UPC's Bilingual N-gram Translation System

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

Abstract

This paper describes UPC's bilingual n-gram approach to statistical machine translation, which consists of a log-linear combination of a bilingual n-gram translation model, along with other six feature functions. Translation results for the Spanish-to-English and English-to-Spanish tasks considered during TC-STAR's second evaluation campaign are presented and discussed.

1 Introduction

The UPC's statistical machine translation approach implements a log-linear combination of feature functions (Och and Ney, 2002) along with a translation model which is based on bilingual n-grams (de Gispert and Mariño, 2002). This translation model differs from the phrase-based translation approach (Koehn *et al.*, 2003) in two basic issues: training data is monotonously segmented and the model considers n-gram probabilities instead of relative frequencies. The original version of the system (Mariño *et al.*, 2005) implemented four feature functions along with the bilingual n-gram model. In addition, some novel feature functions and reordering strategies that consider POS information are presented here.

Translation results for three of the four tasks considered during TC-STAR's second evaluation campaign are presented and discussed. More specifically, these tasks are: EPPS[1] Spanish-to-English, EPPS English-to-Spanish and CORTES[2] Spanish-to-English. For each one of these tasks, three different translation conditions were considered: final text edition, verbatim transcriptions and automatic speech recognition.

The paper is structured as follows. Section 2 describes the bilingual n-gram translation model and section 3 describes all feature functions and reordering strategies implemented. Section 4 presents and discusses the translation experiments and their results and, finally, section 5 presents some conclusions and further work.

2 UPC's Translation Model

The UPC's translation model has been derived from the finite-state perspective; more specifically, from the work of Casacuberta (2001; 2004). However, different from it, where the translation model is implemented by using a finite-state transducer, the UPC's system implements a bilingual 5-gram model. 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 5-grams (de Gispert and Mariño, 2002), such as described by the following equation:

$$p(T,S) \approx \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, \ldots, (t,s)_{k-4}) \tag{1}$$

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

[1] European Parliament Plenary Sessions

[2] Spanish Parliament Speeches

Tuples are extracted from from Viterbi alignments 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 (Crego *et al.*, 2004). According to this, tuples can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual corpus. Figure 1 presents a simple example illustrating the unique tuple segmentation for a given pair of sentences.

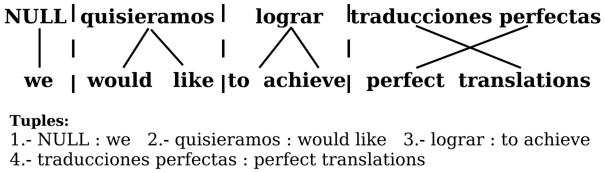

Figure 1: Example of tuple extraction.

Two important issues regarding this translation model must be considered. First, it often occurs that an important amount of single-word translation probabilities are left out of the model. This happens for all those words that appear always embedded into tuples containing two or more words. Consider for example the word “translations” from figure 1. As seen from the figure, “translations” appears embedded into tuple 4. If a similar situation is encountered for all occurrences of “translations” in the training corpus then no translation probability for an independent occurrence of such word will exist.

To overcome this problem, the tuple 5-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words detected during the tuple extraction step (de Gispert *et al.*, 2004). While tuples are extracted from the union of both, source-to-target and target-to-source, these embedded-word translation probabilities are computed from the intersection of alignments.

The second important issue has to do with the fact that some words linked to NULL end up producing tuples with NULL source sides. Consider for example the tuple 1 from figure 1. Since no NULL is actually expected to occur in translation inputs, such a kind of tuple cannot be allowed. This problem is solved by preprocessing the union set of alignments before the tuple extraction is performed. During this preprocessing, any target word that is linked to NULL is attached to either its precedent word or its following word according to a weight based on IBM model1 (Crego *et al.*, 2005b). In this way, no target word remains linked to NULL, and tuples with NULL source side are not extracted.

3 Additional System’s Features

In addition to the bilingual n-gram translation model described in the previous section, the UPC’s translation system implements a log linear combination of six additional feature functions, which are described in detail within this section. All these six feature functions are taken into account, along with the translation model, during the decoding stage.

The search engine for UPC’s translation system was developed by Crego *et al.* (2005). It implements a beam-search strategy based on dynamic programming and allows for threshold pruning and hypothesis recombination. Additionally, an optimization tool, based on a simplex method (Press *et al.*, 2002), was developed and used for weighting each feature function contribution. This algorithm adjusts the log-linear weights so that a non-linear combination of translation *BLEU* (Papineni *et al.*, 2002) and *NIST* is maximized over the provided development set for each task under consideration.

3.1 Target Language Model

This feature provides information about the target language structure and fluency. It favors those partial-translation hypotheses which are more likely to constitute correctly structured target sentences over those which are not. The model is implemented by using a word 4-gram model of the target language, which is computed according to the following expression:

$$h_{TL}(T) = log \prod_{k=1}^{K} p(w_k|w_{k-1}, \ldots, w_{k-3}) \qquad (2)$$

where w_k refers to k^{th} word in the considered partial-translation hypothesis. Notice that this model only depends on the target side of the data, and actually it can be trained by including additional information from other available monolingual corpora.

3.2 Word Bonus Model

This feature introduces a bonus which depends on the partial-translation hypothesis length. This is done in order to compensate the system preference for short translations over large ones. The model is implemented through a bonus factor which directly depends on the total number of words contained in the partial-translation hypothesis, and it is computed as follows:

$$h_{WP}(T) = M \tag{3}$$

where M is the number of words contained in the partial-translation hypothesis.

3.3 Source-to-Target Lexicon Model

This feature actually constitutes a complementary translation model. This model provides, for a given tuple, a translation probability estimate between the source and target sides of it. This feature is implemented by using the IBM-1 lexical parameters (Brown *et al.*, 1993; Och *et al.*, 2004). According to this, the source-to-target lexicon probability is computed for each tuple according to the following equation:

$$h_{LF}(T,S) = log \frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} q(t_i^n | s_j^n) \tag{4}$$

where s_j^n and t_i^n 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 of words in each side of it. In the equation $q(.)$ refers to IBM-1 lexical parameters which are estimated from alignments computed in the source-to-target direction.

3.4 Target-to-Source Lexicon Model

Similar to the previous feature, this feature function constitutes a complementary translation model too. It is computed exactly in the same way the previous model is, with the only difference that IBM-1 lexical parameters are estimated from alignments computed in the target-to-source direction instead.

3.5 Target POS-tag Language Model

This feature implements a 5-gram language model of target POS-tags. This model is trained by considering POS-tags, instead of words, for 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 POS string containing the POS-tags corresponding to the words in the target strings.

It is important to mention that the POS information contained in the triplet is not actually used for computing the bilingual translation model probabilities described in section 2. This information is used only during decoding by the 5-gram language model of target POS-tags in order to score the alternative POS-tag sequences associated to the competing partial-translation hypothesis. In this way, this feature helps, along with the target language model, to provide a better concatenation of tuples during the decoding process. This procedure is illustrated in figure 2.

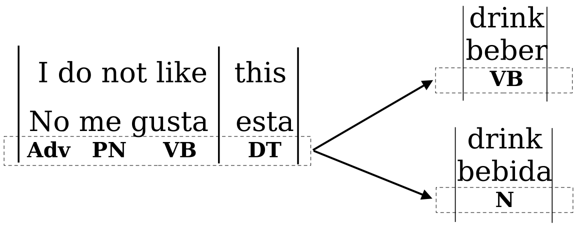

Figure 2: Augmented tuple and target POS-tag language model implementation.

3.6 Source POS-tag Language Model

This feature implements a word reordering strategy that is supported by a 5-gram language model of reordered source POS-tags. As a first step, during training, reordering patterns for the source POS-tags are learned from the aligned corpus. These reordering patterns are identified by looking at the link crossings occurring in the aligned corpus and are classified according to the corresponding POS-tags of the source words involved. Then all identified link crossings are unfolded by reordering the source words and their corresponding POS-tags while keeping the target side of the corpus untouched. From this reordered sequence of POS-tags, a 5-gram language model is trained.

As a second step, during translation, the input sentence to be translated is replaced by a word-graph by adding alternative paths based on the POS reordering patterns learned during training. Notice that this procedure requires tagging the input sentence to be translated. Then, the single sentence word-graph is augmented by adding as many paths as POS reordering patterns can be applied. In this way, the constructed word-graph is used as the decoder's input, and during decoding, the 5-gram language model of reordered source POS-tags is used. This model helps the decoder to select those more appropriate paths among all possible paths in the input word-graph. This procedure is illustrated in figure 3, and further described in Crego (2006).

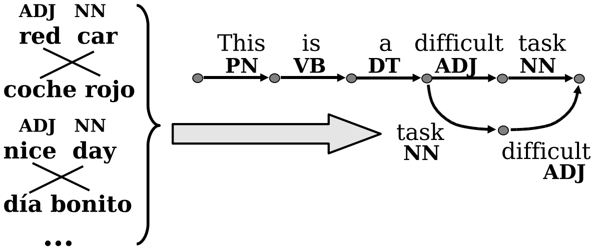

Figure 3: Example of word reordering based on source POS information.

3.7 Word Reordering Based on Alignment Block Classification

This constitutes an alternative word reordering strategy, which is implemented as a preprocessing stage instead of as while-decoding feature function. A detailed description for this method is provided in Costa-jussà and Fonollosa (2006). This strategy is intended to infer those most probable reorderings for sequences of words, which are referred to as blocks, in order to monotonize current data alignments and generalize reorderings for unseen pairs of blocks.

Given a word alignment, a list of Alignment Blocks, which consists of a pair of consecutive source blocks with swapped target translations, is extracted. Then, by using a classification algorithm, the list is processed in order to decide whether two consecutive blocks have to be reordered or not. Based on this information, the source sides of both bilingual training and development corpora, as well as the test data, are reordered. This modified training corpus is aligned once again and the translation model and feature function probabilities are computed from these new alignments. Finally, translations are computed from the modified input test data.

4 TC-STAR Second Evaluation Results

The data sets used for experiments presented here correspond to those provided by ELDA for the 2006 TC-STAR[3] Second Evaluation Campaign, which are available through the ELDA's website at: `http://www.elda.org/tcstar-workshop/2006eval.htm`.

Results for three different tasks are presented here: English-to-Spanish (EPPS), Spanish-to-English (EPPS) and Spanish-to-English (CORTES). For each one of these tasks, three different translation conditions were considered: final text edition, verbatim transcriptions, and automatic speech recognition (ASR). The final text edition condition corresponds to the official transcripts of the respective parliament sessions, so it is actually a written language translation condition. On the other hand, the other two conditions are spoken language translation conditions. More specifically, the verbatim condition corresponds to literal transcriptions of parliamentary speeches, which include hesitations, repeated words and other spontaneous speech effects; and the ASR output condition corresponds to the output of an automatic speech recognition system, so it additionally includes speech recognition errors.

Table 1 presents basic statistics for the training data, which existed only for the case of the EPPS tasks. The CORTES Spanish-to-English task was performed by using the EPPS training corpus. More specifically, the table presents the total number of sentences, the total number of running words and the vocabulary size for each language.

Table 1: *Basic statistics for the EPPS training data. The total number of sentences, words and the vocabulary size are provided in millions.*

Language	Sentences	Words	Vocabulary
English	1.28	34.9	0.106
Spanish	1.28	36.6	0.153

Table 2, on the other hand, presents the total num-

[3] TC-STAR (Technology and Corpora for Speech to Speech Translation) is an European Community project funded by the *Sixth Framework Programme*. More information can be found at the consortium website: `http://www.tc-star.org/`

ber of sentences contained in each condition's development and test data set, as well as the number of references available for computing the automatic error and accuracy measures.

Table 2: *Total number of sentences, and available number of references, for each condition's development and test data set.*

Task	Data	Cond.	Sents.	Ref.
EN→ES EPPS	Dev.	FTE	735	2
		VBT	1194	2
		ASR	863	2
	Test	FTE	1117	2
		VBT	1155	2
		ASR	894	2
ES→EN EPPS	Dev.	FTE	430	2
		VBT	440	2
		ASR	440	2
	Test	FTE	894	2
		VBT	897	2
		ASR	1092	2
ES→EN CORTES	Dev.	FTE	380	2
		VBT	460	2
		ASR	460	2
	Test	FTE	888	2
		VBT	699	2
		ASR	1133	2

Table 3 presents the *BLEU*, *NIST* and *mWER* scores obtained for the test data of each translation task and condition. For all results presented in table 3 in the ES→EN direction, the reordering strategy described in subsection 3.7 was implemented. In the EN→ES direction, no reordering strategy was used, except for the result identified as FTE⋆, for which the reordering strategy described in subsection 3.6 was used.

Table 3: *BLEU, NIST and mWER scores obtained for the test data of each translation task and condition. All* BLEU *scores are case sensitive and use as many translation references as described in table 2.*

Task	Cond.	BLEU	NIST	mWER
EN→ES EPPS	FTE	0.482	9.999	40.89
	FTE⋆	0.488	10.06	40.21
	VBT	0.440	9.500	44.66
	ASR	0.347	8.557	51.78
ES→EN EPPS	FTE	0.552	10.60	36.94
	VBT	0.520	10.61	38.84
	ASR	0.383	9.142	48.66
ES→EN CORTES	FTE	0.415	9.207	47.05
	VBT	0.446	9.640	45.18
	ASR	0.298	7.995	55.87

Notice from table 3, that a better performance is achieved for the EPPS task when translating from Spanish-to-English than when translating from English-to-Spanish. This is clearly due to the more inflected nature of Spanish vocabulary, which makes more difficult the translation in the English-to-Spanish direction. Notice also the significant difference in translation quality between the Spanish-to-English EPPS and CORTES tasks. This result makes a lot of sense since the training data was from the EPPS corpus.

Another important observation from table 3 is that translation accuracy deteriorates when moving from text translations (FTE) to speech translations (VBT and ASR). Nevertheless, in the particular case of the CORTES task, the best result was obtained for the verbatim condition. A more detailed evaluation of translation outputs is required.

Regarding the EPPS English-to-Spanish, for which the reordering strategy presented in subsection 3.6 was tested, it can be seen from table 3 that the FTE⋆ experiment performed slightly better than the other according to all evaluation metrics.

Table 4: *Translation results for the 2005 test data by using the previous year's and current systems. Evaluation measures are case sensitive and were computed by using 2006's evaluation scripts.*

Task	System	BLEU	NIST	mWER
EN→ES	2005	0.456	9.657	41.20
	2006	0.486	9.880	40.66
ES→EN	2005	0.524	10.55	35.11
	2006	0.555	10.79	33.68

Finally, we also evaluate how much improvement, with respect to the previous year system, has been achieved. In this sense, table 4 presents translation results for the test data used in the 2005 TC-STAR first evaluation campaign by using the previous year's system (Mariño *et al.*, 2005) and the current one. Only the final text edition condition for the EPPS task is presented. As seen from the table, performance has been improved in about 0.03 *BLEU* marks (in the 0 to 1 scale) for both translation directions.

5 Conclusions and Further Work

As it can be concluded from the presented results the current system proved to provide better results than the previous year one. However, Spanish-to-English translations continue to be significantly better than English-to-Spanish translations, text translations better than speech translations, and morphology and reordering continue to be important problems.

In this sense, further research should focus on:

- Reordering strategies, as well as non-monotonous decoding schemes.
- Using more complete and diverse linguistic information sources.
- Specific preprocessing strategies for verbatim and ASR output data.

6 Acknowledgments

This work was partly 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`) and the European Social Fund.

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.

F. Casacuberta. 2001. "Finite-state transducers for speech input translation". *Proc. IEEE ASRU*, Madonna di Campiglio, Italy.

F. Casacuberta and E. Vidal. 2004. "Machine translation with inferred stochastic finite-state transducers". *Computational Linguistics*, 30(2):205–225.

Marta R.Costa-jussà and José A. R. Fonollosa. 2006. "Using Reordering in Statistical Machine Translation based on Alignment Block Classification". *Internal Report*.

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". *INTERSPEECH 2005*, Lisbon, September.

Josep M. Crego, Adrià de Gispert and José B. Mariño. 2005. "The TALP Ngram-based SMT System for IWSLT'05". *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT 2005*, Pittsburgh (PA), October.

Josep M. Crego. 2006. "A Reordering Framework for Statitstical Machine Translation". *Internal Report*.

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*.

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

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.

J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa and M. Ruiz. 2005. "Bilingual N-gram Statistical Machine Translation". *Proc. of the tenth Machine Translation Summit*, :275–282. Phuket, Thailand, September.

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.

F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. "A smorgasbord of features for statistical machine translation". *Proc. of the Human Language Technology Conf. NAACL*, :161–168, Boston, MA, May.

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. *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics, ACL'02*,

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.