CHINESE-SPANISH STATISTICAL MACHINE TRANSLATION EXPERIMENTS

Rafael E. Banchs, Josep M. Crego, Patrik Lambert and José B. Mariño

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya

ABSTRACT

This article presents some experimental results on Chinese to Spanish machine translation. The implemented translation system is based on the statistical framework and, more specifically, it implements the bilingual n-gram approach. Since, as far as we know, no Chinese-Spanish parallel corpus is currently available for training purposes, an alternative method for artificially constructing a training corpus is proposed and evaluated. Finally, the obtained translations are compared, in terms of translation quality, with those obtained by using a rule-based translation engine which is available on-line.

1. INTRODUCTION

In spite of the global effort currently invested in machine translation technologies, it is interesting to notice that most of it is currently concentrated in some specific translation pairs such as Spanish-English, Chinese-English, Arabic-English and Japanese-English; while some other language pairs such as Chinese-Spanish are, as far as we know, somehow unattended from both research and commercial perspectives. In response to this, the main goal of this work is to build and evaluate a direct Chinese to Spanish Statistical Machine Translation (SMT) system.

Nevertheless, the main drawback for building such a SMT system is the inexistence, at least as a publicly available resource, of a bilingual Chinese-Spanish parallel corpus large enough to perform an appropriate training of the bilingual translation model. In this way, an alternative experimental method for building the required training corpus is proposed and studied.

This procedure relies on using an English to Spanish SMT system for translating the English side of a Chinese-English parallel corpus into Spanish. The resulting Chinese-Spanish parallel corpus is depurated by an additional filtering stage in order to eliminate possible noisy data resulting from the translation errors implied in its generation.

Finally, the resulting translation system is evaluated by comparing the obtained translations, in terms of translation quality, with those obtained by using a rule-based translation system which is publicly available on-line.

This document is structured as follows. First, section 2 presents a very brief overview of the SMT system used within this work. Next, section 3 describes in detail the method proposed for constructing the Chinese-Spanish parallel training corpus. Then, section 4 describes and discusses the experiments conducted and their corresponding results. Finally, section 5 presents the conclusions, as well as the further work to be performed in the near future for improving the presented Chinese to Spanish SMT system.

2. THE BILINGUAL N-GRAM SMT APPROACH

This section presents a very brief description of the SMT approach that was used for all experiments performed within this work. For a more detailed description see [1]. This approach implements a translation model that has been derived from the finite-state perspective; more specifically, from [2]. However, different from it, where the translation model is implemented by using a finite-state transducer, the SMT approach used here considers a translation model which is based on 3-grams of bilingual units which are referred to as tuples.

Tuples are extracted from Viterbi alignments (more specifically, from the union of source-to-target and target-to-source alignments) in such a way that a maximal monotonic segmentation of each bilingual sentence pair is achieved [1]. Figure 1 illustrates the corresponding tuple segmentation for a given bilingual sentence pair.

In addition to the tuple 3-gram translation model, the considered SMT system implements four additional feature functions which are log-linearly combined with the translation model for decoding purposes. These feature functions are the following: a target language model implemented by means of word 4-grams, a word bonus model that compensates the system preference for short translations over large ones, and two complementary translation models (source-to-target and target-to-source) which are implemented by using the IBM-1 lexical parameters.

This work has been partly funded by TALP (Centre de Tecnologies i Aplicacions del Llenguatge i la Parla) and by the Spanish Department of Education and Science.

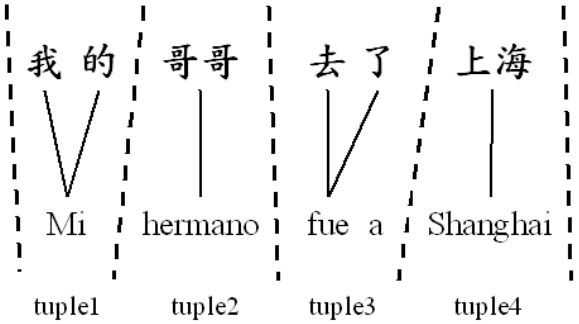

Figure 1. *Example of tuple segmentation for a given bilingual sentence pair.*

Finally, a specific n-gram based SMT search engine was used for decoding. A detailed description can be found in [3]. This decoder implements a beam-search strategy based on dynamic programming and allows for non-monotonic search. However, all experiments performed in this work were performed by using monotonic search. Although word reordering plays a very important role in the translation task under consideration, however as a first approximation and in order to maintain computational time manageable, we have opted for monotonic decoding. Additionally, an optimization tool based on a downhill simplex method was also developed. This algorithm allows for estimating log-linear weights for each feature so that the translation *BLEU* is maximized over a development set.

3. PARALLEL CORPUS CONSTRUCTION

As already mentioned, the main drawback for building the proposed SMT system is the inexistence, at least as a publicly available resource, of a bilingual parallel corpus for performing the translation model training. Although much information in Chinese and Spanish is available through the LDC (Linguistic Data Consortium *http://www.ldc.upenn.edu/*), the intersection of databases containing either Spanish or Chinese is null, so the extraction of a Chinese-Spanish parallel corpus from the existing databases is not possible at all. For this reason, an alternative method for building the required training corpus is proposed and evaluated. The method is depicted in detail within this section.

The proposed method relies on using an English to Spanish SMT system for constructing a Chinese-Spanish parallel corpus by translating the English side of a Chinese-English parallel corpus into Spanish. The Chinese-English parallel corpus used corresponds to a collection of speeches from United Nations, which is available through the LDC. More specifically, a 100K-sentence subset of the *UN Chinese English Parallel Text* (LDC2004E12) was used. On the other hand, for the English-Spanish parallel corpus, the European Parliament data available through the TC-STAR (Technology and Corpora for Speech-to-Speech Translation *http://www.tc-star.org/*) was considered. More specifically, a 100K-sentence subset of the training data from the TC-STAR second evaluation campaign was used. Table 1 presents basic statistics for the Chinese-English and Spanish-English parallel data sets.

Corpus	Language	Sentences	Words	Vocab.
ZH-EN	Chinese	105 K	1.9 M	29.5 K
	English	105 K	2.1 M	34.8 K
ES-EN	Spanish	105 K	2.0 M	40.0 K
	English	105 K	2.0 M	27.0 K

Table 1. *Number of sentences, running words and vocabulary for Chinese-English and Spanish-English corpora (K stands for thousands and M for millions).*

In this way, an English to Spanish translation system, based on the bilingual n-gram approach described in section 2, was trained by using the corresponding 100K-sentence training data presented in table 1 and optimized by maximizing the translation *BLEU* over a development set extracted from the original Spanish-English parallel data. The obtained system was then used to translate into Spanish the English side of the Chinese-English training set presented in table 1.

Finally, Chinese-Spanish development and test sets were manually constructed. The development corpus was created by manually translating into Spanish the English side of a 330-sentence set extracted from the original Chinese-English parallel data. This set was selected such that no overlap occurred with the 100K-sentence training set. Similarly, a 100-sentence test set was built by means of a similar procedure. Table 2 presents the basic statistics for the resulting Chinese-Spanish training set, as well as for the manually constructed development and test sets.

Corpus	Language	Sentences	Words	Vocab.
TRAIN	Chinese	105 K	1.9 M	29.5 K
	Spanish	105 K	2.0 M	34.8 K
DEV	Chinese	330	6.0 K	1.6 K
	Spanish	330	6.8 K	2.0 K
TEST	Chinese	100	1.9 K	813
	Spanish	100	2.1 K	908

Table 2. *Number of sentences, running words and vocabulary for the constructed Chinese-Spanish data sets (K stands for thousands and M for millions).*

4. EXPERIMENTS AND RESULTS

This section presents and discusses the experimental procedures considered in this work and their corresponding results. Two experimental procedures were performed. The first set of experiments attempts to evaluate the possibility of improving translation accuracy for the direct Chinese-Spanish translation system by filtering the artificially constructed Chinese-Spanish training corpus. These

experiments are presented in subsection 4.1. The second experiment is intended to compare the performance of the developed SMT system with a rule-based MT system. This second experiment is presented in subsection 4.2.

4.1. Corpus Filtering

Previous experimentation has confirmed that the artificial construction of a Chinese-Spanish bilingual corpus does not necessarily conduce to significant improvements in translation accuracy with respect to the simpler approach of performing Chinese to Spanish translations by using English as a bridge [4]. According to this result, it seems that the only way to actually exploit a procedure of this kind is by being able to retain the most useful sentence pairs of the artificially generated corpus. In this way, the proposed methodology is complemented with a corpus preprocessing stage in which the constructed Chinese-Spanish bilingual corpus is filtered in order to eliminate possible noisy data resulting from the translation errors implied in its generation.

To this end, a filtering strategy based on language model statistics is implemented. This filtering strategy consists on using a Spanish language model for selecting those best Spanish sentences in the Chinese-Spanish parallel corpus. Notice that this filtering is conducted only in the Spanish side of the corpus because it corresponds to the one which was artificially generated by translating the English side of the original Chinese-English parallel corpus. So, the noise expected to occur in the Chinese-Spanish corpus should be related to the translation errors produced by the English to Spanish translation system.

In order to implement the filter, a 3-gram language model, trained from the Spanish side of the 100K-sentence English-Spanish corpus presented in table 1 was used. Since language model probabilities are affected by sentence length, this filtering was performed independently for each subset of Spanish sentences of equal length.

According to this, the 100K-sentence Chinese-Spanish data set presented in table 2 was filtered by using the proposed language model criterion. Different threshold values were considered in order to generate several training subsets of some predefined different sizes. More specifically, training subsets of 95, 90, 85, 80, 70, 60, 50, 40, 30, 20 and 10K-sentences were generated.

Notice that, although training data size reduction has a negative impact on translation accuracy, it is expected that the noisy data reduction provided by the filtering process prevails over the data reduction effect so the overall system performance is incremented. In order to evaluate the effect of filtering independently from the effect of reducing the size of the training data set, an additional control experiment was performed for each of the eleven filtering experiments under consideration. Such control experiments consisted in training and optimizing a translation system by using a randomly generated equal-size training subset. In this way each filtering experiment could be compared with a control one that was trained with the same amount of data; but different from the filtered one, the control training data was selected at random.

In summary, a total amount of twenty two training subsets were generated. Eleven of them by using the proposed filtering strategy, and the other eleven at random. Accordingly, twenty two Chinese to Spanish translation systems were independently trained and optimized. The corresponding results, in terms of translation *BLEU* measured over the development data set, are presented in figure 2.

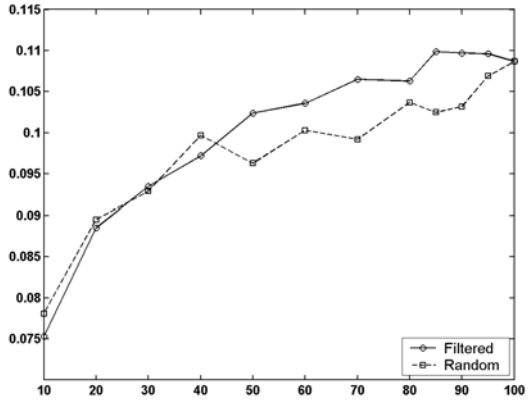

Figure 2. *Translation BLEU, over the development set, for the eleven filtered-data translation systems and their corresponding control systems. The horizontal axis represents the size of the training data set in thousands of words.*

As seen from figure 2, down to the 50K-sentence corpus, filtering does help improving a little bit translation accuracy with respect to those systems which has been trained with the same amount of non-filtered data. Notice also, how this improvement tends to fade out for smaller-sized corpora. Additionally, notice from the figure that no actual translation quality improvement has been achieved by any of the filtered systems with respect to the original 100K-sentence system. This suggests that the negative effect resulting from data training reduction is at least as relevant for the overall system performance as the positive effect resulting from filtering.

In order to alleviate the incidence of data reduction on overall system performance, the use of an independently trained target language model was evaluated. In this way, all experiments presented in figure 2 were repeated by using a Spanish language model (trained from data in table 1) as the decoder's target language model feature, instead of training an individual model for each system from its corresponding training data. These results are presented in figure 3.

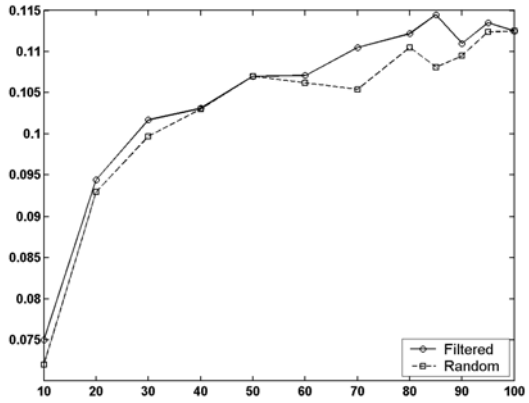

Figure 3. *Translation BLEU, over the development set, for the eleven filtered-data translation systems and their corresponding control systems. A common target language model feature, trained from the 100K-sentence Spanish data in table 1, was used for all systems.*

As seen from figure 3, in this second set of experiments, filtered-data and control systems achieved better *BLEU* scores than their corresponding systems in figure 2. This can be explained in terms of the better target language model that was used for this second set of experiments. Additionally, it can be noticed that similar to figure 2, filtered-data systems performed slightly better than control systems for larger-sized data sets. However, in this case, the observed gap is smaller than the one observed in figure 2.

4.2. Comparison with a Rule Based MT System

As an additional evaluation of the developed translation system, a comparison with a rule-based translation system was performed. In order to do this, a translation system which is publicly available on-line was used (*http://www.worldlingo.com/en/products_services/worldlingo_translator.html*). In order to avoid any possible bias resulting from the optimization process, the test set described in table 2 was used for this evaluation.

According to this, the test data set was translated from Chinese into Spanish by using the best performing system from figure 3 and the on-line translation system. Both outputs were automatically evaluated in terms of translation accuracy (*BLEU* and *NIST*) and error rates (*PER* and *WER*). Results are presented in table 3.

System	*BLEU*	*NIST*	*PER*	*WER*
Developed	0.1336	4.3101	57.99%	79.73%
On-line	0.0697	2.8355	74.15%	93.71%

Table 3. *Translation accuracy scores and error rates computed over the test set translations for both the developed statistical translation system and the on-line translation system.*

As seen from table 3, the developed translation system outperforms the on-line available one according to all evaluation metrics. Although it is known that accuracy scores such as *BLEU* and *NIST* tend to favor statistical systems over rule-based systems, notice that both error rates (*PER* and *WER*) are also significantly better for the developed system than for the on-line one.

5. CONCLUSIONS AND FURTHER WORK

This work presented some experimental results on Chinese to Spanish machine translation. A method for artificially constructing and filtering a Spanish-Chinese parallel corpus was presented and evaluated. Results demonstrated that the negative effect resulting from data reduction is at least as relevant as the positive effect resulting from filtering for the overall system performance. However, this situation can be somehow alleviated by using a larger and independently trained target language model. Additionally, the constructed translation system was compared with a rule-based translation system which is publicly available on-line. The developed system outperformed the on-line one.

For further research we are planning to work in two main directions. First, we will attempt improving the Chinese-Spanish parallel corpus construction technique. In this sense, different alternatives for filtering the artificially constructed data set should be designed and evaluated, such as the use of dictionaries and morpho-syntactic information. The second main direction of work should be related to improvements in the translation system, by including additional features and allowing for non-monotonic search in the translation task under consideration.

6. REFERENCES

[1] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, M. Ruiz, 2005, "Bilingual n-gram statistical machine translation", *in Proc. of the X MT-summit*, pp. 275-282.

[2] F. Casacuberta, E. Vidal, 2004, "Machine translation with inferred stochastic finite-state transducers", *Computational Linguistics*, vol 30, pp. 205-225.

[3] J.M. Crego, J.B. Mariño, A. de Gispert, 2005, "A ngram-based statistical machine translation decoder", *in Proc. of the 9th European Conference on Speech Communication and Tech-nology*, Interspeech.

[4] R.E. Banchs, J.M. Crego, P. Lambert, J.B. Mariño, 2006, "A feasibility study for chinese-spanish statistical machine translation", in *Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing.*