Ongoing Study for Enhancing Chinese-Spanish Translation with Morphology Strategies

Marta R. Costa-jussà[1]
Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[1]marta@nlp.cic.ipn.mx

Abstract

Chinese and Spanish have different morphology structures, which poses a big challenge for translating between this pair of languages. In this paper, we analyze several strategies to better generalize from the Chinese non-morphology-based language to the Spanish rich morphology-based language. Strategies use a first-step of Spanish morphology-based simplifications and a second-step of fullform generation. The latter can be done using a translation system or classification methods. Finally, both steps are combined either by concatenation in cascade or integration using a factored-based style. Ongoing experiments (based on the United Nations corpus) and their results are described.

1 Introduction

The structure of Chinese and Spanish differs at most linguistic levels, e.g. morphology, syntax and semantics. In this paper, we are focusing on reducing the gap between both languages at the level of morphology. On the one hand, Chinese is an isolating language, which means having a low morpheme per word ratio. On the other hand, Spanish is a fusional language, which means having a tendency to overlay many morphemes. The challenge when translating between Chinese and Spanish is bigger in the direction from Chinese to Spanish, given that the same Chinese word can generate multiple Spanish words. For example, the Chinese word *fàn* (in transcribed Pinyin) can be translated by *comer, como, comí, comeré*[1] which correspond to several tense flexions of the same verb and also by *comes, comiste, comerás*[2], all of which also correspond to several person flexions of the same verb. This poses a challenge in Statistical Machine Translation (SMT) because translations are learnt by co-ocurrence of words in both languages. When a word has multiple translations, it generates sparsity in the translation model.

In this study, we experiment with different strategies to add morphology knowledge in a standard phrase-based SMT system (Koehn et al., 2003) for the Chinese-to-Spanish translation direction. However, the presented techniques could be used for other pairs involving isolating and fusional languages. The rest of the paper is organized as follows. Section 2 reports a brief overview of the related work both in using morphology knowledge in SMT and in translating from Chinese-to-Spanish. Section 3 explains the theoretical framework of phrase-based SMT at a high level and the details of each strategy to introduce morphology in the mentioned system. Section 4 describes the experiments and first results obtained for each theoretical strategy presented. Finally, Section 5 concludes this ongoing research and outlines the future research directions.

2 Related Work

There are numerous studies which deal with morphology in the field of SMT. Without aiming at completeness, we cite works that:

- Preprocess the data to make the structure of both languages more similar by means of enriching (Avramidis and Koehn, 2008; Ueffing and Ney, 2003) or segmentation techniques in agglutinative (S.Virpioja et al., 2007) or fusional languages (Costa-jussà, 2015a)
- Modify models (Koehn and Hoang, 2007)
- Post-process the data (Toutanova et al., 2008; Bojar and Tamchyna, 2011; Formiga et al., 2013).

[1] *to eat, I eat, I ate, I will eat*
[2] *you eat, you ate, you will eat*

Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), pages 56–60,
Beijing, China, July 31, 2015.

The research work in this area is being very active, e.g. PhD proposals using strategies based on deep learning (Gutierrez-Vasques, 2015).

Previous works on the Chinese-Spanish language pair focus on compiling corpus and using pivot stategies (Costa-jussà et al., 2012) and on building a Rule-Based Machine Translation (RBMT) system (Costa-jussà and Centelles, In Press 2015). A high-level description of the state-of-the-art of the translation on this language pair is detailed in (Costa-jussà, 2015b).

Our work mixes several strategies but basically it goes in the direction of (Formiga et al., 2013) that focuses on solving the challenge of morphology as a post-processing classification problem. The idea is to translate from Chinese to a morphology-based simplified Spanish and, then, re-generate the morphology by means of classification algorithms. The competitive advantage from this strategy is the rise of algorithms based on deep learning techniques that can achieve high success rates, e.g. (Collobert et al., 2011).

3 Theoretical Framework

The phrase-based SMT system (Koehn et al., 2003) is trained on a parallel corpus at the level of sentences. It learns co-ocurrences and each token in the training set is considered as a different one no matter if it is morphologically related. Therefore, in the extreme case where the word *canto*[3] is in the training set and the inflection of the same verb *cantó*[4] is not, the latter is going to be considered an out-of-vocabulary word.

Strategy 1. One well-known strategy to face this challenge is to add a part-of-speech (POS) language model which evaluates the probability of the POS-sequences instead of the word sequences.

Strategy 2. This second strategy consists on doing a cascade of systems: first, translate from source to morphology-based simplified target; second, translate from this simplified target to fullform target as shown in Figure 3.

One straightforward simplification in morphology can be adopting lemmas as shown in Table 1.

Strategy 3. This third strategy is based on factored-based translation (Koehn and Hoang, 2007), which uses linguistic information of words, e.g. lemmas and POS. The idea is that the translation model based on words is used if the translation of the word is available, and if not, lemmas and POS are used in combination with a model to generate the final word. Figure 3 shows a typical representation of this factored strategy.

[3] *I sing*
[4] *I sang*

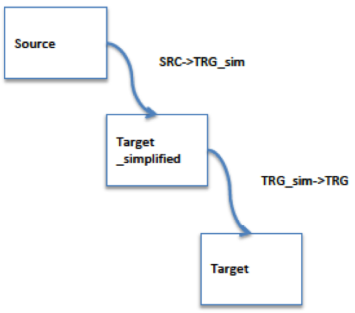

Figure 1: Illustration of the cascade strategy.

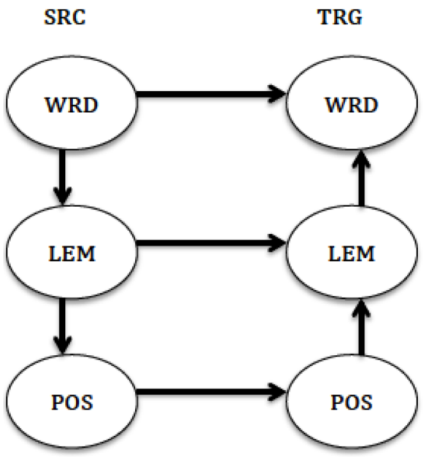

Figure 2: Illustration of the factored strategy.

Strategy 4. This fourth strategy is based on previous work like (Formiga et al., 2013), where the idea is to do a first translation from source to a morphology-based simplified target and then, use a classifier to go from this simplified target to the fullform target. See the schema of this classification-based strategy in 3.

The main challenges in the last strategy are:

1. Explore different simplifications of the target language in order to use the one with a higher trade-off between the highest oracle and the lowest classification complexity.

2. Explore several classification algorithms.

Es_{lemmas}	decidir examinar el cuestión en el período de sesión el tema titular " cuestión relativo a el derecho humano "
Es^{N}_{lemmas}	Decide examinar la cuestión en el período de sesión el tema titulado " cuestión relativas a los derecho humanos " .
Es^{V}_{lemmas}	decidir examinar la cuestión en el período de sesiones el tema titulado " Cuestiones relativas a los derechos humanos " .
Es^{D}_{lemmas}	Decide examinar el cuestión en el período de sesiones el tema titulado " Cuestiones relativas a el derechos humanos " .
Es^{P}_{lemmas}	Decide examinar la cuestión en el período de sesiones el tema titulado " Cuestiones relativas a los derechos humanos " .
Es^{A}_{lemmas}	Decide examinar la cuestión en el período de sesiones el tema titular " Cuestiones relativo a los derechos humano " .
Es_{tags}	VMIP3S0 VMN0000 DA0MS0 NCFS000 SPS00 DA0MS0 NCMS000 SPS00 NCFP000 DA0MS0 NCMS000 AQ0MS0 Fp NCFP000 AQ0FP0 SPS00 DA0MS0 NCMP000 AQ0MP0 Fp Fp
Es_{num}	decidir[VMIP3**N**0] examinar[VMN0000] el[DA0M**N**0] cuestión[NCF**N**000] en[SPS00] el[DA0M**N**0] período[NCM**N**000] de[SPS00] sesión[NCF**N**000] el[DA0M**N**0] tema[NCM**N**000] titular[AQ0M**N**0] "[Fp] cuestión[NCF**N**000] relativo[AQ0F**N**0] a[SPS00] el[DA0M**N**0] derecho[NCM**N**000] humano[AQ0M**N**0] "[Fp] .[Fp]
Es_{gen}	decidir[VMIP3S0] examinar[VMN0000] el[DA0**G**S0] cuestión[NC**G**S000] en[SPS00] el[DA0**G**S0] período[NC**G**S000] de[SPS00] sesión[NC**G**S000] el[DA0**G**S0] tema[NC**G**S000] titular[AQ0**G**S0] "[Fp] cuestión[NC**G**S000] relativo[AQ0**G**S0] a[SPS00] el[DA0**G**S0] derecho[NC**G**S000] humano[AQ0**G**S0] "[Fp] .[Fp]
Es_{numgen}	decidir[VMIP3**N**0] examinar[VMN0000] el[DA0**GN**0] cuestión[NC**GN**000] en[SPS00] el[DA0**GN**0] período[NC**GN**000] de[SPS00] sesión[NC**GN**000] el[DA0**GN**0] tema[NC**GN**000] titular[AQ0**GN**0] "[Fp] cuestión[NC**GN**000] relativo[AQ0**GN**00] a[SPS00] el[DA0**GN**0] derecho[NC**GN**000] humano[AQ0**GN**0] "[Fp] .[Fp]
Es	Decide examinar la cuestión en el período de sesiones el tema titulado " Cuestiones relativas a los derechos humanos " .

Table 1: Example of Spanish simplification into lemmas and different variations

.

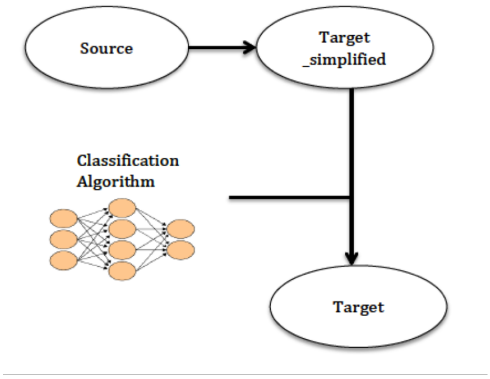

Figure 3: Illustration of the classification-based strategy.

In this paper, we study the first challenge of exploring different simplifications. However, we do not face the classification challenge, which is left to further work. It would be interesting to use deep learning knowledge which is leading to large improvements in natural language processing (Collobert et al., 2011).

4 Ongoing Experiments

In this section we show experiments and results with the four strategies proposed in the previous section.

As discussed in the literature, there are not many parallel corpora available for Chinese-Spanish (Costa-jussà et al., 2012). In this work, we use the data set from the United Nations (Rafalovitch and Dale, 2009). The training corpus contains about 60,000 sentences (and around 2 million words) and the development and test corpus contain 1,000 sentences each one. The baseline system is standard phrase-based SMT trained with Moses (Koehn et al., 2007), with the default parameters.

Table 2 shows results for the strategies 1, 2 and 3 in terms of BLEU (Papineni et al., 2002). From the BLEU scores, we see that strategy 1 gives slight improvements, but strategies 2 and 3 do not.

Strategy	System	BLEU
	Baseline	32.29
1	+LM$_{pos}$	32.54
2	Cascade	31.80
	Zh2Es$_{lemmas}$	36.40
	Es$_{lemmas}$2Es	71.79
3	+Generation	32.11

Table 2: BLEU scores for Zh2Es translation task and different morphology strategies.

Table 3 shows several oracles for strategy 4 with different morphology-based simplifications of Spanish. Best oracles are for lemmas. Then, we explore other simplifications, including lemmatizing only: nouns (N), verbs (V), determiners (D), posesives (P) or adjectives (A). Non of these alternatives approach the best oracle from lemmatizing all words.

However, the interesting results are obtained when simplifying by number (*num*) and/or gender (*gen*). When simplifying number or gender, note that we use the information of lemmas and tags. When generalizing number, note that instead of using the information of singular (*S*) or plural (*P*) in the POS tag with the respective *S* or *P*, we use the generic *N*. Therefore, we generalize the information of number. Similarly when generalizing gender or both (*numgen*).

Oracles get closer to the lemmas simplification when only simplifying both number and gender in Spanish. This finding is relevant in the sense that it simplifies the classification task in the further work that we are considering.

System	Oracles
Baseline	32.29
Zh2Es$_{lemmas}$	36.40
Zh2Es$^{N}_{lemmas}$	32.44
Zh2Es$^{V}_{lemmas}$	33.07
Zh2Es$^{D}_{lemmas}$	33.53
Zh2Es$^{P}_{lemmas}$	32.22
Zh2Es$^{A}_{lemmas}$	24.50
Zh2Es$_{num}$	34.05
Zh2Es$_{gen}$	33.36
Zh2Es$_{numgen}$	**35.80**

Table 3: Oracles for different generalizations. In bold, the most interesting finding.

Table 1 shows examples of all simplifications presented in previous Table 3. Note that simplifications in number and gender use lemmas plus POS tags to omit just the corresponding information that will need to be recovered in the classification stage.

5 Conclusions and Further Work

This paper presents an ongoing work on enhacing a standard phrase-based SMT system by dealing with morphology. We have reported several strategies including adding POS language modeling, experimenting with cascade systems and factored-based translation models. Only the first one reported improvements over the baseline. An additional strategy consists of studying different Spanish simplifications and then, generating the full-form with classification techniques. Experiments show that simplification only in gender and number almost achieves improvements as good as the simplification on lemmas. This is an interesting result that reduces the level of complexity for the classification task. As further work, we will use classification techniques based on deep learning.

Acknowledgements

This work has been supported in part by Spanish Ministerio de Economía y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER) and the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951).

References

E. Avramidis and P. Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*, pages 763–770.

O. Bojar and A. Tamchyna. 2011. Forms wanted: Training smt on monolingual data. In *Workshop of Machine Translation and Morphologically-Rich Languages*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, November.

M. R. Costa-jussà and J. Centelles. In Press, 2015. Description of the chinese-to-spanish rule-based machine translation system developed with a hybrid combination of human annotation and statistical techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

M. R. Costa-jussà, C. A. Henríquez Q, and R. E. Banchs. 2012. Evaluating indirect strategies for chinese-spanish statistical machine translation. *Journal of Artificial Intelligence Research*, 45(1):761–780, September.

M. R. Costa-jussà. 2015a. Segmentation strategies to face morphology challenges in brazilian-portuguese/english statistical machine translation and its integration in cross-language information retrieval. *Computación y Sistemas*, In Press.

M. R. Costa-jussà. 2015b. Traducción automática entre chino y español: dónde estamos? *Komputer Sapiens*, 1.

L. Formiga, M. R. Costa-jussà, J. B. Mariño, J. A. R. Fonollosa, A. Barrón-Cedeño, and L. Márquez. 2013. The TALP-UPC phrase-based translation systems for WMT13: System combination with morphology generation, domain adaptation and corpus filtering. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 134–140, Sofia, Bulgaria, August.

X. Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proc. of the NAACL Student Research Workshop*.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

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

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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

K Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Rafalovitch 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.

S.Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Machine Translation Summit XI*, pages 491–498.

K. Toutanova, H. Suzuki, and A. Ruopp. 2008. Applying morphology generation models to machine translation. In *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*, pages 514–522, Columbus, Ohio.

N. Ueffing and H. Ney. 2003. Using pos information for statistical machine translation into morphologically rich languages. In *Proc. of the 10th conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 347–354, Stroudsburg, PA, USA.