Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. The Galician and Portuguese case.

Abstract:

In order to build many statistically-driven NLP tools, it is essential to use a significantly large amount of data. To overcome the limitation of the scarcity of computational resources for languages such as Galician it is necessary to develop new strategies. In the case of Galician, well-known romanicists have theorized that Galician and Portuguese are two varieties of European Portuguese. From a pragmatic standpoint, this assumption could open up a new line of research to supply Galician with rich computational resources. Drawing from the English-Portuguese Europarl parallel corpus, ***imaxin****|software has compiled an English-Galician parallel corpus that we used to build an English-Galician Statistical Machine Translation prototype whose performance is comparable to Google Translate. We contend that this strategy can be implemented to develop a great variety of computational tools for languages like Galician that are closely related to languages for which there already exist great computational resources.*

Key words: Parallel corpus, English, Galician, Portuguese, Statistical Machine Translation

Resumen:

Para desarrollar muchas herramientas estadísticas de Procesamiento del Lenguaje Natural resulta esencial utilizar grandes cantidades de datos. Para salvar la limitación de la escasez de recursos computacionales para lenguas, como el gallego, es necesario diseñar nuevas estrategias. En el caso del gallego, importantes romanistas han teorizado que gallego y portugués son dos variantes del portugués europeo. Desde un punto de vista pragmático, esta hipótesis podería abrir una nueva línea de investigación para proporcionar al gallego ricos recursos computacionales. Partiendo del corpus paralelo inglés-portugués Europarl, ***imaxin****|software ha compilado un corpus paralelo inglés-gallego que hemos utilizado para crear un prototipo de traductor automático estadístico inglés-gallego, cuyo rendimiento es comparable a Google Translate. Sostenemos que es posible implementar esta estrategia para desarrollar una gran variedad de herramientas computacionales para lenguas, como el gallego, íntimamente relacionadas con lenguas que ya cuentan con un gran repertorio de recursos computacionales.*

Palabras clave: Corpus paralelo, Inglés, Gallego, Portugués, Traducción Automática Estadística

0. PREFACE

From the point of view of Hallidayan Functional-Systemic Linguistics (FSL) theory, language serves, according to Gee (1999, 1) "as both a tool for action and a scaffolding for 'human affiliation within cultures and social groups and institutions"'. In other words, language works as a tool not only for

communication but also for negotiating the relationships and the structures of society itself. It is precisely through this social dimension that language manages to play an extremely crucial symbolic role.

In developing computational tools for particular languages, computational linguists, whether they are primarily computer scientists or linguists, have a responsibility to the language(s) they are working with. It is possible that for high-prestige languages this responsibility is not obvious. In these cases, decisions about which linguistic phenomena are to be studied and, more importantly from the point of view of this paper, which tools are to be developed may seem trivial because they seem to not imply any particular ideological position. However, for those scientists who work with and for minoritized languages, especially if they are speakers of those languages, their decisions are never innocuous.

It is with this responsibility as language researchers and as speakers firmly in mind that this work is has been undertaken.

1. INTRODUCTION

In 2008 and 2009, at **imaxin**|software we carried out a project, subsidized by the *Dirección Xeral de I+D* of the *Xunta de Galicia*, called "RecursOpentrad: Recursos lingüístico-computacionais para a tradución automática avanzada de código aberto para a integración europea da lingua galega"[1]. In this project, in addition to building an English-Galician Ruled-based Machine Translation (RBMT) system, we thought that, given the progress[2] achieved in the field of Statistical Machine Translation (SMT), it was an excellent moment for taking a step forward in the development of Natural Language Processing (NLP) tools for Galician.

1 RecursOpentrad: Linguistic and computational resources for advanced open source machine translation for European integration of the Galician language.

2 Serve, just as an example, how high-quality SMT has become popular with the implementation done by Google of their statistical machine translation system, Google Translated (freely available at http://translate.google.com/).

When we decided to develop a prototype of a SMT system for English and Galician, we knew that "the larger the available training corpus, the better the performance of a translation system" (Popović & Ney: 2006, 25) we would achieve. However, while gathering the necessary resources for the development of such a prototype for the above-mentioned pair of languages we came to same conclusion with what Popović & Ney (2006) began their paper given at Language Resources and Evaluation (LREC) in 2006:

> Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality parallel text for the desired domain and language pair requires a lot of time and effort, and for some languages is not even possible. (25)

It is worth noting that it is possible to find English-Galician parallel corpora[3] published under the General Public License (GPL) on the Internet. Xavier Guinovart's research group at the Universidade de Vigo's Facultade de Traducción e Interpretación has gathered a collection of parallel corpora[4] within which the pair English-Galician is represented with a subcorpus of approximately 9 million words, which is by all means insufficient for the purpose of building a SMT system.

At this point it was clear for us that we needed to take a different route in order to achieve our goal. It is well-known in the linguistic community that important romanicists, such as Coseriu (1987), Cunha & Cintra (2002) and Aracil (1985), have theorized that, from a linguistic point of view, Galician should be considered a variety of Portuguese together with European, Brazilian, Asiatic and African Portuguese varieties. This is exactly what Fernández Rei (1991), member of the *Real Academia Galega*[5], and Coseriu (1987), one of the most important romanicist of the XXth century, point out:

3 Thanks to the localization projects of open source tools and operating systems carried out by the Galician open source community it is possible to manually collect domain-specific English-Galician parallel corpora. However, these corpora lack uniformity and, once again, they are insufficient for the purpose of building a SMT system.

4 This collection of parallel corpora can be consulted at http://sli.uvigo.es/CLUVI/.

5 Royal Academy of Galician Language.

> Na actualidade, desde o punto de vista estrictamente lingüísitico, ás dúas marxes do Miño fálase o mesmo idioma, pois os dialectos miñotos e transmontanos son unha continuación dos falares galegos, cos que comparten trazos comúns que os diferencian dos do centro e sur de Portugal; pero no plano da lingua común, e desde unha perspectiva sociolingüísitca, hai no actual occidente peninsular dúas linguas modernas, con diferencias fonéticas, morfosintácticas e léxicas, que poden non impedi-la intercomprensión ó existir un bilingüismo inherente entre o galego e o portugués, semellante ó existente entre o catalán e o occitano, o danés e o noruegués, o eslovaco e o checo, o feroés e o islandés.[6] (Fernández Rei: 1991, 17-18)

> los romanistas e hispanistas están en general de acuerdo en que el gallego es una forma particular del conjunto dialectal gallego-portugués, en cuanto opuesto al conjunto dialectal español (no "castellano", sino: astur-leonés, castellano, en sus muchas formas, y navarro-aragonés) y al conjunto catalán (o catalán-valenciano)[7] (Cuserio: 1987, 795)

Thus, drawing from the assumption that Galician and Portuguese are very closely related linguistic varieties and trying to take advantage of the privileged position of Portuguese as a computationally-developed language – that is, a language for which many NLP tools an resources have been developed–, we have investigated the possibility of using free English-Portuguese parallel corpora to create an English-Galician parallel corpus that we would use to develop an English-Galician SMT prototype.

2. CORPUS COMPILATION AND PROCESSING

2.1 The source corpus

Since our project was clearly guided by an open-sourced spirit, we wanted as

6 Today, from a strictly linguistic point of view, on both sides of the Miño River the same language is spoken, since Miñoto and Transmontano dialects are a continuation of the Galician dialects with which they share common traits that make them different from the dialects of Midland and Southern Portugal. However, in terms of common language, and from a sociolinguistic perspective, currently in the west of the Iberian Peninsula there are two modern languages, with phonetic, morphosyntactic and lexical differences, which do not prevent mutual understanding because of the inherent bilingualism that exists between Galician and Portuguese, similar to the existing bilingualism between Catalan and Occitan, Danish and Norwegian, Slovak and Czech, Faroese and Icelandic.

7 In general, romanicists and hispanists agree that Galician is a particular form of the Galician-Portuguese dialectological body, as opposed to the Spanish dialectological body (not “Castillian”, but: Asturian-Leonese, Castillian, in its many forms, and Navarrese-Aragonese) and the Catalan body (or Catalan-Valencian)

many components of it as possible to be open source, or at least freely available for non-commercial use.

Because of its large size and liberal copyright license[8] we chose the English-Portuguese Europarl corpus as the source corpus for our project.

The Europarl corpus[9] is a parallel corpus extracted from the European Parliament Proceedings, dating back to 1996, that includes versions of its contents in eleven European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

After an initial processing devoted to cleaning the XML tags that mark the discursive structure of the utterances contained in the corpus, we ended up with an English-Portuguese parallel corpus that contains 58 million words in total.

2.2 From an English-Portuguese to an English-Galician parallel corpus

The conversion of the source parallel corpus into an English-Galician parallel corpus we designed at **imaxin**|software was a semi-automatized process that involved the use of two main pieces of software: a RBMT system and a spelling converter –that is, a transliteration engine[10].

Thus, the simplified workflow we designed was the following:

1) Machine translation of the Portuguese side of the source parallel corpus into Galician using EixOpentrad[11].
2) Identification of the unknown, therefore, untranslated words outputted by EixOpentrad.
3) Transliteration of these untranslated words into Galician

8 "The European Parliament web site states: "Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged."" (Koehn: 2005, 2)
9 Freely available at http://www.statmt.org/europarl/
10 Spelling converters are usually used to write the same language code in two different ways. Such converters do nothing more than replace string patterns of the source language into the corresponding string patterns of the target language. This strategy does not involve morphological, syntactic, or semantic information.
11 EixOpenTrad is a further version of OpenTrad, a platform of open source Machine Translation services (www.opentrad.com). EixOpenTrad is a Galician-Portuguese and Portuguese-Galician MT prototype containing 8,500 words for both directions. This system is based on the Spanish-Portugues Apertium translation engine.

spelling using a simple Portuguese-Galician transliteration Perl script called port2gal[12].

4) Manual correction of the transliteration errors.

This process took over three months to complete, which is insignificant in contrast to the time-consuming and cost-expensive effort that would have been associated with the manual compilation of a English-Galician corpus of this size.

2.3 The target corpus

After processing of the English-Portuguese Europarl parallel corpus we obtained an English-Galician tokenized corpus composed of 34,715,016 tokens in English and 34,688,010 tokens in Galician.

3. ENGLISH-GALICIAN SMT PROTOTYPE

In order to exemplify the utility of the English-Galician parallel corpus we built and also to demonstrate the validity of our strategy, we will show the quality of the translations we have achieved with the SMT prototype we trained[13] using that English-Galician corpus, in comparison to the quality of Google's own SMT service, which in 2008 incorporated Galician in its catalogue of linguistic tools.

The following example shows a sample automatic translation of the

12 port2gal's first version was developed by Alberto Garcia (Igalia Free software Company). It was later improved by Pablo Gamallo (Department of Spanish Language at Universidade de Santiago de Compostela). It simply converts European Portuguese spelling into current Galician spelling and vice versa. It is freely available at http://gramatica.usc.es/~gamallo/port2gal.htm.

13 The English-Galician SMT prototype was built within the paradigm of what is known in the field of NLP as Phrase-based SMT. The two main pieces of software we used for this purpose were Moses and GIZA++, which can be respectively downloaded from http://www.statmt.org/moses/ and http://fjoch.com/GIZA++.html.

wikipedia *Art* entry[14] performed by our system:

> Arte é o proceso ou produto de arranxar deliberadamente elementos dunha forma que apela á sentidos ou emocións. Engloba un diversificado abano de actividades humanas, creacións e modos de expresión, inclusive da música, da literatura, filmes, escultura e pinturas. O significado de arte é explotada en un ramo da filosofía coñecida como aesthetics.

The next example shows the translation of the same wikipedia entry performed by Google Translate on March 2nd 2010:

> A arte é o proceso ou produto de deliberadamente organizar elementos de un modo que pide aos sentidos ou emocións. Engloba unha variada gama de actividades humanas, creacións, e modos de expresión, incluíndo a música, literatura, cine, escultura e pintura. O significado da arte é explotado desde unha rama da filosofía coñecido como estética.[15]

4. CONCLUSION

As shown in the previous section, we can confidently conclude that the strategy of creating NLP tools for Galician drawing from Portuguese resources is not only linguistically justifiable but, given the high quality of the results that can be achieved, is absolute legitimate.

It is, therefore, not adventurous to conclude that the use of resources from a closely related language, especially if this is a computationally-developed language, is extremely important for linguistic varieties, such as

14 The sample English sentence, located at http://en.wikipedia.org/wiki/Art, we used to test these two systems is: "Art is the process or product of deliberately arranging elements in a way that appeals to the senses or emotions. It encompasses a diverse range of human activities, creations, and modes of expression, including music, literature, film, sculpture, and paintings. The meaning of art is explored in a branch of philosophy known as aesthetics."

15 Google Translate was trained using English-Portuguese parallel corpora partially converted into Galician spelling. Until recently, 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, as shown by a translation we performed with this service in April 2009: "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."

Galician, that lack NLP tools due to their minoritized position.

5. REFERENCES

Aracil, Ll. et al. (1985). *Lingüística e sócio-lingüística galaico-portuguesa: reintegracionismo e conflito lingüístico na Galiza*. Ourense: Asociaçom Socio-Pedagógica Galaico-Portuguesa.

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. *Actas do II Congresso Internacional da Língua Galego-Portuguesa* (793-800). A Coruña: AGAL.

Fernández Rei, F. (1991). *Dialectoloxía da lingua galega*. (2nd ed.). Vigo: Edicións Xerais de Galicia.

Gamallo P. & Pichel, J.R. (2007). Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural*, 39, 241-248.

Gamallo P. & Pichel, J.R. (2008). Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *Lecture Notes in Computer Science* (Vol. 4919), 423-433.

Gee, J. P. 1999. *An Introduction to Discourse Analysis: Theory and Method*. London: Routledge.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Paper presented at the *MT Summit 2005*. Pukhet, Thailand, September 12-16.

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. Demonstration session at the *Annual*

Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic, June, 23-30.

Malvar Fernández. P. (2008). *Improving Word-to-Word Alignment using Morphological Information* (Master's Thesis). San Diego State University: San Diego, CA.

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

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

Pichel, J.R. (2009). "Estrategia google". *Galicia Hoxe*. Available at (http://www.galicia-hoxe.com/index_2.php?idMenu=149&idEdicion=1211&idNoticia=414218)

Popović, M. & Hey, H. Statistical Machine Translation with a Small Amount of Bilingual Training Data. Paper at *Language Resources and Evaluation (LREC): 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority language"*. Genova, Italy, May 23rd, 25-29.