Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses

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

Abstract

This paper describes the 2007 Ngram-based statistical machine translation system developed at the TALP Research Center of the UPC (Universitat Politècnica de Catalunya) in Barcelona. Emphasis is put on improvements and extensions of the previous years system, being highlighted and empirically compared. Mainly, these include a novel word ordering strategy based on: (1) statistically monotonizing the training source corpus and (2) a novel reordering approach based on weighted reordering graphs. In addition, this system introduces a target language model based on statistical classes, a feature for out-of-domain units and an improved optimization procedure.

The paper provides details of this system participation in the ACL 2007 SECOND WORKSHOP ON STATISTICAL MACHINE TRANSLATION. Results on three pairs of languages are reported, namely from Spanish, French and German into English (and the other way round) for both the in-domain and out-of-domain tasks.

1 Introduction

Based on estimating a joint-probability model between the source and the target languages, Ngram-based SMT has proved to be a very competitive alternatively to phrase-based and other state-of-the-art systems in previous evaluation campaigns, as shown in (Koehn and Monz, 2005; Koehn and Monz, 2006).

Given the challenge of domain adaptation, efforts have been focused on improving strategies for Ngram-based SMT which could generalize better. Specifically, a novel reordering strategy is explored. It is based on extending the search by using precomputed statistical information. Results are promising while keeping computational expenses at a similar level as monotonic search. Additionally, a bonus for tuples from the out-of-domain corpus is introduced, as well as a target language model based on statistical classes. One of the advantages of working with statistical classes is that they can easily be used for any pair of languages.

This paper is organized as follows. Section 2 briefly reviews last year's system, including tuple definition and extraction, translation model and feature functions, decoding tool and optimization criterion. Section 3 delves into the word ordering problem, by contrasting last year strategy with the novel weighted reordering input graph. Section 4 focuses on new features: both tuple-domain bonus and target language model based on classes. Later on, Section 5 reports on all experiments carried out for WMT 2007. Finally, Section 6 sums up the main conclusions from the paper and discusses future research lines.

2 Baseline N-gram-based SMT System

The translation model is based on bilingual n-grams. It actually constitutes a language model of bilingual units, referred to as tuples, which approximates the joint probability between source and target languages by using bilingual n-grams.

Tuples are extracted from a word-to-word aligned corpus according to the following two constraints: first, tuple extraction should produce a monotonic segmentation of bilingual sentence pairs; and second, no smaller tuples can be extracted without violating the previous constraint.

For all experiments presented here, the translation model consisted of a 4-gram language model of tuples. In addition to this bilingual n-gram translation model, the baseline system implements a log linear combination of four feature functions. These four additional models are: a **target language model** (a 5-gram model of words); a **word bonus**; a **source-to-target lexicon model** and a **target-to-source lexicon model**, both features provide a complementary probability for each tuple in the translation table.

The decoder (called MARIE) for this translation sys-

Proceedings of the Second Workshop on Statistical Machine Translation, pages 167–170,
Prague, June 2007.

tem is based on a beam search [1].

This baseline system is actually the same system used for the first shared task *"Exploiting Parallel Texts for Statistical Machine Translation"* of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond. A more detailed description of the system can be found in (Mariño et al., 2006).

3 Baseline System Enhanced with a Weighted Reordering Input Graph

This section briefly describes the statistical machine reordering (SMR) technique. Further details on the architecture of SMR system can be found on (Costa-jussà and Fonollosa, 2006).

3.1 Concept

The SMR system can be seen as a SMT system which translates from an original source language (*S*) to a reordered source language (*S'*), given a target language (*T*). The SMR technique works with statistical word classes (Och, 1999) instead of words themselves (particularly, we have used 200 classes in all experiments).

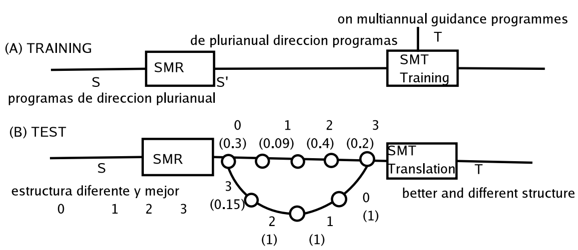

Figure 1: *SMR approach in the (A) training step (B) in the test step (the weight of each arch is in brackets).*

3.2 Using SMR technique to improve SMT training

The original source corpus *S* is translated into the reordered source corpus *S'* with the SMR system. Figure 1 (A) shows the corresponding block diagram. The reordered training source corpus and the original training target corpus are used to build the SMT system.

The main difference here is that the training is computed with the *S'2T* task instead of the *S2T* original task. Figure 2 (A) shows an example of the alignment computed on the original training corpus. Figure 2 (B) shows the same links but with the source training corpus in a different order (this training corpus comes from the SMR output). Although, the quality in alignment is the same, the tuples that can be extracted change (notice that the tuple extraction is monotonic). We are able to extract smaller tuples which reduces the translation vocabulary sparseness. These new tuples are used to build the SMT system.

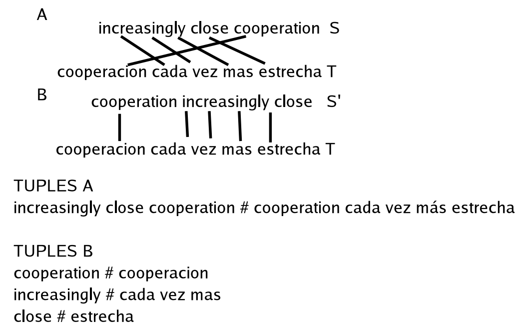

Figure 2: *Alignment and tuple extraction (A) original training source corpus (B) reordered training source corpus.*

3.3 Using SMR technique to generate multiple weighted reordering hypotheses

The SMR system, having its own search, can generate either an output 1-best or an output graph. In decoding, the SMR technique generates an output graph which is used as an input graph by the SMT system. Figure 1 (B) shows the corresponding block diagram in decoding: the SMR output graph is given as an input graph to the SMT system. Hereinafter, this either SMR output graph or SMT input graph will be referred to as (weighted) reordering graph. The monotonic search in the SMT system is extended with reorderings following this reordering graph. This reordering graph has multiple paths and each path has its own weight. This weight is added as a feature function in the log-linear framework. Figure 3 shows the weighted reordering graph.

The main difference with the reordering technique for WMT06 (Crego et al., 2006) lies in (1) the tuples are extracted from the word alignment between the reordered source training corpus and the given target training corpus and (2) the graph structure: the SMR graph provides weights for each reordering path.

4 Other features and functionalities

In addition to the novel reordering strategy, we consider two new features functions.

4.1 Target Language Model based on Statistical Classes

This feature implements a 5-gram language model of target statistical classes (Och, 1999). This model is trained by considering statistical classes, instead of words, for

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

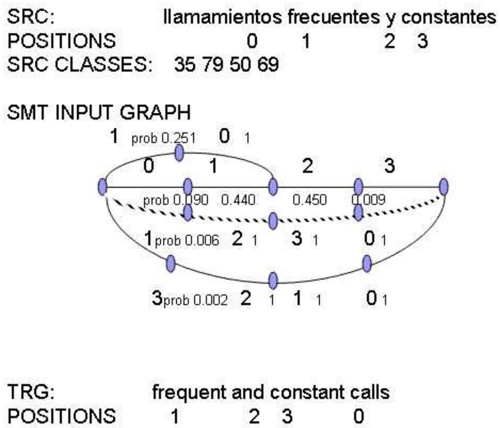

Figure 3: *Weighted reordering input graph for SMT system.*

the target side of the training corpus. Accordingly, the tuple translation unit is redefined in terms of a triplet which includes: a source string containing the source side of the tuple, a target string containing the target side of the tuple, and a class string containing the statistical classes corresponding to the words in the target strings.

4.2 Bonus for out-of-domain tuples

This feature adds a bonus to those tuples which comes from the training of the out-of-domain task. This feature is added when optimizing with the development of the out-of-domain task.

4.3 Optimization

Finally, a n-best re-ranking strategy is implemented which is used for optimization purposes just as proposed in *http://www.statmt.org/jhuws/*. This procedure 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. The current optimization procedure uses the Simplex algorithm.

5 Shared Task Framework

5.1 Data

The data provided for this shared task corresponds to a subset of the official transcriptions of the European Parliament Plenary Sessions [2]. Additionally, there was available a smaller corpus called News-Commentary. For all tasks and domains, our training corpus was the catenation of both.

[2] http://www.statmt.org/wmt07/shared-task/

5.2 Processing details

Word Alignment. The word alignment is automatically computed by using GIZA++ [3] 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* are splited into *de el* or *a el*. As a post-processing, in the En2Es direction we used a POS target language model as a feature (instead of the target language model based on classes) that allowed to recover the segmentations (de Gispert, 2006).

Language Model Interpolation. In other to better adapt the system to the out-of-domain condition, the target language model feature was built by combining two 5-gram target language models (using SRILM [4]). One was trained from the EuroParl training data set, and the other from the available, but much smaller, news-commentary data set. The combination weights for the EuroParl and news-commentary language models were empirically adjusted by following a minimum perplexity criterion. A relative perplexity reduction around 10-15% respect to original EuroParl language model was achieved in all the tasks.

5.3 Experiments and Results

The main difference between this year's and last year's systems are: the amount of data provided; the word alignment; the Spanish morphology reduction; the reordering technique; the extra target language model based on statistical classes (except for the En2Es); and the bonus for the out-of-domain task (only for the En2Es task).

Among them, the most important is the reordering technique. That is why we provide a fair comparison between the reordering patterns (Crego and Mariño, 2006) technique and the SMR reordering technique. Table 1 shows the system described above using either reordering patterns or the SMR technique. The BLEU calculation was case insensitive and sensitive to tokenization.

Table 2 presents the *BLEU* score obtained for the 2006 test data set comparing last year's and this year's systems. The computed *BLEU* scores are case insensitive, sensitive to tokenization and uses one translation reference. The improvement in BLEU results shown from *UPC-jm*

[3] http://www.fjoch.com/GIZA++.html

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

Task	Reordering patterns	SMR technique
es2en	31.21	33.34
en2es	31.67	32.33

Table 1: *BLEU comparison: reordering patterns vs. SMR technique.*

Task	UPC-jm 2006		UPC 2007	
	in-d	out-d	in-d	out-d
es2en	31.01	27.92	33.34	32.85
en2es	30.44	25.59	32.33	33.07
fr2en	30.42	21.79	32.44	26.93
en2fr	31.75	23.30	32.30	27.03
de2en	24.43	17.57	26.54	21.63
en2de	17.73	10.96	19.74	15.06

Table 2: *BLEU scores for each of the six translation directions considered (computed over 2006 test set) comparing last year's and this year's system results (in-domain and out-domain).*

2006 Table 2 and *reordering patterns* Table 1 in the English/Spanish in-domain task comes from the combination of: the additional corpora, the word alignment, the Spanish morphology reduction and the extra target language model based on classes (only in the Es2En direction).

6 Conclusions and Further Work

This paper describes the UPC system for the WMT07 Evaluation. In the framework of Ngram-based system, a novel reordering strategy which can be used for any pair of languages has been presented and it has been showed to significantly improve translation performance. Additionally two features has been added to the log-lineal scheme: the target language model based on classes and the bonus for out-of-domain translation units.

7 Acknowledgments

This work has been funded by the European Union under the TC-STAR project (IST-2002-FP6-506738) and the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project).

References

M.R. Costa-jussà and J.A.R. Fonollosa. 2006. Statistical machine reordering. In *EMNLP*, pages 71–77, Sydney, July. ACL.

J.M. Crego and J.B. Mariño. 2006. Reordering experiments for n-gram-based smt. In *SLT*, pages 242–245, Aruba.

Josep M. Crego, Adrià de Gispert, Patrik Lambert, Marta R. Costa-jussà, Maxim Khalilov, Rafael Banchs, José B. Mariño, and José A. R. Fonollosa. 2006. N-gram-based smt system enhanced with reordering patterns. In *WMT*, pages 162–165, New York City, June. ACL.

Adrià de Gispert. 2006. *Introducing Linguistic Knowledge in Statistical Machine Translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, December.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between european languages. In *WMT*, pages 119–124, Michigan, June. ACL.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *WMT*, pages 102–121, New York City, June. ACL.

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. 1999. An efficient method for determining bilingual word classes. In *EACL*, pages 71–76, Bergen, Norway, June.