

Barcelona Media SMT system description for the IWSLT 2009: introducing source context information

Marta R. Costa-jussà and Rafael E. Banchs

Barcelona Media Research Center
Av Diagonal, 177, 9th floor, 08018 Barcelona
{marta.ruiz|rafael.banchs}@barcelonamedia.org

Abstract

This paper describes the Barcelona Media SMT system in the IWSLT 2009 evaluation campaign. The Barcelona Media system is an statistical phrase-based system enriched with source context information. Adding source context in an SMT system is interesting to enhance the translation in order to solve lexical and structural choice errors. The novel technique uses a similarity metric among each test sentence and each training sentence. First experimental results of this technique are reported in the Arabic and Chinese Basic Traveling Expression Corpus (BTEC) task. Although working in a single domain, there are ambiguities in SMT translation units and slight improvements in BLEU are shown in both tasks (Zh2En and Ar2En).

1. Introduction

This paper describes the phrase-based baseline SMT system and the main innovative ideas of the Barcelona Media research center (BMRC) phrase-based system for the IWSLT 2009, which integrates source context information.

Adding source context in an SMT system may be interesting to enhance the translation in order to deal with polysemic words and similars.

The basic idea of our approach is based on computing a similarity metric among each test sentence and each training sentence. Then, the similarity metric will be added as a feature function in the translation phrase table. This feature function is intended to push the decoder to use the translation units provided by the training sentence which is more similar to the test sentence.

We participated on the Arabic and Chinese to English Basic Traveling Expression Corpus (BTEC) task. Our primary system was a standard phrase-based SMT system enhanced with source context information.

This paper is organized as follows. Section 2 makes a brief description of some related work to the introduction of source context information in a machine translation system. Section 3 describes the baseline system. Then, Section 4 reports the novel technique of adding source context information. As follows, Section 5 shows the experimental details of the system and the experiments performed with the

novel technique. Section 6 discusses the results obtained on the evaluation campaign and, finally, Section 7 presents the conclusions.

2. Related work

The phrase-based translation model allows to introduce both source and target context information in comparison to the word-based translation model. However, the idea of introducing context information is simplified in the phrase-based systems given that all training sentences contribute equally to the final translation.

More complex works which introduce source context information can be found in the SMT literature. For example, [10, 4] incorporate source language context using neighbouring words, part-of-speech tags and/or supertags. They use a memory-based classification approach to obtain the probability for the given additional contexts with the source phrase. Works such as [2] embed context-rich approaches from Word Sense Disambiguation methods. Other related works focus on extending the translation and target language model using neural networks [8] which aims at smoothing both the translation and target language model in order to use the n -grams more adequate in the translated sentence.

3. Phrase-based Baseline System

The basic idea of phrase-based translation is to segment the given source sentence into units (hereinafter called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

Basically, a bilingual phrase is a pair of m source words and n target words. For extraction from a bilingual word aligned training corpus, two additional constraints are considered:

1. the words are consecutive, and,
2. they are consistent with the word alignment matrix.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions.

The translation model is combined together with the following six additional feature models: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. These models are optimized in the decoder following the procedure described in <http://www.statmt.org/moses/>.

4. Introducing source context information

For introducing source context information into the translation system, we redefine the concept of phrase as a translation unit. In our proposed methodology a translation unit should be composed of a conventional phrase plus its corresponding original source context, which is the context of the source language side of the bilingual sentence pair the phrase was originally extracted from. For simplicity, in this first implementation of the proposed methodology, we will restrict the idea of original source context to the whole source sentence the phrase was extracted from. Notice that, by this definition of translation unit, two identical phrases extracted from different aligned sentence pairs will constitute two different translation units.

The similarity metric used as feature function for incorporating the source context information into the translation system is the cosine distance. According to this, the feature is computed for each phrase by considering the cosine distance between the vector models of the input sentence to be translated and the original source sentence the phrase was extracted from. For constructing the vector models, the standard bag of words approach with TFIDF weighting is used [7].

Once the cosine distance is computed for each phrase and each input sentence to be translated, we can add it as feature function (hereinafter, cosine distance feature). Notice that, differently from most of the feature functions commonly implemented by state-of-the-art phrase based systems, the cost of this new feature function depends on the input sentence to be translated, which means that has to be computed during translation time (this, indeed, constitutes a computational overhead that cannot be dealt with beforehand). Because of this, we must keep one translation table for each input sentence to be translated. In the case one phrase table of a specific test sentence contains several identical phrase units with different costs of the cosine distance feature, we keep the one that has the highest cosine distance value.

At the Moses level, the cosine distance feature is added as one *tm* feature model, optimized with a modified *mert* algorithm which translates one sentence at a time. The resulting increment in translation time (i.e. the optimization time as well) is around three times with respect to the translation time of the standard Moses baseline system.

The proposed methodology is graphically illustrated in Figure 1.

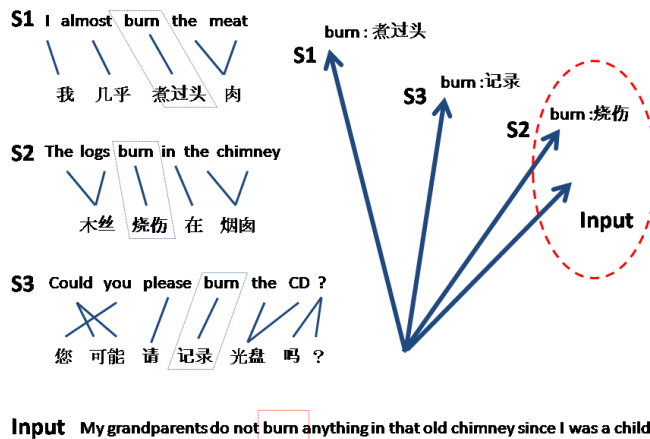


Figure 1: Example of source context information methodology.

5. Experiments

We participated on the Arabic and Chinese to English BTEC task (correct recognition results).

Experiments with the Arabic and Chinese to English MT were carried out on the BTEC data [11]. Corpus statistics are shown Tables 1, 2 and 3. Model weights were tuned with the 2006 development corpus (Dev6), containing 489 sentences and 6 reference translations. The internal test set was the 2007 development set (486 sentences and 16 reference translations), according to which we make a decision about better or worse system performance. Weights obtained in the optimization were used as well for the evaluation test. However, in the evaluation campaign, we concatenated the training, development and test sets from Table 1, 2, 3 and we used the concatenation as training data for translating the evaluation set.

5.1. Arabic data

One first run we participated was the Arabic to English BTEC translation task. We used a similar approach to that shown in [3], namely the MADA+TOKAN system for disambiguation and tokenization. For disambiguation only diacritic unigram statistics were employed. For tokenization we used the D3 scheme with -TAGBIES option. The scheme splits the following set of enclitics: w+, f+, b+, k+, l+, Al+ and pronominal enclitics. The -TAGBIES option produces Bies POS tags on all taggable tokens.

Table 1 gives details about the training, development and test set that we used to make experiments. The first column shows the Arabic corpus statistics without processing and the second column shows the Arabic corpus statistics after using the MADA+TOKAN tool.

		Arabic	Arabic'
Training	Sentences	21,484	21,484
	Words	168,5k	216,9k
	Vocabulary	18,591	11,038
Development	Sentences	489	489
	Words	2,989	3,806
	Vocabulary	1,168	980
Test	Sentences	507	507
	Words	3,224	4,132
	Vocabulary	1,209	1,002
Evaluation	Sentences	469	469
	Words	2,289	3,760
	Vocabulary	1,217	948

Table 1: Arabic training, development, test and evaluation sets before the preprocessing (Arabic) and after (Arabic')

5.2. Chinese data

One second run we participated was the Chinese to English BTEC translation task.

Table 2 gives details about the training, development and test set that we used to make experiments. No preprocessing was done in this case.

		Chinese
Training	Sentences	21,484
	Words	182,2k
	Vocabulary	8,773
Development	Sentences	489
	Words	3,169
	Vocabulary	881
Test	Sentences	507
	Words	3,352
	Vocabulary	888
Evaluation	Sentences	469
	Words	3,019
	Vocabulary	859

Table 2: Chinese training, development, test and evaluation sets.

5.3. English data

Table 3 gives details about the training, development and test set that we used to make the experiments before the evaluation. We tokenized punctuation marks and contractions and all words were lowercase, both in the training and development sets.

5.4. Primary and contrastive submission

As a primary system we submitted the MOSES-based system enhanced with the source context information technique. As

		English	English'
Training	Sentences	21,484	21,484
	Words	162,3k	200,4k
	Vocabulary	13,666	7,334
Development	Sentences	489	489
	Words	2,969	3721
	Vocabulary	1,101	820
Test	Sentences	507	-
	Words	3,042	-
	Vocabulary	1,097	-

Table 3: English training, development, test and evaluation sets before the preprocessing (English) and after (English')

a contrastive system we submitted the MOSES-based system.

Both Machine Translation Systems were phrase-based systems as described in Section 3 and both were based on MOSES open source package [6]. IBM word reordering constraints [1] were applied during decoding to reduce the computational complexity. The other models and feature functions employed by MOSES decoder were:

- TM(s), direct and inverse phrase/word based TM (10 words as maximum length per phrase).
- Distortion model, which assigns a cost linear to the re-ordering distance, while the cost is based on the number of source words which are skipped when translating a new source phrase.
- Lexicalized word reordering model [5, 12].
- Word and phrase penalties, which count the number of words and phrases in the target string.
- Target-side LM (4-gram).

The TM and reordering model were trained using the standard MOSES tools. Weights of feature functions were tuned by using the optimization tools from the MOSES package. The search operation was accomplished by MOSES decoder.

In the primary submission, we introduced context information as explained in Section 4. Several tools provided in the Moses package were modified in order to introduce the novel technique.

5.5. Postprocessing

We used a strategy for restoring punctuation and case information as proposed on the IWSLT'08 web page, using standard SRI LM[9] tools: *disambig* to restore case information and *hidden-ngram* to insert missing punctuation marks.

5.6. Experimental results

Results in the test set are shown in Table 4 and 5. The baseline system enhanced with the context technique produces

slightly better translations in terms of BLEU. The low increase of performance, which is not statistically significant, may be explained by either one, or the combination, of the following two facts: (1) IWSLT corpus sentence lengths are comparable to the maximum size of phrases used by the system, so considering whole sentences as source context information might not provide additional information to the translation system, and (2) IWSLT corpus is restricted to a very specific domain, so there is not actually any context variation in the corpus for our proposed method to provide a significant benefit in terms of translation quality.

	Test
Baseline	54.47
Baseline+Context	54.59

Table 4: BLEU results for Arabic-English test set.

	Test
Baseline	41.32
Baseline+Context	41.38

Table 5: BLEU results for Chinese-English test set.

Finally, Figure 2 shows some translation examples with and without the context technique, drawn from IWSLT internal test sets. As shown previously in [2], these examples illustrate that, even in a single domain, there are sense ambiguities in SMT translation units (i.e. *see* and *say*), which can be solved by adding extra-information of the source context.

Baseline: Please bring me a .
Baseline+Context: Give me another one , please .
REF: I would like one more , please .
Baseline: You see me ?
Baseline+Context: Do you understand what I'm saying ?
REF: Do you understand me ?
Baseline: What time does this train to ?
Baseline+Context: What time will the train arrive ?
REF: What time does the train arrive in Dover ?
Baseline: Got medicine without a prescription .
Baseline+Context: I got medicine without a prescription .
REF: I bought over-the-counter drugs .

Figure 2: Translation examples from the BASELINE and BASELINE+CONTEXT systems: Zh2En and Ar2En (from top to bottom).

6. Evaluation results and discussion

Results in the evaluation campaign are shown in Table 6 and 7. The primary system was the baseline system enhanced with the context technique and the contrastive system was the

baseline system. Results are not coherent with the results that we obtained in the internal test set, which were commented in the section above. The observed differences with respect to the internal test evaluation might be a consequence of incorporating both, development and test, datasets into the training set without performing any new optimization. In such a case, this would suggest that incorporating the cosine distance feature makes the translation system more sensible to optimization parameters. However, more research is necessary to confirm this assumption.

	Evaluation	Position
Primary	49.51	6/9
Contrastive	50.64	6/9

Table 6: BLEU results for Arabic-English evaluation set (case+punctuation). Additionally we show the position compared to the other participants.

	Evaluation	Position
Primary	39.55	6/12
Contrastive	39.66	6/12

Table 7: BLEU results for Chinese-English evaluation set (case+punctuation). Additionally we show the position compared to the other participants.

7. Conclusions

This paper presented a novel technique which allows to introduce source context information into a phrase-based SMT system. The technique is based on using a new concept of translation unit which is composed of a conventional phrase plus its corresponding original source context. The cosine distance is used as a measure of similarity between the source language side of the bilingual sentence pair and the input sentence.

Preliminary results on the internal test set shows that this approach slightly helps to improve translation when working on a single domain like the IWSLT task. This means that even working on a single domain, test sentence translation can be further improved if using the translation unit which have been extracted from a more similar training sentence (similarity measured with the cosine distance).

The presented technique of adding source context information can be further improved in the near future. At the moment, we are using the entire sentence as source context. The novel technique may be further improved by: (1) using shorter or variable source context lengths; (2) using lemmas instead of words; and/or (3) using syntactic categories. Finally, this type of technique may be more useful when working on tasks which include different domains.

8. Acknowledgements

This work has been partially funded by Barcelona Media Innovation Center and the Spanish Ministry of Education and Science through the *Juan de la Cierva* research program.

9. References

- [1] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer. Language translation apparatus and method using context-based translation models. (5510981), 1996.
- [2] M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–72, Prague, June 2007.
- [3] N. Habash and F. Sadat. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL'06)*, New York, USA, June 2006.
- [4] R. Haque, S. Kumar Naskar, Y Ma, and A. Way. Using supertags as source language context in smt. In *13th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 234–241, Barcelona, 2009.
- [5] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the Int. Workshop on Spoken Language Translation (IWSLT'05)*, Pittsburg, USA, October 2005.
- [6] 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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 177–180, Prague, Czech Republic, June 2007.
- [7] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [8] H. Schwenk, M.R. Costa-jussà, and J.A.R. Fonollosa. Smooth bilingual translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 430–438, Prague, June 2007.
- [9] A. Stolcke. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO, September 2002.
- [10] N. Stroppa, A. van de Bosch, and A. Way. Exploiting source similarity for smt using context-informed features. In *11th Conference on Theoretical and Methodological Issues in Machine TRanslation (TMI)*, pages 231–240, Skövde, 2007.
- [11] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceeding of LREC-2002: Third International Conference on Language Resources and Evaluation*, pages 147–152, Las Palmas, Spain, May 2002.
- [12] C. Tillmann. A unigram orientation model for statistical machine translation. In *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pages 101–104, Boston, May 2004.