The TALP-UPC phrase-based translation system for EACL-WMT 2009

José A.R. Fonollosa and **Maxim Khalilov** and **Marta R. Costa-jussà** and
José B. Mariño and **Carlos A. Henríquez Q.** and **Adolfo Hernández H.** and
Rafael E. Banchs
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona 08034
{adrian,khalilov,mruiz,canton,carloshq,adolfohh,rbanchs}@talp.upc.edu

Abstract

This study presents the TALP-UPC submission to the EACL Fourth Worskhop on Statistical Machine Translation 2009 evaluation campaign. It outlines the architecture and configuration of the 2009 phrase-based statistical machine translation (SMT) system, putting emphasis on the major novelty of this year: combination of SMT systems implementing different word reordering algorithms.

Traditionally, we have concentrated on the Spanish-to-English and English-to-Spanish *News Commentary* translation tasks.

1 Introduction

TALP-UPC (Center of Speech and Language Applications and Technology at the Universitat Politècnica de Catalunya) is a permanent participant of the ACL WMT shared translations tasks, traditionally concentrating on the Spanish-to-English and vice versa language pairs. In this paper, we describe the 2009 system's architecture and design describing individual components and distinguishing features of our model.

This year's system stands aside from the previous years' configurations which were performed following an N-gram-based (tuple-based) approach to SMT. By contrast to them, this year we investigate the translation models (TMs) interpolation for a state-of-the-art phrase-based translation system. Inspired by the work presented in (Schwenk and Estève, 2008), we attack this challenge using the coefficients obtained for the corresponding monolingual language models (LMs) for TMs interpolation.

On the second step, we have performed additional word reordering experiments, comparing the results obtained with a statistical method (R. Costa-jussà and R. Fonollosa, 2009) and syntax-based algorithm (Khalilov and R. Fonollosa, 2008). Further the outputs of the systems were combined selecting the translation with the Minimum Bayes Risk (MBR) algorithm (Kumar, 2004) that allowed significantly outperforming the baseline configuration.

The remainder of this paper is organized as follows: Section 2 presents the TALP-UPC'09 phrase-based system, along with the translation models interpolation procedure and other minor novelties of this year. Section 3 reports on the experimental setups and outlines the results of the participation in the EACL WMT 2009 evaluation campaign. Section 4 concludes the paper with discussions.

2 TALP-UPC phrase-based SMT

The system developed for this year's shared task is based on a state-of-the-art SMT system implemented within the open-source MOSES toolkit (Koehn et al., 2007). A phrase-based translation is considered as a three step algorithm: (1) the source sequence of words is segmented in phrases, (2) each phrase is translated into target language using translation table, (3) the target phrases are reordered to be inherent in the target language.

A bilingual phrase (which in the context of SMT do not necessarily coincide with their linguistic analogies) is any pair of m source words and n target words that satisfies two basic constraints: (1) words are consecutive along both sides of the bilingual phrase and (2) no word on either side of the phrase is aligned to a word outside the phrase. Given a sentence pair and a corresponding word-to-word alignment, phrases are extracted following the criterion in (Och and Ney, 2004). The probability of the phrases is estimated by relative frequencies of their appearance in the training corpus.

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 85–89,
Athens, Greece, 30 March – 31 March 2009.

Classically, a phrase-based 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 $e_1^I = e_1, e_2, ..., e_I$ 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.

2.1 Translation models interpolation

We implemented a TM interpolation strategy following the ideas proposed in (Schwenk and Estève, 2008), where the authors present a promising technique of target LMs linear interpolation; in (Koehn and Schroeder, 2007) where a log-linear combination of TMs is performed; and specifically in (Foster and Kuhn, 2007) where the authors present various ways of TM combination and analyze in detail the TM domain adaptation.

In the framework of the evaluation campaign, there were two Spanish-to-English parallel training corpora available: *Europarl v.4* corpus (about 50M tokens) and *News Commentary* (NC) corpus (about 2M tokens). The test dataset provided by the organizers this year was from the news domain, so we considered the Europarl training corpus as "out-of-domain" data and the News Commentary as "in-domain" training material. Unfortunately, the in-domain corpus is much smaller in size, however the Europarl corpus can be also used to increase the final translation and reordering tables in spite of its different nature.

A straightforward approach to the TM interpolation would be an iterative TM reconstruction adjusting scale coefficients on each step of the loop with use of the highest BLEU score as a maximization criterion.

However, we did not expect a significant gain from this time-consumption strategy and we decided to follow a simpler approach. In the presented results, we obtained the best interpolation weight following the standard entropy-based optimization of the target-side LM. We adjust the weight coefficient $\lambda_{Europarl}$ ($\lambda_{NC} = 1 - \lambda_{Europarl}$) of the linear interpolation of the target-side LMs:

$$P(w) = \lambda_{Europarl} \cdot P^w_{Europarl} + \lambda_{NC} \cdot P^w_{NC} \quad (1)$$

where $P^w_{Europarl}$ and P^w_{NC} are probabilities assigned to the word sequence w by the LM estimated on Europarl and NC data, respectively.

The scale factor values are automatically optimized to obtain the lowest perplexity $ppl(w)$ produced by the interpolated LM $P(w)$. We used the standard script $compute - best - mix$ from the SRI LM package (Stolcke, 2002) for optimization.

On the next step, the optimized coefficients $\lambda_{Europarl}$ and λ_{NC} are generalized on the interpolated translation and reordering models. In other words, reordering and translation models are interpolated using the same weights which yield the lowest perplexity for LM interpolation.

The word-to-word alignment was obtained from the joint (merged) database (*Europarl + NC*). Then, we separately computed the translation and reordering tables corresponding to the in- and out-of-domain parts of the joint alignment. The final tables, as well as the final target LM were obtained using linear interpolation. The weights were selected using a minimum perplexity criterion estimated on the corresponding interpolated combination of the target-side LMs.

The optimized coefficient values are: for Spanish: NC weight = 0.526, Europarl weight = 0.474; for English: NC weight = 0.503, Europarl weight = 0.497. The perplexity results obtained using monolingual LMs and the 2009 development set (English and Spanish references) can be found in Table 1, while the corresponding improvement in BLEU score is presented in Section 3.3 and summary of the obtained results (Table 4).

	Europarl	NC	Interpolated
English	463.439	489.915	353.305
Spanish	308.802	347.092	246.573

Table 1: *Perplexity results obtained on the Dev 2009 corpus and the monolingual LMs.*

Note that the corresponding reordering models are interpolated with the same weights.

2.2 Statistical Machine Reordering

The idea of the Statistical Machine Reordering (SMR) stems from the idea of using the powerful techniques developed for SMT and to translate

the source language (S) into a reordered source language (S'), which more closely matches the order of the target language. To infer more reorderings, it makes use of word classes. To correctly integrate the SMT and SMR systems, both are concatenated by using a word graph which offers weighted reordering hypotheses to the SMT system. The details are described in (**?**).

2.3 Syntax-based Reordering

Syntax-based Reordering (SBR) approach deals with the word reordering problem and is based on non-isomorphic parse subtree transfer as described in details in (Khalilov and R. Fonollosa, 2008).

Local and long-range word reorderings are driven by automatically extracted permutation patterns operating with source language constituents. Once the reordering patterns are extracted, they are further applied to monotonize the bilingual corpus in the same way as shown in the previous subsection. The target-side parse tree is considered as a filter constraining reordering rules to the set of patterns covered both by the source- and target-side subtrees.

2.4 System Combination

Over the past few years the MBR algorithm utilization to find the best consensus outputs of different translation systems has proved to improve the translation accuracy (Kumar, 2004). The system combination is performed on the 200-best lists which are generated by the three systems: (1) MOSES-based system without pre-translation monotonization (baseline), (2) MOSES-based SMT enhanced with SMR monotonization and (3) MOSES-based SMT augmented with SBR monotonization. The results presented in Table 4 show that the combined output significantly outperforms the baseline system configuration.

3 Experiments and results

We followed the evaluation baseline instructions [1] to train the MOSES-based translation system.

In some experiments we used MBR decoding (Kumar and Byrne, 2004) with the smoothed BLEU score as a similarity criteria, that allowed gaining 0.2 BLEU points comparing to the standard procedure of outputting the translation with the highest probability (HP). We applied the Moses implementation of this algorithm to the list of 200 best translations generated by the TALP-UPC system. The results obtained over the official 2009 Test dataset can be found in Table 2.

Task	HP	MBR
EsEn	24.48	24.62
EnEs	23.46	23.64

Table 2: *MBR versus MERT decoding.*

The *"recase"* script provided within the baseline was supplemented with and additional module, which restore the original case for unknown words (many of them are proper names and loosing of case information leads to a significant performance degradation).

3.1 Language models

The target-side language models were estimated using the SRILM toolkit (Stolcke, 2002). We tried to use all the available in-domain training material: apart from the corresponding portions of the bilingual NC corpora we involved the following monolingual corpora:

- News monolingual corpus (49M tokens for English and 49M for Spanish)
- Europarl monolingual corpus (about 504M tokens for English and 463M for Spanish)
- A collection of News development and test sets from previous evaluations (151K tokens for English and 175K for Spanish)
- A collection of Europarl development and test sets from previous evaluations (295K tokens for English and 311K for Spanish)

Five LMs per language were estimated on the corresponding datasets and interpolated following the maximum perplexity criteria. Hence, the larger LMs incorporating in- and out-of-domain data were used in decoding.

3.2 Spanish enclitics separation

For the Spanish portion of the corpus we implemented an enclitics separation procedure on the preprocessing step, i.e. the pronouns attached to the verb were separated and contractions as *del* or *al* were splitted into *de el* or *a el*. Consequently, training data sparseness due to Spanish morphology was reduced improving the performance of the overall translation system. As a

[1] http://www.statmt.org/wmt09/baseline.html

post-processing, the segmentation was recovered in the English-to-Spanish direction using target-side Part-of-Speech tags (de Gispert, 2006).

3.3 Results

The automatic scores provided by the WMT'09 organizers for TALP-UPC submissions calculated over the News 2009 dataset can be found in Table 3. BLEU and NIST case-insensitive (CI) and case-sensitive (CS) metrics are considered.

Task	Bleu CI	Bleu CS	NIST CI	NIST CS
EsEn	25.93	24.54	7.275	7.017
EnEs	24.85	23.37	6.963	6.689

Table 3: *BLEU and NIST scores for preliminary official test dataset 2009 (primary submission) with 500 sentences excluded.*

The TALP-UPC primary submission was ranked the 3rd among 28 presented translations for the Spanish-to-English task and the 4th for the English-to-Spanish task among 9 systems.

The following system configurations and the internal results obtained are reported:

- *Baseline*: Moses-based SMT, as proposed on the web-page of the evaluation campaign with Spanish enclitics separation and modified version of "recase" tool,
- *Baseline+TMI*: *Baseline* enhanced with TM interpolation as described in subsection 2.1,
- *Baseline+TMI+MBR*: the same as the latter but with MBR decoding,
- *Baseline+TMI+SMR*: the same as *Baseline+TMI* but with SMR technique applied to monotonize the source portion of the corpus, as described in subsection 2.2,
- *Baseline+SBR*: the same as *Baseline* but with SBR algorithm applied to monotonize the source portion of the corpus, as described in subsection 2.3,
- *System Combination*: a combined output of the 3 previous systems done with the MBR algorithm, as described in subsection 2.4.

Impact of TM interpolation and MBR decoding is more significant for the English-to-Spanish translation task, for which the target-side monolingual corpus is smaller than for the Spanish-to-English translation.

We did not have time to meet the evaluation deadline for providing the system combination output. Nevertheless, during the post-evaluation period we performed the experiments reported in the last three lines of Table 4 (*Baseline+TMI+SMR*, *Baseline+SBR* and *System combination*).

Note that the results presented in Table 4 differ from the ones which can be found the Table 3 due to selective conditions of preliminary evaluation done by the Shared Task organizers.

System	News 2009 Test CI	News 2009 Test CS
Spanish-to-English		
Baseline	25.82	24.37
Baseline+TMI	25.84	24.47
Baseline+TMI+MBR (Primary)	26.04	24.62
Baseline+SMR	24.95	23.62
Baseline+SBR	24.24	22.89
System combination	26.44	25.00
English-to-Spanish		
Baseline	24.56	23.05
Baseline+TMI	25.01	23.41
Baseline+TMI+MBR (Primary)	25.16	23.64
Baseline+SMR	24.09	22.65
Baseline+SBR	23.52	22.05
System combination	25.39	23.86

Table 4: Experiments summary.

4 Conclusions

In this paper, we present the TALP-UPC phrase-based translation system developed for the EACL-WMT 2009 evaluation campaign. The major novelties of this year are translation models interpolation done in linear way and combination of SMT systems implementing different word reordering algorithms. The system was ranked pretty well for both translation tasks in which our institution has participated.

Unfortunately, the promising reordering techniques and the combination of their outputs were not applied within the evaluation deadline, however we report the obtained results in the paper.

5 Acknowledgments

This work has been funded by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project).

References

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.

A. de Gispert. 2006. *Introducing linguistic knowledge into Statistical Machine Translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, December.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *In Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT)*, pages 128–135, Prague, Czech Republic, June.

M. Khalilov and J. R. Fonollosa. 2008. A new subtree-transfer approach to syntax-based reordering for statistical machine translation. Technical report, Universitat Politècnica de Catalunya.

Ph. Koehn and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *In Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT)*, pages 224–227, Prague, Czech Republic, June.

Ph. 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 Association for Computational Linguistics (ACL) 2007*, pages 177–180.

Sh. Kumar and W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *In HLTNAACL'04*, pages 169–176.

Sh. Kumar. 2004. *Minimum Bayes-Risk Techniques in Automatic Speech Recognition and Statistical Machine Translation*. Ph.D. thesis, Johns Hopkins University.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 3(4):417–449, December.

M. R. Costa-jussà and J. R. Fonollosa. 2009. An Ngram reordering model. *Computer Speech and Language*. ISSN 0885-2308, accepted for publication.

H. Schwenk and Y. Estève. 2008. Data selection and smoothing in an open-source system for the 2008 nist machine translation evaluation. In *Proceedings of the Interspeech'08*, pages 2727–2730, Brisbane, Australia, September.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904.