N-gram-based SMT System Enhanced with Reordering Patterns

Josep M. Crego
Marta R. Costa-jussà
José B. Mariño

Adrià de Gispert
Maxim Khalilov
José A. R. Fonollosa

Patrik Lambert
Rafael E. Banchs

Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
{jmcrego,agispert,lambert,mruiz,khalilov,rbanchs,canton,adrian}@gps.tsc.upc.edu

Abstract

This work presents translation results for the three data sets made available in the shared task *"Exploiting Parallel Texts for Statistical Machine Translation"* of the HLT-NAACL 2006 Workshop on Statistical Machine Translation. All results presented were generated by using the N-gram-based statistical machine translation system which has been enhanced from the last year's evaluation with a tagged target language model (using Part-Of-Speech tags). For both Spanish-English translation directions and the English-to-French translation task, the baseline system allows for linguistically motivated source-side reorderings.

1 Introduction

The statistical machine translation approach used in this work implements a log-linear combination of feature functions along with a translation model which is based on bilingual n-grams (de Gispert and Mariño, 2002).

This translation model differs from the well known phrase-based translation approach (Koehn et al., 2003) in two basic issues: first, training data is monotonously segmented into bilingual units; and second, the model considers n-gram probabilities instead of relative frequencies. This translation approach is described in detail in (Mariño et al., 2005).

For those translation tasks with Spanish or English as target language, an additional tagged (using POS information) target language model is used. Additionally a reordering strategy that includes POS information is described and evaluated.

Translation results for all six translation directions proposed in the shared task are presented and discussed. Both translation directions are considered for the pairs: **English-Spanish**, **English-French**, and **English-German**.

The paper is structured as follows: Section 2 briefly outlines the baseline system. Section 3 describes in detail the implemented POS-based reordering strategy. Section 4 presents and discusses the shared task results and, finally, section 5 presents some conclusions and further work.

2 Baseline N-gram-based SMT System

As already mentioned, the translation model used here 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 (de Gispert and Mariño, 2002).

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. See (Crego et al., 2004) for further details.

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 five feature functions.

These five additional models are:

- A **target language model**. 5-gram of the target side of the bilingual corpus.
- A **word bonus**. Based on the number of target words in the partial-translation hypothesis, to compensate the LM preference for short sentences.
- A **Source-to-target lexicon model**. Based on IBM Model 1 lexical parameters(Brown et al., 1993), providing a complementary probability for each tuple in the translation table. These parameters are obtained from source-to-target alignments.
- A **Target-to-source lexicon model**. Analogous to the previous feature, but obtained from target-to-source alignments.
- A **Tagged (POS) target language model**. This feature implements a 5-gram language model of target POS-tags. In this case, each translation unit carried the information of its target side POS-tags, though this is not used for translation model estimation (only in order to evaluate the target POS language model at decoding time). Due to the non-availability of POS-taggers for French and German, it was not possible to incorporate this feature in all translation tasks considered, being only used for those translation tasks with Spanish and English as target languages.

The search engine for this translation system is described in (Crego et al., 2005) and implements a beam-search strategy based on dynamic programming, taking into account all feature functions described above, along with the bilingual n-gram translation model. Monotone search is performed, including histogram and threshold pruning and hypothesis recombination.

An optimization tool, which is based on a downhill simplex method was developed and used for computing log-linear weights for each of the feature functions. This algorithm adjusts the weights so that a non-linear combination of BLEU *and* NIST scores is maximized over the development set for each of the six translation directions considered.

This baseline system is actually very similar to the system used for last year's shared task *"Exploiting Parallel Texts for Statistical Machine Translation"* of ACL'05 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond (Banchs et al., 2005), whose results are available at: `http://www.statmt.org/wpt05/mt-shared-task/`. A more detailed description of the system can be found in (2005).

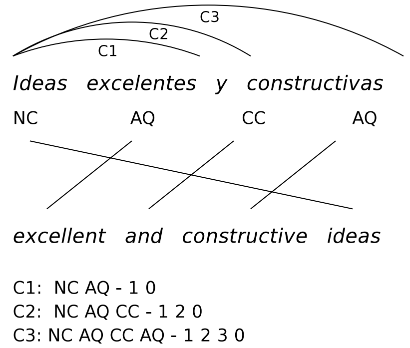
The tools used for POS-tagging were Freeling (Carreras et al., 2004) for Spanish and TnT (Brants, 2000) for English. All language models were estimated using the SRI language modeling toolkit. Word-to-word alignments were extracted with GIZA++. Improvements in word-to-word alignments were achieved through verb group classification as described in (de Gispert, 2005).

3 Reordering Framework

In this section we outline the reordering framework used for the experiments (Crego and Mariño, 2006). 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.

To extract patterns, we use the word-to-word alignments (the union of both alignment directions) and source-side POS tags. The main procedure consists of identifying all crossings produced in the

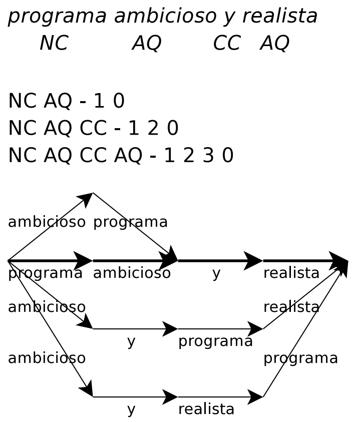
Figure 1: *Reordering patterns are extracted using word-to-word alignments. The generalization power is achieved through the POS tags. Three instances of different patterns are extracted using the sentences in the example.*



word-to-word alignments. Once a crossing has been detected, its source POS tags and alignments are used to account for a new instance of pattern. The target side of a pattern (source-side positions after reordering), is computed using the original order of the target words to which the source words are aligned. See figure 1 for a clarifying example of pattern extraction.

The monotone search graph is extended with reorderings following the patterns found in training. 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 learnt in the pattern). However, this addition of a new arc is not performed if a translation unit with the same source-side words already exists in the training. Figure 2 shows an example of the procedure.

Figure 2: *Three additional arcs have been added to the original monotone graph (bold arcs) given the reordering patterns found matching any of the source POS tags sequence.*



Once the search graph is built, the decoder traverses the graph looking for the best translation. Hence, the winner hypothesis is computed using all the available information (the whole SMT models). The reordering strategy is additionally supported by a **5-gram language model of reordered source POS-tags**. In training, POS-tags are reordered according with the extracted reordering patterns and word-to-word links. The resulting sequence of source POS-tags are used to train the n-gram LM.

Notice that this reordering framework has only been used for some translation tasks (Spanish-to-English, English-to-Spanish and English-to-French). The reason is double: first, because we did not have available a French POS-tagger. Second, because the technique used to learn reorderings (detailed below) does not seem to apply for language pairs like German-English, because the agglutinative characteristic of German (words are formed by joining morphemes together).

Table 1: *BLEU, NIST and mWER scores (computed using **two** reference translations) obtained for both translation directions (Spanish-to-English and English-to-Spanish).*

Conf	BLEU	NIST	mWER
Spanish-to-English			
base	55.23	10.69	34.40
+rgraph	55.59	10.70	34.23
+pos	56.39	10.75	33.75
English-to-Spanish			
base	48.03	9.84	41.18
+rgraph	48.53	9.81	41.15
+pos	48.91	9.91	40.29

Table 1 shows the improvement of the original baseline system described in section 2 (**base**), enhanced using reordering graphs (**+rgraph**) and provided the tagged-source language model (**+pos**). The experiments in table 1 were not carried out over the official corpus of this shared task. The Spanish-English corpus of the TC-Star 2005 Evaluation was used. Due to the high similarities between both corpus (this shared task corpus consists of a subset of the whole corpus used in the TC-Star 2005 Evaluation), it makes sense to think that comparable results would be obtained.

It is worth mentioning that the official corpus of the shared task (HLT-NAACL 2006) was used when building and tuning the present shared task system.

4 Shared Task Results

The data provided for this shared task corresponds to a subset of the official transcriptions of the European Parliament Plenary Sessions. 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.

Table 2 presents the *BLEU*, *NIST* and *mWER* scores obtained for the development-test data set. The last column shows whether the target POS language model feature was used or not. Computed scores are case sensitive and compare to **one** reference translation. Tasks in bold were conducted allowing for the reordering framework. For French-to-English task, block reordering strategy was used, which is described in (Costa-jussà et al., 2006). As it can be seen, for the English-to-German task we did not use any of the previous enhancements.

Table 2: *Translation results*

Task	BLEU	NIST	mWER	tPOS
en → es	29.50	7.32	58.95	yes
es → en	30.29	7.51	57.72	yes
en → fr	30.23	7.40	59.76	no
fr → en	30.21	7.61	56.97	yes
en → de	17.40	5.61	71.18	no
de → en	23.78	6.70	65.83	yes

Important differences can be observed between the German-English and the rest of translation tasks. They result from the greater differences in word order present in this language pair (the German-English results are obtained under monotone decoding conditions). Also because the greater vocabulary of words of German, which increases sparseness in any task where German is envolved. As expected, differences in translation accuracy between Spanish-English and French-English are smaller.

5 Conclusions and Further Work

As it can be concluded from the presented results, although in principle some language pairs (Spanish-English-French) seem to have very little need for reorderings (due to their similar word order), the use of linguistically-based reorderings proves to be useful to improve translation accuracy.

Additional work is to be conducted to allow for reorderings when translating from/to German.

6 Acknowledgments

This work was partly funded by the European Union under the integrated project TC-STAR[1]: Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738) and the European Social Fund.

[1] http://www.tc-star.org

References

R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, and J. B. Mariño. 2005. Statistical machine translation of euparl data by using bilingual n-grams. *Proc. of the ACL Workshop on Building and Using Parallel Texts (ACL'05/Wkshp)*, pages 67–72, June.

T. Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

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*, May.

M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, R. Banchs, J.B. Mariño, and J.A.R. Fonollosa. 2006. Talp phrase-based statistical translation system for european language pairs. *Proc. of the HLT/NAACL Workshop on Statistical Machine Translation*, June.

J. M. Crego and J. Mariño. 2006. A reordering framework for statistical machine translation. *Internal Report*.

J. M. Crego, J. Mariño, and A. de Gispert. 2004. Finite-state-based and phrase-based statistical machine translation. *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, pages 37–40, October.

J. M. Crego, J. Mariño, and A. Gispert. 2005. An ngram-based statistical machine translation decoder. *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, September.

A. de Gispert and J. Mariño. 2002. Using X-grams for speech-to-speech translation. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

A. de Gispert. 2005. Phrase linguistic classification and generalization for improving statistical machine translation. *Proc. of the ACL Student Research Workshop (ACL'05/SRW)*, 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.B. Mariño, R Banchs, J.M. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, and J.A.R. Fonollosa. 2005. Bilingual n–gram statistical machine translation. *Proc. of the MT Summit X*, September.