The TALP-UPC Ngram-based statistical machine translation system for ACL-WMT 2008

Maxim Khalilov, Adolfo Hernández H., Marta R. Costa-jussà, Josep M. Crego, Carlos A. Henríquez Q., Patrik Lambert, José A. R. Fonollosa, José B. Mariño and **Rafael E. Banchs**
Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
(khalilov, adolfohh, mruiz, jmcrego, carloshq, lambert, adrian, canton, rbanchs)@gps.tsc.upc.edu

Abstract

This paper reports on the participation of the TALP Research Center of the UPC (Universitat Politècnica de Catalunya) to the ACL WMT 2008 evaluation campaign.

This year's system is the evolution of the one we employed for the 2007 campaign. Main updates and extensions involve linguistically motivated word reordering based on the reordering patterns technique. In addition, this system introduces a target language model, based on linguistic classes (Part-of-Speech), morphology reduction for an inflectional language (Spanish) and an improved optimization procedure.

Results obtained over the development and test sets on Spanish to English (and the other way round) translations for both the traditional Europarl and a challenging News stories tasks are analyzed and commented.

1 Introduction

Over the past few years, the Statistical Machine Translation (SMT) group of the TALP-UPC has been developing the Ngram-based SMT system (Mariño et al., 2006). In previous evaluation campaigns the Ngram-based approach has proved to be comparable with the state-of-the-art phrase-based systems, as shown in Koehn and Monz(2006), Callison-Burch et al. (2007).

We present a summary of the TALP-UPC Ngram-based SMT system used for this shared task. We discuss the system configuration and novel features, namely linguistically motivated reordering technique, which is applied on the decoding step. Additionally, the reordering procedure is supported by an Ngram language model (LM) of reordered source Part-of-Speech tags (POS).

In this year's evaluation we submitted systems for Spanish-English and English-Spanish language pairs for the traditional (*Europarl*) and challenging (*News*) tasks. In each case, we used only the supplied data for each language pair for models training and optimization.

This paper is organized as follows. Section 2 briefly outlines the 2008 system, including tuple definition and extraction, translation model and additional feature models, decoding tool and optimization procedure. Section 3 describes the word reordering problem and presents the proposed technique of reordering patterns learning and application. Later on, Section 4 reports on the experimental setups of the WMT 2008 evaluation campaign. In Section 5 we sum up the main conclusions from the paper.

2 Ngram-based SMT System

Our translation system implements a log-linear model in which a foreign language sentence $f_1^J = f_1, f_2, ..., f_J$ is translated into another language $e_1^I = f_1, f_2, ..., e_I$ by searching for the translation hypothesis $\hat{e}_1^I$ maximizing a log-linear combination of several feature models (Brown et al., 1990):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

The core part of the system constructed in that way is a translation model, which is based on bilingual *n-grams*. It actually constitutes an Ngram-based LM of bilingual units (called tuples), which approximates the joint probability between the languages under consideration. The procedure of tuples extraction from a word-to-word alignment according to certain constraints is explained in detail in Mariño et al. (2006).

The Ngram-based approach differs from the phrase-based SMT mainly by distinct representating of the bilingual units defined by word alignment and using a higher

Proceedings of the Third Workshop on Statistical Machine Translation, pages 127–130,
Columbus, June 2008.

order HMM of the translation process. While regular phrase-based SMT considers context only for phrase reordering but not for translation, the N-gram based approach conditions translation decisions on previous translation decisions.

The TALP-UPC 2008 translation system, besides the bilingual translation model, which consists of a 4-*gram* LM of tuples with *Kneser-Ney discounting* (estimated with SRI Language Modeling Toolkit[1]), implements a log-linear combination of five additional feature models:

- a **target language model** (*a* 4-*gram* model of words, estimated with *Kneser-Ney smoothing*);
- a **POS target language model** (a 4-*gram* model of tags with *Good-Turing discounting* (TPOS));
- a **word bonus model**, which is used to compensate the system's preference for short output sentences;
- a **source-to-target lexicon model** and a **target-to-source lexicon model**, these models use word-to-word IBM Model 1 probabilities (Och and Ney, 2004) to estimate the lexical weights for each tuple in the translation table.

Decisions on the particular LM configuration and smoothing technique were taken on the minimal-perplexity and maximal-BLEU bases.

The decoder (called MARIE), an open source tool[2], implementing a beam search strategy with distortion capabilities was used in the translation system.

Given the development set and references, the log-linear combination of weights was adjusted using a simplex optimization method (with the optimization criteria of the highest BLEU score) and an n-best re-ranking just as described in *http://www.statmt.org/jhuws/*. This strategy allows for a faster and more efficient adjustment of model weights by means of a double-loop optimization, which provides significant reduction of the number of translations that should be carried out.

3 Reordering framework

For a great number of translation tasks a certain reordering strategy is required. This is especially important when the translation is performed between pairs of languages with non-monotonic word order. There are various types of distortion models, simplifying bilingual translation. In our system we use an extended monotone reordering model based on automatically learned reordering rules. A detailed description can be found in Crego and Mariño (2006).

[1] http://www.speech.sri.com/projects/srilm/

[2] http://gps-tsc.upc.es/veu/soft/soft/marie/

Apart from that, tuples were extracted by an unfolding technique: this means that the tuples are broken into smaller tuples, and these are sequenced in the order of the target words.

3.1 Reordering patterns

Word movements are realized according to the reordering rewrite rules, which have the form of:

$$t_1, ..., t_n \mapsto i_1, ..., i_n$$

where $t_1, ..., t_n$ is a sequence of POS tags (relating a sequence of source words), and $i_1, ..., i_n$ indicates which order of the source words generate monotonically the target words.

Patterns are extracted in training from the crossed links found in the word alignment, in other words, found in translation tuples (as no word within a tuple can be linked to a word out of it (Crego and Mariño, 2006)).

Having all the instances of rewrite patterns, a score for each pattern on the basis of relative frequency is calculated as shown below:

$$p(t_1, ..., t_n \mapsto i_1, ..., i_n) = \frac{N(t_1, ..., t_n \mapsto i_1, ..., i_n)}{NN(t_1, ..., t_n)}$$

3.2 Search graph extension and source POS model

The monotone search graph is extended with reorderings following the patterns found in training. Once the search graph is built, the decoder traverses the graph looking for the best translation. Hence, the winning hypothesis is computed using all the available information (the whole SMT models).

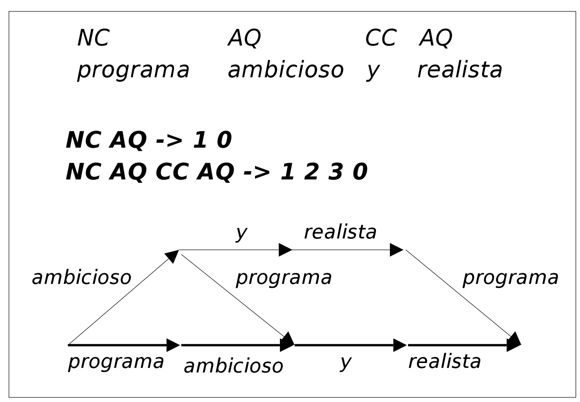

Figure 1: Search graph extension. *NC, CC and AQ stand respectively for name, conjunction and adjective.*

The procedure identifies first the sequences of words in the input sentence that match any available pattern. Then, each of the matchings implies the addition of an arc into the search graph (encoding the reordering learned in the pattern). However, this addition of a new arc is not

Task	BL		BL+SPOS	
	Europarl	News	Europarl	News
es2en	32.79	36.09	32.88	36.36
en2es	32.05	33.91	32.10	33.63

Table 1: BLEU comparison demonstrating the impact of the source-side POS tags model.

performed if a translation unit with the same source-side words already exists in the training. Figure 1 shows how two rewrite rules applied over an input sentence extend the search graph given the reordering patterns that match the source POS tag sequence.

The reordering strategy is additionally supported by a *4-gram language model* (estimated with *Good-Turing smoothing*) of reordered **source POS tags** (SPOS). In training, POS tags are reordered according with the extracted reordering patterns and word-to-word links. The resulting sequence of source POS tags is used to train the Ngram LM.

Table 1 presents the effect of the source POS LM introduction to the reordering module of the Ngram-based SMT. As it can be seen, the impactya le h of the source-side POS LM is minimal, however we decided to consider the model aiming at improving it in future. The reported results are related to the *Europarl* and News Commentary (*News*) development sets. BLEU calculation is case insensitive and insensitive to tokenization. *BL* (baseline) refers to the presented Ngram-based system considering all the features, apart from the target and source POS models.

4 WMT 2008 Evaluation Framework

4.1 Corpus

An extraction of the official transcriptions of the 3rd release of the European Parliament Plenary Sessions[3] was provided for the ACL WMT 2008 shared translation task. About 40 times smaller corpus from news domain (called News Commentary) was also available. For both tasks, our training corpus was the catenation of the Europarl and News Commentary corpora.

TALP UPC participated in the constraint to the provided training data track for Spanish-English and English-Spanish translation tasks. We used the same training material for the traditional and challenging tasks, while the development sets used to tune the system were distinct (**2000** sentences for **Europarl task** and **1057** for **News Commentary**, **one reference translation** for each of them). A brief training and development corpora statistics is presented in Table 2.

[3] http://www.statmt.org/wmt08/shared-task.html

	Spanish	English
Train		
Sentences	1.3 M	1.3 M
Words	38.2 M	35.8 K
Vocabulary	156 K	120 K
Development Europarl		
Sentences	2000	2000
Words	61.8 K	58.7 K
Vocabulary	8 K	6.5 K
Development News Commentary		
Sentences	1057	1057
Words	29.8 K	25.8 K
Vocabulary	5.4 K	4.9 K

Table 2: Basic statistics of ACL WMT 2008 corpus.

4.2 Processing details

The training data was preprocessed by using provided tools for tokenizing and filtering.

POS tagging. POS information for the source and the target languages was considered for both translation tasks that we have participated. The software tools available for performing POS-tagging were Freeling (Carreras et al., 2004) for Spanish and TnT (Brants, 2000) for English. The number of classes for English is 44, while Spanish is considered as a more inflectional language, and the tag set contains 376 different tags.

Word Alignment. The word alignment is automatically computed by using GIZA++[4](Och and Ney, 2000) in both directions, which are symmetrized by using the union operation. Instead of aligning words themselves, stems are used for aligning. Afterwards case sensitive words are recovered.

Spanish Morphology Reduction. We implemented a morphology reduction of the Spanish language as a preprocessing step. As a consequence, training data sparseness due to Spanish morphology was reduced improving the performance of the overall translation system. In particular, the pronouns attached to the verb were separated and contractions as *del* or *al* were splitted into *de el* or *a el*. As a post-processing, in the En2Es direction we used a POS target LM as a feature (instead of the target language model based on classes) that allowed to recover the segmentations (de Gispert, 2006).

4.3 Experiments and Results

In contrast to the last year's system where statistical classes were used to train the target-side tags LM, this year we used **linguistically motivated word classes**

[4] http://code.google.com/p/giza-pp/

Task	BL+SPOS		BL+SPOS+TPOS (UPC 2008)	
	Europarl	News	Europarl	News
es2en	32.88	36.36	32.89	36.31
en2es	31.52	34.13	30.72	32.72
en2es "clean"[5]	32.10	33.63	32.09	35.04

Table 3: BLEU scores for Spanish-English and English-Spanish 2008 development corpora (Europarl and News Commentary).

Task	UPC 2008	
	Europarl	News
es2en	32.80	19.61
en2es	31.31	19.28
en2es "clean"[5]	32.34	20.05

Table 4: BLEU scores for official tests 2008.

(POS) which were considered to train the POS target LM and extract the reordering patterns. Other characteristics of this year's system are:

- **reordering patterns** technique;
- **source POS model**, supporting word reordering;
- **no LM interpolation**. For this year's evaluation, we trained two separate LMs for each domain-specific corpus (i.e., Europarl and News Commentary tasks).

It is important to mention that 2008 training material is identical to the one provided for the 2007 shared translation task.

Table 3 presents the *BLEU* score obtained for the 2008 development data sets and shows the impact of the target-side POS LM introduction, which can be characterized as highly corpus- and language-dependent feature. *BL* refers to the same system configuration as described in subsection 3.2. The computed *BLEU* scores are case insensitive, insensitive to tokenization and use one translation reference.

After submitting the systems we discovered a bug related to incorrect implementation of the target LMs of words and tags for Spanish, it caused serious reduction of translation quality (1.4 BLEU points for development set in case of English-to-Spanish Europarl task and 2.3 points in case of the corresponding News Commentary task). The last raw of table 3 (*en2es "clean"*) represents the results corresponding to the UPC 2008 post-evaluation system, while the previous one (*en2es*) refers to the "bugged" system submitted to the evaluation.

The experiments presented in Table 4 correspond to the 2008 test evaluation sets.

[5]Corrected post-evaluation results (see subsection 4.3.)

5 Conclusions

In this paper we introduced the TALP UPC Ngram-based SMT system participating in the WMT08 evaluation. Apart from briefly summarizing the decoding and optimization processes, we have presented the feature models that were taken into account, along with the bilingual Ngram translation model. A reordering strategy based on linguistically-motivated reordering patterns to harmonize the source and target word order has been presented in the framework of the Ngram-based system.

6 Acknowledgments

This work has been funded by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project). The authors want to thank Adrià de Gispert (Cambridge University) for his contribution to this work.

References

T. Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. D. Lafferty, R. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the ACL 2007 Workshop on Statistical and Hybrid methods for Machine Translation (WMT)*, pages 136–158.

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th Int. Conf. on Language Resources and Evaluation (LREC'04)*.

J. M. Crego and J. B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

A. de Gispert. 2006. *Introducing linguistic knowledge into statistical machine translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, December.

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the ACL 2006 Workshop on Statistical and Hybrid methods for Machine Translation (WMT)*, pages 102–121.

J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the the 38th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 440–447.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. 30(4):417 – 449, December.