Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 379-381 recibido 1-05-2009; aceptado 5-06-2009

Carvalho: English-Galician SMT system from EuroParl English-Portuguese parallel corpus

Carvalho: Un sistema de traducción estadística inglés-galego construído a partir del corpus paralelo inglés-portugués EuroParl

José Ramom Pichel Campos, Paulo Malvar Fernández, Oscar Senra Gómez
Language technology department
imaxin|software
Rua Salgueirinhos de abaixo N11 L-6
Santiago de Compostela, Galiza, Spain
jramompichel@imaxin.com

Pablo Gamallo Otero, Alberto García
Department of Spanish Language, Faculdade de Filologia, Compostela, Galiza, Spain
pablogam@usc.es
Engineering department, Igalia Free Software Company, Corunha, Galiza, Spain
agarcia@igalia.com

Resumen: Para poder construír sistemas de traducción estadística es preciso contar con corpora paralelos suficientemente relevantes. No existe en estos momentos suficientes corpus paralelos entre el par de lenguas inglés-gallego. Siguiendo las teorías de importantes romanistas como Eugene Coseriu o Cunha & Cintra que gallego, portugués y brasileño son tres variedades del mismo sistema lingüístico y puesto que la variante portuguesa si que tiene estos corpus, en este proyecto investigamos si podemos usar el corpus EUROPARL inglés-portugués para conseguir un ingenio de traducción estadística entre el inglés-galego. Para conseguir esto, convertimos los corpus inglés-portugués a inglés-gallego usando un traductor RBMT Opentrad portugués-gallego. Las palabras no detectadas por el traductor son enviadas a un conversor ortográfico entre la grafía etimológica e histórica que usa el portugués y la grafía castellanizada del gallego. Posteriormente mediante Moses y Giza++ conseguimos modelos de lenguaje de nuestro prototipo. Los resultados obtenidos nos permiten pensar en la posibilidad de usar recursos lingüístico-computacionais del portugués para construír recursos, herramientasy aplicaciones para el gallego normativo ILG-RAG.

Palabras clave: SMT, gallego, portugués, Coseriu, Moses, Giza++, imaxin, Opentrad, apertium.

Abstract: In order to build reliable Statistical Machine Translation (SMT) engines between two languages it is essential to use a significantly big amount of parallel corpora. Since available English-Galician parallel corpora are not yet sufficient, it is obvious that other strategies must be followed. Important Romanicists, such as Coseriu (1987) or Cunha & Cintra (2002) have theorized that Galician and Portuguese are two varieties of European Portuguese. From a Computational Linguistics practical stand point, this assumption opens a new line of research that potentially supplies Galician with huge amount of computational resources from both European and Brazilian Portuguese. Thus, drawing from the English-Portuguese Europarl parallel corpus, **imaxin|**software has built a English-Galician Phrase-based Statistical Machine Translation prototype. To achieve that, the English-Portuguese parallel corpus was first converted into English-Galician using a Opentrad Portuguese-Galician Rule-based Machine Translation (RBMT) engine and a spelling converter. Secondly, using Moses, Kohen et al. (2007), and GIZA++, Och & Ney (2003) we built the English-Galician translations and language models of our prototype. The results obtained allow us to conclude that SMT tools based on Galician can be drawn from Portuguese resources, which otherwise would have been an unthinkable task due to the lack of English-Galician parallel corpora. We can also conclude that this strategy can be implemented to develop a great variety of computational tools for Galician language.

Keywords: SMT, Galician, Portuguese, Coseriu, Moses, GIZA++, imaxin, Opentrad, apertium.

1 EuroParl

The Europarl parallel corpus has been extracted from the European Parliament Proceedings, which includes versions of its contents in eleven European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

Given huge parallel corpora are absolute necessary to build reliable SMT systems and that there are not available English-Galician corpora of the size required from that purpose, we have assumed that it would be reasonable to build an English-Galician SMT system drawing from the English-Portuguese Europarl corpus. This corpus contains about 58 million English and Portuguese words.

2 SMT en-gl prototype

The construction process of this prototype involved the conversion of the English-Portuguese Europarl Corpus into English-Galician. For this purpose, we used both EixOpentrad (EixOpentrad is a Galician-Portuguese and Portuguese-Galician MT prototype containing 8.500 words for both directions based on Opentrad-Apertium system) and the spelling converter developed by Alberto García and Pablo Gamallo. First, words were translated by our Portuguese-Galician RBMT system.

Then, words not included in the EixOpentrad dictionaries were transliterated into Galician correct spelling. Finally, using Moses anda Giza++ we built an English-Galician SMT prototype. The following example shows a sample automatic translation of the wikipedia Art entry performed by our system:

“A arte é o proceso de obras de arte ou de elementos efectivamente dunha forma que os chamamentos á razón ou de emocións. Ela abrangue unha gama diversificada de actividades humanas, para crear e medios de expresión, inclusive, música, literatura. O sentido de arte é explotada para un ramo da filosofía apelidados de estética.”

3 Google

Google has recently incorporated Galician in its catalogue of linguistic tools. The Google translator was trained using English-Portuguese parallel corpora partially converted into Galician spelling. Unlike **imaxin**software's strategy, Google did not seem to use spelling converters. Thus, Portuguese words which were not in their dictionaries remained in their original spelling. To compare both systems, we show below a sample translation ot that same wikipedia Art entry performed by Google's SMT system:

“A arte é o proceso ou produto de deliberadamente organizar elementos dun modo que apelido aos sentidos ou emoções. Engloba un conxunto diversificado de actividades humanas, criações, e modos de expresión, incluíndo a música e a literatura. O significado da arte é explorador no ramo da filosofía coñecido como estética.”

References

Cunha, C. & Cintra, L. (2002). Nova Gramática do Português Comtemporâneo. Lisboa: Edições João Sá da Costa.

Coseriu, E. (1987). “El gallego en la historia y en la actualidad”. In Actas do II Congresso Internacional da Língua Galego-Portuguesa.

Gamallo P. & Pichel, J.R. "Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary", Lecture Notes in Computer Science, vol. 4919, Springer-Verlag, (423-433).

Koehn, P, Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, M., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.

Och, F.J. & Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models". In Computational Linguistics, 29(1): 19-51.

Pichel, J.R. (7-04-2009). "Estratégia google". In Galicia Hoxe (http://www.galicia-hoxe.com/index_2.php?idMenu=149&idEdicion=1211&idNoticia=414218)

Pichel, J.R. (27-11-2007). "Falta de corpus". In Galicia Hoxe (http://www.galicia-hoxe.com/index_2.php?idMenu=153&idNoticia=236722)