Chinese-to-Spanish rule-based machine translation system

Jordi Centelles[1] **and Marta R. Costa-jussà**[2]
[1]Centre de Tecnologies i Aplicacions del llenguatge i la Parla (TALP),
Universitat Politècnica de Catalunya (UPC), Barcelona
[2]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico
[1]jordi.centelles.sabater@alu-etsetb.upc.edu,[2]marta@nlp.cic.ipn.mx

Abstract

This paper describes the first freely available Chinese-to-Spanish rule-based machine translation system. The system has been built using the Apertium technology and combining manual and statistical techniques. Evaluation in different test sets shows a high coverage between 82-88%.

1 Introduction

Chinese and Spanish are two of the most spoken languages in the world and they are gaining interest in the actual information society. In this sense, machine translation (MT) between these two languages would be clearly of interest for companies, tourists, students and politicians. Eventhough the necessity is a fact, there are not many Chinese-to-Spanish MT systems available in the Internet. In addition, the translation quality is quite behind the standards. Most of the approaches are corpus-based and they seem to produce translation through English pivoting to compensate the lack of Chinese-Spanish parallel corpora.

When it comes to academic research, there have been few works in these pair of languages which mainly are reviewed in Costa-jussà et al (2012b) and they also rely on the pivoting procedure. The linguistic differences (mainly at the level of morphology) between the two languages makes the training of data-driven systems rather difficult. Actually, Chinese and Spanish are languages with many linguistic differences, especially at the level of morphology and semantics. Chinese is an isolating language, which means that there is a one-to-one correspondence between words and morphemes. Whereas, Spanish is a fusional language, which means that words and morphemes are mixed together without clear limits. Regarding semantics, Chinese is a language that has a massive number of homophonyms at the lexical level (Zhang et al., 2006). Therefore, lexical semantic disambiguation towards Spanish will be harder.

Given these challenges, we decided to build a Chinese-to-Spanish rule-based machine translation (RBMT) system. These types of systems provide a translation based on linguistic knowledge in contrast to the existing and popular corpus-based approaches. The translation process is divided in: analysis, transfer and generation. Analysis and generation cover mainly the morphological and semantic variations of the languages, the transfer phase is in charge of the grammatical aspects (Hutchins and Sommers, 1992). The main advantages of RBMT are that they use linguistic knowledge and the produced errors can be traced.

Among the different linguistic motivations to build a Chinese-to-Spanish RBMT, we can list the following: (1) the proposed system will coherently manage the difference in morphology from Chinese to Spanish; (2) and the RBMT approach is able to exploit the use of linguistic tools which are available separately for Chinese and Spanish.

The main drawback of a RBMT system is that it requires a lot of human dedication and years of development (Costa-Jussà et al., 2012a) and that they exhibit weakness in lexical selection transfer, which is quite relevant in this pair of languages. However, in our case, we are using the Apertium platform (Forcada et al., 2011) that eases the process. In addition, when building the proposed RBMT approach, we use automatic techniques to feed the system from parallel corpus.

The rest of the paper is organized as follows. Section 2 reports a detailed description of the Chinese-to-Spanish RBMT architecture including the procedure of compiling monolingual and bilingual dictionaries as well as the computation of grammatical transfer rules. Section 3 reports an evaluation of the system in terms of coverage. Finally, Section 4 discusses the results and it draws the final conclusions.

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 82–86,
Gothenburg, Sweden, April 27, 2014.

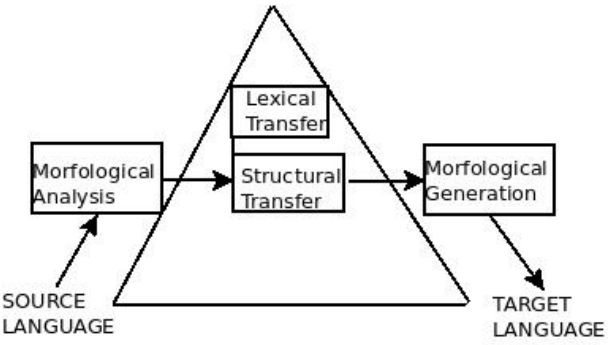

Figure 1: Block diagram of a typical rule-based machine translation system.

2 Rule-based machine translation architecture

The general architecture of a RBMT translation architecture has been defined in the literature in works such as (Hutchins and Sommers, 1992) or (Forcada et al., 2011), which is the open-source toolbox that we are using in this paper. In this section, we describe in detail how the system has been developed following similar procedures as (Cortés et al., 2012). Novelties in our work are that we are aiming a really challenging language pair with few bilingual speakers capable of developing the resources required to compile the targeted system.

Human annotation counted with two bilingual English-Spanish annotators and one trilingual annotator Chinese-English-Spanish, who was in charge of checking every step out.

2.1 System architecture

The system is based on the Apertium platform (Forcada et al., 2011) which is a free/open-source toolbox for shallow transfer machine translation. As well as the platform, the linguistic data for the MT systems are also available under the terms of the GNU GPL.

The platform was originally designed for the Romance languages of Spain, but it is moving away from those initial objectives (see the list of available languages in *wiki.apertium.org*. In practice, we use the architecture of the system, but, differently, we are using statistical techniques to complete our system.

Figure 1 shows the representative block diagram modules of the RBMT system. In this first description of the system, the only step that is not addressed is the lexical transfer.

Development to date has consisted of: feeding monolingual and bilingual dictionaries, to extend coverage, with statistical methods and with human annotation; filtering and cleaning monolingual and bilingual dictionaries to make them consistent; and computing grammatical transfer rules. Although the monolingual and bilingual dictionaries require the same entries, the function of each one is different. The monolingual dictionary contains morphological information and the bilingual dictionary contains the translation entry itself.

This first track of development has taken place in over the course of five months, which contrasts with the long time required to develop classical RBMT systems. The key point here is that our system has been developed using a hybrid approach and that, although the system is capable of achieving state-of-the-art translations, it is still under construction. The last version of the system is available for download at the Apertium site[1].

2.2 Bilingual dictionary

The bilingual dictionary was computed following two methodologies or procedures.

The first one is manual by using the Yellow Bridge resource[2]. This web is, as mentioned by the authors, the premier guide to Chinese language and culture for English speakers. They provide comprehensive tools for learning the Chinese language. Although there are many Chinese-related websites, this one is well-organized and complete. For Chinese, they provide a list of words classified following grammatical categories, including: adjectives, adverbs, conjunctions, interjections, measure words, nouns, numerals, onomatopoeia, particles, prefixes, prepositions, pronouns, question words, suffixes, time words and different types of verbs. For each category, each word has its corresponding translation into English. Then, this dictionary was used to feed the dictionary. But to double-check the translations provided, each word was translated using another on-line dictionary[3] and Google Translate. This procedure allowed to add several hundreds of numerals, conjunctions, adverbs, pronouns, determinants, adjectives, 3,000 nouns and 2,000 verbs.

The second procedure is statistical-based. The parallel corpus of the United Nations (UN) (Rafalovitch and Dale, 2009) was aligned at the

[1] http://sourceforge.net/projects/

[2] http://www.yellowbridge.com/chinese/chinese-parts-of-speech.php

[3] http://www.chinese-tools.com/

level of word by using the standard GIZA++ (Och and Ney, 2003) software. Alignment was performed from source to target and target to source. Symmetrization was done using intersection because it provides the most reliable links. Then, we extracted phrases of length one, which means that we extracted translations from word-to-word. This dictionary was manually filtered to eliminate incorrect entries. This procedure allowed to add around 3,500 words in the dictionaries. Our dictionary has around 9,000 words.

2.3 Chinese monolingual dictionary

The Chinese monolingual dictionary was extracted from the source part of the bilingual dictionary. Additionally, it was filtered with regular expressions to avoid repeated entries.

Regarding the morphological analysis, Chinese is an isolating language, which in brief means that words (or symbols) cannot be segmented in submorphemes. In this sense, no morphological analysis is required. However, the main challenge of Chinese is that most of the time symbols appear concatenated and sentences are not segmented into words as it is most common in other languages. Therefore, Chinese requires to be segmented. We used the ZhSeg (Dyer, 2013) programmed in C++. We evaluated the performance of this segmenter in comparison to the Left to Right Longest Match (LRLM), which is the parsing strategy used by Apertium in analysis mode. This procedure read tokens from left to right, matching the longest sequence that is in the dictionary (like "greedy" matching of regular expressions). Both ZhSeg and LRLM were compared using an in-house segmented test set of 456 words as a reference. The Word Error Rate (WER) measure for the ZhSeg was 16.56% and 16.89% for LRLM. Given that results were comparable, we decided to use the Apertium LRLM strategy.

It is mandatory that the monolingual and the bilingual dictionary are coherent, which means that they have to have the same entries. Both dictionaries were cleaned up with different regular expressions. Therefore, we have to ensure that there are not situations like there is a word in the monolingual dictionary, which is not in the bilingual dictionary and the other way round. In order to check out this, we used testvoc. As mentioned in the Apertium documentation [4], a testvoc is literally a test of vocabulary. At the most basic level, it just expands the monolingual dictionary, and runs each possibly analyzed lexical form through all the translation stages to see that for each possible input, a sensible translation in the target language is generated. This tool was used to clean up dictionaries.

[4] http://wiki.apertium.org/wiki/Testvoc

2.4 Spanish generation

This part of the translator was taken from the repository of Apertium given that is has been developed during years. Some previous publications that explain Spanish generation can by found in (Armentano-Oller et al., 2006; Corbí-Bellot et al., 2005). Basically, it consists of the three modules of Apertium: morphological generator that delivers a surface (inflected) form for each transferred word. This is done using the generation dictionary, which for each lemma and part-of-speech tag is able to generate the final form. Then, the post-generator that performs orthographic operations such as contractions (e.g. *de el* and *del*).

2.5 Transfer-rules

Grammatical transfer-rules were extracted following a manual procedure, which consisted in performing a translation of a source text and contrasting the output translation, the source and the reference. From this observation, manual patterns were extracted in order to design a rule that could cover the necessary modifications to be done. Following this procedure, there were 28 rules extracted intrasyntagms, which modify inside a syntagm, and 34 intersyntagms, which modify among different sytnagms.

As follows we show an example of rule extracted intrasyntagm.

```
< rule comment = RULE : adj nom >
 < pattern >
  < pattern − itemn = adj/ >
  < pattern − itemn = nom/ >
 < /pattern >
 < action >
 < call − macron = f − concord2 >
  < with − parampos = 2/ >
  < with − parampos = 1/ >
 < /call − macro >
 < out >
  < chunkname = j_ncase = caseFirstWord >
   < tags >
    < tag >< lit − tagv = SN/ >< /tag >
    < tag >< clip pos = 2side = tlpart = gen/ >< /tag >
    < tag >< clip pos = 2side = tlpart = nbr/ >< /tag >
    < tag >< lit − tagv = p3/ >< /tag >
   < /tags >
  < lu >
   < clip pos = 2side = tlpart = whole/ >
  < /lu >
  < b pos = 1/ >
  < lu >
   < clip pos = 1side = tlpart = lem/ >
```

```
    < clip pos = 1side = tlpart = a_adj/ >
    < clip pos = 1side = tlpart = gen/ >
    < clip pos = 1side = tlpart = nbr/ >
   < /lu >
  < /chunk >
 < /out >
 < /action >
< /rule >
```

This rule reorders *adjective* + *noun* into *noun* + *adjective*. Moreover, this rule ensures that the number and gender of the noun and the adjective agree.

3 Evaluation framework

This section reports the evaluation framework we have used to analyze the quality of the Chinese-to-Spanish RBMT described.

Dataset	Domain	Words	Coverage
Dev	News	1,651	88.7
Test	UN	35,914	83.8
	In-house	10,361	82.8

Table 1: Coverage results.

We can evaluate the rule-based MT systems in terms of coverage. We are using texts from different domains to perform the evaluation. Domains include news (extracted from the web[5][6]) for checking the evolution of the rule-based system; a subcorpus of UN (Rafalovitch and Dale, 2009); and an in-house developed small corpus in the transportation and hospitality domains. To do the evaluation of coverage we do not need a reference of translation. Table 1 shows the coverage results of our system.

This rule-based MT approach can be the baseline system towards a hybrid architecture. Inspired in previous promising works (España-Bonet et al., 2011), we have identified some ways of building a hybrid architecture given a rule-based MT system and available parallel and monolingual corpus:

- Starting with the core of a rule-based system, there is the necessity of extracting transfer-rules from parallel corpus and offering a translation probability to each one. This would allow to building rule-based MT systems by a monolingual human linguist. At the moment, rule-based MT systems have to be developed by bilingual native linguists or at least people who are proficient in the source and target language.

- In order to help rule-based MT systems be more fluent and natural, it would be nice to integrate a language model in the generation step. The language model could be n-gram-based, syntax-based or trained on neural-based. In each case, a different decoding would be required to be integrated in the system.

- Additional feature functions as the popular lexical ones or others that introduce source context information can be used together with the above language model.

4 Conclusions and further work

This paper has described the construction of the first Chinese-to-Spanish open-source RBMT system;. Particularly, the human knowledge has been used for providing exhaustive monolingual and bilingual dictionaries as well as for defining grammatical transfer rules. The statistical knowledge has complemented the creation of dictionaries. Therefore, we have shown effective techniques of building dictionaries using hybrid techniques. The new RBMT system has shown a high coverage in different domains.

As future work, the RBMT has to be improved mainly with new dictionary entries and more complex transfer rules. Both enhancements can be done combining human and statistical knowledge.

5 Acknowledgements

This work has been partially supported by the Google Summer of Code and the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951). Authors want to thank Apertium experts that believed in this project and helped a lot during the development, specially Francis Tyers, Víctor Sánchez-Cartagena, Filip Petkovsky, Gema Ramírez and Mikel Forcada.

[5]http://politics.people.com.cn/n/2013/0709/c1001-22134594.html

[6]http://finance.people.com.cn/n/2013/0722/c1004-22275982.html

References

C. Armentano-Oller, R. C. Carrasco, A. M. Corb-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, and M. A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, and M.C. Dias, editors, *Computational Processing of the Portuguese Language, Proc. of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR*, volume 3960 of *Lecture Notes in Computer Science*, pages 50–59. Springer-Verlag, May.

A. M. Corbí-Bellot, M. L. Forcada, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor, and K. Sarasola. 2005. An open-source shallow-transfer machine translation engine for the romance languages of spain. In *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pages 79–86, May.

J. P. Martínez Cortés, J. O'Regan, and F. M. Tyers. 2012. Free/open source shallow-transfer based machine translation for spanish and aragonese. In *LREC*, pages 2153–2157.

M. R. Costa-Jussà, M. Farrús, J. B. Mariño, and J. A. R. Fonollosa. 2012a. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and Informatics*, 31(2):245–270.

M. R. Costa-jussà, C. A. Henríquez Q, and R. E. Banchs. 2012b. Evaluating indirect strategies for Chinese-Spanish statistical machine translation. *J. Artif. Int. Res.*, 45(1):761–780.

C. Dyer. 2013. http://code.google.com/p/zhseg/.

C. España-Bonet, G. Labaka, A. Díaz de Ilarraza, L. Màrquez, and K. Sarasola. 2011. Hybrid machine translation guided by a rule-based system. In *Proc of the 13th Machine Translation Summit*, pages 554–561, Xiamen, China, sep.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

W. J. Hutchins and L. Sommers. 1992. An introduction to machine translation. *Academic Press*, 362.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

A. Rafalovitch and R. Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proc. of the MT Summit XII*, pages 292–299, Ottawa.

Y. Zhang, N. Wu, and M. Yip. 2006. Lexical ambiguity resolution in chinese sentence processing. *Handbook of East Asian Psycholinguistics*, 1:268–278.