Evaluating Indirect Strategies for Chinese–Spanish Statistical Machine Translation: Extended Abstract *

Marta R. Costa-jussà†, Carlos A. Henríquez‡ and Rafael E. Banchs†
†Institute for Infocomm Research, Singapore 138632
{vismrc,rembanchs}@i2r.a-star.edu.sg
‡Universitat Politècnica de Catalunya, 08034 Barcelona
carlos.henriquez@upc.edu

Abstract

Although, Chinese and Spanish are two of the most spoken languages in the world, not much research has been done in machine translation for this language pair. This paper focuses on investigating the state-of-the-art of Chinese-to-Spanish statistical machine translation (SMT), which nowadays is one of the most popular approaches to machine translation. We conduct experimental work with the largest of these three corpora to explore alternative SMT strategies by means of using a pivot language. Three alternatives are considered for pivoting: cascading, pseudo-corpus and triangulation. As pivot language, we use either English, Arabic or French. Results show that, for a phrase-based SMT system, English is the best pivot language between Chinese and Spanish. We propose a system output combination using the pivot strategies which is capable of outperforming the direct translation strategy. The main objective of this work is motivating and involving the research community to work in this important pair of languages given their demographic impact.

1 Introduction

Chinese and Spanish are very distant languages in many aspects. However, they come close together in the ranking of most spoken languages in the world [Ethnologue, 2012]. In the Web 2.0 era, in which most of the content is produced by the users, the number of native speakers is an excellent indicator of the actual relevance of machine translation between two languages. Of course, other factors such as literacy, amount of text published and strength of commercial relationships are also to be taken into account, but these factors will actually support further our idea of the strategic importance of developing machine translation technologies between Chinese and Spanish. The huge increase in volume of online contents in Chinese during the last years, as well as the steady increase of commercial relationships between Spanish speaking Latin American countries and China are just two basic examples supporting this fact. Needless to say, these languages involve many economical interests [Zapatero, 2010]. Nevertheless, these two languages seem to become far apart again when looking for bilingual resources.

We have been recently interested in gathering and collecting Chinese–Spanish bilingual resources for research and machine translation application purposes. The amount of bilingual resources that are currently available for this specific language pair is surprisingly low. Similarly, the related amount of work we have found, within the computational linguistic community, can be reduced to a very small set of references [Banchs *et al.*, 2006; Banchs and Li, 2008; Bertoldi *et al.*, 2008; Wang *et al.*, 2008]. Apart from the BTEC[1] corpus available through International Workshop on Spoken Language Translation (IWSLT) competition [Bertoldi *et al.*, 2008] and *Holy Bible* datasets [Banchs and Li, 2008], we were not aware of any other Chinese–Spanish parallel corpus suitable for training phrase-based [Koehn *et al.*, 2003][2] statistical machine translation systems between these two languages, until a six-language parallel corpus from United Nations was released for research purposes [Rafalovitch and Dale, 2009].

Using the recently released United Nations parallel corpus as a starting point, this work focuses on the problem of developing Chinese-to-Spanish phrase-based machine translation technologies with a limited set of bilingual resources. We explore and evaluate different alternatives for the problem in hand by means of pivot-language strategies through other languages available in the United Nations parallel corpus, such as Arabic, English and French [3]. Existing strategies such as system cascading, pseudo-corpus generation and triangulation are implemented and compared against a baseline system built with a direct translation approach. As follows, we briefly describe these pivot approaches:

*This paper is an extended abstract of the JAIR publication [Costa-jussà et al., 2012]

[1] Basic Traveller Expressions Corpus.

[2] Note that *phrase-based* is commonly used to refer to statistical machine translation systems, in which the term *phrase* refers to segments of one or more than one word and it does not have the usual meaning of *multi-word syntactical consitutent,* as it has in linguistics. as it has in linguistics.

[3] Although Russian is available in the UN corpus, we discard to use it because we do not have the proper preprocessing tools for it.

- The cascaded approach generates Chinese-to-Spanish translations by concatenating a system that translates Chinese into a pivot language with a system that translates from the pivot language into Spanish.
- The pseudo-corpus approach builds a synthetic Chinese–Spanish corpus either by translating into Spanish the pivot side of a Chinese–pivot corpus or by translating into Chinese the pivot side of a Pivot–Spanish corpus.
- The triangulation approach implements a Chinese-to-Spanish translation system by combining the translation table probabilities of a Chinese–pivot system and a Pivot–Spanish system.

Additionally, we implement and evaluate a system combination of the three pivot strategies based on the minimum Bayes risk (MBR) [Kumar and Byrne, 2004] technique. Such a combination strategy is capable of outperforming the direct system.

2 Direct and Pivot Statistical Machine Translation Approaches

There are several strategies that we can follow when translating a pair of languages in statistical machine translation (SMT). In this section we present the details of the ones we are using in this work.

2.1 Direct System

Our direct system uses the phrase-based translation approach [Koehn *et al.*, 2003]. The basic idea is to segment the given source sentence s into segments of one or more words, then each source segment is translated using a bilingual phrase obtained from the training corpus and finally compose the target sentence from these phrase translations. A bilingual phrase is a pair of m source words and n target words extracted from a parallel sentence that belongs to a bilingual corpus previously aligned by words. For extraction, we consider the words that are consecutive in both source and target sides and which are consistent with the word alignment. We consider a phrase is consistent with the word alignment if no word inside the phrase is aligned with one word outside the phrase.

2.2 Pivot-Based System

The cascaded approach handles the source–pivot and the pivot–target system independently. They are both built and tuned to improve their local translation quality and then composed to translate from the source language to the target language in two steps: first, the translation output from source to pivot is computed and then it is used to obtain the target translation output.

The pseudo-corpus approach translates the pivot section of the source–pivot parallel corpus to the target language using a pivot–target system built previously. Then, a source–target SMT system is built using the source side and the translated pivot side of the source–pivot corpus. The pseudo-corpus system is tuned using an original source–target development corpus, since we have it available.

The triangulation approach combines the source–pivot ($P(s|p)$ and $P(p|s)$) and pivot–target ($P(p|t)$ and $P(t|p)$) relative frequencies following the strategy proposed by Cohn & Lapata 2007 in order to build a source–target translation model. The translation probabilities are computed assuming the independence between the source and target phrases when given the pivot phrase.

$$P(s|t) = \sum_p P(s|p)P(p|t) \quad (1)$$

$$P(t|s) = \sum_p P(t|p)P(p|s) \quad (2)$$

where s, t, and p represent phrases in the source, target and pivot language respectively.

The lexical weights are computed in a similar manner 2007. This approach does not handle the lexicalized reordering and the other pivot strategies and therefore represents a limitation in its potential. Instead, a simple distance-based reordering is applied during decoding. This model gives a cost linear to the reordering distance. For instance, skipping over two words costs twice as much as skipping over one word.

Once the corresponding translation model have been obtained, the source–target system is tuned using the same original source–target development corpus mentioned in the previous approach.

3 Evaluation framework

The following section introduces the details of the evaluation framework. We prepared the training, development and test set from UN corpus as described in [Costa-jussà *et al.*, 2012]. The training set contains around 58 thousand sentences and the development and test sets contain one thousand sentences each.

We built and compared several translation approaches in order to study the impact of the different pivot languages when translating from Chinese into Spanish. Moreover, we evaluated how the quality of pivot approaches differs from a direct system. We built the pivot systems using five of the languages available in the UN parallel corpus: English, Spanish, Chinese, Arabic and French, and we built the direct system on a Chinese–Spanish parallel corpus.

3.1 Pivot results

Table 1 shows the results for our Chinese-to-Spanish configurations with the UN corpus. We can see there that the best pivot system used the pseudo-corpus approach with English as the pivot language.

In order to observe the benefits of the pivot language against the direct translation, table 2 presents three examples where the BLEU scores of the pivot approach were better than those of the direct approach. Notice how some phrases that disappeared from the direct translation correctly appear on the pseudo-corpus approach.

3.2 Pivot Combination

Using the 1-best translation output from the different pivot strategies, we built an n-best list and computed the final trans-

Pivot	System	BLEU	Pivot vocab.
–	direct	33.06	-
En	cascaded	32.90	14k
Fr	cascaded	30.37	18k
Ar	cascaded	28.88	17k
En	pseudo	32.97	14k
Fr	pseudo	32.61	18k
Ar	pseudo	32.23	17k
En	triangulation	32.05	14k
Fr	triangulation	30.41	18k
Ar	triangulation	30.61	17k

Table 1: Chinese-to-Spanish cascaded, pseudo-corpus and triangulation approaches.

DIRECT	cuestiones como a que consideren seriamente la posibilidad de ratificar la tortura y otros tratos o penas crueles , inhumanos
PSEUDO	como cuestiones a que consideren seriamente la posibilidad de ratificar **la convención contra** la tortura y otros tratos o penas crueles , inhumanos
REF	considere seriamente la posibilidad de ratificar , con carácter prioritario , **la convención contra** la tortura y otros tratos o penas crueles , inhumanos
EN REF	to seriously consider ratifying , as a matter of priority , **the convention against** torture and other cruel , inhuman treatment or punishment
DIRECT	habiendo examinado el segundo informe de la **comisión** y la recomendación
PSEUDO	habiendo examinado el segundo informe de la **comisión de verificación de poderes** y las recomendaciones
REF	habiendo examinado el segundo informe de la **comisión de verificación de poderes** y la recomendación
EN REF	having considered the second report of the **credentials committee** and the recommendation
DIRECT	pide al secretario general que prepare un informe sobre la aplicación de esta resolución **a la asamblea general**
PSEUDO	pide al secretario general que prepare un informe sobre la aplicación de la presente resolución **para su examen por la asamblea general**
REF	pide al secretario general que prepare un informe sobre la aplicación de la presente resolución , **que será examinado por la asamblea general**
EN REF	requests the secretary-general to prepare a report on the implementation of the present resolution **for consideration by the general assembly**

Table 2: Chinese-to-Spanish examples for which the pseudo-corpus system (through English) is better than the direct system. EN REF is the English reference of the sentence

lation using minimum Bayes risk (MBR) [Kumar and Byrne, 2004]. Table 3 combines all the outputs from table 1.

	Cascaded	Pseudo	Triangulation	All
A	32.66*	33.30*	31.84	33.97*
D+A	33.60*	33.77*	32.90	34.09*

Table 3: Chinese-to-Spanish percent BLEU score for system combinations of En + Fr + Ar languages (A), direct system (D) and pivot approaches using MBR. (*) statistically significant better BLEU than the direct system.

4 Conclusions

This work provided a brief survey in the state-of-the-art of Chinese–Spanish SMT. First of all, this language pair is of great interest both economically and culturally if we take into account the high number of Chinese and Spanish speakers.

Besides, statistical machine translation is the most popular approach in the field of MT given that has shown great quality in all the international evaluation campaigns such as NIST 2009 and WMT 2012.

The main points covered in our study were:

- English is the best pivot language for conducting Chinese-to-Spanish translations compared to languages such as French or Arabic. The system built using English as pivot was significantly better than the ones built with French or Arabic, with a 99% confidence in both comparisons.
- No significant difference is found among the best cascaded and pseudo-corpus pivot approaches, but the pseudo-corpus strategy is the best pivot strategy for Chinese-to-Spanish. Additionally, pseudo-corpus and cascaded approaches are significantly better than the triangulation approach.
- The output combination using MBR is able to improve the direct system in 1 BLEU point in the best case. This improvement is significantly better with a 99% confidence and is coherent with improvements in all other evaluation metrics studied.

Further experiments, descriptions and conclusions can be found at [Costa-jussà *et al.*, 2012].

Acknowledgments

The authors would like to specially thank Umberto Grandi for his invitation to participate in the IJCAI journal track. Additionally, the authors would like to thank the Universitat Politècnica de Catalunya and the Institute for Infocomm Research for their support and permission to publish this research.

This work has been partially funded by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951); and by the Spanish Ministry of Economy and Competitiveness through the FPI Scholarship BES-2008-003851 for Ph.D. students under the AVIVAVOZ project (TEC2006-13694-C03-01); and the BUCEADOR project (TEC2009-14094-C04-01).

References

[Banchs and Li, 2008] R. E. Banchs and H. Li. Exploring Spanish Morphology effects on Chinese-Spanish SMT. In *MATMT 2008: Mixing Approaches to Machine Translation*, pages 49–53, Donostia-San Sebastian, Spain, February 2008.

[Banchs *et al.*, 2006] R. E. Banchs, J. M. Crego, P. Lambert, and J. B. Mariño. A Feasibility Study For Chinese-Spanish Statistical Machine Translation. In *Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)CONLL*, pages 681–692, Kent Ridge, Singapore, December 13–16 2006.

[Bertoldi *et al.*, 2008] N. Bertoldi, R. Cattoni, M. Federico, and M. Barbaiani. FBK @ IWSLT-2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 34–38, Hawaii, USA, 2008.

[Callison-Burch *et al.*, 2012] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of

the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012.

[Cohn and Lapata, 2007] T. Cohn and M. Lapata. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proc. of the ACL*, 2007.

[Costa-jussà *et al.*, 2012] M. R. Costa-jussà, C. A. Henríquez Q, and R. E. Banchs. Evaluating indirect strategies for chinese-spanish statistical machine translation. *J. Artif. Int. Res.*, 45(1):761–780, September 2012.

[Ethnologue, 2012] Ethnologue. Ranking of most spoken languages, 2012. [Online; accessed 12-December-2012].

[Koehn *et al.*, 2003] P. Koehn, F.J. Och, and D. Marcu. Statistical Phrase-Based Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, 2003.

[Kumar and Byrne, 2004] S. Kumar and W. Byrne. Minimum Bayes-Risk Decoding For Statistical Machine Translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL'04)*, pages 169–176, Boston, USA, May 2004.

[Nist, 2009] Nist. NIST machine translation evaluation campaign, 2009. [Online; accessed 12-December-2012].

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

[Wang *et al.*, 2008] H. Wang, H. Wu, X. Hu, Z. Liu, J. Li, D. Ren, and Z. Niu. The TCH Machine Translation System for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 124–131, Hawaii, USA, 2008.

[Zapatero, 2010] J. R. Zapatero. China is a top priority for the spanish economy; our companies are well aware of that, 2010. [Online; accessed 12-December-2012].