Sistema Estadístico de Reordenamiento de Palabras en Traducción Automática

Marta R. Costa-jussà y José A. R. Fonollosa
Centro de investigación TALP
Universidad Politécnica de Cataluña
Campus Nord UPC 08034-Barcelona
{mruiz,adrian}@gps.tsc.upc.edu

Resumen: Actualmente los errores debidos al cambio de orden de las palabras son una de las principales fuentes de error en los sistemas de traducción automática estocástica (TAE) basados en frases. Esta comunicación propone una nueva estrategia estadística para afrontar los reordenamientos que denominaremos RAE (*Reordenamiento automático estocástico*). El método propuesto aprovecha la poderosas técnicas de aprendizaje estadístico desarrolladas en traducción estadística para traducir el lenguaje fuente (*S*) en un lenguaje fuente reordenado (*S'*), que nos permita mejorar la traducción final al lenguaje destino (*T*). Por lo tanto, el lenguaje fuente de la tarea de traducción en sí pasa a ser *S'*, y esto nos permite generar un alineado más monótono entre las palabras de ambos lenguajes y unas unidades de traducción menores. Además, el uso de clases de palabras en la estrategia RAE ayuda a generalizar reordenamientos. En este artículo se presentan resultados en la tarea de ZhEn de la evaluación IWSLT05 que muestran una mejora significativa en la calidad de la traducción.
Palabras clave: traducción estadística, reordenamiento, tuplas.

Abstract: Nowadays, reordering is one of the most important problems in Statistical Machine Translation (SMT) systems. This paper exposes a novel strategy to face it: Statistical Machine Reordering (SMR). It consists of using the powerful techniques developed for Statistical Machine Translation (SMT) in order to translate the source language (*S*) into a reordered source language (*S'*), which allows for an improved translation into the target language (*T*). Then, the SMT task changes from *S2T* to *S'2T* which leads to a monotonized word alignment and shorter translation units. In addition, the use of classes in SMR helps to generalize word reorderings. Experiments are reported in the ZhEn IWSLT05 task showing significant improvement in translation quality.
Keywords: statistical machine translation, reordering, tuples

1. Introducción

La traducción automática estocástica (TAE) considera que una oración f de una lengua fuente (oración a traducir) puede ser traducida en cualquier oración e del lenguaje destino (en el que se desea la traducción) con probabilidad no nula. La traducción consiste precisamente en determinar la oración con mayor probabilidad de constituir una traducción para la oración original. Estas probabilidades se aprenden principalmente a partir textos paralelos bilingües.

Los primeros sistemas de TAE seguían la aproximación del canal ruidoso y trabajaban a nivel de palabras (Brown et al., 1990) (las unidades bilingües se componían de palabras aisladas).

Recientemente, los sistemas de TAE tienden a utilizar secuencias de palabras, denominadas frases (Koehn, Och, y Marcu, 2003), como unidades básicas del modelo de traducción, con el objetivo de introducir el contexto en dicho modelo. En paralelo, al modelo de frases, también se ha propuesto el uso de n-gramas de tuplas bilingües (Crego et al., 2005) (Mariño et al., 2005) como una alternativa para tener en cuenta el contexto con unidades bilingües más pequeñas. Ambos sistemas llevan a cabo la traducción mediante una búsqueda que maximiza una combinación loglineal (como alternativa al modelo de canal ruidoso) de las probabilidades asignadas a la traducción por el modelo de traducción en sí y otras características (Berger, Della Pietra, y Della Pietra, 1996) adicionales.

Tanto en los sistemas de TAE basados en frases como el los basados en n-gramas, la introducción de reordenamientos es crucial si los pares de lenguas que intervienen en la traducción tienen una estructura diferente (e.g. chino/inglés). Recientemente han apareciendo diversas estrategias de reordenamiento de palabras que intentan modificar el orden de la oración fuente para que se corresponda con el orden de la oración destino, ver (Kanthak et al., 2005). En (Mauser, Matusov, y Ney, 2006), por ejemplo, se describe un método que simultáneamente alinea y monotoniza el corpus de entrenamiento. Sin embargo, los proble-

mas más importantes de estas estrategias son: (1) la monotonización propuesta se basa en el alineado entre el corpus de entrenamiento del lenguaje fuente y el del lenguaje destino, con lo cual no se puede aplicar al traducir el corpus de test; y (2) no hay una generalización del reordenamiento que permita extenderlo a secuencias de palabras no observadas en las bases de datos de entrenamiento. Otro enfoque, como el descrito en (Crego, Mariño, y de Gispert, 2005), consiste en permitir que las unidades de traducción no sigan el orden del lenguaje fuente cuando se traduce. De esta manera el decodificador permite reordenamientos según los criterios de diversos modelos estocásticos, pero con el inconveniente de incrementar sensiblemente el coste computacional.

La aproximación propuesta en esta comunicación (RAE) para el reordenamiento de las palabras se basa en los mismos principios que la traducción automática estadística (TAE) y comparte el mismo tipo de decodificador. El reordenamiento se trata como una traducción estocástica del lenguaje fuente (*S*) al lenguaje fuente reordenado (*S'*) y se entrena a partir de la información de alineado. Sin embargo, una vez estimadas las probabilidades de reordenado, sólo necesitamos como entrada la oración fuente para reordenar y esto nos permitirá aplicar a las oraciones de test el mismo proceso de reordenamiento que a las oraciones de entrenamiento. Además, para mejorar la capacidad de generalización del sistema propuesto, se usarán clases de palabras en vez de palabras como entrada al sistema de reordenamineto RAE.

La comunicación se organiza de la siguiente manera. En la Sección 2 se describe brevemente el sistema de referencia. En la siguiente sección se describe con detalle la estrategia de reordenamiento propuesta. En la Sección 4 se presentan y se discuten los resultados, y finalmente en la Sección 5 se presentan las conclusiones y el trabajo futuro.

2. *Sistema de referencia basado en n-gramas*

El modelo de traducción puede entenderse como un modelo de lenguaje de unidades bilingües (llamadas tuplas). Dichas tuplas, definen una segmentación monótona de los pares de oraciones utilizadas en el entrenamiento del sistema (f_1^J, e_1^I), en K unidades $(t_1, ..., t_K)$.

En la extracción de las unidades bilingües, cada par de oraciones da lugar a una secuencia de tuplas que sólo depende de los alineamientos internos entre las palabras de la oración.

La figura 1 muestra un ejemplo de extracción de tuplas.

En la traducción de una oración de entrada, el decodificador debe encontrar la secuencia de tuplas asociada a una segmentación de la oración de entrada que produzca probabilidad máxima. Tal probabilidad máxima, se calcula como combinación lineal de los modelos utilizados en el sistema de traducción.

El modelo de traducción se ha implementado utilizando un modelo de lenguaje (bilingüe) basado en N-gramas (B), (con $N = 4$):

$$p(d, f) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k \,|\, t_{k-2}, t_{k-1}) \quad (1)$$

Usando el enfoque de la combinación loglineal, el decodificador utiliza cinco funciones características (definidas como probabilidades):

- Un **modelo de lenguaje** basado en N-gramas del idioma destino (LM). Particularmente, utilizamos un 4-grama.
- Una **bonificación** basada en el número de palabras de la traducción, usada para compensar la preferencia del decodificador por las traducciones cortas (WB).
- Un **modelo de traducción** calculado utilizando las probabilidades **léxicas** del modelo IBM1, para ambas direcciones (IBM1).
- Un **modelo de distorsión** basado en la distancia entre palabras (R).

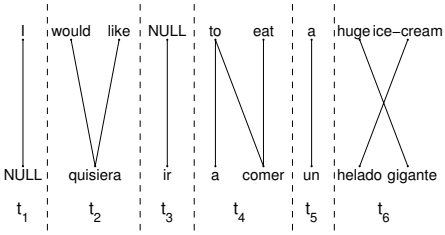

Figura 1: *Extracción de tuplas a partir de un par de oraciones alineadas palabra a palabra.*

El sistema de búsqueda o decodificación utilizado, denominado MARIE, está descrito en (Crego, Mariño, y de Gispert, 2005).

Para calcular los pesos de la combinación loglineal se usa una herramienta de optimización, que se basa en el método (Nelder y Mead, 1965).

3. *Reordenamiento automático estocástico*

3.1. Concepto

El sistema de Reordenamiento Automático Estadístico (RAE) se basa en utilizar un sistema de Traducción Automática Estadística (TAE) para solventar los problemas de distorsión o reordenamiento. Por lo tanto, un sistema de RAE se puede ver como un sistema de TAE que traduce de un lenguaje fuente (*S*) a un lenguaje fuente modificado (*S'*), dado un lenguaje destino (*T*). Con lo cual la estrategia de reordenamiento se enfoca como una tarea de traducción *S2S'*. Y en consecuencia, la tarea de traducción en sí cambia de *S2T* a *S'2T*. La principal diferencia entre ambas tareas radica en que la segunda permite: (1) un alineamiento de palabras más monótono; (2) unas unidades de traducción más pequeñas; y (3) una traducción monótona correcta.

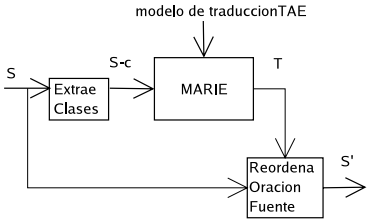

Figura 2: *Diagrama de bloques del sistema RAE*

(a) tupla bilingue S2T

only possible compromise # compromiso solo podria # 0-1 1-1 1-2 2-0
(destino) (fuente) (alineamiento de palabras)
(palabra fuente-palabra destino)

(b) alineado multiple-a-multiple → alineado multiple-a-una

P_ibm (only, solo)>P_ibm(possible, solo)

only possible compromise # compromiso solo podria # 0-1 1-2 2-0

(c) tupla bilingue S2S'

compromiso solo podria # 1 2 0
(fuente) (nuevo orden)

(e) sustitucion de clases

C43 C49 C42 # 1 2 0

Figura 4: *Ejemplo de extracción de tuplas reordenadas*

3.2. Descripción

La figura 2 muestra un diagrama de bloques de descripción del sistema RAE. La entrada es una oración fuente (*S*) y la salida el una oración fuente reordenada (*S'*). El sistema RAE se basa en tres bloques: (1) el extractor de clases; (2) el decodificador MARIE que requiere un RAE-LM, i.e. un modelo de traducción; y, (3) el bloque que reordena la oración original usando los índices a la salida del decodificador.

El siguiente ejemplo especifica la entrada y la salida de cada bloque perteneciente al TAE.

1. Oración fuente (*S*):
   El compromiso sólo podría mejorar
2. Clases de la oración fuente (*S-c*):
   C38 C43 C49 C42 C22
3. Traducción reordenada (*R*):
   0 | 1 2 0 | 0
   donde | indica la segmentación en unidades de traducción.
4. Oración fuente reordenada (*S'*):
   El sólo podría compromiso mejorar

3.3. Entrenamiento

Para traducir de *S* a *S'* utilizamos un sistema de TAE basado en n-gramas de tuplas, considerando únicamente el modelo básico de traducción. El entrenamiento de este sistema, tal como se muestra en el diagrama de bloques de la figura 3, consta de los siguiente pasos:

1. Obtener clases de palabras del lenguaje fuente y del destino. Para ello se ha utilizado el programa 'mkcls', una herramienta libre adjunta al GIZA++ (Och, 2003).
2. Utilizar la herramienta de alineado GIZA++ (con las iteraciones $H^5 1^4 2^0 3^0 4^4$) para alinear a nivel de palabra en ambas direcciones de traducción (utilizando las clases del paso anterior para mejorar la convergencia). Simetrizar el alineado mediante la unión de ambos alineamientos. El resultado es un alineado de múltiples-a-múltiples palabras.
3. Extraer tuplas de reordenamiento, ver figura 4.

   a) Partiendo del alineamiento unión, extraer tuplas bilingües *S2T* (i.e. fragmentos fuente y destino) manteniendo la información de alineado. Un ejemplo de tupla bilingüe *S2T* es: *only possible compromise # compromiso sólo podría # 0-1 1-1 1-2 2-0*, ver figura 4, donde los diferentes campos están separados por # y se corresponden a: (1) fragmento destino; (2) fragmento fuente; y (3) alineado de palabras (aquí, los campos están separados por − y se corresponden a la palabra destino y fuente, respectivamente).

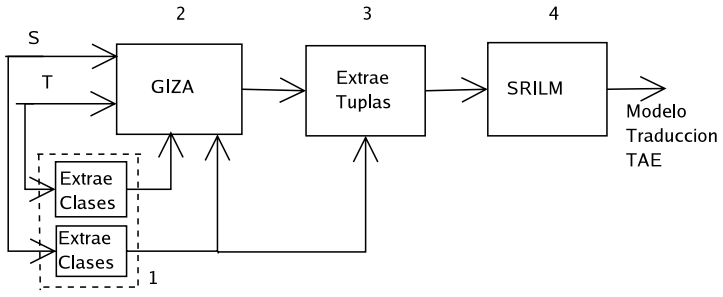

Figura 3: *Diagrama de bloques del entrenamiento del sistema RAE*

b) Pasar de un alineado de múltiples-a-múltiples palabras dentro de cada tupla a un alineado de múltiples-a-una palabra. Si una palabra fuente está alineada con dos o más palabras destino, se escoge el vínculo más probable según el modelo IBM 1, y los otros vínculos se omiten (i.e. el número de palabras fuente se mantiene antes y después de la traducción de reordenamiento). En el ejemplo anterior, la tupla cambiará a: *only possible compromise # compromiso sólo podría # 0-1 1-2 2-0*, porque P_{ibm1}(only, sólo) es mayor que P_{ibm1}(possible, sólo).

c) A partir de las *S2T* tuplas bilingües (con el alineado de palabras múltiples-a-una), extraer *S2S'* tuplas bilingües (i.e. fragmento fuente y su reordenamiento). Siguiendo el ejemplo: *compromiso sólo podría # 1 2 0*, donde el primer campo es el fragmento fuente, y el segundo el reordenamiento de estas palabras fuente.

d) Eliminar aquellas tuplas cuyo fragmento fuente es la palabra NULL.

e) Sustituir las palabras de cada fragmento fuente por las clases calculadas en el paso 1.

4. Utilizar la herramienta SRILM (Stolcke, 2002) para calcular el modelo de lenguaje de la secuencia de tuplas bilingües *S2S'* compuestas por el fragmento fuente (en clases) y su reordenamiento.

El proceso de RAE se puede iterar para obtener una mayor monotonización de la oración. Una vez entrenado, todo el corpus fuente original *S* se traduce para obtener el corpus fuente reordenado *S'* con el sistema RAE, ver la figura 2. A continuación, el corpus de entrenamiento (fuente reordenado y destino original) se usa para entrenar el sistema de traducción (tal y como se explica en 2). Y finalmente, con este sistema de traducción, se traduce el corpus de test previamente reordenado.

BTEC	*Chino*	*Inglés*
Oraciones de entrenamiento	20 k	20 k
Palabras	176.2 k	182.3 k
Vocabulario	8.7 k	7.3 k
Oraciones de desarrollo	506	506
Palabras	3.5 k	3 k
Vocabulario	870	799
Oraciones de test	500	500
Palabras	3.8 k	3 k
Vocabulario	893	840

Tabla 1: *Corpus BTEC. Para el conjunto de test en inglés se utilizaron 16 referencias.*

4. Experimentos

4.1. Corpus

Los experimentos se efectuaron utilizando el corpus IWSLT 2005 BTEC[1] (chino-inglés).

La tabla 1 muestra las estadísticas básicas de dicho corpus, es decir, número de oraciones, palabras y vocabulario.

4.2. Unidades

El reordenamiento del corpus fuente permite mejorar la extracción de unidades de traducción. En esta sección presentamos diferentes estadísticas de unidades de ambas sistemas: (*S2T* y *S'2T*).

La tabla 3 muestra el vocabulario de n-gramas bilingües en el modelo de traducción. Después de hacer la traducción *S2S'* de reordenamiento, el alineado *S'2T* resulta más monótono que el alineado original *S2T* y esto, en general, facilita la tarea al sistema de alineado. Por lo tanto, es de esperar que el alineado mejore tanto

[1] www.slt.atr.jp/IWSLT2005

Sistema	Total	Vocabulario
NB	4151	915
RAE + NB	4374	1041

Tabla 2: *Número de tuplas en el corpus de test y vocabulario. El primer caso es el sistema de referencia. El segundo caso se aplica el sistema RAE previo al sistema de referencia. NB significa sistema TAE n-grama.*

en monotonización como en calidad y, en consecuencia, mejoremos la calidad de la traducción. Aquí, podemos observar un notable crecimiento del número de unidades de traducción, que conduce a un crecimiento de vocabulario de traducción.

La figura 5 muestra el histograma del tamaño de las tuplas en ambos sistemas. El cambio de orden en las palabras de las oraciones fuente (de acuerdo con el orden del lenguaje destino) permite mejorar el modelado mediante n-gramas de las unidades de traducción. En la figura 5 se observa que por encima de la longitud 5 el número de tuplas es similar, pero en cambio, el número de unidades de longitud menor es considerablemente mayor en el caso de utilizar el sistema de RAE previamente a la traducción automática estocástica.

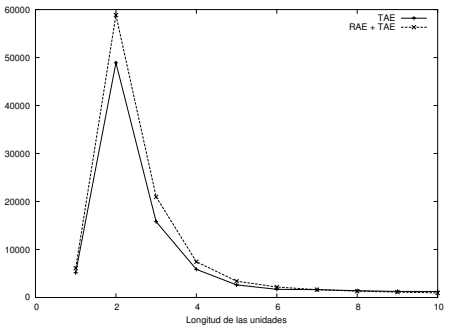

Figura 5: *Comparación de histogramas de las longitudes de tuplas: del sistema RAE+TAE y del sistema TAE.*

La tabla 2 muestra las tuplas utilizadas para traducir el corpus de test (el número total y el vocabulario). De acuerdo con la tabla 3, el número de tuplas y el vocabulario usado para traducir el corpus de test es significativamente mayor después de la traducción de reordenamiento.

Sistema	BLEU	NIST
NB	30.37	6.5284
RAE + NB	32.47	7.2058

Tabla 4: *Resultados en el corpus de test usando la configuración de referencia (modelo de traducción, modelo de lenguaje, bonificación de palabras y modelos léxicos de traducción)*

4.3. Resultados

En este apartado se exponen los experimentos realizados y los resultados obtenidos al evaluar la influencia de la estrategia de reordenamiento propuesta (RAE) en la calidad de la traducción. En todos los experimentos, hemos utilizado el algoritmo Simplex (Nelder y Mead, 1965) para optimizar los pesos de la combinación loglineal de características del TAE, con la medida BLEU (Papineni et al., 2002) como función objetivo. La evaluación se ha llevado a cabo convirtiendo a minúsculas tanto las referencias y como la traducción y sin símbolos de puntuación.

Hemos estudiado la influencia del reordenamiento RAE propuesto en el sistema de traducción basado en n-gramas. Para ello se ha considerado por un lado, un decodificador de traducción monótono (NB o configuración básica de referencia), y por otro lado, un decodificador de traducción que permite ciertos reordenamientos en las palabras de la oración fuente (Crego, Mariño, y de Gispert, 2005) (NBR o configuración completa de referencia). Cuando se permite reordenamiento en el decodificador del sistema TAE el límite de la longitud de la distorsión (m) y el límite en el número de reordenamientos (j) se han establecido empíricamente en 5 y 3, respectivamente, como un compromiso entre la eficiencia y la calidad. Ambos sistemas incluyen además las cuatro funciones características descritas en las Sección 2: el modelo de lenguaje, la bonificación de palabras, y los modelos de lexición IBM1 en ambos sentidos.

Las tablas 4 y 5 muestran los resultados en el corpus de test. La primera muestra la influencia del sistema RAE en la configuración básica de referencia NB y, la segunda, en la configuración completa NBR.

Los algoritmos utilizados para calcular las medidas de evaluación (BLEU y NIST) provienen del conjunto oficial de herramientas de la evaluación 2006 del proyecto TC-STAR distribuidas por ELDA (http://www.elda.org/).

Sistema	1gr	2gr	3gr	4gr
NB	26876	58898	3696	2050
RAE + NB	29482	67744	4894	2684

Tabla 3: *Vocabulario de n-gramas en el modelo de traducción. El primer caso es el sistema de referencia. El segundo caso se aplica el sistema RAE previo al sistema de referencia*

Sistema	BLEU	NIST
NBR	32.82	7.1602
RAE + NBR	34.36	7.855

Tabla 5: *Resultados en el corpus de test usando la configuración completa (configuración básica de referencia más reordenamiento en el decodificador ponderado con la correspondiente función característica de distorsión)*

4.4. Discusión

Ambas medidas BLEU y NIST mejoran con la incorporación de la técnica RAE. La mejora en calidad de traducción es consecuencia de la mejora en la calidad del sistema de TAE. En tareas de traducción más monótonas, las unidades de traducción tienden a ser más cortas y los sistemas de TAE obtienen mejores resultados.

Ambas configuraciones de referencia, la básica (NB) y la completa (NBR), presentan ganancias similares cuando se usa el corpus fuente reordenado que proporciona el bloque RAE. La ganancia obtenida en el caso RAE+NBR indica que ambos reordenamientos, el proporcionado por el sistema RAE y el introducido en el decodificador, son complementarios. También nos indica que la salida RAE todavía podría ser más monótona. Una razón puede ser la complejidad de la tarea ZhEn a nivel de reordenamientos de palabras.

Estos resultados preliminares también muestran que el sistema RAE por si solo proporciona mejoras equivalentes a las que proporciona el reordenamiento en la búsqueda del decodificador, pero con la ventaja de ser computacionalmente mucho menos costoso.

5. Conclusiones y trabajo futuro

En esta comunicación hemos propuesto una solución para el problema del reordenamiento de las palabras en un sistema de traducción automática estocástica (TAE). El sistema propuesto ha sido descrito y probado en un sistema de TAE basado en n-gramas de tuplas, pero se puede aplicar de manera similar a un sistema de TAE basado en frases. En la aproximación propuesta, el problema del reordenamiento ha sido modelado como una traducción de un lenguaje fuente a un lenguaje fuente monotonizado, dado un lenguaje destino. El sistema de reordenamiento automático estocástico (RAE) se aplica previamente al sistema de traducción estocástica. Ambos sistemas, RAE y TAE se basan en los mismos principios y comparten el mismo tipo de decodificador.

Cuando se extraen las unidades bilingües del TAE, el cambio de orden que se realiza en la oración fuente permite mejorar el modelado de las unidades de traducción, dado las unidades de traducción son ahora más cortas. Además, la estrategia de reordenamiento propuesta se puede aplicar por igual tanto al corpus de entrenamiento como al de test, mejorando así la coherencia de los parámetros estimados.

Por otro lado, el hecho de realizar el reordenamiento como un preproceso y de manera independiente al sistema propiamente dicho de traducción permite obtener un sistema final eficiente y una traducción más rápida. Además, la estrategia propuesta permite utilizar clases de palabras en el reordenamiento para inferir reordenamientos no vistos durante el entrenamiento del sistema.

Los resultados preliminares muestran mejoras consistentes e interesantes en la calidad de la traducción. Como trabajo futuro, tenemos previsto la construcción de un sistema más completo de RAE (añadiendo funciones características adicionales), y el estudio de diferentes alternativas para la generación de clases de palabras.

6. Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el gobierno español (beca FPU), y la Unión Europea, FP6-506738 (proyecto TC-STAR).

Bibliografía

Berger, A., S. Della Pietra, y V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, y P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Crego, J. M., M. R. Costa-jussà, J. Mariño, y J. A. Fonollosa. 2005. Ngram-based versus phrase-based statistical machine translation. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, October.

Crego, J.M., J. Mariño, y A. de Gispert. 2005. An Ngram-based statistical machine translation decoder. *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'05*.

Kanthak, S., D. Vilar, E. Matusov, R. Zens, y H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, páginas 167–174, June.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, May.

Mariño, J.B., R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, y M. Ruiz. 2005. Bilingual n-gram statistical machine translation. En *Proc. of the MT Summit X*, páginas 275–82, Pukhet (Thailand), May.

Mauser, A., E. Matusov, y H.Ñey. 2006. Training a statistical machine translation system without giza++. *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, May.

Nelder, J.A. y R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Och, F.J. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/˜och/ software/giza++.html.

Papineni, K.A., S. Roukos, R.T. Ward, y W-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, páginas 311–318, July.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.