Exploring Spanish-morphology effects on Chinese–Spanish SMT

Rafael E. Banchs
Barcelona Media Innovation Centre
Ocata #1, Barcelona
08003 Spain
rafael.banchs@barcelonamedia.org

Haizhou Li
Institute for Infocomm Research
Heng Mui Keng Terrace #21
119613 Singapore
hli@i2r.a-star.edu.sg

Abstract

This paper presents some statistical machine translation results among English, Spanish and Chinese, and focuses on exploring Spanish-morphology effects on the Chinese to Spanish translation task. Although not strictly comparable, it is observed that by reducing Spanish morphology the accuracy achieved in the Chinese to Spanish translation task becomes comparable to the one achieved in the Chinese to English task. Further experimentation on approaching the problem of generating Spanish morphology as a translation task by itself is also performed, and results discussed. All experiments have been carried out by using a trilingual parallel corpus extracted from the Bible.

1 Introduction

The Chinese–Spanish translation task has been recently explored by Banchs et al. (2006). As far as we know, no Chinese–Spanish parallel corpus large enough for training a statistical machine translation system is available, at least as a public resource. For this reason, in that previous work, the artificial generation of the required Chinese–Spanish parallel corpus was attempted in order to pursue machine translation experimentation for this specific language pair.

From that work it was concluded that artificial generation of the bilingual corpus did not provide better translation accuracy than cascading two independent translation systems by using English as a bridge. Even more, filtering the artificially generated corpus aiming at improving translation results did not help at all, because the negative effect of reducing corpus size was more influential than the positive effect of improving corpus quality, at least for the corpus size considered in that opportunity.

In the present work, we present some experimentation results with a small parallel corpus we have extracted from the Bible. The collected corpus includes English, as well as Spanish and Chinese. The corpus collection and preparation, as well as its statistics are presented in section 2. Then, some baseline experimentation is carried out among the three languages in order to determine the best alignment set, phrase size and language model order for each of the six possible translation tasks. These results are presented in section 3. Then, the effect of Spanish morphology is explored for the particular case of the Chinese to Spanish translation task. In this sense, Spanish morphology is reduced by using a morphological analyzer and a Chinese to Spanish-without-morphology translation system is constructed. The problem of Spanish morphology generation is also approached as a translation task, and the Chinese to Spanish translation problem is attempted in a two step procedure in order to alleviate the translation task complexity by decoupling the translation task from the morphology generation task. These results are presented and discussed in section 4. Finally, some conclusions are presented, and future research strategies in this area are depicted.

2 Corpus collection and preparation

The trilingual parallel corpus used in this work has been extracted from three versions of the Bible Chinese (ZH), English (EN) and Spanish (ES). The original documents have been obtained from the web in digital format.[1] In the case of the Chinese and English versions, the complete text was available in a single document; while in the case of the Spanish version, each of the 66 books was in a separated file. The collected corpus was preprocessed and prepared for SMT experimentation by using the procedures described below.

Alignment: alignment at the sentence level was performed. In this particular case, this step was actually a simple one since the original text included annotation marks for chapters and verses. However some manual verification and edition was required since some missing verses and annotation inconsistencies were detected among the three different versions.

Tokenization: each data file was tokenized. In the case of Spanish and English, this implies the separation between punctuation marks and words. For the case of Chinese, for which word segmentation is not obvious, automatic word segmentation was performed by using the freely available tool ICTCLAS (Zhang et al, 2003).[2]

Morphology reduction: Morphological analysis and preprocessing of Spanish data was carried out. Such a preprocessing produces a slightly different tokenization for the Spanish data mainly because some multi-word units are reduced to single lexical forms. Because of this, four different Spanish data sets are considered: the original tokenized data, a lowercased version of the original tokenized data (lwc), re-tokenized data resulting from applying morphological analysis to the lowercased data set (rtk), and a lemmatized version of the re-tokenized one (lem). The lemmatized data corresponds to a morphologically reduced corpus in which all full forms have been replaced by their corresponding lemma forms. The morphological analysis was performed by using the freely available tool FreeLing (Carreras et al, 2002).[3]

Length restriction: all sentences (in any of the three languages) containing more than 80 tokens were removed from the corpus along with their corresponding other-two-language sentences. This restriction was mainly adopted in order to avoid possible alignment problems.

Fertility filtering: all trilingual sentence sets, for which any pair of them presented a token ratio equal to or higher than 9, were removed from the corpus. This avoids symmetrization errors due to fertility filtering implemented by the word to word alignment tool used for training the models, which also considers a token ratio of 9.

Corpus segmentation: finally, the corpus was divided into three trilingual parallel data sets: training, development and test.

Table 1 presents the main corpus statistics for all data set considered in the experiments. These statistics include the total number of sentences, the total number of words, the size of the vocabulary and the average sentence length. The out-of-vocabulary rates for development and test data are, respectively, *3.7%* and *4.2%* for Chinese, *9.3%* and *8.9%* for Spanish, and *5.3%* and *4.2%* for English.

Training data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	28,887	848,776	13,216	29.38
ZH	28,887	760,451	12,670	26.33
ES	28,887	781,113	28,178	27.04
ES-lwc	28,887	781,113	26,251	27.04
ES-rtk	28,887	784,398	25,240	27.15
ES-lem	28,887	784,398	14,229	27.15
Development data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	1,033	30,199	3,234	29.23
ZH	1,033	27,235	3,404	26.37
ES	1,033	27,862	4,634	26.97
ES-lwc	1,033	27,862	4,413	26.97
ES-rtk	1,033	27,986	4,403	27.09
ES-lem	1,033	27,986	2,882	27.09
Test data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	1,035	30,008	3,158	28.99
ZH	1,035	26,794	3,396	25.89
ES	1,035	27,368	4,652	26.44
ES-lwc	1,035	27,368	4,428	26.44
ES-rtk	1,035	27,452	4,426	26.52
ES-lem	1,035	27,452	2,864	26.52

Table 1: Main corpus statistics

[1] The Spanish version was downloaded from http://es.catholic.net/biblia/, the Chinese version from http://www.o-bible.org/download/hgb.txt and the English version from http://www.o-bible.com/dlb.html

[2] Available at http://www.nlp.org.cn/project/project.php?proj_id=6

[3] Available at http://garraf.epsevg.upc.es/freeling/

3 Baseline experimentation

For all experiments presented in this work, a very basic phrase-based SMT system is used. Word to word alignments are computed for the training data sets by using GIZA++ (Och & Ney, 2003).[4] Phrases are extracted from alignments and the translation probabilities are estimated by using relative frequencies. Language models are computed by using the SRILM toolkit (Stolcke, 2002),[5] and decoding is carried out by using Pharaoh (Koehn, 2004),[6] for which only four basic feature functions are considered: the translation model, the language model, the distortion model and the word penalty factor. Model weight optimization is performed by using the standard minimum-error-training procedure (Och, 2003) which was implemented by using the Simplex algorithm for maximizing translation BLEU over the development data set.

Some baseline experimentation was carried out among the three languages in order to determine the best alignment set, phrase size and language model order for each of the six possible translation tasks. In these baseline experiments, four different alignment sets were considered for phrase extraction: source to target (sr2tg), intersection (inter), union (union) and symmetrized alignments (sym) (Matusov et al, 2004). Regarding phrase lengths and target language models, two maximum phrase lengths were considered for translation model computation: *3* and *4* tokens; and three maximum *n*–gram sizes were considered for target language model computation: *2*–, *3*– and *4*–grams.

According to results from these baseline experiments, the optimal maximum phrase length for translation model computation was consistently *4* tokens for all translation tasks; and, similarly, the optimal language model order for target language model computation was consistently *3*. However, in the case of the alignment set considered for phrase extraction interesting differences could be observed. Table 2 presents BLEU scores over the test set for all of the six possible translation tasks when extracting phrases from each of the four different alignment sets considered.[7] For all results presented in table 2, the maximum phrase lengths considered were *4* tokens, and the maximum *n*–gram sizes considered were *3*–grams.

Note from table 2 that, although in many cases performances are relatively similar, in the cases were Spanish is the target language the intersection clearly offers the best performance. In all other cases, with the exception of the English to Chinese task for which the source to target seems to be performing better, the symmetrized set of alignments performs slightly better.

Task	Sr2tg	Inter	Union	Sym
ZH-ES	14.1	**14.3**	12.9	13.8
ES-ZH	16.7	17.4	15.2	**17.5**
ZH-EN	18.6	19.2	17.2	**19.6**
EN-ZH	**20.2**	20.1	19.3	19.7
EN-ES	30.6	**31.5**	30.6	30.5
ES-EN	34.4	34.2	34.3	**34.5**

Table 2: Translation BLEU over the test set for all six tasks and the four alignment sets considered.

Note also from table 2, how the lowest translation qualities are obtained for the Chinese–Spanish language pair, and the highest qualities are obtained for the English–Spanish language pair. Moreover, if we take a closer look at the table, these results suggest that having Spanish as the target language seems to add a significant degree of complexity to the translation task, and the most suspicious element for explaining this behavior is, for sure, its high morphological variations.

4 Effects of Spanish morphology

Previous works have shown how morphological information can be used to improve statistical machine translation results, especially when a limited amount of training data is available (Nießen and Ney, 2004; Popovic and Ney, 2004). In this section we explore the effects of reducing the Spanish morphology on the Chinese–Spanish translation tasks. For all experiments presented here, phrases extracted from the intersection set of alignments were used, and the four different Spanish data sets described in section 2 were considered.

Table 3 presents BLEU scores for both Chinese to Spanish and Spanish to Chinese translation tasks when using the four different Spanish data sets.

From table 3 it can be seen that reducing Spanish morphology by using lemmas instead of full

[4] Available at http://www.fjoch.com/GIZA++.html
[5] Available at http://www.speech.sri.com/projects/srilm/
[6] Available at http://www.isi.edu/publications/licensed-sw/pharaoh/
[7] Note that in these experiments only one translation reference is available for computing BLEU scores, in both the optimization procedure and the evaluation procedure.

forms definitively improves the translation system performance; and, as it would be logically expected, the greater impact occurs when Spanish is the target language. In this case an absolute improvement of more than four BLEU points was achieved. In this sense, note from tables 2 and 3 that, although not strictly comparable, translation quality achieved for the Chinese to lemmatized-Spanish task seems to be similar to the quality achieved for the Chinese to English translation task. On the other hand, for the Spanish to Chinese translation task, the improvement obtained by reducing the Spanish morphology was only a little bit more than a half BLEU point.

Spanish set	ES to ZH	ZH to ES
Baseline	17.4	14.3
Lowercased	17.3	16.1
Re-tokenized	17.6	15.5
Lemmatized	17.9	18.9

Table 3: Translation BLEU over the test set for Chinese–Spanish tasks and the four Spanish sets.

It can also be observed from table 3 that the effects of lowercasing and the re-tokenization generated by the analyzer seem to have opposite effects in both translation tasks. While lowercasing helps the Chinese to Spanish task, this is not the case for opposite direction; and re-tokenization seems to be producing an opposite effect.

Additionally, the problem of Spanish morphology generation was also approached as a translation task. In this sense, a translation system was implemented by using the lemmatized Spanish data set as source language and the original Spanish data set as the target language. By training and optimizing such a system, a BLEU score of *67.4* was measured over the corresponding test set, which happens to be a very high BLEU score due to the fact that a single translation reference was used. Then, by cascading the two systems: Chinese to lemmatized-Spanish and the lemmatized-Spanish to full-Spanish, a BLEU score of *14.3* was measured over the test set. Note that this result is basically the same as the one reported in table 2 for the direct Chinese to full-Spanish translation system. At a first glance, this seems to suggest that translating from Chinese into a lemmatized version of Spanish and the subsequent generation of the Spanish final forms are independent components of the Chinese to Spanish translation task, because training and optimizing a direct system provides exactly the same translation accuracy that training and optimizing both components separately.[8]

In order to explore in more detail the possible independence of the lemma translation and the final form generation processes, we decided to perform a simultaneous optimization of both systems in the cascade. In this sense we optimized the model weights of both components in the cascade (Chinese to lemmatized-Spanish and lemmatized-Spanish to full-Spanish) with respect to the BLEU score of the overall output of the cascade. In the case both components were indeed independent, we would expect exactly the same translation accuracy that was obtained when optimizing each component independently from the other. But this was not the case because a small, but statistically significant, improvement of more than a half BLEU point was achieved when performing the simultaneous optimization (a score of 14.9 was measured over the test set). This reveals that some interactions exist among the models in both components of the cascade system. Further study is necessary in order to better understand such interactions.

5 Conclusions and future research

This paper presented some statistical machine translation results among English, Spanish and Chinese, focusing on the exploration of Spanish-morphology effects on Chinese to Spanish translation tasks. In this sense, the reduction of Spanish morphology produced an absolute improvement of more than four BLEU points in the Chinese to Spanish direction; and only produced an improvement of a half BLEU point for the opposite translation direction. Although not strictly comparable, it was also observed that the accuracy achieved in the Chinese to Spanish translation task becomes comparable to the one achieved in the Chinese to English task when Spanish morphology is dropped.

Further experimentation on approaching the problem of generating Spanish morphology as a

[8] Another interesting observation is the fact that the cascade system is actually behaving in an analog manner to a series connection of two conductances: the cascade connection will perform poorer than the poorer of the two components. As an interesting fact, the reader can verify that the series combination of BLEUs holds approximately: *67.4 x18.9 / (67.4+18.9) = 14.7 ≈ 14.3.*

translation task by itself was also performed, and a small improvement over the direct Chinese to Spanish task was achieved by jointly optimizing a cascade system of two SMT components: the first one dealing with the problem of Chinese to lemmatized-Spanish translation, and the second one dealing with Spanish-morphology generation.

According to this, further research in Chinese–Spanish SMT must consider as important issues the design and evaluation of strategies for handling Spanish morphology in the particular case of Chinese to Spanish translation tasks. In this sense, better understanding of model interactions and their implications in the translation task should be performed. We will continue exploring new strategies in the direction presented in section 4. Additionally, alternative means for Spanish morphology generation which are independent from the translation task should be considered and studied.

Nevertheless, the actual drawback of the Chinese–Spanish translation task is the lack of a parallel corpus large enough for training a state-of-the-art SMT system. Most of the problems identified in this work, which are related to the richness of Spanish morphology, can be counteracted by means of a larger data set. In this sense, the development of bilingual Chinese–Spanish resources is also another important issue to deal with. In order to pursue research in this direction, the development of Chinese–Spanish translation models by combining translation models that involve intermediate languages should be explored (Wu & Wang, 2007; Cohn & Lapata, 2007). Additionally, methods for extracting parallel corpus from comparable corpora could also be an option for the automatic generation of parallel data sets for SMT purposes (Munteanu & Marcu, 2005).

In the next future, we intend to explore in more detail some of these options in the specific context of Chinese–Spanish statistical machine translation tasks.

Acknowledgements

This work has been made possible by a grant from AGAUR (*Agència de Gestió d'Ajuts Universitaris i de Recerca*) of the Catalonian Department of Innovation, Universities and Enterprises; and by the collaboration of the Human Language Technology Department of the Institute for Infocomm Research in Singapore.

References

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. Sym. on Chinese Spoken Language Processing*.

X. Carreras, I. Chao, L. Padró, M. Padró, 2002, "FreeLing: an open-source suite of language analyzers", in *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation*.

T. Cohn, M. Lapata, 2007, "Machine translation by triangulation: making effective use of multi-parallel corpora", in *Proc. of the 45th Ann. Meeting of the Association of Computational Linguistics*, pp. 728-735.

P. Koehn, 2004, "Pharaoh: a beam search decoder for phrase-based statistical machine translation alignment models", in *Proc. of AMTA*.

E. Matusov, R. Zens, H. Ney, 2004, "Symmetric word alignments for statistical machine translation", in *Proc. of the 20th Int. Conf. on Computational Linguistics*.

D.S. Munteanu, D. Marcu, 2005, "Improving machine translation performance by exploiting non-parallel corpora", *Computational Linguistics*, vol 31, no 4, pp. 477-504.

S. Nießen, H. Ney, 2004, "Statistical machine translation with scarce resources using morpho-syntactic information", *Computational Linguistics*, vol 30, no 2, pp. 181-204.

F.J. Och, 2003, "Minimum error rate training in statistical machine translation", in *Proc. of ACL*.

F.J. Och, H. Ney, 2003, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, vol 29, no 1, pp. 19-51.

M. Popovic, H. Ney, 2004, "Towards the use of word stems and suffixes for statistical machine translation", in *Proc. of the 4th Int. Conf. on Language Resources and Evaluation*, pp. 1585-1588.

A. Stolcke, 2002, "SRILM - an extensible language modeling toolkit", in *Proceedings of the International Conference on Spoken Language Processing*.

H. Wu, H. Wang, 2007, "Pivot language approach for phrase-based statistical machine translation", in *Proc. of the 45th Ann. Meeting of the Association of Computational Linguistics*, pp. 856-863.

H. Zhang, H. Yu, D. Xiong, Q. Liu, 2003, "HHMM-based Chinese lexical analyzer ICTCLAS", in *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing*, pp. 184-187.