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

José Ramom Pichel Campos, Paulo Malvar Fernández, Oscar Senra Gómez

Área de tecnologia linguistica, **imaxin**|software, Santiago de Compostela, Galiza, Spain.

jramompichel@imaxin.com

Pablo Gamallo Otero

Departamento de Língua Espanhola, Faculdade de Filologia, Universidade de Santiago de Compostela, Galiza, Spain.

pablogam@usc.es

Alberto García

agarcia@igalia.com

Engineering department, Igalia Free software company, A Coruña, Galiza, Spain.

Abstract

Since building a statistical translation engine between two languages is essential to obtain a significant amount of parallel corpus, and given that available parallel corpus English-Galician are not yet sufficient, it may seem obvious to follow other strategies. Galician and portuguese are considered by the major theorists of Romanic Linguistics (i.e.: Eugene Coseriu) as two variants of the same language. This assumption might open a new line of research which could provide Galician language with computational linguistics resources from both European Portuguese and Brazilian. In the field of machine translation, imaxin|software has built a prototype between English and Galician using English-Portuguese Europarl parallel corpus. First to achieve that, English-Portuguese corpus was converted into English-Galician using Portuguese-Galician RBMT engines and spelling converters, and second language and translation models were built by using and configuring Moses and Giza++. The results we have obtained allow us to conclude SMT machine translation tools based on Galician can be designed from Portuguese resources, a task otherwise unthinkable because of lack of corpus. We also assume this strategy can also be implemented to elaborate a great variety of linguistic technologies tools.

Introduction

Galician language is considered by most theorists in Linguistics as a lusophone language variety, such as European, Brazilian, Asiatic and African Portuguese. In fact it was in Galiza and Northern Portugal up to the Douro river where the language known internationally as Portuguese was born. According to this theory, we investigated whether it could be possible to make use of free English-Portuguese parallel corpus, in order to elaborate an English-Galician statistical machine translation prototype, through the application of both a Portuguese-Galician RBMT, and a spelling converter (i.e., a transliterate engine). We use the spelling converter to turn those Portuguese words which are not in the main dictionaries into galician spelling (very similar to Spanish). Let's notice that, shortly after launching this research project, Google included galician as reference language in his main tools. Surprisingly Google uses the same strategy as that described in this paper, however without using a transliterate engine.

EuroParl

The Europarl parallel corpus is extracted from the proceedings of the European Parliament. It

includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Given that this parallel corpus is often used for the construction of statistical machine translation engines and there's no enough English-Galician corpus to train a SMT system, we assume that it would be reasonable to build a SMT system for Galician from English-Portuguese texts. This corpus contains about 2x29 million English and Portuguese words.

EixOpentrad

EixOpenTrad is a further version of OpenTrad, an open source machine translation system (www.opentrad.com). EixOpenTrad is a Galician-Portuguese and Portuguese-Galician machine translation prototype containing 8500 words in both directions. This system is based on the translation engine apertium es-pt.

Pt-gl spelling converters

Spelling converters are usually used to write the same language code in two different ways. Such converters do nothing more than replace strings of letters and patterns of source language into other patterns of letters of the target language. This strategy does not involve morphological, syntactic, or semantic information. We have used a pt-gl spelling converter whose first version was developed by Berto Garcia from Igalia Free Company and later improved by Pablo Gamallo from Spanish department (University of Santiago de Compostela). It converts european Portuguese spelling into current Galician spelling (coinciding mostly with Spanish).

SMT en-gl prototype

The construction process of this prototype was done by converting the English-Portuguese Europarl Corpus to English-Galician. For this purpose, we made use of both EixOpentrad and the spelling converter. First, words were translated by our Portuguese-Galician RBMT machine translation, then words lacking in the EixOpentrad dictionaries were transliterated into the Galician spelling. Finally using Moses and Giza ++ software a SMT prototype English-Galician was designed. Let's see the following example, which is the automatic translation of the wikipedia entry Art 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.”

Google

Google has recently incorporated Galician language in its linguistic tools. For this purpose, the Google translator were trained with English-Portuguese parallel corpus partially converted into Galician spelling. Unlike imaxin|software strategy, Google does not use spelling converters. Thus, those Portuguese words which are not in the dictionaries remain in their original spelling. To compare both systems, you can see below the translation performed over the same example by the SMT system of Google:

”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 ea literatura. O significado da arte é explorador no ramo da filosofía coñecido como estética”.

Further research

The results of our SMT system may be improved by making use of two strategies: first, by increasing the Portuguese-Galician dictionaries integrated in the RBTM system called EixOpentrad, and second, by making the Portuguese corpus bigger. This way, we will focus on how to obtain free parallel corpus from Brazil and Portugal, and measure results.

Bibliography

- Franz Josef Och, Hermann Ney. "**A Systematic Comparison of Various Statistical Alignment Models**", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.
- 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, Evan Herbst, "**Moses: Open Source Toolkit for Statistical Machine Translation"**, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
- Eugene Coseriu, "**El gallego en la historia y en la actualidad**". Actas do II Congresso Internacional da Língua Galego-Portuguesa.
- Coromines J., Chaves de Melo, G., Alonso Estravis, I. "**Lingüística e lingüística galaico-portuguesa**". Editorial Irmandades da Fala de Galiza e Portugal.
- Gamallo P., and J-R. Pichel "Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary", Lecture Notes in Computer Science, vol. 4919, Springer-Verlag, (423-433). ISNN: 0302-9743.
- Gamallo P. and J.R. Pichel (2007) "Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego", Procesamiento del Lenguaje Natural, 39, pp. 241-248.
- "Estrategia google", José Ramom Pichel, 7-04-2009, Galicia Hoxe http://www.galicia-hoxe.com/index_2.php?idMenu=149&idEdicion=1211&idNoticia=414218
- "Falta de corpus", José Ramom Pichel, 27-11-2007, Galicia Hoxe http://www.galicia-hoxe.com/index_2.php?idMenu=153&idNoticia=236722