. Both methods build upon the recent work on unsupervised cross-lingual embedding mappings, which independently train word embeddings in two languages and learn a linear transformation to map them to a shared space through self-learning (Artetxe et al., 2017, 2018a) or adversarial training (Conneau et al., 2018). The resulting cross-lingual embeddings are used to initialize a shared encoder for both languages, and the entire system is trained using a combination of denoising autoencoding, back-translation and, in the case of Lample et al. (2018a), adversarial training. This method was further improved by Yang et al. (2018), who use two language-specific encoders sharing only a subset of their parameters, and incorporate a local and a global generative adversarial network. Concurrent to our work, Lample and Conneau (2019) report strong results initializing an unsupervised NMT system with a cross-lingual language model.

Following the initial work on unsupervised NMT, it was argued that the modular architecture of phrase-based SMT was more suitable for this problem, and Lample et al. (2018b) and Artetxe et al. (2018b) adapted the same principles discussed above to train an unsupervised SMT model, obtaining large improvements over the original unsupervised NMT systems. More concretely, both approaches learn cross-lingual n-gram embeddings from monolingual corpora based on the mapping method discussed earlier, and use them to induce an initial phrase-table that is combined with an n-gram language model and a distortion model. This initial system is then refined through iterative back-translation (Sennrich et al., 2016) which, in the case of Artetxe et al. (2018b), is preceded by an unsupervised tuning step. Our work identifies some deficiencies in these previous systems, and proposes a more principled approach to unsupervised SMT that incorporates subword information, uses a theoretically better founded unsupervised tuning method, and applies a joint refinement procedure, outperforming these previous systems by a substantial margin.

Very recently, some authors have tried to combine both SMT and NMT to build hybrid unsupervised machine translation systems. This idea was already explored by Lample et al. (2018b), who aided the training of their unsupervised NMT system by combining standard back-translation with synthetic parallel data generated by unsupervised SMT. Marie and Fujita (2018) go further and use synthetic parallel data from unsupervised SMT to train a conventional NMT system from scratch. The resulting NMT model is then used to augment the synthetic parallel corpus through back-translation, and a new NMT model is trained on top of it from scratch, repeating the process iteratively. Ren et al. (2019) follow a similar approach, but use SMT as posterior regularization at each iteration. As shown later in our experiments, our proposed NMT hybridization obtains substantially larger absolute gains than all these previous approaches, even if our initial SMT system is stronger and thus more challenging to improve upon.

3 Principled unsupervised SMT

Phrase-based SMT is formulated as a log-linear combination of several statistical models: a translation model, a language model, a reordering model and a word/phrase penalty. As such, building an unsupervised SMT system requires learning these different components from monolingual corpora. As it turns out, this is straightforward for most of them: the language model is learned from monolingual corpora by definition; the word and phrase penalties are parameterless; and one

can drop the standard lexical reordering model at a small cost and do with the distortion model alone, which is also parameterless. This way, the main challenge left is learning the translation model, that is, building the phrase-table.

Our proposed method starts by building an initial phrase-table through cross-lingual embedding mappings (Section 3.1). This initial phrase-table is then extended by incorporating subword information, addressing one of the main limitations of previous unsupervised SMT systems (Section 3.2). Having done that, we adjust the weights of the underlying log-linear model through a novel unsupervised tuning procedure (Section 3.3). Finally, we further improve the system by jointly refining two models in opposite directions (Section 3.4).

3.1 Initial phrase-table

So as to build our initial phrase-table, we follow Artetxe et al. (2018b) and learn n-gram embeddings for each language independently, map them to a shared space through self-learning, and use the resulting cross-lingual embeddings to extract and score phrase pairs.

More concretely, we train our n-gram embeddings using *phrase2vec*[1], a simple extension of skip-gram that applies the standard negative sampling loss of Mikolov et al. (2013) to bigram-context and trigram-context pairs in addition to the usual word-context pairs.[2] Having done that, we map the embeddings to a cross-lingual space using VecMap[3] with *identical* initialization (Artetxe et al., 2018a), which builds an initial solution by aligning identical words and iteratively improves it through self-learning. Finally, we extract translation candidates by taking the 100 nearest-neighbors of each source phrase, and score them by applying the softmax function over their cosine similarities:

$$\phi(\bar{f}|\bar{e}) = \frac{\exp\left(\cos(\bar{e},\bar{f})/\tau\right)}{\sum_{\bar{f}'}\exp\left(\cos(\bar{e},\bar{f}')/\tau\right)}$$

where the temperature τ is estimated using maximum likelihood estimation over a dictionary induced in the reverse direction. In addition to the phrase translation probabilities in both directions, the forward and reverse lexical weightings are also estimated by aligning each word in the target phrase with the one in the source phrase most likely generating it, and taking the product of their respective translation probabilities. The reader is referred to Artetxe et al. (2018b) for more details.

[1] https://github.com/artetxem/phrase2vec

[2] So as to keep the model size within a reasonable limit, we restrict the vocabulary to the most frequent 200,000 unigrams, 400,000 bigrams and 400,000 trigrams.

[3] https://github.com/artetxem/vecmap

3.2 Adding subword information

An inherent limitation of existing unsupervised SMT systems is that words are taken as atomic units, making it impossible to exploit character-level information. This is reflected in the known difficulty of these models to translate named entities, as it is very challenging to discriminate among related proper nouns based on distributional information alone, yielding to translation errors like *"Sunday Telegraph"* → *"The Times of London"* (Artetxe et al., 2018b).

So as to overcome this issue, we propose to incorporate subword information once the initial alignment is done at the word/phrase level. For that purpose, we add two additional weights to the initial phrase-table that are analogous to the lexical weightings, but use a character-level similarity function instead of word translation probabilities:

$$\text{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \text{sim}(\bar{f}_i, \bar{e}_j)\right)$$

where $\epsilon = 0.3$ guarantees a minimum similarity score, as we want to favor translation candidates that are similar at the character level without excessively penalizing those that are not. In our case, we use a simple similarity function that normalizes the Levenshtein distance $\text{lev}(\cdot)$ (Levenshtein, 1966) by the length of the words $\text{len}(\cdot)$:

$$\text{sim}(f,e) = 1 - \frac{\text{lev}(f,e)}{\max(\text{len}(f),\text{len}(e))}$$

We leave the exploration of more elaborated similarity functions and, in particular, learnable metrics (McCallum et al., 2005), for future work.

3.3 Unsupervised tuning

Having trained the underlying statistical models independently, SMT tuning aims to adjust the weights of their resulting log-linear combination to optimize some evaluation metric like BLEU in a parallel validation corpus, which is typically done through Minimum Error Rate Training or MERT (Och, 2003). Needless to say, this cannot be done in strictly unsupervised settings, but we argue that

it would still be desirable to optimize some unsupervised criterion that is expected to correlate well with test performance. Unfortunately, neither of the existing unsupervised SMT systems do so: Artetxe et al. (2018b) use a heuristic that builds two initial models in opposite directions, uses one of them to generates a synthetic parallel corpus through back-translation (Sennrich et al., 2016), and applies MERT to tune the model in the reverse direction, iterating until convergence, whereas Lample et al. (2018b) do not perform any tuning at all. In what follows, we propose a more principled approach to tuning that defines an unsupervised criterion and an optimization procedure that is guaranteed to converge to a local optimum of it.

Inspired by the previous work on CycleGANs (Zhu et al., 2017) and dual learning (He et al., 2016), our method takes two initial models in opposite directions, and defines an **unsupervised optimization objective** that combines a cyclic consistency loss and a language model loss over the two monolingual corpora E and F:

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

The cyclic consistency loss captures the intuition that the translation of a translation should be close to the original text. So as to quantify this, we take a monolingual corpus in the source language, translate it to the target language and back to the source language, and compute its BLEU score taking the original text as reference:

$$L_{cycle}(E) = 1 - \mathrm{BLEU}(\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E)), E)$$

At the same time, the language model loss captures the intuition that machine translation should produce fluent text in the target language. For that purpose, we estimate the per-word entropy in the target language corpus using an n-gram language model, and penalize higher per-word entropies in machine translated text as follows:[4]

$$L_{lm}(E) = \mathrm{LP} \cdot \max(0, \mathrm{H}(F) - \mathrm{H}(\mathrm{T}_{E\to F}(E)))^2$$

where the length penalty $\mathrm{LP} = \mathrm{LP}(E) \cdot \mathrm{LP}(F)$ penalizes excessively long translations:[5]

$$\mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}(\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E)))}{\mathrm{len}(E)}\right)$$

So as to minimize the combined loss function, we **adapt MERT to jointly optimize** the parameters of the two models. In its basic form, MERT approximates the search space for each source sentence through an n-best list, and performs a form of coordinate descent by computing the optimal value for each parameter through an efficient line search method and greedily taking the step that leads to the largest gain. The process is repeated iteratively until convergence, augmenting the n-best list with the updated parameters at each iteration so as to obtain a better approximation of the full search space. Given that our optimization objective combines two translation systems $\mathrm{T}_{F\to E}(\mathrm{T}_{E\to F}(E))$, this would require generating an n-best list for $\mathrm{T}_{E\to F}(E)$ first and, for each entry on it, generating a new n-best list with $\mathrm{T}_{F\to E}$, yielding a combined n-best list with N^2 entries. So as to make it more efficient, we propose an alternating optimization approach where we fix the parameters of one model and optimize the other with standard MERT. Thanks to this, we do not need to expand the search space of the fixed model, so we can do with an n-best list of N entries alone. Having done that, we fix the parameters of the opposite model and optimize the other, iterating until convergence.

3.4 Joint refinement

Constrained by the lack of parallel corpora, the procedure described so far makes important simplifications that could compromise its potential performance: its phrase-table is somewhat unnatural (e.g. the translation probabilities are estimated from cross-lingual embeddings rather than actual frequency counts) and it lacks a lexical reordering model altogether. So as to overcome this issue, existing unsupervised SMT methods generate a synthetic parallel corpus through back-translation and use it to train a standard SMT system from scratch, iterating until convergence.

[4] We initially tried to directly minimize the entropy of the generated text, but this worked poorly in our preliminary experiments on English-Spanish (note that we used this language pair exclusively for development to be faithful to our unsupervised scenario at test time). More concretely, the behavior of the optimization algorithm was very unstable, as it tended to excessively focus on either the cyclic consistency loss or the language model loss at the cost of the other, and we found it very difficult to find the right balance between the two factors.

[5] Without this penalization, the system tended to produce unnecessary tokens (e.g. quotes) that looked natural in their context, which served to minimize the per-word perplexity of the output. Minimizing the overall perplexity instead of the per-word perplexity did not solve the problem, as the opposite phenomenon arose (i.e. the system tended to produce excessively short translations).

An obvious drawback of this approach is that the back-translated side will contain ungrammatical n-grams and other artifacts that will end up in the induced phrase-table. One could argue that this should be innocuous as long as the ungrammatical n-grams are in the source side, as they should never occur in real text and their corresponding entries in the phrase-table should therefore not be used. However, ungrammatical source phrases do ultimately affect the estimation of the backward translation probabilities, including those of grammatical phrases.[6] We argue that, ultimately, the backward probability estimations can only be meaningful when all source phrases are grammatical (so the probabilities of all plausible translations sum to one) and, similarly, the forward probability estimations can only be meaningful when all target phrases are grammatical.

Following the above observation, we propose an alternative approach that jointly refines both translation directions. More concretely, we use the initial systems to build two synthetic corpora in opposite directions.[7] Having done that, we independently extract phrase pairs from each synthetic corpus, and build a phrase-table by taking their intersection. The forward probabilities are estimated in the parallel corpus with the synthetic source side, while the backward probabilities are estimated in the one with the synthetic target side. This does not only guarantee that the probability estimates are meaningful as discussed previously, but it also discards the ungrammatical phrases altogether, as both the source and the target n-grams must have occurred in the original monolingual texts to be present in the resulting phrase-table. This phrase-table is then combined with a lexical reordering model learned on the synthetic parallel corpus in the reverse direction, and we apply the unsupervised tuning method described in Section 3.3 to adjust the weights of the resulting system. We repeat this process for a total of 3 iterations.[8]

[6]For instance, let's say that the target phrase *"dos gatos"* has been aligned 10 times with *"two cats"* and 90 times with *"two cat"*. While the ungrammatical phrase-table entry *two cat- dos gatos* should never be picked, the backward probability estimation of *two cats - dos gatos* is still affected by it (it would be 0.1 instead of 1.0 in this example).

[7]For efficiency purposes, we restrict the size of each synthetic parallel corpus to 10 million sentence pairs.

[8]For the last iteration, we do not perform any tuning and use default Moses weights instead, which we found to be more robust during development. Note, however, that using unsupervised tuning during the previous steps was still strongly beneficial.

4 NMT hybridization

While the rigid and modular design of SMT provides a very suitable framework for unsupervised machine translation, NMT has shown to be a fairly superior paradigm in supervised settings, outperforming SMT by a large margin in standard benchmarks. As such, the choice of SMT over NMT also imposes a hard ceiling on the potential performance of these approaches, as unsupervised SMT systems inherit the very same limitations of their supervised counterparts (e.g. the locality and sparsity problems). For that reason, we argue that SMT provides a more appropriate architecture to find an initial alignment between the languages, but NMT is ultimately a better architecture to model the translation process.

Following this observation, we propose a hybrid approach that uses unsupervised SMT to warm up a dual NMT model trained through iterative back-translation. More concretely, we first train two SMT systems in opposite directions as described in Section 3, and use them to assist the training of another two NMT systems in opposite directions. These NMT systems are trained following an iterative process where, at each iteration, we alternately update the model in each direction by performing a single pass over a synthetic parallel corpus built through back-translation (Sennrich et al., 2016).[9] In the first iteration, the synthetic parallel corpus is entirely generated by the SMT system in the opposite direction but, as training progresses and the NMT models get better, we progressively switch to a synthetic parallel corpus generated by the reverse NMT model. More concretely, iteration t uses $N_{smt} = N \cdot \max(0, 1 - t/a)$ synthetic parallel sentences from the reverse SMT system, where the parameter a controls the number of transition iterations from SMT to NMT back-translation. The remaining $N - N_{smt}$ sentences are generated by the reverse NMT model. Inspired by Edunov et al. (2018), we use greedy decoding for half of them, which produces more fluent and predictable translations, and random sampling for the other half, which produces more varied translations. In our experiments, we use $N = 1,000,000$ and $a = 30$, and perform a total of 60 such iterations. At test time, we use beam search decoding with an ensemble of all check-

[9]Note that we do not train a new model from scratch each time, but continue training the model from the previous iteration.

		WMT-14				WMT-16	
		fr-en	en-fr	de-en	en-de	de-en	en-de
NMT	Artetxe et al. (2018c)	15.6	15.1	10.2	6.6	-	-
	Lample et al. (2018a)	14.3	15.1	-	-	13.3	9.6
	Yang et al. (2018)	15.6	17.0	-	-	14.6	10.9
	Lample et al. (2018b)	24.2	25.1	-	-	21.0	17.2
SMT	Artetxe et al. (2018b)	25.9	26.2	17.4	14.1	23.1	18.2
	Lample et al. (2018b)	27.2	28.1	-	-	22.9	17.9
	Marie and Fujita (2018)*	-	-	-	-	20.2	15.5
	Proposed system	28.4	30.1	20.1	15.8	25.4	19.7
	*detok. SacreBLEU**	27.9	27.8	19.7	14.7	24.8	19.4
SMT + NMT	Lample et al. (2018b)	27.7	27.6	-	-	25.2	20.2
	Marie and Fujita (2018)*	-	-	-	-	26.7	20.0
	Ren et al. (2019)	28.9	29.5	20.4	17.0	26.3	21.7
	Proposed system	**33.5**	**36.2**	**27.0**	**22.5**	**34.4**	**26.9**
	*detok. SacreBLEU**	33.2	33.6	26.4	21.2	33.8	26.4

Table 1: Results of the proposed method in comparison to previous work (BLEU). Overall best results are in bold, the best ones in each group are underlined.
*Detokenized BLEU equivalent to the official `mteval-v13a.pl` script. The rest use tokenized BLEU with `multi-bleu.perl` (or similar).

points from every 10 iterations.

5 Experiments and results

In order to make our experiments comparable to previous work, we use the French-English and German-English datasets from the WMT 2014 shared task. More concretely, our training data consists of the concatenation of all News Crawl monolingual corpora from 2007 to 2013, which make a total of 749 million tokens in French, 1,606 millions in German, and 2,109 millions in English, from which we take a random subset of 2,000 sentences for tuning (Section 3.3). Preprocessing is done using standard Moses tools, and involves punctuation normalization, tokenization with aggressive hyphen splitting, and truecasing.

Our SMT implementation is based on Moses[10], and we use the KenLM (Heafield et al., 2013) tool included in it to estimate our 5-gram language model with modified Kneser-Ney smoothing. Our unsupervised tuning implementation is based on Z-MERT (Zaidan, 2009), and we use FastAlign (Dyer et al., 2013) for word alignment within the joint refinement procedure. Finally, we use the big transformer implementation from fairseq[11] for our NMT system, training with a total batch size of 20,000 tokens across 8 GPUs with the exact same hyperparameters as Ott et al. (2018).

We use newstest2014 as our test set for French-English, and both newstest2014 and newstest2016 (from WMT 2016[12]) for German-English. Following common practice, we report tokenized BLEU scores as computed by the `multi-bleu.perl` script included in Moses. In addition to that, we also report detokenized BLEU scores as computed by SacreBLEU[13] (Post, 2018), which is equivalent to the official `mteval-v13a.pl` script.

We next present the results of our proposed system in comparison to previous work in Section 5.1. Section 5.2 then compares the obtained results to those of different supervised systems. Finally, Section 5.3 presents some translation examples from our system.

5.1 Main results

Table 1 reports the results of the proposed system in comparison to previous work. As it can be seen, our full system obtains the best published results in all cases, outperforming the previous state-of-the-art by 5-7 BLEU points in all datasets and translation directions.

A substantial part of this improvement comes from our more principled unsupervised SMT ap-

[10] http://www.statmt.org/moses/
[11] https://github.com/pytorch/fairseq

[12] Note that it is only the test set that is from WMT 2016. All the training data comes from WMT 2014 News Crawl, so it is likely that our results could be further improved by using the more extensive monolingual corpora from WMT 2016.
[13] SacreBLEU signature: BLEU+case.mixed+lang.LANG +numrefs.1+smooth.exp+test.TEST+tok.13a+version.1.2.11, with LANG ∈ {fr-en, en-fr, de-en, en-de} and TEST ∈ {wmt14/full, wmt16}

		WMT-14		WMT-16	
		fr-en	en-fr	de-en	en-de
Lample et al. (2018b)	Initial SMT	27.2	28.1	22.9	17.9
	+ NMT hybrid	27.7 (+0.5)	27.6 (-0.5)	25.2 (+2.3)	20.2 (+2.3)
Marie and Fujita (2018)	Initial SMT	-	-	20.2	15.5
	+ NMT hybrid	-	-	26.7 (+6.5)	20.0 (+4.5)
Proposed system	Initial SMT	28.4	30.1	25.4	19.7
	+ NMT hybrid	**33.5 (+5.1)**	**36.2 (+6.1)**	**34.4 (+9.0)**	**26.9 (+7.2)**

Table 2: NMT hybridization results for different unsupervised machine translation systems (BLEU).

		WMT-14			
		fr-en	en-fr	de-en	en-de
Unsupervised	Proposed system	33.5	36.2	27.0	22.5
	*detok. SacreBLEU**	33.2	33.6	26.4	21.2
Supervised	WMT best*	35.0	35.8	29.0	20.6†
	Vaswani et al. (2017)	-	41.0	-	28.4
	Edunov et al. (2018)	-	45.6	-	35.0

Table 3: Results of the proposed method in comparison to different supervised systems (BLEU).
*Detokenized BLEU equivalent to the official `mteval-v13a.pl` script. The rest use tokenized BLEU with `multi-bleu.perl` (or similar).
†Results in the original test set from WMT 2014, which slightly differs from the full test set used in all subsequent work. Our proposed system obtains 22.4 BLEU points (21.1 detokenized) in that same subset.

proach, which outperforms all previous SMT-based systems by around 2 BLEU points. Nevertheless, it is the NMT hybridization that brings the largest gains, improving the results of this initial SMT systems by 5-9 BLEU points. As shown in Table 2, our absolute gains are considerably larger than those of previous hybridization methods, even if our initial SMT system is substantially better and thus more difficult to improve upon. This way, our initial SMT system is about 4-5 BLEU points above that of Marie and Fujita (2018), yet our absolute gain on top of it is around 2.5 BLEU points higher. When compared to Lample et al. (2018b), we obtain an absolute gain of 5-6 BLEU points in both French-English directions while they do not get any clear improvement, and we obtain an improvement of 7-9 BLEU points in both German-English directions, in contrast with the 2.3 BLEU points they obtain.

More generally, it is interesting that pure SMT systems perform better than pure NMT systems, yet the best results are obtained by initializing an NMT system with an SMT system. This suggests that the rigid and modular architecture of SMT might be more suitable to find an initial alignment between the languages, but the final system should be ultimately based on NMT for optimal results.

5.2 Comparison with supervised systems

So as to put our results into perspective, Table 3 reports the results of different supervised systems in the same WMT 2014 test set. More concretely, we include the best results from the shared task itself, which reflect the state-of-the-art in machine translation back in 2014; those of Vaswani et al. (2017), who introduced the now predominant transformer architecture; and those of Edunov et al. (2018), who apply back-translation at a large scale and, to the best of our knowledge, hold the current best results in the test set.

As it can be seen, our unsupervised system outperforms the WMT 2014 shared task winner in English-to-German, and is around 2 BLEU points behind it in the other translation directions. This shows that unsupervised machine translation is already competitive with the state-of-the-art in supervised machine translation in 2014. While the field of machine translation has undergone great progress in the last 5 years, and the gap between our unsupervised system and the current state-of-the-art in supervised machine translation is still large as reflected by the other results, this suggests that unsupervised machine translation can be a usable alternative in practical settings.

Source	Reference	Artetxe et al. (2018b)	Proposed system
D'autres révélations ont fait état de documents divulgués par Snowden selon lesquels la NSA avait intercepté des données et des communications émanant du téléphone portable de la chancelière allemande Angela Merkel et de ceux de 34 autres chefs d'État.	Other revelations cited documents leaked by Snowden that the NSA monitored German Chancellor Angela Merkel's cellphone and those of up to 34 other world leaders.	Other disclosures have reported documents disclosed by Snowden suggested the NSA had intercepted communications and data from the mobile phone of German Chancellor Angela Merkel and those of 32 other heads of state.	Other revelations have pointed to documents disclosed by Snowden that the NSA had intercepted data and communications emanating from German Chancellor Angela Merkel's mobile phone and those of 34 other heads of state.
La NHTSA n'a pas pu examiner la lettre d'information aux propriétaires en raison de l'arrêt de 16 jours des activités gouvernementales, ce qui a ralenti la croissance des ventes de véhicules en octobre.	NHTSA could not review the owner notification letter due to the 16-day government shutdown, which tempered auto sales growth in October.	The NHTSA could not consider the letter of information to owners because of halting 16-day government activities, which slowed the growth in vehicle sales in October.	NHTSA said it could not examine the letter of information to owners because of the 16-day halt in government operations, which slowed vehicle sales growth in October.
Le M23 est né d'une mutinerie, en avril 2012, d'anciens rebelles, essentiellement tutsi, intégrés dans l'armée en 2009 après un accord de paix.	The M23 was born of an April 2012 mutiny by former rebels, principally Tutsis who were integrated into the army in 2009 following a peace agreement.	M23 began as a mutiny in April 2012, former rebels, mainly Tutsi integrated into the national army in 2009 after a peace deal.	The M23 was born into a mutiny in April 2012, of former rebels, mostly Tutsi, embedded in the army in 2009 after a peace deal.
Tunks a déclaré au Sunday Telegraph de Sydney que toute la famille était «extrêmement préoccupée» du bien-être de sa fille et voulait qu'elle rentre en Australie.	Tunks told Sydney's Sunday Telegraph the whole family was "extremely concerned" about his daughter's welfare and wanted her back in Australia.	Tunks told The Times of London from Sydney that the whole family was "extremely concerned" of the welfare of her daughter and wanted it to go in Australia.	Tunks told the Sunday Telegraph in Sydney that the whole family was "extremely concerned" about her daughter's well-being and wanted her to go into Australia.

Table 4: Randomly chosen translation examples from French→English newstest2014 in comparison of those reported by Artetxe et al. (2018b).

5.3 Qualitative results

Table 4 shows some translation examples from our proposed system in comparison to those reported by Artetxe et al. (2018b). We choose the exact same sentences reported by Artetxe et al. (2018b), which were randomly taken from newstest2014, so they should be representative of the general behavior of both systems.

While not perfect, our proposed system produces generally fluent translations that accurately capture the meaning of the original text. Just in line with our quantitative results, this suggests that unsupervised machine translation can be a usable alternative in practical settings.

Compared to Artetxe et al. (2018b), our translations are generally more fluent, which is not surprising given that they are produced by an NMT system rather than an SMT system. In addition to that, the system of Artetxe et al. (2018b) has some adequacy issues when translating named entities and numerals (e.g. *34* → *32*, *Sunday Telegraph* → *The Times of London*), which we do not observe for our proposed system in these examples.

6 Conclusions and future work

In this paper, we identify several deficiencies in previous unsupervised SMT systems, and propose a more principled approach that addresses them by incorporating subword information, using a theoretically well founded unsupervised tuning method, and developing a joint refinement procedure. In addition to that, we use our improved SMT approach to initialize a dual NMT model that is further improved through on-the-fly back-translation. Our experiments show the effectiveness of our approach, as we improve the previous state-of-the-art in unsupervised machine translation by 5-7 BLEU points in French-English and German-English WMT 2014 and 2016. Our code is available as an open source project at `https://github.com/artetxem/monoses`.

In the future, we would like to explore learnable similarity functions like the one proposed by (McCallum et al., 2005) to compute the character-level scores in our initial phrase-table. In addition to that, we would like to incorporate a language modeling loss during NMT training similar to He

et al. (2016). Finally, we would like to adapt our approach to more relaxed scenarios with multiple languages and/or small parallel corpora.

Acknowledgments

This research was partially supported by the Spanish MINECO (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, cofunded by EU FEDER), the BigKnowledge project (BBVA foundation grant 2018), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe was supported by a doctoral grant from the Spanish MECD.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*, pages 820–828.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b.

Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2005. A conditional random field for discriminatively-trained finite-state string edit distance. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 388–395.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *arXiv preprint arXiv:1901.04112*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.

Omar Zaidan. 2009. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Bilingual Lexicon Induction through Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

A recent research line has obtained strong results on bilingual lexicon induction by aligning independently trained word embeddings in two languages and using the resulting cross-lingual embeddings to induce word translation pairs through nearest neighbor or related retrieval methods. In this paper, we propose an alternative approach to this problem that builds on the recent work on unsupervised machine translation. This way, instead of directly inducing a bilingual lexicon from cross-lingual embeddings, we use them to build a phrase-table, combine it with a language model, and use the resulting machine translation system to generate a synthetic parallel corpus, from which we extract the bilingual lexicon using statistical word alignment techniques. As such, our method can work with any word embedding and cross-lingual mapping technique, and it does not require any additional resource besides the monolingual corpus used to train the embeddings. When evaluated on the exact same cross-lingual embeddings, our proposed method obtains an average improvement of 6 accuracy points over nearest neighbor and 4 points over CSLS retrieval, establishing a new state-of-the-art in the standard MUSE dataset.

1 Introduction

Cross-lingual word embedding mappings have attracted a lot of attention in recent times. These methods work by independently training word embeddings in different languages, and mapping them to a shared space through linear transformations. While early methods required a training dictionary to find the initial alignment (Mikolov et al., 2013), fully unsupervised methods have managed to obtain comparable results based on either adversarial training (Conneau et al., 2018) or self-learning (Artetxe et al., 2018b).

A prominent application of these methods is Bilingual Lexicon Induction (BLI), that is, using the resulting cross-lingual embeddings to build a bilingual dictionary. For that purpose, one would typically induce the translation of each source word by taking its corresponding nearest neighbor in the target language. However, it has been argued that this basic approach suffers from the hubness problem[1], which has motivated alternative retrieval methods like inverted nearest neighbor[2] (Dinu et al., 2015), inverted softmax (Smith et al., 2017), and Cross-domain Similarity Local Scaling (CSLS) (Conneau et al., 2018).

In this paper, we go one step further and, rather than directly inducing the bilingual dictionary from the cross-lingual word embeddings, we use them to build an unsupervised machine translation system, and extract a bilingual dictionary from a synthetic parallel corpus generated with it. This allows us to take advantage of a strong language model and naturally extract translation equivalences through statistical word alignment. At the same time, our method can be used as a drop-in replacement of traditional retrieval techniques, as it can work with any cross-lingual word embeddings and it does not require any additional resource besides the monolingual corpus used to train them. Our experiments show the effectiveness of this alternative approach, which outperforms the previous best retrieval method by 4 accuracy points on average, establishing a new state-of-the-art in the standard MUSE dataset. As such, we conclude that, contrary to recent trend, future research in BLI should not focus exclusively on direct retrieval methods.

[1] Hubness (Radovanović et al., 2010a,b) refers to the phenomenon of a few points being the nearest neighbors of many other points in high-dimensional spaces, which has been reported to severely affect cross-lingual embedding mappings (Dinu et al., 2015).

[2] The original paper refers to this method as *globally corrected* retrieval.

2 Proposed method

The input of our method is a set of cross-lingual word embeddings and the monolingual corpora used to train them. In our experiments, we use fastText embeddings (Bojanowski et al., 2017) mapped through VecMap (Artetxe et al., 2018b), but the algorithm described next can also work with any other word embedding and cross-lingual mapping method.

The general idea of our method is to to build an unsupervised phrase-based statistical machine translation system (Lample et al., 2018; Artetxe et al., 2018c, 2019), and use it to generate a synthetic parallel corpus from which to extract a bilingual dictionary. For that purpose, we first derive phrase embeddings from the input word embeddings by taking the 400,000 most frequent bigrams and and the 400,000 most frequent trigrams in each language, and assigning them the centroid of the words they contain. Having done that, we use the resulting cross-lingual phrase embeddings to build a phrase-table as described in Artetxe et al. (2018c). More concretely, we extract translation candidates by taking the 100 nearest-neighbors of each source phrase, and score them with the softmax function over their cosine similarities:

$$\phi(\bar{f}|\bar{e}) = \frac{\exp\left(\cos(\bar{e},\bar{f})/\tau\right)}{\sum_{\bar{f}'}\exp\left(\cos(\bar{e},\bar{f}')/\tau\right)}$$

where the temperature τ is estimated using maximum likelihood estimation over a dictionary induced in the reverse direction. In addition to the phrase translation probabilities in both directions, we also estimate the forward and reverse lexical weightings by aligning each word in the target phrase with the one in the source phrase most likely generating it, and taking the product of their respective translation probabilities.

We then combine this phrase-table with a distortion model and a 5-gram language model estimated in the target language corpus, which results in a phrase-based machine translation system. So as to optimize the weights of the resulting model, we use the unsupervised tuning procedure proposed by Artetxe et al. (2019), which combines a cyclic consistency loss and a language modeling loss over a subset of 2,000 sentences from each monolingual corpora.

Having done that, we generate a synthetic parallel corpus by translating the source language monolingual corpus with the resulting machine translation system.[3] We then word align this corpus using FastAlign (Dyer et al., 2013) with default hyperparameters and the *grow-diag-final-and* symmetrization heuristic. Finally, we build a phrase-table from the word aligned corpus, and extract a bilingual dictionary from it by discarding all non-unigram entries. For words with more than one entry, we rank translation candidates according to their direct translation probability.

3 Experimental settings

In order to compare our proposed method head-to-head with other BLI methods, the experimental setting needs to fix the monolingual embedding training method, as well as the cross-lingual mapping algorithm and the evaluation dictionaries. In addition, in order to avoid any advantage, our method should not see any further monolingual corpora than those used to train the monolingual embeddings. Unfortunately, existing BLI datasets distribute pre-trained word embeddings alone, but not the monolingual corpora used to train them. For that reason, we decide to use the evaluation dictionaries from the standard MUSE dataset (Conneau et al., 2018) but, instead of using the pre-trained Wikipedia embeddings distributed with it, we extract monolingual corpora from Wikipedia ourselves and train our own embeddings trying to be as faithful as possible to the original settings. This allows us to compare our proposed method to previous retrieval techniques in the exact same conditions, while keeping our results as comparable as possible to previous work reporting results for the MUSE dataset.

More concretely, we use WikiExtractor[4] to extract plain text from Wikipedia dumps, and preprocess the resulting corpus using standard Moses tools (Koehn et al., 2007) by applying sentence splitting, punctuation normalization, tokenization with aggressive hyphen splitting, and lowercasing. We then train word embeddings for each language using the skip-gram implementation of fastText (Bojanowski et al., 2017) with default hyperparameters, restricting the vocabulary to the 200,000 most frequent tokens. The official embeddings in

[3] For efficiency purposes, we restricted the size of the synthetic parallel corpus to a maximum of 10 million sentences, and use cube-pruning for faster decoding. As such, our results could likely be improved by translating the full monolingual corpus with standard decoding.

[4] https://github.com/attardi/wikiextractor

	en-es		en-fr		en-de		en-ru		avg.
	→	←	→	←	→	←	→	←	
Nearest neighbor	81.9	82.8	81.6	81.7	73.3	72.3	44.3	65.6	72.9
Inv. nearest neighbor (Dinu et al., 2015)	80.6	77.6	81.3	79.0	69.8	69.7	43.7	54.1	69.5
Inv. softmax (Smith et al., 2017)	81.7	82.7	81.7	81.7	73.5	72.3	44.4	65.5	72.9
CSLS (Conneau et al., 2018)	82.5	84.7	83.3	83.4	75.6	75.3	47.4	67.2	74.9
Proposed method	**87.0**	**87.9**	**86.0**	**86.2**	**81.9**	**80.2**	**50.4**	**71.3**	**78.9**

Table 1: P@1 of proposed system and previous retrieval methods, using the same cross-lingual embeddings.

the MUSE dataset were trained using these exact same settings, so our embeddings only differ in the Wikipedia dump used to extract the training corpus and the pre-processing applied to it, which is not documented in the original dataset.

Having done that, we map these word embeddings to a cross-lingual space using the unsupervised mode in VecMap (Artetxe et al., 2018b), which builds an initial solution based on the intralingual similarity distribution of the embeddings and iteratively improves it through self-learning. Finally, we induce a bilingual dictionary using our proposed method and evaluate it in comparison to previous retrieval methods (standard nearest neighbor, inverted nearest neighbor, inverted softmax[5] and CSLS). Following common practice, we use precision at 1 as our evaluation measure.[6]

4 Results and discussion

Table 1 reports the results of our proposed system in comparison to previous retrieval methods. As it can be seen, our method obtains the best results in all language pairs and directions, with an average improvement of 6 points over nearest neighbor and 4 points over CSLS, which is the best performing previous method. These results are very consistent across all translation directions, with an absolute improvement between 2.7 and 6.3 points over CSLS. Interestingly, neither inverted nearest neighbor nor inverted softmax are able to outperform standard nearest neighbor, presumably because our cross-lingual embeddings are less sensitive to hubness thanks to the symmetric re-weighting in VecMap (Artetxe et al., 2018a). At the same time, CSLS obtains an absolute improvement of 2 points over nearest neighbor, only a third of what our method achieves. This suggests that, while previous retrieval methods have almost exclusively focused on addressing the hubness problem, there is a substantial margin of improvement beyond this phenomenon.

So as to put these numbers into perspective, Table 2 compares our method to previous results reported in the literature.[7] As it can be seen, our proposed method obtains the best published results in all language pairs and directions, outperforming the previous state-of-the-art by a substantial margin. Note, moreover, that these previous systems mostly differ in their cross-lingual mapping algorithm and not the retrieval method, so our improvements are orthogonal.

We believe that, beyond the substantial gains in this particular task, our work has **important implications** for future research in cross-lingual word embedding mappings. While most work in this topic uses BLI as the only evaluation task, Glavas et al. (2019) recently showed that BLI results do not always correlate well with downstream performance. In particular, they observe that some mapping methods that are specifically designed for BLI perform poorly in other tasks. Our work shows that, besides their poor performance in those tasks, these BLI-centric mapping methods might not even be the optimal approach to BLI, as our alternative method, which relies on unsupervised machine translation instead of direct

[5] Inverted softmax has a temperature hyperparameter T, which is typically tuned in the training dictionary. Given that we do not have any training dictionary in our fully unsupervised settings, we use a fixed temperature of $T = 30$, which was also used by some previous authors (Lample et al., 2018). While we tried other values in our preliminary experiments, but we did not observe any significant difference.

[6] We find a few out-of-vocabularies in the evaluation dictionary that are likely caused by minor pre-processing differences. In those cases, we use copying as a back-off strategy (i.e. if a given word is not found in our induced dictionary, we simply leave it unchanged). In any case, the percentage of out-of-vocabularies is always below 1%, so this has a negligible effect in the reported results.

[7] Note that previous results are based on the pre-trained embeddings of the MUSE dataset, while we had to train our embeddings to have a controlled experiment (see Section 3). In any case, our embeddings are trained following the official dataset setting, using Wikipedia, the same system and hyperparameters, so our results should be roughly comparable.

	en-es		en-fr		en-de		en-ru		avg.
	→	←	→	←	→	←	→	←	
Conneau et al. (2018)	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	72.3
Hoshen and Wolf (2018)	82.1	84.1	82.3	82.9	74.7	73.0	47.5	61.8	73.6
Grave et al. (2018)	82.8	84.1	82.6	82.9	75.4	73.3	43.7	59.1	73.0
Alvarez-Melis and Jaakkola (2018)	81.7	80.4	81.3	78.9	71.9	72.8	45.1	43.7	69.5
Yang et al. (2018)	79.9	79.3	78.4	78.9	71.5	70.3	-	-	-
Mukherjee et al. (2018)	84.5	79.2	-	-	-	-	-	-	-
Alvarez-Melis et al. (2018)	81.3	81.8	82.9	81.6	73.8	71.1	41.7	55.4	71.2
Xu et al. (2018)	79.5	77.8	77.9	75.5	69.3	67.0	-	-	-
Proposed method	**87.0**	**87.9**	**86.0**	**86.2**	**81.9**	**80.2**	**50.4**	**71.3**	**78.9**

Table 2: Results of the proposed method in comparison to previous work (P@1). All systems are fully unsupervised and use fastText embeddings trained on Wikipedia with the same hyperparameters.

retrieval over mapped embeddings, obtains substantially better results without requiring any additional resource. As such, we argue that 1) future work in cross-lingual word embeddings should consider other evaluation tasks in addition to BLI, and 2) future work in BLI should consider other alternatives in addition to direct retrieval over cross-lingual embedding mappings.

5 Related work

While BLI has been previously tackled using count-based vector space models (Vulić and Moens, 2013) and statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012), these methods have recently been superseded by cross-lingual embedding mappings, which work by aligning independently trained word embeddings in different languages. For that purpose, early methods required a training dictionary, which was used to learn a linear transformation that mapped these embeddings into a shared cross-lingual space (Mikolov et al., 2013; Artetxe et al., 2018a). The resulting cross-lingual embeddings are then used to induce the translations of words that were missing in the training dictionary by taking their nearest neighbor in the target language.

The amount of required supervision was later reduced through self-learning methods (Artetxe et al., 2017), and then completely eliminated through adversarial training (Zhang et al., 2017a; Conneau et al., 2018) or more robust iterative approaches combined with initialization heuristics (Artetxe et al., 2018b; Hoshen and Wolf, 2018). At the same time, several recent methods have formulated embedding mappings as an optimal transport problem (Zhang et al., 2017b; Grave et al., 2018; Alvarez-Melis and Jaakkola, 2018).

In addition to that, a large body of work has focused on addressing the hubness problem that arises when directly inducing bilingual dictionaries from cross-lingual embeddings, either through the retrieval method (Dinu et al., 2015; Smith et al., 2017; Conneau et al., 2018) or the mapping itself (Lazaridou et al., 2015; Shigeto et al., 2015; Joulin et al., 2018). While all these previous methods directly induce bilingual dictionaries from cross-lingually mapped embeddings, our proposed method combines them with unsupervised machine translation techniques, outperforming them all by a substantial margin.

6 Conclusions and future work

We propose a new approach to BLI which, instead of directly inducing bilingual dictionaries from cross-lingual embedding mappings, uses them to build an unsupervised machine translation system, which is then used to generate a synthetic parallel corpus from which to extract bilingual lexica. Our approach does not require any additional resource besides the monolingual corpora used to train the embeddings, and outperforms traditional retrieval techniques by a substantial margin. We thus conclude that, contrary to recent trend, future work in BLI should not focus exclusively in direct retrieval approaches, nor should BLI be the only evaluation task for cross-lingual embeddings. Our code is available at `https://github.com/artetxem/monoses`.

In the future, we would like to further improve our method by incorporating additional ideas from unsupervised machine translation such as joint refinement and neural hybridization (Artetxe et al., 2019). In addition to that, we would like to integrate our induced dictionaries in other downstream

tasks like unsupervised cross-lingual information retrieval (Litschko et al., 2018).

Acknowledgments

This research was partially supported by the Spanish MINECO (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, co-funded by EU FEDER), the BigKnowledge project (BBVA foundation grant 2018), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe was supported by a doctoral grant from the Spanish MECD.

References

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. 2018. Towards optimal transport with global invariances. *arXiv preprint arXiv:1806.09277*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of*

the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280. Association for Computational Linguistics.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. *arXiv preprint arXiv:1805.00879*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. *Ridge Regression, Hubness, and Zero-Shot Learning*, pages 135–151. Springer International Publishing.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR 2017)*.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Shuangzhi Wu, Jingjing Xu, Dongdong Zhang, and Xu Sun. 2018. Learning unsupervised word mapping by maximizing mean discrepancy. *arXiv preprint arXiv:1811.00275*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

A Call for More Rigor in Unsupervised Cross-lingual Learning

Mikel Artetxe[†*], **Sebastian Ruder**[‡*], **Dani Yogatama**[‡], **Gorka Labaka**[†], **Eneko Agirre**[†]
[†]HiTZ Center, University of the Basque Country (UPV/EHU)
[‡]DeepMind
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus
{ruder,dyogatama}@google.com

Abstract

We review motivations, definition, approaches, and methodology for unsupervised cross-lingual learning and call for a more rigorous position in each of them. An existing rationale for such research is based on the lack of parallel data for many of the world's languages. However, we argue that a scenario without *any* parallel data and abundant monolingual data is unrealistic in practice. We also discuss different training signals that have been used in previous work, which depart from the pure unsupervised setting. We then describe common methodological issues in tuning and evaluation of unsupervised cross-lingual models and present best practices. Finally, we provide a unified outlook for different types of research in this area (i.e., cross-lingual word embeddings, deep multilingual pretraining, and unsupervised machine translation) and argue for comparable evaluation of these models.

1 Introduction

The study of the connection among human languages has contributed to major discoveries including the evolution of languages, the reconstruction of proto-languages, and an understanding of language universals (Eco and Fentress, 1995). In natural language processing, the main promise of multilingual learning is to bridge the digital language divide, to enable access to information and technology for the world's 6,900 languages (Ruder et al., 2019). For the purpose of this paper, we define *"multilingual learning"* as learning a common model for two or more languages from raw text, without any downstream task labels. Common use cases include translation as well as pretraining multilingual representations. We will use the term interchangeably with *"cross-lingual learning"*.

[*]Equal contribution.

Recent work in this direction has increasingly focused on purely unsupervised cross-lingual learning (UCL)—i.e., cross-lingual learning without any parallel signal across the languages. We provide an overview in §2. Such work has been motivated by the apparent dearth of parallel data for most of the world's languages. In particular, previous work has noted that *"data encoding cross-lingual equivalence is often expensive to obtain"* (Zhang et al., 2017a) whereas *"monolingual data is much easier to find"* (Lample et al., 2018a). Overall, it has been argued that unsupervised cross-lingual learning *"opens up opportunities for the processing of extremely low-resource languages and domains that lack parallel data completely"* (Zhang et al., 2017a).

We challenge this narrative and argue that the scenario of no parallel data *and* sufficient monolingual data is unrealistic and not reflected in the real world (§3.1). Nevertheless, UCL is an important research direction and we advocate for its study based on an inherent scientific interest (to better understand and make progress on general language understanding), usefulness as a lab setting, and simplicity (§3.2).

Unsupervised cross-lingual learning permits no supervisory signal by definition. However, previous work implicitly includes monolingual and cross-lingual signals that constitute a departure from the pure setting. We review existing training signals as well as other signals that may be of interest for future study (§4). We then discuss methodological issues in UCL (e.g., validation, hyperparameter tuning) and propose best evaluation practices (§5). Finally, we provide a unified outlook of established research areas (cross-lingual word embeddings, deep multilingual models and unsupervised machine translation) in UCL (§6), and conclude with a summary of our recommendations (§7).

2 Background

In this section, we briefly review existing work on UCL, covering cross-lingual word embeddings (§2.1), deep multilingual pre-training (§2.2), and unsupervised machine translation (§2.3).

2.1 Cross-lingual word embeddings

Cross-lingual word embedding methods traditionally relied on parallel corpora (Gouws et al., 2015; Luong et al., 2015). Nonetheless, the amount of supervision required was greatly reduced via cross-lingual word embedding mappings, which work by separately learning monolingual word embeddings in each language and mapping them into a shared space through a linear transformation. Early work required a bilingual dictionary to learn such a transformation (Mikolov et al., 2013a; Faruqui and Dyer, 2014). This requirement was later reduced with self-learning (Artetxe et al., 2017), and ultimately removed via unsupervised initialization heuristics (Artetxe et al., 2018a; Hoshen and Wolf, 2018) and adversarial learning (Zhang et al., 2017a; Conneau et al., 2018a). Finally, several recent methods have formulated cross-lingual embedding alignment as an optimal transport problem (Zhang et al., 2017b; Grave et al., 2019; Alvarez-Melis and Jaakkola, 2018).

2.2 Deep multilingual pretraining

Following the success in learning shallow word embeddings (Mikolov et al., 2013b; Pennington et al., 2014), there has been an increasing interest in learning contextual word representations (Dai and Le, 2015; Peters et al., 2018; Howard and Ruder, 2018). Recent research has been dominated by BERT (Devlin et al., 2019), which uses a bidirectional transformer encoder trained on masked language modeling and next sentence prediction, which led to impressive gains on various downstream tasks.

While the above approaches are limited to a single language, a multilingual extension of BERT (mBERT) has been shown to also be effective at learning cross-lingual representations in an unsupervised way.[1] The main idea is to combine monolingual corpora in different languages, upsampling those with less data, and training a regular BERT model on the combined data. Conneau and Lample (2019) follow a similar approach but perform a more thorough evaluation and report substantially stronger results,[2] which was further scaled up by Conneau et al. (2019). Several recent studies (Wu and Dredze, 2019; Pires et al., 2019; Artetxe et al., 2020b; Wu et al., 2019) analyze mBERT to get a better understanding of its capabilities.

2.3 Unsupervised machine translation

Early attempts to build machine translation systems using monolingual data alone go back to statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012, 2013). However, this approach was only shown to work in limited settings, and the first convincing results on standard benchmarks were achieved by Artetxe et al. (2018c) and Lample et al. (2018a) on unsupervised Neural Machine Translation (NMT). Both approaches rely on cross-lingual word embeddings to initialize a shared encoder, and train it in conjunction with the decoder using a combination of denoising autoencoding, back-translation, and optionally adversarial learning.

Subsequent work adapted these principles to unsupervised phrase-based Statistical Machine Translation (SMT), obtaining large improvements over the original NMT-based systems (Lample et al., 2018b; Artetxe et al., 2018b). This alternative approach uses cross-lingual n-gram embeddings to build an initial phrase table, which is combined with an n-gram language model and a distortion model, and further refined through iterative back-translation. There have been several follow-up attempts to combine NMT and SMT based approaches (Marie and Fujita, 2018; Ren et al., 2019; Artetxe et al., 2019b). More recently, Conneau and Lample (2019), Song et al. (2019) and Liu et al. (2020) obtain strong results using deep multilingual pretraining rather than cross-lingual word embeddings to initialize unsupervised NMT systems.

3 Motivating fully unsupervised learning

In this section, we challenge the narrative of motivating UCL based on a lack of parallel resources. We argue that the strict unsupervised scenario cannot be motivated from an immediate practical perspective, and elucidate what we believe should be the true goals of this research direction.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

[2] The full version of their model (XLM) requires parallel corpora for their translation language modeling objective, but the authors also explore an unsupervised variant using masked language modeling alone.

3.1 How practical is the strict unsupervised scenario?

Monolingual resources subsume parallel resources. For instance, each side of a parallel corpus effectively serves as a monolingual corpus. From this argument, it follows that monolingual data is cheaper to obtain than parallel data, so unsupervised cross-lingual learning should in principle be more generally applicable than supervised learning.

However, we argue that the common claim that the requirement for **parallel data** *"may not be met for many language pairs in the real world"* (Xu et al., 2018) is largely inaccurate. For instance, the JW300 parallel corpus covers 343 languages with around 100,000 parallel sentences per language pair on average (Agić and Vulić, 2019), and the multilingual Bible corpus collected by Mayer and Cysouw (2014) covers 837 language varieties (each with a unique ISO 639-3 code). Moreover, the PanLex project aims to collect multilingual lexica for all human languages in the world, and already covers 6,854 language varieties with at least 20 lexemes, 2,364 with at least 200 lexemes, and 369 with at least 2,000 lexemes (Kamholz et al., 2014). While 20 or 200 lexemes might seem insufficient, weakly supervised cross-lingual word embedding methods already proved effective with as little as 25 word pairs (Artetxe et al., 2017). More recent methods have focused on completely removing this weak supervision (Conneau et al., 2018a; Artetxe et al., 2018a), which can hardly be justified from a practical perspective given the existence of such resources and additional training signals stemming from a (partially) shared script (§4.2). Finally, given the availability of sufficient monolingual data, noisy parallel data can often be obtained by mining bitext (Schwenk et al., 2019a,b).

In addition, large **monolingual data** is difficult to obtain for low-resource languages. For instance, recent work on cross-lingual word embeddings has mostly used Wikipedia as its source for monolingual corpora (Gouws et al., 2015; Vulić and Korhonen, 2016; Conneau et al., 2018a). However, as of November 2019, Wikipedia exists in *only* 307 languages[3] of which nearly half have less than 10,000 articles. While one could hope to overcome this by taking the entire web as a corpus, as facilitated by Common Crawl[4] and similar initiatives, this is not always feasible for low-resource languages. First, the presence of less resourced languages on the web is very limited, with only a few hundred languages recognized as being used in websites.[5] This situation is further complicated by the limited coverage of existing tools such as language detectors (Buck et al., 2014; Grave et al., 2018), which only cover a few hundred languages. Alternatively, speech could also serve as a source of monolingual data (e.g., by recording public radio stations). However, this is an unexplored direction within UCL, and collecting, processing and effectively capitalizing on speech data is far from trivial, particularly for low-resource languages.

All in all, we conclude that the alleged scenario involving no parallel data *and* sufficient monolingual data is not met in the real world in the terms explored by recent UCL research. Needless to say, effectively exploiting unlabeled data is important in any low-resource setting. However, refusing to use an informative training signal—which parallel data is—when it does indeed exist, cannot be justified from a practical perspective if one's goal is to build the strongest possible model. For this reason, we believe that semi-supervised learning is a more suitable paradigm for truly low-resource languages, and UCL should not be motivated from an immediate practical perspective.

3.2 A scientific motivation

Despite not being an entirely realistic setup, we believe that UCL is an important research direction for the reasons we discuss below.

Inherent scientific interest. The extent to which two languages can be aligned based on independent samples—without any cross-lingual signal—is an open and scientifically relevant problem *per se*. In fact, it is not entirely obvious that UCL should be possible at all, as humans would certainly struggle to align two unknown languages without any grounding. Exploring the limits of UCL could help to understand the limits of the principles that the corresponding methods are based on, such as the distributional hypothesis. Moreover, this research line could bring new insights into the properties and inner workings of both language acquisition and the underlying computational models that ultimately make UCL possible. Finally, such methods may be useful in areas where supervision is impos-

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias

[4] https://commoncrawl.org/

[5] https://w3techs.com/technologies/overview/content_language

sible to obtain, such as when dealing with unknown or even non-human languages.

Useful as a lab setting. The strict unsupervised scenario, although not practical, allows us to isolate and better study the use of monolingual corpora for cross-lingual learning. We believe lessons learned in this setting can be useful in the more practical semi-supervised scenario. In a similar vein, monolingual language models, although hardly useful on their own, have contributed to large improvements in other tasks. From a research methodology perspective, unsupervised systems also set a competitive baseline, which any semi-supervised method should improve upon.

Simplicity as a value. As we discussed previously, refusing to use an informative training signal when it does exist can hardly be beneficial, so we should not expect UCL to perform better than semi-supervised learning. However, simplicity is a value in its own right. Unsupervised approaches could be preferable to their semi-supervised counterparts if the performance gap between them is small enough. For instance, unsupervised cross-lingual embedding methods have been reported to be competitive with their semi-supervised counterparts in certain settings (Glavaš et al., 2019), while being easier to use in the sense that they do not require a bilingual dictionary.

4 What does *unsupervised* mean?

In its most general sense, unsupervised cross-lingual learning can be seen as referring to any method relying *exclusively* on monolingual text data in two or more languages. However, there are different training signals—stemming from common assumptions and varying amounts of linguistic knowledge—that one can potentially exploit under such a regime. This has led to an inconsistent use of this term in the literature. In this section, we categorize different training signals available both from a monolingual and a cross-lingual perspective and discuss additional scenarios enabled by multiple languages.

4.1 Monolingual training signals

From a computational perspective, text is modeled as a sequence of discrete symbols. In UCL, the training data consists of a set of such sequences in each of the languages. In principle, without any knowledge about the languages, one would have no prior information of the nature of such sequences or the possible relations between them. In practice, however, sets of sequences are assumed to be independent, and existing work differs whether they assume document-level sequences (Conneau and Lample, 2019) or sentence-level sequences (Artetxe et al., 2018c; Lample et al., 2018a).

Nature of atomic symbols. A more important consideration is the nature of the atomic symbols in such sequences. To the best of our knowledge, previous work assumes some form of word segmentation or tokenization (e.g., splitting by whitespaces or punctuation marks). Early work on cross-lingual word embeddings considered such tokens as atomic units. However, more recent work (Hoshen and Wolf, 2018; Glavaš et al., 2019) has primarily used fastText embeddings (Bojanowski et al., 2017) which incorporate subword information into the embedding learning, although the vocabulary is still defined at the token level. In addition, there have also been approaches that incorporate character-level information into the alignment learning itself (Heyman et al., 2017; Riley and Gildea, 2018). In contrast, most work on contextual word embeddings and unsupervised machine translation operates with a subword vocabulary (Devlin et al., 2019; Conneau and Lample, 2019).

While the above distinction might seem irrelevant from a practical perspective, we think that it is important from a more fundamental point of view (e.g. in relation to the distributional hypothesis as discussed in §3.2). Moreover, some of the underlying assumptions might not generalize to different writing systems (e.g. logographic instead of alphabetic). For instance, subword tokenization has been shown to perform poorly on reduplicated words (Vania and Lopez, 2017). In relation to that, one could also consider the text in each language as a stream of discrete character-like symbols without any notion of tokenization. Such a *tabula rasa* approach is potentially applicable to any arbitrary language, even when its writing system is not known, but has so far only been explored for a limited number of languages in a monolingual setting (Hahn and Baroni, 2019).

Linguistic information. Finally, one can exploit additional linguistic knowledge through linguistic analysis such as lemmatization, part-of-speech tagging, or syntactic parsing. For instance, before the advent of unsupervised NMT, statistical deci-

pherment was already shown to benefit from incorporating syntactic dependency relations (Dou and Knight, 2013). For other tasks such as unsupervised POS tagging (Snyder et al., 2008), monolingual tag dictionaries have been used. While such approaches could still be considered unsupervised from a cross-lingual perspective, we argue that the interest of this research direction is greatly limited by two factors: (i) from a theoretical perspective, it assumes some fundamental knowledge that is not directly inferred from the raw monolingual corpora; and (ii) from a more practical perspective, it is not reasonable to assume that such resources are available in the less resourced settings where this research direction has more potential for impact.

4.2 Cross-lingual training signals

Pure UCL should not use any cross-lingual signal by definition. When we view text as a sequence of discrete atomic symbols (either characters or tokens), a strict interpretation of this principle would consider the set of atomic symbols in different languages to be disjoint, without prior knowledge of the relationship between them.

Needless to say, any form of learning requires making assumptions, as one needs some criterion to prefer one mapping over another. In the case of UCL, such assumptions stem from the structural similarity across languages (e.g. semantically equivalent words in different languages are assumed to occur in similar contexts). In practice, these assumptions weaken as the distribution of the datasets diverges, and some UCL models have been reported to break under a domain shift (Søgaard et al., 2018; Guzmán et al., 2019; Marchisio et al., 2020). Similarly, approaches that leverage linguistic features such as syntactic dependencies may assume that these are similar across languages.

In addition, one can also assume that the sets of symbols that are used to represent different languages have some commonalities. This departs from the strict definition of UCL above, establishing some prior connections between the sets of symbols in different languages. Such an assumption is reasonable from a practical perspective, as there are a few scripts (e.g. Latin, Arabic or Cyrillic) that cover a large fraction of languages. Moreover, even when two languages use different writing systems or scripts, there are often certain elements that are still shared (e.g. Arabic numerals, named entities written in a foreign script, URLs, certain punctuation marks, etc.). In relation to that, several models have relied on identically spelled words (Artetxe et al., 2017; Smith et al., 2017; Søgaard et al., 2018) or string-level similarity across languages (Riley and Gildea, 2018; Artetxe et al., 2019b) as training signals. Other methods use a joint subword vocabulary for all languages, indirectly exploiting the commonalities in their writing system (Lample et al., 2018b; Conneau and Lample, 2019).

However, past work greatly differs on the nature and relevance that is attributed to such a training signal. The reliance on identically spelled words has been considered as a weak form of supervision in the cross-lingual word embedding literature (Søgaard et al., 2018; Ruder et al., 2018), and significant effort has been put into developing strictly unsupervised methods that do not rely on such signal (Conneau et al., 2018a). In contrast, the unsupervised machine translation literature has not payed much attention to this factor, and has often relied on identical words (Artetxe et al., 2018c), string-level similarity (Artetxe et al., 2019b), or a joint subword vocabulary (Lample et al., 2018b; Conneau and Lample, 2019) under the unsupervised umbrella. The same is true for unsupervised deep multilingual pretraining, where a shared subword vocabulary has been a common component (Pires et al., 2019; Conneau and Lample, 2019), although recent work shows that it is not important to share vocabulary across languages (Artetxe et al., 2020b; Wu et al., 2019).

Our position is that making assumptions on linguistics universals is acceptable and ultimately necessary for UCL. However, we believe that any connection stemming from a (partly) shared writing system belongs to a different category, and should be considered a separate cross-lingual signal. Our rationale is that a given writing system pertains to a specific form to encode a language, but cannot be considered to be part of the language itself.[6]

4.3 Multilinguality

While most work in unsupervised cross-lingual learning considers two languages at a time, there have recently been some attempts to extend these methods to multiple languages (Duong et al., 2017; Chen and Cardie, 2018; Heyman et al., 2019), and most work on unsupervised cross-lingual pretraining is multilingual (Pires et al., 2019; Conneau

[6] As a matter of fact, languages existed well before writing was invented, and a given language can have different writing systems or new ones can be designed.

Monolingual signal	Cross-lingual signal
Sequence of symbols	Shared writing system
Sets of sentences/documents	Identical words
Tokens/subwords	String similarity
Linguistic analysis	

Table 1: Different types of monolingual and cross-lingual signals that have been used for unsupervised cross-lingual learning, ordered roughly from least to most linguistic knowledge (top to bottom).

and Lample, 2019). When considering parallel data across a subset of the language pairs, multilinguality gives rise to additional scenarios. For instance, the scenario where two languages have no parallel data between each other but are well connected through a third (pivot) language has been explored by several authors in the context of machine translation (Cheng et al., 2016; Chen et al., 2017). However, given that the languages in question are still indirectly connected through parallel data, this scenario does not fall within the *unsupervised* category, and is instead commonly known as *zero-resource* machine translation.

An alternative scenario explored in the contemporaneous work of Liu et al. (2020) is where a set of languages are connected through parallel data, and there is a separate language with monolingual data only. We argue that, when it comes to the isolated language, such a scenario should still be considered as UCL, as it does not rely on any parallel data for that particular language nor does it assume any previous knowledge of it. This scenario is easy to justify from a practical perspective given the abundance of parallel data for high-resource languages, and can also be interesting from a more theoretical point of view. This way, rather than considering two unknown languages, this alternative scenario would assume some knowledge of how one particular language is connected to other languages, and attempt to align it to a separate unknown language.

4.4 Discussion

As discussed throughout the section, there are different training signals that we can exploit depending on the available resources of the languages involved and the assumptions made regarding their writing system, which are summarized in Table 1. Many of these signals are not specific to work on UCL but have been observed in the past in allegedly language-independent NLP approaches, as discussed by Bender (2011). Others, such as a reliance on subwords or shared symbols are more recent phenomena.

While we do not aim to open a terminological debate on what UCL encompasses, we advocate for future work being more aware and explicit about the monolingual and cross-lingual signals they employ, what assumptions they make (e.g. regarding the writing system), and the extent to which these generalize to other languages.

In particular, we argue that it is critical to consider the assumptions made by different methods when comparing their results. Otherwise the blind chase for state-of-the-art performance may benefit models making stronger assumptions and exploiting all available training signals, which could ultimately conflict with the eminently scientific motivation of this research area (see §3.2).

5 Methodological issues

In this section, we describe methodological issues that are commonly encountered when training and evaluating unsupervised cross-lingual models and propose measures to ameliorate them.

5.1 Validation and hyperparameter tuning

In conventional supervised or semi-supervised settings, we use a separate validation set for development and hyperparameter tuning. However, this becomes tricky in unsupervised cross-lingual learning, where we ideally should not use any parallel data other than for testing purposes.

Previous work has not paid much attention to this aspect, and different methods are evaluated with different validation schemes. For instance, Artetxe et al. (2018b,c) use a **separate language pair** with a parallel validation set to make all development and hyperparameter decisions. They test their final system on other language pairs without any parallel data. This approach has the advantage of being strictly unsupervised with respect to the test language pairs, but the optimal hyperparameter choice might not necessarily transfer well across languages. In contrast, Conneau et al. (2018a) and Lample et al. (2018a) propose an **unsupervised validation criterion** that is defined over monolingual data and shown to correlate well with test performance. This enables systematic tuning on the language pair of interest, but still requires parallel data to guide the development of the unsupervised validation criterion itself. A **parallel validation set** has also been used for systematic tuning in

the context of unsupervised machine translation (Marie and Fujita, 2018; Marie et al., 2019; Stojanovski et al., 2019). While this is motivated as a way to abstract away the issue of unsupervised tuning—which the authors consider to be an open problem—we argue that any systematic use of parallel data should not be considered UCL. Finally, previous work often does not report the validation scheme used. In particular, unsupervised cross-lingual word embedding methods have almost exclusively been evaluated on bilingual lexicons that do not have a validation set, and presumably use the **test set** to guide development to some extent.

Our position is that a completely blind development model without any parallel data is unrealistic. Some cross-lingual signals to guide development are always needed. However, this factor should be carefully controlled and reported with the necessary rigor as a part of the experimental design. We advocate for using one language pair for development and evaluating on others when possible. If parallel data in the target language pair is used, the test set should be kept blind to avoid overfitting, and a separate validation should be used. In any case, we argue that the use of parallel data in the target language pair should be minimized if not completely avoided, and it should under no circumstances be used for extensive tuning. Instead, we recommend to use unsupervised validation criteria for systematic tuning in the target language.

5.2 Evaluation practices

We argue that there are also several issues with common evaluation practices in UCL.

Evaluation on favorable conditions. Most work on UCL has focused on relatively close languages with large amounts of high-quality parallel corpora from similar domains. Only recently have approaches considered more diverse languages as well as language pairs that do not involve English (Glavaš et al., 2019; Vulić et al., 2019), and some existing methods have been shown to completely break in less favorable conditions (Guzmán et al., 2019; Marchisio et al., 2020). In addition, most approaches have focused on learning from similar domains, often involving Wikipedia and news corpora, which are unlikely to be available for low-resource languages. We believe that future work should pay more attention to the effect of the typology and linguistic distance of the languages involved, as well as the size, noise and domain similarity of the training data used.

Over-reliance on translation tasks. Most work on UCL focuses on translation tasks, either at the word level (where the problem is known as *bilingual lexicon induction*) or at the sentence level (where the problem is known as *unsupervised machine translation*). While translation can be seen as the ultimate application of cross-lingual learning and has a strong practical interest on its own, it only evaluates a particular facet of a model's cross-lingual generalization ability. In relation to that, Glavaš et al. (2019) showed that bilingual lexicon induction performance does not always correlate well with downstream tasks. In particular, they observe that some mapping methods that are specifically designed for bilingual lexicon induction perform poorly on other tasks, showing the risk of relying excessively on translation benchmarks for evaluating cross-lingual models.

Moreover, existing translation benchmarks have been shown to have several issues on their own. In particular, bilingual lexicon induction datasets have been reported to misrepresent morphological variations, overly focus on named entities and frequent words, and have pervasive gaps in the gold-standard targets (Czarnowska et al., 2019; Kementchedjhieva et al., 2019). More generally, most of these datasets are limited to relatively close languages and comparable corpora.

Lack of an established cross-lingual benchmark. At the same time, there is no *de facto* standard benchmark to evaluate cross-lingual models beyond translation. Existing approaches have been evaluated in a wide variety of tasks including dependency parsing (Schuster et al., 2019), named entity recognition (Rahimi et al., 2019), sentiment analysis (Barnes et al., 2018), natural language inference (Conneau et al., 2018b), and document classification (Schwenk and Li, 2018). XNLI (Conneau et al., 2018b) and MLDoc (Schwenk and Li, 2018) are common choices, but they have their own problems: MultiNLI, the dataset from which XNLI was derived, has been shown to contain superficial cues that can be exploited (Gururangan et al., 2018), while MLDoc can be solved by keyword matching (Artetxe et al., 2020b). There are non-English counterparts for more challenging tasks such as question answering (Cui et al., 2019; Hsu et al., 2019), but these only exist for a handful of languages. More recent datasets such as XQuAD

Methodological issues	Examples
Validation and hyperparameter tuning	Systematic tuning with parallel data or on test data
Evaluation on favorable conditions	Typologically similar languages; always including English; training on the same domain
Over-reliance on translation tasks	Overfitting to bilingual lexicon induction; known issues with existing datasets
Lack of an established benchmark	Evaluation on many different tasks; problems with common tasks (MLDoc and XNLI)

Table 2: Methodological issues pertaining to validation and hyperparameter tuning and evaluation practices in current work on unsupervised cross-lingual learning.

(Artetxe et al., 2020b), MLQA (Lewis et al., 2019) and TyDi QA (Clark et al., 2020) cover a wider set of languages, but a comprehensive benchmark that evaluates multilingual representations on a diverse set of tasks—in the style of GLUE (Wang et al., 2018)—*and* languages has been missing until very recently. The contemporaneous XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) benchmarks try to close this gap, but they are still restricted to languages where existing labelled data is available. Finally, an additional issue is that a large part of these benchmarks were created through translation, which was recently shown to introduce artifacts (Artetxe et al., 2020a).

We present a summary of the methodological issues discussed in Table 2.

6 Bridging the gap between unsupervised cross-lingual learning flavors

The three categories of UCL (§2) have so far been treated as separate research topics by the community. In particular, cross-lingual word embeddings have a long history (Ruder et al., 2019), while deep multilingual pretraining has emerged as a separate line of research with its own best practices and evaluation standards. At the same time, unsupervised machine translation has been considered a separate problem in its own right, where cross-lingual word embeddings and deep multilingual pretraining have just served as initialization techniques.

While each of these families have their own defining features, we believe that they share a strong connection that should be considered from a more holistic perspective. In particular, both cross-lingual word embeddings and deep multilingual pretraining share the goal of learning (sub)word representations, and essentially differ on whether such representations are static or context-dependent. Similarly, in addition to being a downstream application of the former, unsupervised machine translation can also be useful to develop other multilingual applications or learn better cross-lingual representations. This has previously been shown for *supervised* machine translation (McCann et al., 2017; Siddhant et al., 2019) and recently for bilingual lexicon induction (Artetxe et al., 2019a). In light of these connections, we call for a more holistic view of UCL, both from an experimental and theoretical perspective.

Evaluation. Most work on cross-lingual word embeddings focuses on bilingual lexicon induction. In contrast, deep multilingual pretraining has not been tested on this task, and is instead typically evaluated on zero-shot cross-lingual transfer. We think it is important to evaluate both families—cross-lingual word embeddings and deep multilingual representations—in the same conditions to better understand their strengths and weaknesses. In that regard, Artetxe et al. (2020b) recently showed that deep pretrained models are much stronger in some downstream tasks, while cross-lingual word embeddings are more efficient and sufficient for simpler tasks. However, this could partly be attributed to a particular integration strategy, and we advocate for using a common evaluation framework in future work to allow a direct comparison between the different families.

Theory. From a more theoretical perspective, it is still not well understood in what ways cross-lingual word embeddings and deep multilingual pretraining differ. While one could expect the latter to be learning higher-level multilingual abstractions, recent work suggests that deep multilingual models might mostly be learning a lexical-level alignment (Artetxe et al., 2020b). For that reason, we believe that further research is needed to understand the relation between both families of models.

7 Recommendations

To summarize, we make the following practical recommendations for future cross-lingual research:

- Be rigorous when motivating UCL. Do not present it as a practical scenario unless supported by a real use case.

- Be explicit about the monolingual and cross-lingual signals used by your approach and the assumptions it makes, and take them into considerations when comparing different models.
- Report the validation scheme used. Minimize the use of parallel data by preferring an unsupervised validation criterion and/or using only one language for development. Always keep the test set blind.
- Pay attention to the conditions in which you evaluate your model. Consider the impact of typology, linguistic distance, and the domain similarity, size and noise of the training data. Be aware of known issues with common benchmarks, and favor evaluation on a diverse set of tasks.
- Keep a holistic view of UCL, including cross-lingual word embeddings, deep multilingual pretraining and unsupervised machine translation. To the extent possible, favor a common evaluation framework for these different families.

8 Conclusions

In this position paper, we review the status quo of unsupervised cross-lingual learning—a relatively recent field. UCL is typically motivated by the lack of cross-lingual signal for many of the world's languages, but available resources indicate that a scenario with no parallel data and sufficient monolingual data is not realistic. Instead, we advocate for the importance of UCL for scientific reasons.

We also discuss different monolingual and cross-lingual training signals that have been used in the past, and advocate for carefully reporting them to enable a meaningful comparison across different approaches. In addition, we describe methodological issues related to the unsupervised setting and propose measures to ameliorate them. Finally, we discuss connections between cross-lingual word embeddings, deep multilingual pre-training, and unsupervised machine translation, calling for an evaluation on an equal footing.

We hope that this position paper will serve to strengthen research in UCL, providing a more rigorous look at the motivation, definition, and methodology. In light of the unprecedented growth of our field in recent times, we believe that it is essential to establish a rigorous foundation connecting past and present research, and an evaluation protocol that carefully controls for the use of parallel data and assesses models in diverse, challenging settings.

Acknowledgments

This research was partially funded by a Facebook Fellowship, the Basque Government excellence research group (IT1343-19), the Spanish MINECO (UnsupMT TIN2017-91692-EXP MCIU/AEI/FEDER, UE) and Project BigKnowledge (Ayudas Fundación BBVA a equipos de investigación científica 2018).

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493, Melbourne, Australia. Association for Computational Linguistics.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1586–1595, Hong Kong, China. Association for Computational Linguistics.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 973–982, Hong Kong, China. Association for Computational Linguistics.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28*, pages 3079–3087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904, Valencia, Spain. Association for Computational Linguistics.

Umberto Eco and James Fentress. 1995. *The search for the perfect language*. Blackwell Oxford.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of Machine Learning Research*, volume 89, pages 1880–1890. PMLR.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6097–6110, Hong Kong, China. Association for Computational Linguistics.

Michael Hahn and Marco Baroni. 2019. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1085–1095, Valencia, Spain. Association for Computational Linguistics.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv preprint arXiv:2003.11080*.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3327–3332, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating Cross-lingual Extractive Question Answering. *arXiv preprint arXiv:1910.07475*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems 30*, pages 6294–6305.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with SMT as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248.

Parker Riley and Daniel Gildea. 2018. Orthographic features for bilingual lexicon induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–394, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M Parallel Sentences. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. *arXiv preprint arXiv:1911.04944*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2019. Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation. *arXiv preprint arXiv:1909.00437*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050, Honolulu, Hawaii. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5926–5936, Long Beach, California, USA. PMLR.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2019. The LMU munich unsupervised machine translation system for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 393–399, Florence, Italy. Association for Computational Linguistics.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.