Data Inferred Multi-word Expressions for Statistical Machine Translation

Patrik Lambert
TALP Research Center,
Jordi Girona Salgado, 1-3
08034 Barcelona,
Spain,
lambert@gps.tsc.upc.edu

Rafael Banchs
TALP Research Center,
Jordi Girona Salgado, 1-3
08034 Barcelona,
Spain,
rbanchs@gps.tsc.upc.edu

Abstract

This paper presents a strategy for detecting and using multi-word expressions in Statistical Machine Translation. Performance of the proposed strategy is evaluated in terms of alignment quality as well as translation accuracy. Evaluations are performed by using the Verbmobil corpus. Results from translation tasks from English-to-Spanish and from Spanish-to-English are presented and discussed.

1 Introduction

Statistical machine translation (SMT) was originally focused on word to word translation and was based on the noisy channel approach (Brown et al., 1993). Present SMT systems have evolved from the original ones in such a way that mainly differ from them in two issues: first, word-based translation models have been replaced by phrase-based translation models (Zens et al., 2002) and (Koehn et al., 2003); and second, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och and Ney, 2002).

Nevertheless, it is interesting to call the attention about one important fact. Despite the change from a word-based to a phrase-based translation approach, word to word approaches for inferring translation probabilities from bilingual data (Vogel et al., 1996; Och and Ney, 2003) continue to be widely used.

On the other hand, from observing bilingual data sets, it becomes evident that in some cases it is just impossible to perform a word to word alignment between two phrases that are translations of each other. For example, certain combination of words might convey a meaning which is somehow independent from the words it contains. This is the case of bilingual pairs such as "fire engine" and "camión de bomberos".

Notice, from the example presented above, that a word-to-word alignment strategy would most probably[1] provide the following Viterbi alignments for words contained in the previous example:

- "camión:truck",
- "bomberos:firefighters",
- "fuego:fire", and
- "máquina:engine".

Of course, it cannot be concluded from these examples that a SMT system which uses a word to word alignment strategy will not be able to handle properly the kind of word expression described above. This is because there are other models and feature functions involved which can actually *help* the SMT system to get the right translation. However these ideas motivate for exploring alternatives of using multi-word expression information in order to improve alignment quality and consequently translation accuracy.

This paper presents a technique for extracting bilingual multi-word expressions (BMWE) from parallel corpora.

This technique will be explained in section 3, after presenting the baseline translation system used (section 2). The proposed bilingual multi-word extraction technique is applied to the Verbmobil corpus, which is described in section 4.1. The impact of using the extracted BMWE on both alignment quality and translation accuracy, is evaluated and studied in sections 4.2 and 4.3. Finally some conclusions are presented and further work in this area is depicted.

2 Baseline Translation Model

This section describes the SMT approach that is used in this work. A more detailed description

[1] Of course, alignment results strongly depends on corpus statistics.

of the presented translation model is available in Mariño *et al.* (2005).

This approach implements a translation model which is based on bilingual n-grams, and was developed by de Gispert and Mariño (2002). 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.

The bilingual n-gram translation model actually constitutes a language model of bilingual units which are referred to as tuples. 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. More specifically, word-to-word alignments are performed in both directions, source-to-target and target-to-source, by using GIZA++ (Och and Ney, 2003), and tuples are extracted from the union set of alignments according to the following constraints (de Gispert and Mariño, 2004):

- a monotonous segmentation of each bilingual sentence pairs is produced,
- no word inside the tuple is aligned to words outside the tuple, and
- no smaller tuples can be extracted without violating the previous constraints.

As a consequence of these constraints, only one segmentation is possible for a given sentence pair. Figure 1 presents a simple example illustrating the tuple extraction process.

Two important issues regarding this translation model must be mentioned:

- When tuples are extracted, some words always appear embedded into tuples containing two or more words, so no translation probability for an independent occurrence of such words exists (consider for example the words "perfect" and "translations" contained in tuple t_4 of Figure 1). To overcome this problem, the tuple 3-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words (de Gispert and Mariño, 2004), which are extracted from the intersection set of alignments.

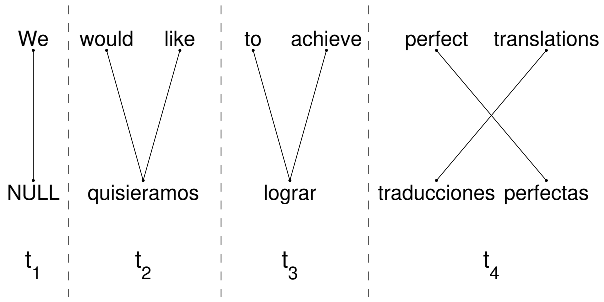

Figure 1: Example of tuple extraction from an aligned bilingual sentence pair.

- It occurs very often that 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.

A tuple set for each translation direction, Spanish-to-English and English-to-Spanish, is extracted from the union set of alignments. Then the tuple 3-gram models are trained by using the SRI Language Modelling toolkit (Stolcke, 2002); and finally, the obtained models are enhanced by incorporating the 1-gram probabilities for the embedded word tuples.

The search engine for this translation system was developed by Crego *et al.* (2005). It implements a beam-search strategy based on dynamic programming.

This decoder was designed to take into account various different feature functions simultaneously, so translation hypotheses are evaluated by considering a log-linear combination of feature functions. However, for all the results presented in this work, the translation model was used alone, without any additional feature function, not even a target language model. Actually, as shown in (Mariño et al., 2005), since the translation model is a bilingual language model, adding as only feature a target language model has little effect on the translation quality.

Additionally, the decoder's monotonic search modality was used.

3 Experimental Procedure

In this section we describe the technique used to see the effect of multi-words information on the translation model described in section 2.

First, BMWE were automatically extracted from the parallel training corpus and the most relevant ones were stored in a dictionary. More details on this stage of the process are given in section 3.1. In a second stage, BMWE present in the dictionary were detected in the training corpus in order to modify the word alignment (see section 3.2 for more details). Every word of the source side of the BMWE was linked to every word of the target side.

Then the source words and target words of each detected BMWE were grouped in a unique "super-token" and this modified training corpus was aligned again with GIZA++, in the same way as explained in section 2. By grouping multi-words, we increased the size of the vocabulary and thus the sparseness of data. However, we expect that if the meaning of the multi-words expressions we grouped is effectively different from the meaning of the words it contains, the individual word probabilities should be improved. After re-aligning, we unjoined the super-tokens that had been grouped in the previous stage, correcting the alignment set accordingly. More precisely, if two super-tokens A and B were linked together, after ungrouping them into various tokens, every word of A was linked to every word of B. Note that even after re-aligning, the vocabulary and the sequence of words to train the translation model were the same as in the baseline model, since we unjoined the super-tokens. The difference comes from the alignment, and thus from the translation units and from their corresponding n-grams.

3.1 Bilingual Multi-words Extraction

Various methods to extract BMWE were experimented.

3.1.1 Asymmetry Based Extraction

Multi-word expressions were extracted with the method proposed by Lambert and Castell (2004). This method is based on word-to-word alignments which are different in the source-target and target-source directions. Such alignments can be produced with the IBM Translation Models (Brown et al., 1993). We used GIZA++ (Och and Ney, 2003), which implements these models, to perform word-to-word alignments in both directions, source-target and target-source. Multi-words like idiomatic expressions or collocations can typically not be aligned word-to-word, and cause an asymmetry in the (source-target and target-source) alignment sets. An asymmetry in the alignment sets is a subset where source-target and target-source links are different. An example is depicted in figure 2. A word does not belong to an asymmetry if it is linked to exactly one word, which is linked to the same word and is not linked to any other word. In the method proposed by Lambert and Castell, asymmetries in the training corpus are detected and stored as bilingual multi-words, along with their number of occurrences.

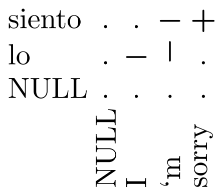

Figure 2: Asymmetry in the word-to-word alignments of an idiomatic expression. Source-target and target-source links are represented respectively by horizontal and vertical dashes.

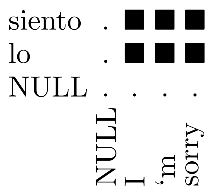

Figure 3: A multi-word expression has been detected in the asymmetry depicted in figure 2 and aligned as a group. Each word of the source side is linked to each word of the target side.

These asymmetries can be originated by idiomatic expressions, but also by translation errors or omissions. The method relies on the idea that if the asymmetry is caused by a language feature, it will be repeated various times in the corpus, otherwise it will occur only once. Thus only those bilingual multi-words which appeared at least twice are selected. Still, some bilingual multi-words, whose source side is not the translation of the target side, can appear various times. An example is "de que - you". To minimise this type of errors, we wanted to be able to select the N best asymmetry based BMWE, and ranked them according to their number of occurences.

3.1.2 Bilingual Phrase Extraction

Here we refer to *Bilingual Phrase* (BP) as the bilingual phrases used by Och and Ney (2004). The BP are pairs of word groups which are supposed to be the translation of each other. The set of BP is consistent with the alignment and consists of all phrase pairs in which all words within the target language are only aligned to the words of the source language and vice versa. At least one word of the target language phrase has to be aligned with at least one word of the source language phrase. Finally, the algorithm takes into account possibly unaligned words at the boundaries of the target or source language phrases.

We extracted all BP of length up to three words, with the algorithm described by Och and Ney (2004). Again, we established a ranking between them. In that purpose, we estimated the phrase translation probability distribution by relative frequency:

$$p(t|s) = \frac{N(t,s)}{N(s)} \tag{2}$$

In equation 2, s and t stand for the source and target side of the BP, respectively. $N(t,s)$ is the number of times the phrase s is translated by t, and $N(s)$ is the number of times s occurs in the corpus. Data sparseness can cause probabilities estimated in this way to be overestimated, and the inverse probability ($p(s|t)$) has proved to contribute to a better estimation (Ruiz and Fonollosa, 2005). To increase reliability, we took the minimum of both relative frequencies as probability of a BP, as shown in equation 3:

$$p(s,t) = \min(p(t|s), p(s|t)) \tag{3}$$

Many phrases occur very few times but always appear as the translation of the same phrase in the other language, so that their mutual probability as given by equation 3 is 1. However, this does not necessarily imply that they are a good translation of each other. To avoid to give a high score to these entries, we took as final score the minimum of the relative frequencies multiplied by the number of occurrences of this phrase pair in the whole corpus.

3.1.3 Intersection

Taking the intersection between the asymmetry based multi-word expressions and the BP presents the following advantages:

- BP imply a stronger constraint on the alignment between source and target side than asymmetries. In particular, entries which appear various times and whose source and target sides are not aligned together can't be selected as bilingual phrases and disappear from the intersection.
- Statistics of the BP set, which come from counting occurrences in the whole corpus, are more reliable than the statistics which come from counting occurrences in alignment asymmetries only. Thus, scoring asymmetry based BMWE with the BP statistics should be more reliable than with the number of occurrences in alignment asymmetries.
- Finally, since BP are extracted from all parts of the alignment (and not in asymmetries only), most BP are not BMWE but word sequences that can be decomposed word to word. For example, in the 10 best BP, we find "una reunión - a meeting", which is naturally aligned word to word. So if we want to use a BMWE dictionary of N entries (i.e. the N best scored), in the case of BP this dictionary would contain, let's say, only a 60% of actual BMWE. In the case of the asymmetry based multi-words, it would contain a much higher percentage of actual BMWE, which are the only "usefull" entries for our purpose.

So we performed the intersection between the entire BP set and the entire asymmetry based multi-words set, keeping BP scores.

3.1.4 Extraction Method Evaluation

To compare these three methods, we evaluated the links corresponding to the BMWE grouped in the detection process, with a manual alignment reference (which is described in section 4.1). Table 1 shows the precision and recall for the multi-words detected in the corpus when the three different dictionaries where used. Precision is defined as the number of proposed links that are correct, and recall is defined as the number of links in the reference that were proposed. Here the proposed links are only those of the BMWE detected. However the reference links are not restricted to multi-words. So the recall gives the proportion of detected multi-words links in the total set of links. In all three cases, only the best 650 entries of the dictionary were used.

We see from table 1 that taking the intersection with the BP set allows a nearly 6% improvement in precision with respect to the asymmetry based BMWE. The best precision is reached with the BP dictionary, which suggests than a better precision could be obtained for the intersection, for instance establishing a threshold pruning condition. Note that using a (manually built) dictionary of idiomatic expressions and with verb phrases detected with (manually specified) rules, de Gispert *et al.* (2004) achieved a much higher precision.

Recall scores reflect in a way the number of actual BMWE present in the 650 entries of the dictionary, and how frequent they are in the alignment asymmetries, which are where the BMWE are searched (see section 3.2). Logically, the asymmetry based dictionary, ranked according to the occurrence number, has got the higher recall. As explained in subsection 3.1.3, many high score BP are not multi-words expressions. So in the particular 650 entries we selected, there are less BMWE than in the intersection and the asymmetry based selections, and the recall is much lower. Thus the impact of multi-words information is expected to be lower.

Finally, the intersection dictionary allows to detect BMWE with a high precision and a high recal (compared to the two other methods), so it is the dictionary we used.

	Precision	Recall
Asymmetry based	85.39	20.21
Bilingual phrases	92.98	13.41
Intersection	91.26	18.82

Table 1: Multi-word expressions quality.

3.2 Multi-Words Detection and Grouping

Multi-words detection and grouping was performed with the symmetrisation algorithm described by Lambert and Castell (2004). The dictionary described in the previous subsection was used to detect the presence of BMWE(s) in each asymmetry. First, the best BMWE found is aligned as a group, as shown in figure 3. This process is repeated until all word positions are covered in the asymmetry, or until no multi-word expression matches the positions remaining to cover. The intersection of alignment sets in both directions was applied in the case no BMWE matched uncovered word positions.

4 Experimental Results

4.1 Training and Test Data

Training and test data come from a selection of spontaneous speech databases available from the Verbmobil project[2]. The databases have been selected to contain only recordings in US-English and to focus on the appointment scheduling domain. Then their counterparts in Catalan and Spanish have been generated by means of human translation (Arranz et al., 2003). Dates and times were categorised automatically (and revised manually). A test corpus of 2059 sentences has been separated for the training corpus.

The alignment reference corpus consists of 400 sentence pairs manually aligned by a single annotator, with no distinction between ambiguous or unambiguous links, i.e. with only one type of links.

See the statistics of the data in table 2.

		Spanish	English
Training	Sentences	28000	
	Words	201893	209653
	Vocabulary	4894	3167
Test	Sentences	2059	
	Words	19696	20585
Align. Ref.	Sentences	400	
	Words	3124	3188

Table 2: Characteristics of Verbmobil corpus: training and translation test as well as alignment reference.

4.2 Alignment and Translation Results

The effect of grouping multi-words before aligning the corpus is shown in figures 4 and 5, and in tables 3 and 4.

Figure 4 represents the Alignment Error Rate (AER) versus the number of times BMWE are grouped during our process. Equation 4 gives the expression of the AER in function of the precision P and recall R, defined in section 3.1. Nevertheless, here the whole set of links is evaluated, not only the links restricted to grouped

[2]http://verbmobil.dfki.de/verbmobil

BMWE.

$$AER = 1 - \frac{2PT}{P+T} \quad (4)$$

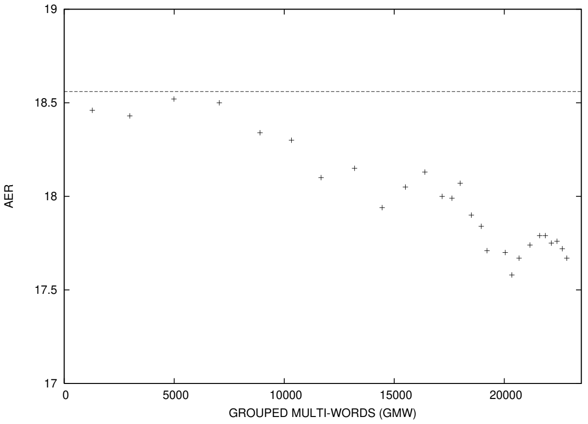

Figure 4: *Alignment Error Rate versus the number of multi-words grouped (GMW).*

In figure 4, the horizontal line represents the AER of the baseline system. We see a clear tendency in lowering the AER while more multi-words are grouped, although the total improvement is slightly less than one percent AER. An analysis of the precision and recall curves (not shown here) reveals that this AER improvement is due to a constant recall increase without precision loss.

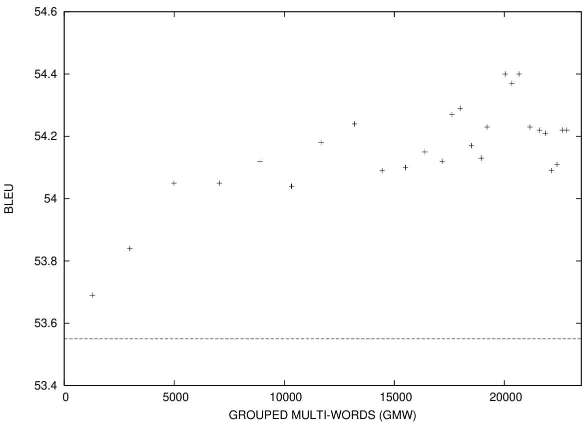

Figure 5: *BLEU score versus the number of multi-words grouped (GMW), in the translation from Spanish to English.*

In figure 5, BLEU score for the Spanish to English translation is plotted against the number of multi-words grouped. The horizontal line is the baseline value. Again, it is clear that while more multi-words are grouped, the translation quality is improved, the overall effect being of 0.85% in absolute BLEU score. However, there is an inflexion point (occuring around 22000 GMW, which corresponds to a dictionary of 1000 BMWE entries), after which there is a saturation and even a decrease of BLEU score. This inflexion could be caused by the lower quality of the worst ranked BMWE entries in the dictionary.

Tables 3 and 4 show experimental results obtained for a size of the BMWE dictionary of 650 entries. In this case, 20671 multi-words were grouped before re-aligning.

As can be seen in table 3, the size of the Spanish and English vocabularies were increased respectively a 5% and 13%, while the number of running words was decreased respectiveley a 3.7% and 10.5%.

	Voc. Size		Running Words	
	Spa	Eng	Spa	Eng
Baseline	4851	3139	198245	207743
650 MW	5089	3551	190730	186066

Table 3: Effect on the vocabulary size and the number of running words with a dictionary of 650 bilingual multi-words.

		S→E		E→S	
	AER	WER	BLEU	WER	BLEU
Bas.	18.6	30.0	53.5	35.2	48.2
Sym.	18.0	30.0	53.7	35.1	48.3
650	17.7	29.6	54.4	34.8	48.5

Table 4: Effect on Alignment and Translation with a dictionary of 650 bilingual multi-words.

Table 4 shows the alignment and translation results. “Bas.” stands for the baseline system, and “Sym.” is the system trained with the alignment set calculated after symmetrising (with a dictionary of 650 BMWE), but before grouping and re-aligning. Because the other systems are trained on the union of alignment sets in both directions, for this particular result, when no multi-word matched uncovered positions, the union was taken instead of the intersection (see section 3.2). “650” stands for the system obtained with the dictionary of 650 BMWE entries. Results are shown for both translation directions, Spanish to English (S→E) and English to Spanish (E→S). First, it can be observed that the symmetrising process doesn’t permit to improve significantly translation results. So the effect is due to the grouping of multi-word expressions and the improvement of individual word alignment probabilities it implies. Secondly, the effect is smaller when translating from English to Spanish than in the other direction.

4.3 Linear Regressions and Significance Analysis

In order to study in more detail the incidence of the proposed multi-word extraction technique on both alignment quality and translation accuracy, linear regressions were computed among some variables of interest. This analysis allows to determine if the variations observed in AER, WER and BLEU are actually due to variations in the number of BMWE used during the alignment procedure; or, on the other hand, if such variations are just random noise.

We were actually interested in checking for two effects:

- the incidence of the total number of bilingual multi-words grouped in the training corpus (GMW) on the resulting quality measurement variations (AER, WER and BLEU), and
- the incidence of alignment quality variations (AER) on translation accuracy variations (WER and BLEU).

A total of nine regression analysis, which are defined in Table 5, were required to evaluate the mentioned effects. More specifically, Table 5 presents the translation direction, a reference number, and the independent and dependent variables considered for each of the nine regressions. For all regression analysis, only variable values corresponding to a maximum of 900 BMWE entries were considered. As seen from figure 5 the behaviour of variables changes drastically when more that 1000 BMWE entries (around 22000 GMW) in the dictionary are considered.

Table 6 presents the regression coefficients obtained, as well as the linear correlation coefficients and the significance test results, for each of the considered regressions.

From the significance analysis results presented in Table 6, it is observed that all regressions performed can be considered statistically significant; i.e. the probabilities for such value distributions occurring by pure chance are extremely low.

These results allow us to conclude that the proposed technique for extracting and using

Dir.	Ref.	Dependent variable	Independent variable
–	reg1	AER	GMW
S → E	reg2	BLEU	GMW
	reg3	WER	GMW
	reg4	BLEU	AER
	reg5	WER	AER
E → S	reg6	BLEU	GMW
	reg7	WER	GMW
	reg8	BLEU	AER
	reg9	WER	AER

Table 5: Linear regressions performed.

	β_1	β_0	ρ	F	p-value
reg1	-0.04	18.7	-0.93	159.9	$0.00\ 10^{-5}$
reg2	0.02	53.8	0.85	58.69	$0.01\ 10^{-5}$
reg3	-0.01	30.0	-0.84	53.31	$0.02\ 10^{-5}$
reg4	-0.45	62.3	-0.79	37.95	$0.28\ 10^{-5}$
reg5	0.31	24.2	0.84	54.18	$0.02\ 10^{-5}$
reg6	0.02	48.0	0.88	79.80	$0.00\ 10^{-5}$
reg7	-0.03	35.5	-0.89	91.64	$0.00\ 10^{-5}$
reg8	-0.45	57.3	-0.81	45.04	$0.08\ 10^{-5}$
reg9	0.62	23.7	0.82	49.43	$0.04\ 10^{-5}$

Table 6: Regression coefficients (β_1: slope, and β_0: intercept), linear correlation coefficients (ρ) and significance analysis results for the regression coefficients (F-test). In this table GMW unit was 1000 GMW.

multi-word expressions has a positive incidence on both alignment quality and translation accuracy. However, as can be verified from slope values (β_1) presented in Table 6, this incidence is actually small. Although increasing the number of multi-words reduces AER and WER, and increases the BLEU, the absolute gains are lower as what we expected.

5 Conclusions and Further work

We proposed a technique for extracting and using BMWE in Statistical Machine Translation. This technique is based on grouping BMWE before performing statistical alignment. It permits to improve both alignment quality and translation accuracy. We showed that the more BMWE are used, the larger the improvement, until some saturation point is reached. These results are encouraging and motivate to do further research in this area, in order to increase the impact of multi-word information.

In the presented work the use of multi-words was actually restricted to the statistical alignment step. Experiments should be performed such that the BMWE are kept linked for tuple extraction and translation, to evaluate the direct impact of using multi-words in translation.

Different methods for extracting and identifying multi-word expressions must be developed and evaluated.

The proposed method considers the bilingual multi-words as units ; the use of each side of the BMWE as independent monolingual multi-words must be considered and evaluated.

6 Acknowledgements

This work has been partially 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 B. Mariño for all his comments and suggestions related to this work.

References

V. Arranz, N. Castell, and J. Giménez. 2003. Development of language resources for speech-to-speech translation. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgary, September, 10-12.

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

J.M. Crego, J. Mariño, and A. de Gispert. 2005. A ngram-based statistical machine translation decoder. In *Submitted to INTERSPEECH 2005*.

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 and J. Mariño. 2004. Talp: Xgram-based spoken language translation system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 85–90, October.

A. de Gispert, J. Mariño, and J.M. Crego. 2004. Phrase-based alignment combining corpus cooccurrences and linguistic knowledge. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 107–114, October.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

P. Lambert and N. Castell. 2004. Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, Lisbon, Portugal, May 25.

J. Mariño, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A. Fonollosa, and M. Ruiz. 2005. Bilingual n-gram statistical machine translation. In *Submitted to MT Summit X*.

F.J. Och and H. Ney. 2002. Dicriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, July.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

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

M. Ruiz and J.A. Fonollosa. 2005. Improving phrase-based statistical translation by modifying phrase extraction and including several features. In *(to be published) ACL05 workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING'96: The 16thInt. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In Springer Verlag, editor, *Proc. German Conference on Artificial Intelligence (KI)*, september.