Statistical Machine Translation of Euparl Data by using Bilingual N-grams

Rafael E. Banchs **Josep M. Crego** **Adrià de Gispert** **Patrik Lambert** **José B. Mariño**
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona 08034, Spain
{rbanchs,jmcrego,agispert,lambert,canton}@gps.tsc.upc.edu

Abstract

This work discusses translation results for the four Euparl data sets which were made available for the shared task *"Exploiting Parallel Texts for Statistical Machine Translation"*. All results presented were generated by using a statistical machine translation system which implements a log-linear combination of feature functions along with a bilingual n-gram translation model.

1 Introduction

During the last decade, statistical machine translation (SMT) systems have evolved from the original word-based approach (Brown *et al.*, 1993) into phrase-based translation systems (Koehn *et al.*, 2003). Similarly, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple models is implemented (Och and Ney, 2002).

The SMT 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. This translation model was developed by de Gispert and Mariño (2002), and it differs from the well known phrase-based translation model 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 model is described in section 2.

Translation results from the four source languages made available for the shared task (es: Spanish, fr: French, de: German, and fi: Finnish) into English (en) are presented and discussed.

The paper is structured as follows. Section 2 describes the bilingual n-gram translation model. Section 3 presents a brief overview of the whole SMT procedure. Section 4 presents and discusses the shared task results and other interesting experimentation. Finally, section 5 presents some conclusions and further work.

2 Bilingual N-gram Translation Model

As already mentioned, the translation model used here is based on bilingual n-grams. It actually constitutes a language model of bilingual units which are referred to as tuples (de Gispert and Mariño, 2002). This model approximates the joint probability between source and target languages by using 3-grams as it is described in the following equation:

$$p(T,S) \approx \prod_{n=1}^{N} p((t,s)_n | (t,s)_{n-2}, (t,s)_{n-1}) \quad (1)$$

where t refers to target, s to source and $(t,s)_n$ to the n^{th} tuple of a given bilingual sentence pair.

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, the produced segmentation is maximal in the sense that no smaller tuples can be extracted without violating the previous constraint (Crego *et al.*, 2004). According to this, tuple extraction provides a unique segmentation for a given bilingual sentence pair alignment. Figure 1 illustrates this idea with a simple example.

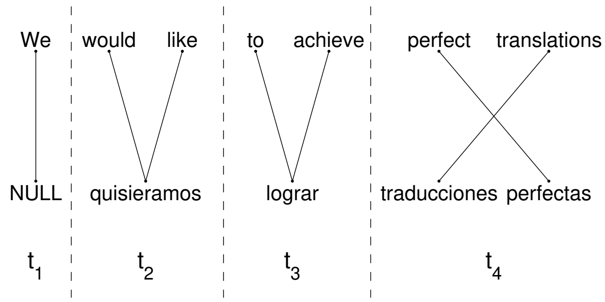

Figure 1: *Example of tuple extraction from an aligned sentence pair.*

Two important issues regarding this translation model must be mentioned. First, when extracting tuples, some words always appear embedded into tuples containing two or more words, so no translation probability for an independent occurrence of such words exists. To overcome this problem, the tuple 3-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words (de Gispert *et al.*, 2004).

Second, some words linked to NULL end up producing tuples with NULL source sides. This cannot be allowed since no NULL is expected to occur in a translation input. This problem is solved by preprocessing alignments before tuple extraction such that any target word that is linked to NULL is attached to either its precedent or its following word.

3 SMT Procedure Description

This section describes the procedure followed for preprocessing the data, training the models and optimizing the translation system parameters.

3.1 Preprocessing and Alignment

The Euparl data provided for this shared task (Euparl, 2003) was preprocessed for eliminating all sentence pairs with a word ratio larger than 2.4. As a result of this preprocessing, the number of sentences in each training set was slightly reduced. However, no significant reduction was produced.

In the case of French, a re-tokenizing procedure was performed in which all apostrophes appearing alone were attached to their corresponding words. For example, pairs of tokens such as *l ’* and *qu ’* were reduced to single tokens such as *l’* and *qu’*.

Once the training data was preprocessed, a word-to-word alignment was performed in both directions, source-to-target and target-to-source, by using GIZA++ (Och and Ney, 2000). As an approximation to the most probable alignment, the Viterbi alignment was considered. Then, the intersection and union of alignment sets in both directions were computed for each training set.

3.2 Feature Function Computation

The considered translation system implements a total of five feature functions. The first of these models is the tuple 3-gram model, which was already described in section 2. Tuples for the translation model were extracted from the union set of alignments as shown in Figure 1. Once tuples had been extracted, the tuple vocabulary was pruned by using histogram pruning. The same pruning parameter, which was actually estimated for Spanish-English, was used for the other three language pairs. After pruning, the tuple 3-gram model was trained by using the SRI Language Modeling toolkit (Stolcke, 2002). Finally, the obtained model was enhanced by incorporating 1-gram probabilities for the embedded word tuples, which were extracted from the intersection set of alignments.

Table 1 presents the total number of running words, distinct tokens and tuples, for each of the four training data sets.

Table 1: *Total number of running words, distinct tokens and tuples in training.*

source language	running words	distinct tokens	tuple vocabulary
Spanish	15670801	113570	1288770
French	14844465	78408	1173424
German	15207550	204949	1391425
Finnish	11228947	389223	1496417

The second feature function considered was a target language model. This feature actually consisted of a word 3-gram model, which was trained from the target side of the bilingual corpus by using the SRI Language Modeling toolkit.

The third feature function was given by a word penalty model. This function introduces a sentence length penalization in order to compensate the system preference for short output sentences. More specifically, the penalization factor was given by the total number of words contained in the translation hypothesis.

Finally, the fourth and fifth feature functions corresponded to two lexicon models based on IBM Model 1 lexical parameters $p(t|s)$ (Brown *et al.*, 1993). These lexicon models were calculated for each tuple according to the following equation:

$$p_{lexicon}((t,s)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i | s_n^j) \quad (2)$$

where s_n^j and t_n^i are the j^{th} and i^{th} words in the source and target sides of tuple $(t,s)_n$, being J and I the corresponding total number words in each side of it.

The forward lexicon model uses IBM Model 1 parameters obtained from source-to-target alignments, while the backward lexicon model uses parameters obtained from target-to-source alignments.

3.3 Decoding and Optimization

The search engine for this translation system was developed by Crego *et al.* (2005). It implements a beam-search strategy based on dynamic programming and takes into account all the five feature functions described above simultaneously. It also allows for three different pruning methods: threshold pruning, histogram pruning, and hypothesis recombination. For all the results presented in this work the decoder's monotonic search modality was used.

An optimization tool, which is based on a simplex method (Press *et al.*, 2002), was developed and used for computing log-linear weights for each of the feature functions described above. This algorithm adjusts the log-linear weights so that *BLEU* (Papineni *et al.*, 2002) is maximized over a given development set. One optimization for each language pair was performed by using the 2000-sentence development sets made available for the shared task.

4 Shared Task Results

Table 2 presents the *BLEU* scores obtained for the shared task test data. Each test set consisted of 2000 sentences. The computed *BLEU* scores were case insensitive and used one translation reference.

Table 2: *BLEU scores (shared task test sets).*

es - en	fr - en	de - en	fi - en
0.3007	0.3020	0.2426	0.2031

As can be seen from Table 2 the best ranked translations were those obtained for French, followed by Spanish, German and Finnish. A big difference is observed between the best and the worst results.

Differences can be observed from translation outputs too. Consider, for example, the following segments taken from one of the test sentences:

es-en: *We know very well that the present Treaties are not enough and that , in the future , it will be necessary to develop a structure better and different for the European Union...*

fr-en: *We know very well that the Treaties in their current are not enough and that it will be necessary for the future to develop a structure more effective and different for the Union...*

de-en: *We very much aware that the relevant treaties are inadequate and , in future to another , more efficient structure for the European Union that must be developed...*

fi-en: *We know full well that the current Treaties are not sufficient and that , in the future , it is necessary to develop the Union better and a different structure...*

It is evident from these translation outputs that translation quality decreases when moving from Spanish and French to German and Finnish. A detailed observation of translation outputs reveals that there are basically two problems related to this degradation in quality. The first has to do with reordering, which seems to be affecting Finnish and, specially, German translations.

The second problem has to do with vocabulary. It is well known that large vocabularies produce data sparseness problems (Koehn, 2002). As can be confirmed from Tables 1 and 2, translation quality decreases as vocabulary size increases. However, it is not clear yet, in which degree such degradation is due to monotonic decoding and/or vocabulary size.

Finally, we also evaluated how much the full feature function system differs from the baseline tuple 3-gram model alone. In this way, *BLEU* scores were computed for translation outputs obtained for the baseline system and the full system. Since the English reference for the test set was not available, we computed translations and *BLEU* scores over development sets. Table 3 presents the results for both the full system and the baseline.[1]

Table 3: *Baseline- and full-system BLEU scores (computed over development sets).*

language pair	baseline	full
es - en	0.2588	0.3004
fr - en	0.2547	0.2938
de - en	0.1844	0.2350
fi - en	0.1526	0.1989

From Table 3, it is evident that the four additional feature functions produce important improvements in translation quality.

5 Conclusions and Further Work

As can be concluded from the presented results, performance of the translation system used is much better for French and Spanish than for German and Finnish. As some results suggest, reordering and vocabulary size are the most important problems related to the low translation quality achieved for German and Finnish.

It is also evident that the bilingual n-gram model used requires the additional feature functions to produce better translations. However, more experimentation is required in order to fully understand each individual feature's influence on the overall log-linear model performance.

6 Acknowledgments

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, `http://www.tc-star.org`).

The authors also want to thank José A. R. Fonollosa and Marta Ruiz Costa-jussà for their participation in discussions related to this work.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. "The mathematics of statistical machine translation: parameter estimation". *Computational Linguistics*, 19(2):263–311.

Josep M. Crego, José B. Mariño, and Adrià de Gispert. 2004. "Finite-state-based and phrase-based statistical machine translation". *Proc. of the 8th Int. Conf. on Spoken Language Processing*, :37–40, October.

Josep M. Crego, José B. Mariño, and Adrià de Gispert. 2005. "A Ngram-based Statistical Machine Translation Decoder". Submitted to INTERSPEECH 2005.

Adrià de Gispert, and José B. Mariño. 2002. "Using X-grams for speech-to-speech translation". *Proc. of the 7th Int. Conf. on Spoken Language Processing*.

Adrià de Gispert, José B. Mariño, and Josep M. Crego. 2004. "TALP: Xgram-based spoken language translation system". *Proc. of the Int. Workshop on Spoken Language Translation*, :85–90. Kyoto, Japan, October.

EUPARL: European Parliament Proceedings Parallel Corpus 1996-2003. Available on-line at: `http://people.csail.mit.edu/people/koehn/publications/europarl/`

Philipp Koehn. 2002. "Europarl: A Multilingual Corpus for Evaluation of Machine Translation". Available on-line at: `http://people.csail.mit.edu/people/koehn/publications/europarl/`

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. "Statistical phrase-based translation". *Proc. of the 2003 Meeting of the North American chapter of the ACL*, Edmonton, Alberta.

Franz J. Och and Hermann Ney. 2000. "Improved statistical alignment models". *Proc. of the 38th Ann. Meeting of the ACL*, Hong Kong, China, October.

Franz J. Och and Hermann Ney. 2002. "Discriminative training and maximum entropy models for statistical machine translation". *Proc. of the 40th Ann. Meeting of the ACL*, :295–302, Philadelphia, PA, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: a method for automatic evaluation of machine translation". *Proc. of the 40th Ann. Conf. of the ACL*, Philadelphia, PA, July.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++: the Art of Scientific Computing*, Cambridge University Press.

Andreas Stolcke. 2002. "SRLIM: an extensible language modeling toolkit". *Proc. of the Int. Conf. on Spoken Language Processing* :901–904, Denver, CO, September. Available on line at: `http://www.speech.sri.com/projects/srilm/`

[1]Differently from *BLEU* scores presented in Table 2, which are case insensitive, *BLEU* scores presented in Table 3 are case sensitive.