Barcelona Media SMT system description for the IWSLT 2009

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

- Statistical machine translation (SMT)

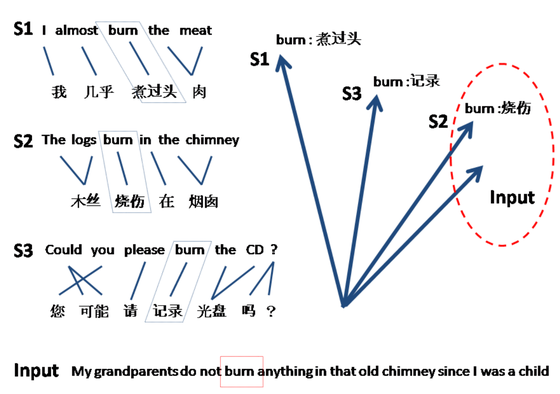
$$\mathbf{e}^* = \arg\max p(\mathbf{e}|\mathbf{f}) = \arg\max_e \{exp(\sum_i \lambda_i h_i(\mathbf{e},\mathbf{f}))\}$$

2 BASELINE SYSTEM

- Bilingual Phrase Translation Model [Och et al, 99; Koehn et al, 03]
  - The translation model is based on phrases.
  - Bilingual units, i.e. phrases, are extracted from a word-to-word aligned corpus according to:
    1. Words are consecutive along both sides of the bilingual phrase,
    2. No word on either side of the phrase is aligned to a word out of the phrase.
- Feature funcions
  - In addition to the translation model, the baseline system implements a log linear combination of feature functions:a **target language model**, a **word bonus**, a **source-to-target lexicon model**, a **target-to-source lexicon model**, a **lexicalized reordering**.

3 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.
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 more, 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 the following Figure.



Example of source context information methodology.

4 EXPERIMENTS

4.1 DATA AND PREPROCESSING

		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

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

- For Arabic, the MADA+TOKAN system was used for disambiguation and tokenization.

		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

Chinese training, development, test and evaluation sets.

- For Chinese, no tokenization was performed.

		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	-

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

- For English the tokenization was performed on punctuation marks and contractions. Additionallay all words were lowercase, both in the training and development sets.

4.2 OFFICIAL SUBMISSIONS

- Primary system: we submitted the MOSES-based system enhanced with the source context information technique. As a contrastive system we submitted the MOSES-based system.
- Secondary system: the above MOSES-based system with the following models and feature functions:
  - 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 reordering 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.
  - Word and phrase penalties, which count the number of words and phrases in the target string.
  - Target-side LM (4-gram).

4.3 POSTPROCESSING

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

4.4 EXPERIMENTAL RESULTS

	Test
Baseline	54.47
Baseline+Context	54.59

BLEU results for Arabic-English test set.

	Test
Baseline	41.32
Baseline+Context	41.38

BLEU results for Chinese-English test set.

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 .

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

4.5 EVALUATION RESULTS

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

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

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

5 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.

6 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.