TALP Phrase-based statistical translation system for European language pairs

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

Abstract

This paper reports translation results for the "Exploiting Parallel Texts for Statistical Machine Translation" (HLT-NAACL Workshop on Parallel Texts 2006). We have studied different techniques to improve the standard Phrase-Based translation system. Mainly we introduce two reordering approaches and add morphological information.

1 Introduction

Nowadays most Statistical Machine Translation (SMT) systems use phrases as translation units. In addition, the decision rule is commonly modelled through a log-linear maximum entropy framework which is based on several feature functions (including the translation model), h_m. Each feature function models the probability that a sentence e in the target language is a translation of a given sentence f in the source language. The weights, λ_i, of each feature function are typically optimized to maximize a scoring function. It has the advantage that additional features functions can be easily integrated in the overall system.

This paper describes a Phrase-Based system whose baseline is similar to the system in Costa-jussà and Fonollosa (2005). Here we introduce two reordering approaches and add morphological information. Translation results for all six translation directions proposed in the shared task are presented and discussed. More specifically, four different languages are considered: English (en), Spanish (es), French (fr) and German (de); and both translation directions are considered for the pairs: **EnEs**, **EnFr**, and **EnDe**. The paper is organized as follows: Section 2 describes the system; Section 3 presents the shared task results; and, finally, in Section 4, we conclude.

[0]This work has been supported by the European Union under grant FP6-506738 (TC-STAR project) and the TALP Research Center (under a TALP-UPC-Recerca grant).

2 System Description

This section describes the system procedure followed for the data provided.

2.1 Alignment

Given a bilingual corpus, we use GIZA++ (Och, 2003) as word alignment core algorithm. During word alignment, we use 50 classes per language estimated by 'mkcls', a freely-available tool along with GIZA++. Before aligning we work with lowercase text (which leads to an Alignment Error Rate reduction) and we recover truecase after the alignment is done.

In addition, the alignment (in specific pairs of languages) was improved using two strategies:

Full verb forms The morphology of the verbs usually differs in each language. Therefore, it is interesting to classify the verbs in order to address the rich variety of verbal forms. Each verb is reduced into its base form and reduced POS tag as explained in (de Gispert, 2005). This transformation is only done for the alignment, and its goal is to simplify the work of the word alignment improving its quality.

Block reordering (*br*) The difference in word order between two languages is one of the most significant sources of error in SMT. Related works either deal with reordering in general as (Kanthak et al., 2005) or deal with local reordering as (Tillmann and Ney, 2003). We report a local reordering technique, which is implemented as a preprocessing stage, with two applications: (1) to improve only alignment quality, and (2) to improve alignment quality and to infer reordering in translation. Here, we present a short explanation of the algorithm, for further details see Costa-jussà and Fonollosa (2006).

Proceedings of the Workshop on Statistical Machine Translation, pages 142–145,
New York City, June 2006.

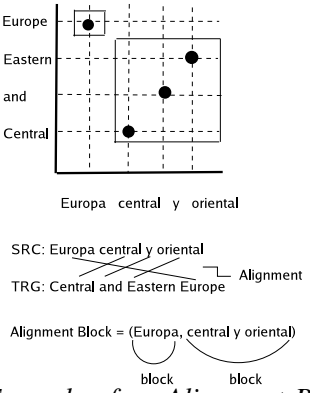

Figure 1: *Example of an Alignment Block, i.e. a pair of consecutive blocks whose target translation is swapped*

This reordering strategy is intended to infer the most probable reordering for sequences of words, which are referred to as blocks, in order to monotonize current data alignments and generalize reordering for unseen pairs of blocks.

Given a word alignment, we identify those pairs of consecutive source blocks whose translation is swapped, i.e. those blocks which, if swapped, generate a correct monotone translation. Figure 1 shows an example of these pairs (hereinafter called Alignment Blocks).

Then, the list of Alignment Blocks (LAB) is processed in order to decide whether two consecutive blocks have to be reordered or not. By using the classification algorithm, see the Appendix, we divide the LAB in groups ($G_n, n = 1 \dots N$). Inside the same group, we allow new internal combination in order to generalize the reordering to unseen pairs of blocks (i.e. new Alignment Blocks are created). Based on this information, the source side of the bilingual corpora are reordered.

In case of applying the reordering technique for purpose (1), we modify only the source training corpora to realign and then we recover the original order of the training corpora. In case of using Block Reordering for purpose (2), we modify all the source corpora (both training and test), and we use the new training corpora to realign and build the final translation system.

2.2 Phrase Extraction

Given a sentence pair and a corresponding word alignment, phrases are extracted following the criterion in Och and Ney (2004). A phrase (or bilingual phrase) is any pair of m source words and n target words that satisfies two basic constraints: words are consecutive along both sides of the bilingual phrase, and no word on either side of the phrase is aligned to a word out of the phrase. We limit the maximum size of any given phrase to 7. The huge increase in computational and storage cost of including longer phrases does not provide a significant improvement in quality (Koehn et al., 2003) as the probability of reappearance of larger phrases decreases.

2.3 Feature functions

Conditional and posterior probability (*cp, pp*) Given the collected phrase pairs, we estimate the phrase translation probability distribution by relative frequency in both directions.

The target language model (*lm*) consists of an n-gram model, in which the probability of a translation hypothesis is approximated by the product of word n-gram probabilities. As default language model feature, we use a standard word-based 5-gram language model generated with Kneser-Ney smoothing and interpolation of higher and lower order n-grams (Stolcke, 2002).

The POS target language model (*tpos*) consists of an N-gram language model estimated over the same target-side of the training corpus but using POS tags instead of raw words.

The forward and backwards lexicon models (*ibm1, ibm1*$^{-1}$) provide lexicon translation probabilities for each phrase based on the word IBM model 1 probabilities. For computing the forward lexicon model, IBM model 1 probabilities from GIZA++ source-to-target alignments are used. In the case of the backwards lexicon model, target-to-source alignments are used instead.

The word bonus model (*wb*) introduces a sentence length bonus in order to compensate the system preference for short output sentences.

The phrase bonus model (*pb*) introduces a constant bonus per produced phrase.

2.4 Decoding

The search engine for this translation system is described in Crego et al. (2005) which takes into account the features described above.

Using reordering in the decoder (*rgraph*) A highly constrained reordered search is performed by means of a set of reordering patterns (linguistically motivated rewrite patterns) which are used to

extend the monotone search graph with additional arcs. See the details in Crego et al. (2006).

2.5 Optimization

It is based on a simplex method (Nelder and Mead, 1965). This algorithm adjusts the log-linear weights in order to maximize a non-linear combination of translation BLEU and NIST: $10 * log10((BLEU * 100) + 1) + NIST$. The maximization is done over the provided development set for each of the six translation directions under consideration. We have experimented an improvement in the coherence between all the automatic figures by integrating two of these figures in the optimization function.

3 Shared Task Results

3.1 Data

The data provided for this shared task corresponds to a subset of the official transcriptions of the European Parliament Plenary Sessions, and it is available through the shared task website at: `http://www.statmt.org/wmt06/shared-task/`. The development set used to tune the system consists of a subset (500 first sentences) of the official development set made available for the Shared Task.

We carried out a morphological analysis of the data. The English POS-tagging has been carried out using freely available TNT tagger (Brants, 2000). In the Spanish case, we have used the $Freeling$ (Carreras et al., 2004) analysis tool which generates the POS-tagging for each input word.

3.2 Systems configurations

The baseline system is the same for all tasks and includes the following features functions: *cp, pp, lm, ibm1,* $ibm1^{-1}$*, wb, pb*. The POStag target language model has been used in those tasks for which the tagger was available. Table 1 shows the reordering configuration used for each task.

The Block Reordering (application 2) has been used when the source language belongs to the Romanic family. The length of the block is limited to 1 (i.e. it allows the swapping of single words). The main reason is that specific errors are solved in the tasks from a Romanic language to a Germanic language (as the common reorder of $Noun + Adjective$ that turns into $Adjective + Noun$). Although the Block Reordering approach does not depend on the task, we have not done the corresponding experiments to observe its efficiency in all the pairs used in this evaluation.

Task	Reordering Configuration
Es2En	*br*2
En2Es	*br*1 + *rgraph*
Fr2En	*br*2
En2Fr	*br*1 + *rgraph*
De2En	-
En2De	-

Table 1: *Additional reordering models for each task: br1 (br2) stands for Block Reordering application 1 (application 2); and rgraph refers to the reordering integrated in the decoder*

The *rgraph* has been applied in those cases where: we do not use *br*2 (there is no sense in applying them simultaneously); and we have the tagger for the source language model available.

In the case of the pair GeEn, we have not experimented any reordering, we left the application of both reordering approaches as future work.

3.3 Discussion

Table 2 presents the BLEU scores evaluated on the test set (using TRUECASE) for each configuration. The official results were slightly better because a lowercase evaluation was used, see (Koehn and Monz, 2006).

For both, Es2En and Fr2En tasks, *br* helps slightly. The improvement of the approach depends on the quality of the alignment. The better alignments allow to extract higher quality Alignment Blocks (Costa-jussà and Fonollosa, 2006).

The En2Es task is improved when adding both *br*1 and *rgraph*. Similarly, the En2Fr task seems to perform fairly well when using the *rgraph*. In this case, the improvement of the approach depends on the quality of the alignment patterns (Crego et al., 2006). However, it has the advantage of delaying the final decision of reordering to the overall search, where all models are used to take a fully informed decision.

Finally, the *tpos* does not help much when translating to English. It is not surprising because it was used in order to improve the gender and number agreement, and in English there is no need. However, in the direction to Spanish, the *tpos* added to the corresponding reordering helps more as the Spanish language has gender and number agreement.

Task	Baseline	+tpos	+rc	+tpos+rc
Es2En	29.08	29.08	29.89	**29.98**
En2Es	27.73	27.66	28.79	**28.99**
Fr2En	27.05	27.06	27.43	**27.23**
En2Fr	26.16	-	27.80	-
De2En	21.59	**21.33**	-	-
En2De	**15.20**	-	-	-

Table 2: *Results evaluated using* TRUECASE *on the test set for each configuration: rc stands for Reordering Configuration and refers to Table 1. The bold results were the configurations submitted.*

4 Conclusions

Reordering is important when using a Phrase-Based system. Although local reordering is supposed to be included in the phrase structure, performing local reordering improves the translation quality. In fact, local reordering, provided by the reordering approaches, allows for those generalizations which phrases could not achieve. Reordering in the DeEn task is left as further work.

References

T. Brants. 2000. Tnt - a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing*.

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

M. R. Costa-jussà and J.A.R. Fonollosa. 2005. Improving the phrase-based statistical translation by modifying phrase extraction and including new features. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*.

M. R. Costa-jussà and J.A.R. Fonollosa. 2006. Using reordering in statistical machine translation based on alignment block classification. *Internal Report*.

J.M. Crego, J. Mariño, and A. de Gispert. 2005. An Ngram-based statistical machine translation decoder. *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'05*.

J. M. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, M. Khalilov, J. Mariño, J. A. Fonollosa, and R. Banchs. 2006. Ngram-based smt system enhanced with reordering patterns. *HLT-NAACL06 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, June.

A. de Gispert. 2005. Phrase linguistic classification for improving statistical machine translation. *ACL 2005 Students Workshop*, June.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, June.

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. June.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, May.

J.A. Nelder and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

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

F.J. Och. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

C. Tillmann and H. Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, March.

A Appendix

Here we describe the classification algorithm used in Section 1.

1. Initialization: set $n \leftarrow 1$ and $LAB' \leftarrow LAB$.
2. Main part: while LAB' is not empty do
   - $G_n = \{(\alpha_k, \beta_k)\}$ where (α_k, β_k) is any element of LAB', i.e. α_k is the first block and β_k is the second block of the Alignment Block k of the LAB'.
   - Recursively, move elements (α_i, β_i) from LAB' to G_n if there is an element $(\alpha_j, \beta_j) \in G_n$ such that $\alpha_i = \alpha_j$ or $\beta_i = \beta_j$
   - Increase n (i.e. $n \leftarrow n + 1$)
3. Ending: For each G_n, construct the two sets A_n and B_n which consists on the first and second element of the pairs in G_n, respectively.