Enhancing scarce-resource language translation through pivot combinations

Marta R. Costa-jussà
Barcelona Media Innovation Center
Av. Diagonal, 177
08018 Barcelona
marta.ruiz@barcelonamedia.org

Carlos Henríquez
TALP-UPC
Jordi Girona, s/n
08034 Barcelona
carlos.henriquez@upc.edu

Rafael E. Banchs
Institute for Infocomm Research
1 Fusionopolis Way 21-01
Singapore 138632
rembanchs@i2r.a-star.sg.edu

Abstract

Chinese and Spanish are the most spoken languages in the world. However, there is not much research done in machine translation for this language pair. We experiment with the parallel Chinese-Spanish corpus (United Nations) to explore alternatives of SMT strategies which consist on using a pivot language. Particularly, two well-known alternatives are shown for pivoting: the cascade system and the pseudo-corpus. As Pivot language we use English, Arabic and French. Results show that English is the best pivot language between Chinese and Spanish. As a new strategy, we propose to perform a combination of the pivot strategies which is capable to highly outperform the direct translation strategy.

1 Introduction

Although they are very distant languages, Chinese and Spanish are very close to each other in the ranking of the most spoken languages in the world[1]. Nevertheless, when interested in bilingual resources between these two languages they become far apart again. Similarly, the related amount of work we have found within the computational linguistic community, can be reduced to a very small set of references.The most popular research event recently performed was the 2008 IWSLT evaluation campaign [2]. This evaluation organized two Chinese-to-Spanish tracks. One of them was focused on direct translation and the other one on pivot translation through English. Best translation results were obtained by far in the pivot task. The best system in the pivot task (Wang et al., 2008) compared two different approaches: The first one, training two translation models on the Chinese-English corpus and English-Spanish corpus, and then building a pivot translation model for Chinese-Spanish translation using English as a pivot language as proposed in (Wu and Wang, 2007); the second one obtained better results and it was based on a cascade approach. The idea here is to translate from Chinese into English and then from English to Spanish, which means performing two translations. Besides the research mentioned above, which directly addressed the Chinese-Spanish language pair, we may also find in the literature another approach similar to Wu's (2007) authored by Cohn and Lapata (2007). Basically, they also used several intermediate pivot language to create source-to-target phrases that are lately interpolate with a direct system build with a source-to-target parallel corpus.

Apart from the BTEC[3] corpus available through the IWSLT[4] competition and *Holy Bible* datasets described in (Paul, 2008) and (Banchs and Li, 2008), respectively, there is a recent release of a six language parallel corpus (including both Chinese and Spanish) from United Nations (UN) for research purposes (Rafalovith and Dale, 2009). Using the recently released UN parallel corpus as a starting point, this work focuses on the problem of developing Chinese-Spanish phrase-based SMT technologies with a limited set of bilingual resources. We explore and evaluate different alternatives for the

[1] www.ethnologue.org/ethno_docs/distribution.asp?by=size
[2] http://mastarpj.nict.go.jp/IWSLT2008/
[3] Basic Traveller Expressions Corpus
[4] International Workshop on Spoken Language Translation

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1361–1365,
Chiang Mai, Thailand, November 8 – 13, 2011.

problem in hand by means of pivot-language strategies through the other languages available in the UN parallel corpus, such as Arabic, English and French. More specifically, strategies such as system cascading and pseudo-corpus generation are implemented and compared against a baseline system implementing a direct translation approach. We propose a system combination different from previous ones (Wu and Wang, 2009) and based on the Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004) technique using both pivot strategies which is capable to highly outperform the direct system. To the best of our knowledge, this idea was not explored before and it is a way of increasing the quality of translation between languages with scarce bilingual resources. In addition, we are performing a combination of the same system but introducing new information through the pivot language.

The paper is structured as follows. Section 2 describes the main strategies for performing Chinese-to-Spanish translation which are tested in this work. Section 3 presents the evaluation framework. Then, section 4 reports the experiments (including the system combination) and the results. Finally, section 5 concludes and proposes new research directions.

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. The next three sub-sections present the details of the ones we are using in this work.

2.1 Direct system

Our direct system uses the phrase-based translation system (Koehn et al., 2003). This popular system implements a log-linear model in which a source language sentence $f^J = f_1, f_2, \ldots, f_J$ is translated into another language (target) sentence $e^I = e_1, e_2, \ldots, e_I$ by searching for the translation hypothesis $\hat{e}^I$ maximizing a log-linear combination of several feature models (Och, 2003).

The main system models are the translation model and the language model. The first one deals with the issue of which target language phrase f_j translates a source language phrase e_i and the latter model estimates the probability of translation hypothesis. Apart from these two models, there are other standard models such as the lexical models, the word bonus, and the reordering model.

For decoding, we used the MOSES toolkit (Koehn et al., 2007) with the option of Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004) decoding. Therefore the 1best translation obtained is not the one with highest priority but the one that is most similar to the most likely translation. The option was activated with its default parameters so it considered the top 200 distinct hypothesis to compute the 1best.

2.2 Cascade System

This 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 joined to translate from the source language to the target language in two steps: first, the 1best translation output from source to pivot is computed and, second, it is used to obtain the 1best target translation output as the final translation.

There is an alternative approach that considers the nbest list in each step instead of the 1best. For instance, it was used in (Khalilov et al., 2008) with their cascade approach in order to obtain the best Chinese-Spanish translation. We also implemented it but the results were similar that those using MBR decoding in each system and keeping the 1best translation. Therefore we maintained MBR decoding for the rest of the experiments, which is also easier to work with.

2.3 Pseudo-Corpus System

This 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 a direct source-target development corpus.

2.4 Pivot combination

Using the 1-best translation output from the different pivot strategies, we built an N-best list and computed the final translation using MBR. MBR has been used both during decoding (Kumar and Byrne, 2004; Ehling et al., 2007) and as a postprocess over an N-best list. The current version of the MOSES

toolkit includes both MBR implementations. For the system combinations we used the second one.

The MBR algorithm implemented in MOSES uses $(1 - BLEU)\beta$ as the Loss Function. The value β weights the hypothesis proportionally to its translation score, but we considered all our hypothesis as equal so β was a constant and therefore could be discarded. At the end, MBR choose the hypotheses E' that fulfills:

$$E' = \arg\min_{\hat{E}'} \left(\sum_{E \neq \hat{E}'} 1 - BLEU(E, \hat{E}') \right) \quad (1)$$

It is important to mention that all N-best list must have at least 3 hypothesis per sentence. Having only two hypothesis would not work as expected because the Loss Function would always choose the longest one, which can be explained by the definition of BLEU:

$$BLEU(E, E') = \exp\left(\sum_{n=1}^{N} \log \frac{p_n(E, E')}{N} \right) * \gamma(E, E') \quad (2)$$

where $p_n(E, E')$ is the precision of n-grams in the hypothesis E' with reference E; and $\gamma(E, E')$ is a brevity penalty if the hypothesis E' is shorter than the reference E. Then $p_n(E, E') = p_n(E', E)$ and

$$\forall E, E' : length(E) > length(E') :$$
$$1 - BLEU(E, E') \geq 1 - BLEU(E', E) \quad (3)$$

3 Evaluation Framework

This section introduces the details of the evaluation framework used. We report the UN corpus statistics, a description of how we built the systems and the evaluation details.

3.1 Corpus statistics

In this study we use the UN corpus taking advantage of the fact that (as far as we are concerned) it is the biggest parallel corpus freely-available in Chinese-Spanish and it contains the same sentences in six other languages, therefore we can experiment with different pivot languages.

When experimenting with different pivot languages, in order to make the systems as comparable as possible, we first did a sentence selection over the corpus so all systems were built exactly with the same training, tuning and testing sets. All corpora were tokenized, using the standard MOSES tokenizer for Spanish, English and French; ictclass (Zhang et al., 2003) for Chinese; and MADA+TOKAN (Habash and Rambow, 2005) for Arabic. The Spanish, English and French corpora were lowercased. If a sentence had more than 100 words in any language, it was deleted from all corpora. If a sentence pair had a word ratio bigger than three for any Chinese-Pivot or Pivot-Spanish parallel corpora, it was deleted from all corpora. For all languages, we identify all sentences that occur only once in the corpora. The tuning and testing sets where drawn from the available multilingual corpus by using a maximum perplexity and lowest out-of-vocabulary word criterion over the English part of the dataset. In order to do this, perplexity was computed on a sentence-by-sentence basis by using a leave-one-out strategy; then, we selected the two thousand sentences which had the highest perplexity and the lowest out-of-vocabulary words for constructing the tuning and testing sets. Table 1 shows the main statistics for all corpora.

	training		development		test	
	s	w	s	w	s	w
Zh	58.6k	1.6M	1k	30.9k	1k	32.6k
Es	58.6k	2.3M	1k	42.2k	1k	44.0k
En	58.6k	2.0M	1k	36.7k	1k	38.3k
Ar	58.6k	2.6M	1k	47.9k	1k	49.9k
Fr	58.6k	2.3M	1k	42.1k	1k	43.9k

Table 1: UN Corpus Statistics (s stands for number of sentences and w for number of words)

3.2 System details

Our systems were build using MOSES. For all systems, we used the default MOSES parameters, which includes the grow-final-diagonal alignment symmetrization, the lexicalized reordering, a 5-gram language model using Kneser-Ney smoothing and phrases up to length 10. The optimization was done using MERT (Och, 2003). The decoding was done using MBR.

4 Chinese-to-Spanish MT strategies

Given the different languages available in the UN corpora, we tested three different pivot languages. Additionally, we compared the cascade and the pseudo-corpus pivot strategies. Finally, we combined the system outputs.

4.1 Experimenting with different pivot languages

Using most of the languages available in the UN parallel corpora (English, Spanish, Chinese, Arabic and French) we built and compare several translation systems in order to study the impact of the different pivot languages when translating from Chinese to Spanish.

Specifically, we built seven Chinese-Spanish systems: the direct Chinese-Spanish system as a quality upper bound; three cascade approach and three pseudo-corpus, using English, Arabic and French as pivots.

Therefore, the first step was to build the different Chinese-Pivot and Pivot-Spanish systems. Table 2 shows the BLEU achieved with the intermediate systems trained with the UN Corpus. These systems are used in the next section when experimenting with different pivot languages.

	BLEU
Chinese-English	35.67
Chinese-Arabic	**46.11**
Chinese-French	28.31
English-Spanish	**51.64**
Arabic-Spanish	41.79
French-Spanish	46.42

Table 2: UN Pivot Systems

As we can see in Table 2 the best Chinese-Pivot system is the Chinese-Arabic system. As for the Pivot-Spanish system, the one that achieved the best BLEU score was the English-Spanish system.

Tables 3 shows the results for our Chinese-Spanish configurations with the UN corpus. We can see there that the best pivot system used the cascade approach with English as the pivot language.

The fact that the pseudo-corpus through English outperforms cascade through English is not statistically significant, with a 95% confidence

Languages	System	BLEU
Chinese-Spanish	direct	**33.06**
Chinese-English-Spanish	cascade	**32.90**
Chinese-French-Spanish	cascade	30.37
Chinese-Arabic-Spanish	cascade	28.88
Chinese-English-Spanish	pseudo	**32.97**
Chinese-French-Spanish	pseudo	32.61
Chinese-Arabic-Spanish	pseudo	32.23

Table 3: UN pivot languages. Best results in bold.

(Koehn, 2004). These results, however, are coherent with previous works using the same language pair (Bertoldi et al., 2008; Henríquez Q. et al., 2010) that also reported the pseudo-corpus strategy was better than the cascade strategy.

In all cases English is statistically significant the best pivot language, with a 99% confidence, which is coherent with the Pivot-Spanish results in table 2. Further analysis is required in order to understand why the cascade through English is able to help so much in the Chinese-to-Spanish translation.

4.2 Pivot combination

Table 4 shows the results of the different output systems combined (from table 3) with the MBR technique. $En + Ar + Fr\ Cascade + Pseudo$ (which combines all system outputs from table 3 except the direct approach) is better than the Chinese-to-Spanish direct system and it is significant with a 99% of confidence. When adding the direct approach (dir) it increases the translation performance slightly and we obtain the best Chinese-to-Spanish translation.

	casc	pseudo	casc+pseudo
En+Ar+Fr	32.66	33.30*	**33.97***
dir+En+Ar+Fr	33.60*	33.77*	**34.09***

Table 4: Output system combination using MBR. * shows statistically significantly better results than the direct system (with a 99% of confidence). Best results in bold. Casc stands for cascade.

5 Conclusions

This work has presented experimental research for the Chinese-Spanish translation pair. The main con-

clusions derived from our study are:

- English is the best Pivot language for Chinese-to-Spanish compared to languages such as French or Arabic. The system built using English as Pivot was significantly better than the ones built with either French or Arabic, with a 99% confidence in both cases.
- There is not a significant difference among the best cascade and pseudo-corpus pivot approaches.
- 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.

6 Acknowledgment

The authors would like to thank Barcelona Media Innovation Center and Institute for Infocomm Research for its support and permission to publish this research. This work has been partially funded by the Spanish Department of Science and Innovation through the *Juan de la Cierva* fellowship program.

References

R. Banchs and H. Li. 2008. Exploring spanish morphology effects on chinese-spanish smt. In *MATMT 2008: Mixing Approaches to Machine Translation*, pages 49–53, Donostia-San Sebastian, Spain, February.

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

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

N. Ehling, R. Zens, and H. Ney. 2007. Minimum bayes risk decoding for bleu. In *Proc. of the ACL*, pages 101–104, Prague, Czech Republic, June. Association for Computational Linguistics.

N. Habash and O. Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 573–580, Ann Arbor, MI, June.

C. A. Henríquez Q., R. E. Banchs, and J. B. Mariño. 2010. Learning reordering models for statistical machine translation with a pivot language. Internal Report TALP-UPC.

M. Khalilov, M. R. Costa-Jussà, C. A. Henríquez, J. A. R. Fonollosa, A. Hernández, J. B. Mariño, R. E. Banchs, B. Chen, M. Zhang, A. Aw, and H. Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 116–123, Hawaii, USA.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 48–54.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the ACL*, pages 177–180, Prague, Czech Republic.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395.

S. Kumar and W. Byrne. 2004. 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.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

M. Paul. 2008. Overview of the IWSLT 2008 Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–17, Hawaii, USA.

A. Rafalovith 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.

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

H. Wu and H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proc. of the ACL*, pages 856–863, Prague.

H. Wu and H. Wang. 2009. Revisiting Pivot Language Approach for Machine Translation. . In *Proc. of the ACL-IJCNLP*, pages 154–162, Singapore.

H. Zhang, H. Yu, D. Xiong, and Q. Liu. 2003. HHMM-based chinese lexical analyzer ICTCLAS. In *Proc. of the 2nd SIGHAN Workshop on Chinese language processing*, pages 184–187, Sapporo, Japan, July.