Learning Reordering Models for Statistical Machine Translation with a Pivot Language

Carlos A. Henríquez Q.[*] **Rafael E. Banchs**[†] **José B. Mariño**[*]

[*]TALP Research Centre, Universitat Politècnica de Catalunya, Barcelona
carlos.henriquez@upc.edu
jose.marino@upc.edu
[†]Barcelona Media Innovation Centre, Barcelona
rafael.banchs@barcelonamedia.org

Abstract

This paper presents a work related to reordering when dealing with translation using pivot languages. Different pivot strategies are presented in order to compare their translation quality on a Chinese-Spanish task. A novel method to generate reordering weights automatically for a language pair that do not share parallel corpus is presented. Experiments which show that the strategy outperforms the cascade approach for pivot translation are reviewed.

1 Introduction

Previous works have already address the problem of translating from a source language to a target language using a third language as pivot. This has given us the oportunity to build translation models for language pairs which do not have parallel corpus for building them directly.

For this matter, different approaches have been proposed like having a system for each pair and perform a cascade translation, translating from the pivot language to the target language and then building a system with the translated corpus as target side (Banchs et al., 2006) and multiplying translation models to generate and artificial model (Wu and Wang, 2007).

More recently (Bertoldi et al., 2008) made a comparision between those approaches and concluded that the second one is the one that provides better translation quality among the three.

Here we worked with a trilingual corpus extracted from the bible. A Chinese, Spanish and English version are available. We compared the cascade approach, the synthetized approach, and the table combination approach with a variation that allowed it to learn an artificial reordering for the pair Chinese-Spanish. The idea behind this variation is to extend the methodology in (Wu and Wang, 2007) over the reordering weights of both Source-Pivot and Pivot-Target SMT system, in order to obtain new reordering weights for the Source-Target language pair.

The document is organized as follows: section 2 describes the phrased based translation system and the tools that were used during the experiments. Section 3 describes the corpus used in all the experiments. Section 4 describes the different approaches that were considered. Section 5 explains the process followed to learn reordering weights using a pivot language. Section 6 explains the experiments made and the main results obtained with the different reordering strategies. The last section gives the conclusions extracted from the results on the different experiments.

2 Phrase-based Translation Systems

The phrase-based translation system (Koehn et al., 2003) implements a log-linear model in which a foreign language sentence $f^J = f_1, f_2, \ldots, f_J$ is translated into another language 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 (Brown et al., 1990):

$$\hat{e}_I = \arg\max_{e^I} \{ \sum_{m=1}^{M} \lambda_m h_m(e^I, f^J) \} \quad (1)$$

where the feature function h_m refers to the system models and λ_m refers to the corresponding optimized model weights.

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, a reordering model based on (Tillman, 2004) is used during the process.

The development of all the systems is based on the MOSES toolkit (Koehn et al., 2007).

3 The Corpus

The corpus used on the different experiments is a trinlingual translation of the Bible. A Chinese, English and a Spanish version of it were used. The corpus was originally presented in (Banchs and Li, 2008). Main statistics for the Bible corpus can be seen in Table 1.

4 Pivot Approaches

4.1 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.

4.2 Synthesized System

The Synthesized System 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.

4.3 Generating Phrase Probabilities with Table Combination

The table combination approach is based on (Wu and Wang, 2007). To obtain the translation probabilities for each Chinese-Spanish phrase, the probabilities from the Chinese-English phrases and the English-Spanish phrases are combined. The final phrase probabilities are calculated as followed:

$$\phi(f_i|e_i) = \sum_{p_i} \phi(f_i|p_i)\phi(p_i|e_i) \qquad (2)$$

where $\phi(f_i|e_i)$ corresponds to the translation probability of the Chinese phrase f_i given the Spanish phrase e_i, $\phi(f_i|p_i)$ stands for the translation probability of the Chinese phrase f_i given the English phrase p_i and $\phi(p_i|e_i)$ stands for the translation probability of the English phrase p_i given the Spanish phrase e_i.

These two scores are supported by a Spanish language model, a word and phrase penalty feature and a distortion model. The system presented here adds reordering weights following the strategy described in section 5.

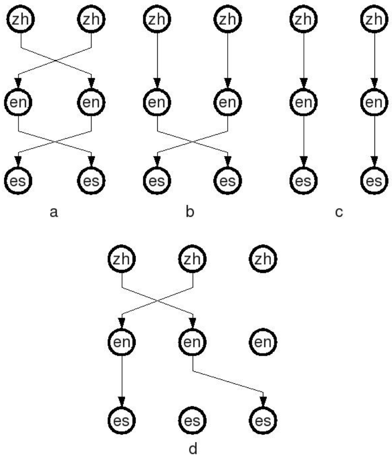

Figure 1: a) Two consecutive swap moves. b) A monotonous with the previous phrase followed by a swap with the next phrase. c) Two monotonous moves. d) A swap move followed by a discontinous move.

5 Learning Reordering for Table Combination

Motivated by equation 2 and the reordering strategy followed in MOSES, a table combination strategy was developed to generate these weights for the Chinese-Spanish system, using the models obtained for the Chinese-English and English-Spanish systems.

The reordering used in MOSES is based on (Tillman, 2004). It considers three different moves a phrase can make related to the previous and following phrase: monotonous move, swap move and discontinous move.

There are three consideration to have in mind to calculate the reordering weights for a system using a pivot language:

- A swap move on the Chinese-English system is dissolved if the same phrase is swapped again on the English-Spanish system. Therefore it is a monotonous move.

- A monotonous move followed by a swap means a swap from Chinese to Spanish. It is the same if the swap if performed first and then the monotonous move.

- A discontinous moves always generates a final discontinous move not matter which move is performed before it.

Training Corpus			
Language	Sentences	Tokens	Vocabulary
Chinese	28,887	760,451	12,670
English	28,887	848,776	13,216
Spanish	28,887	784,398	25,240
Development Corpus			
Language	Sentences	Tokens	Vocabulary
Chinese	1,033	27,235	3,404
English	1,033	30,199	3,234
Spanish	1,033	27,986	4,403
Test Corpus			
Language	Sentences	Tokens	Vocabulary
Chinese	1,035	26,794	3,396
English	1,035	30,008	3,158
Spanish	1,035	27,452	4,426

Table 1: Main Statistics from the Bible Corpus

Figure 1 shows a graphical example of the rules explained above. Following these rules, the monotonous weights for the Chinese-Spanish system can be calculated like this:

$$\begin{aligned} m(f_i|e_i) &= \sum_{p_i} m(f_i|p_i)m(p_i|e_i) + \\ &\quad \sum_{p_i} s(f_i|p_i)s(p_i|e_i) \end{aligned} \tag{3}$$

the swap weights can be calculated using this formula:

$$\begin{aligned} s(f_i|e_i) &= \sum_{p_i} m(f_i|p_i)s(p_i|e_i) + \\ &\quad \sum_{p_i} s(f_i|p_i)m(p_i|e_i) \end{aligned} \tag{4}$$

and the discontinous weights can be calculated the following way

$$\begin{aligned} d(f_i|e_i) &= \sum_{p_i} m(f_i|p_i)d(p_i|e_i) + \\ &\quad \sum_{p_i} s(f_i|p_i)d(p_i|e_i) + \\ &\quad \sum_{p_i} d(f_i|p_i)d(p_i|e_i) \end{aligned} \tag{5}$$

where f_i represents a phrase on the source language, e_i represents a phrase on the target side and p_i represents a phrase on the pivot language.

This formulas give us a way to generate reordering weights that could help to improve the translation quality comparing to the reordering that just considers movement distance.

Reordering Strategy	BLEU
Without reordering	21.80
Distance based	22.19
Lexical reordering	22.61
Lexical and Distance based	**22.78**

Table 2: BLEU obtained for the direct configuration

6 Experiments on Reordering

In order to study the effects of our reordering strategy we designed two different experiments. First, using the Chinese and Spanish version of the corpus, a Chinese-Spanish system was built with the standard configuration from MOSES' scripts. This system was built to have a ideal BLEU to aim for. In practice, this system can never be a reference to compare to because the idea is to built systems for language pairs that do not share parallel corpus. Nevertheless it is usefull to see how far we are from a BLEU obtained with a direct translation.

Appart from the standard configuration, a system without reordering and another one using only a reordering based on movement distance were developed. In this way, we can see the benefits of the reordering strategy for this task.

Table 2 shows the results for the different reordering strategies over the test set.

As we can see from this configuration the lexical reordering joined with distance based reordering outperforms the other strategies. Therefore it was the only reordering option for our follow-

Pivot Strategy	BLEU
Cascade	21.87
Table Combination (with new weights)	21.99
Synthetized	22.34

Table 3: BLEU obtained for different pivot approaches

ing experiment and it was also the reason to believed that an effort to enhanced the approach from (Wu and Wang, 2007) with generated reordering weights was worth it.

The last experiment compares the different strategies explained on section 4. It can be seen that the Table Combination approach with new reordering weights outperformed the Cascade strategy. Despite of that, the synthetized approach performs better than the other two.

7 Conclusions

Translation for laguage pairs that do not share parallel corpus can be address with different approaches. Three different strategies for translating using a third language as pivot were reviewed. The approach presented by (Wu and Wang, 2007) is a valid approximation to generate a translation model on these cases. Nevertheless an additional reordering model should be used when translating from Chinese to Spanish. We showed that reordering for this task was important and that the model presented in (Tillman, 2004) was the best reordering model from the reviewed in this work. For that reason, a new method which estimates this reordering model for a pair that do not share parallel corpus was presented. Experiments showed that this approach outperforms the cascade strategy even though the synthetized method still offered better translation quality.

References

Rafael Banchs and Haizhou 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.

Rafael E. Banchs, Josep M. Crego, Patrik Lambert, and José B. Mariño. 2006. 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.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *Proc. of the Int. Workshop on Spoken Language Translation*, Hawai,USA, October.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics.*, 16(2):79–85.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180.

Christoph Tillman. 2004. A block orientation model for statistical machine translation. In *HLT-NAACL.*

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 856–863, Prague, Czech Republic, June.