Ngram-based versus Phrase-based Statistical Machine Translation

Josep M. Crego, Marta R. Costa-jussà, José B. Mariño and José A. R. Fonollosa

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{jmcrego,mruiz,canton,adrian}@gps.tsc.upc.edu

Abstract

This work summarizes a comparison between two approaches to Statistical Machine Translation (SMT), namely Ngram-based and Phrase-based SMT.

In both approaches, the translation process is based on bilingual units related by word-to-word alignments (pairs of source and target words), while the main differences are based on the extraction process of these units and the statistical modeling of the translation context. The study has been carried out on two different translation tasks (in terms of translation difficulty and amount of available training data), and allowing for distortion (reordering) in the decoding process. Thus it extends a previous work were both approaches were compared under monotone conditions.

We finally report comparative results in terms of translation accuracy, computation time and memory size. Results show how the ngram-based approach outperforms the phrase-based approach by achieving similar accuracy scores in less computational time and with less memory needs.

1. Introduction

From the initial word-based translation models [1], research on statistical machine translation has been strongly boosted. At the end of the last decade the use of context in the translation model (phrase-based approach) lead to a clear improvement in translation quality ([2], [3], [4]). Nowadays the introduction of some reordering abilities is of crucial importance for some language pairs and is an important focus of research in the area of SMT.

In parallel to the phrase-based approach, the ngram-based approach [5] also introduces the word context in the translation model, what allows to obtain comparable results under monotone conditions (as shown in [6]). The addition of reordering abilities in the phrase-based approach is achieved by enabling a certain level of reordering in the source sentence. Though, the translation process consists of a composition of phrases, where the sequential composition of the phrases source words corresponds to the source sentence reordered. This procedure poses additional difficulties when applied to the ngram-based approach, because the characteristics of the ngram-based translation model. Despite of this, recent works ([7], [8]) have shown how applying a reordering schema in the training process the ngram-based approach can also take advantage of the distortion capabilities.

In this paper we study the differences and similarities of both approaches (ngram-based and phrase-based), focusing on the translation model, where the translation context is differently taken into account. We also investigate the differences in the translation (bilingual) units (tuples and phrases) and show efficiency results in terms of computation time and memory size for both systems. We have extended the comparison in [6] to a Chinese to English task (where the use of distortion capabilities implies a clear improvement in translation quality), and using a much larger Spanish to English task corpus.

In section 2 we introduce the modeling underlying both SMT systems, the additional models taken into account in the log-linear combination of features (see equation 1), and the bilingual units extraction methods (namely tuples and phrases). In section 3 is discussed the decoder used in both systems (MARIE) [9], giving details of pruning and reordering techniques. The comparison framework, experiments and results are shown in section 4, while conclusions are detailed in section 5.

2. Modeling

Alternatively to the classical source channel approach, statistical machine translation models directly the posterior probability $p(e_1^I|f_1^J)$ as a log-linear combination of feature models [10], based on the maximum entropy framework, as shown in [11]. This simplifies the introduction of several additional models explaining the translation process, as the search becomes:

$$\arg\max_{e_1^I}\{exp(\sum_i \lambda_i h_i(e,f))\} \quad (1)$$

where the feature functions h_i are the system models (translation model, language model, reordering model, ...), and the λ_i weights are typically optimized to maximize a scoring function on a development set.

The Translation Model is based on bilingual units (here called tuples and phrases). A bilingual unit consists of two monolingual fragments, where each one is supposed to be the translation of its counterpart. During training, the system learns a dictionary of these bilingual fragments, the actual core of the translation systems.

2.1. Ngram-based Translation Model

The Translation Model can be thought of a Language Model of bilingual units (here called tuples). These tuples define a monotonous segmentation of the training sentence pairs (f_1^J, e_1^I), into K units $(t_1, ..., t_K)$.

The Translation Model is implemented using an Ngram language model, (for $N = 3$):

$$p(e, f) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k \mid t_{k-2}, t_{k-1}) \qquad (2)$$

Figure 1 shows an example of tuples extraction from a word-to-word aligned sentence pair.

Bilingual units (tuples) are extracted from any word-to-word alignment according to the following constraints [6]:

- 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.

Resulting from this procedure, some tuples consist of a monolingual fragment linked to the NULL word (words#NULL and NULL#words). Those tuples with a NULL word in its source side are not kept as bilingual units. To use these tuples in decoding it should appear a NULL word in the input sentence (test to translate). Though, we assign the target words of these tuples to the next tuple in the tuples sequence of the sentence (training). In the example of figure1, if the NULL word would be contained in the source side, its counterpart (does) would be assigned to the next tuple (does the flight last#dura el vuelo).

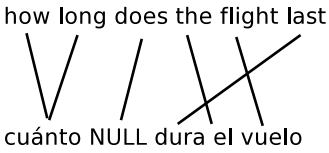
A complementary approach to translation with reordering can be followed if we allow for a certain reordering in the training data. This means that the translation units are modified so that they are not forced to sequentially produce the source and target sentences anymore. The reordering procedure in training tends to monotonize the word-to-word alignment through changing the word order of the source sentences.

The rationale of this approach is double, on the one hand, it makes sense when applied into a decoder with reordering capabilities as the one presented in the following section, and on the other hand, the unfolding technique generates shorter tuples, alleviating the problem of embedded units (tuples only appearing within long distance alignments, not having any translation in isolation). A very relevant problem in a Chinese to English task.

The unfolding technique is here outlined:

It uses the word-to-word alignments obtained by any alignment procedure. It is decomposed in two steps:

- First an iterative procedure, where words in one side are grouped when linked to the same word (or group) in the other side. The procedure loops grouping words in both sides until no new groups are obtained.
- The second step consists of outputting the resulting groups (unfolded tuples), keeping the word order of target sentece words. Though, the tuples sequence modifies the source sentence word order.



TUPLES:
how long#cuánto
does#NULL
the flight last#dura el vuelo

UNFOLDED TUPLES:
how long#cuánto
does#NULL
last#dura
the#el
flight#vuelo

Figure 1: *Different bilingual units (tuples) are extracted using the extract-tuples and extract-unfold-tuples methods. As can be seen,to produce the source sentence, the extracted unfolded tuples must be reordered. It is not the case of the target sentence, as it can be produced in order using both sequence of units.*

Figure 1 shows the bilingual units extracted using the extract-tuples and extract-unfold-tuples methods, for a given word-to-word aligned sentence pair.

2.2. Phrase-based Translation Model

The basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations [12].

Given a sentence pair and a corresponding word alignment, phrases are extracted following the criterion in [13] and the modification in phrase length in [14]. A phrase (or bilingual phrase) is any pair of m source words and n target words that satisfies two basic constraints:

1. Words are consecutive along both sides of the bilingual phrase,
2. No word on either side of the phrase is aligned to a word out of the phrase.

It is infesible to build a dictionary with all the phrases (recent papers show related work to tackle this problem, see [15]). That is why we limit the maximum size of any given phrase. Also, the huge increase in computational and storage cost of including longer phrases does not provide a significant improve in quality [16] as the probability of reappearence of larger phrases decreases.

In our system we considered two length limits. We first extract all the phrases of length X or less (usually X equal to 3 or 4). Then, we also add phrases up to length Y (Y greater than X) if they cannot be generated by smaller phrases. Basically, we select additional phrases with source words that otherwise would be missed because of cross or long alignments [14].

Given the collected phrase pairs, we estimate the phrase translation probability distribution by relative frecuency.

$$P(f|e) = \frac{N(f,e)}{N(e)} \tag{3}$$

where N(f,e) means the number of times the phrase *f* is translated by *e*. If a phrase *e* has $N > 1$ possible translations, then each one contributes as *1/N* [12].

2.3. Additional features

Both systems share the additional features which follows.

- Firstly, we consider the target language model. It actually consists of an n-gram model, in which the probability of a translation hypothesis is approximated by the product of word 3-gram probabilities:

  $$p(T_k) \approx \prod_{n=1}^{k} p(w_n|w_{n-2}, w_{n-1}) \tag{4}$$

  where T_k refers to the partial translation hypothesis and w_n to the n^{th} word in it.

  As default language model feature, we use a standard word-based trigram language model generated with smoothing Kneser-Ney and interpolation of higher and lower order ngrams (by using SRILM [17]).

- The following two feature functions correspond to a forward and backwards lexicon models. These models provides lexicon translation probabilities for each tuple based on the word-to-word IBM model 1 probabilities [18]. These lexicon models are computed according to the following equation:

  $$p((t,s)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p_{IBM1}(t_n^i|s_n^j) \tag{5}$$

  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.

  For computing the forward lexicon model, IBM model 1 probabilities from GIZA++ [19] source-to-target alignments are used. In the case of the backwards lexicon model, GIZA++ target-to-source alignments are used instead.

- The last feature in common we consider corresponds to a word penalty model. This function introduces a sentence length penalization in order to compensate the system preference for short output sentences. This penalization depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

  $$wp(T_k) = exp(\text{number of words in } T_k) \tag{6}$$

  where, again, T_k refers to the partial translation hypothesis.

2.4. Phrase features

In addition to the features from the section above, we use two more functions which get better scores in the phrase-based translation.

- As translation model in the phrase-based system we use the conditional probability. Note that no smoothing is performed, which may cause an overestimation of the probability of rare phrases. This is specially harmful given a bilingual phrase where the source part has a big frecuency of appearence but the target part appears rarely. That is why we use the posterior phrase probability, we compute again the relative frequency but replacing the count of the target phrase by the count of the source phrase [18].

  $$P(e|f) = \frac{N'(f,e)}{N(f)} \tag{7}$$

  where N'(f,e) means the number of times the phrase *e* is translated by *f*. If a phrase *f* has $N > 1$ possible translations, then each one contributes as *1/N*.

  Adding this feature function we reduce the number of cases in which the overall probability is overestimated.

- Finally, the last feature is the widely used phrase penalty [12] which is a constant cost per produced phrase. Here, a negative weight, which means reducing the costs per phrase, results in a preference for adding phrases. Alternatively, by using a positive scaling factors, the system will favor less phrases.

3. Decoding

In SMT decoding, translated sentences are built incrementally from left to right in form of hypotheses, allowing for discontinuities in the source sentence.

A Beam search algorithm with pruning is used to find the optimal path. The search is performed by building partial translations (hypotheses), which are stored in several lists. These lists are pruned out according to the accumulated probabilities of their hypotheses.

Worst hypotheses with minor probabilities are discarded to make the search feasible.

3.1. Search Graph Structure

Hypotheses are stored in different lists depending on the number of source and target words already covered.

Figure 2 shows an example of the search graph structure. It can be decomposed into three levels:

- Hypotheses. In figure 2, represented using '*'.
- Lists. In figure 2, the boxes with a tag corresponding to its covering vector. Every list contains an ordered set of hypotheses (all the hypotheses in a list have translated the same words of the source sentence).
- Groups (of lists). In figure 2, delimited using dotted lines. Every group contains an ordered set of lists, corresponding to the lists of hypotheses covering the same number of source words (to order the lists in one group the cost of their best hypothesis is used). When the search is restricted to monotonous translations, only one list is allowed on each group of lists.

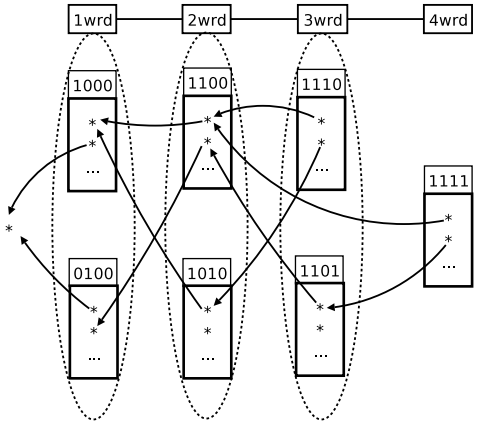


Figure 2: *Search graph corresponding to a source sentence with four words. Details of constraints are given in following sections.*

The search loops expanding available hypotheses. The expansion proceeds incrementally starting in the group of lists covering 1 source word, ending with the group of lists covering $J - 1$ source words (J is the size in words of the source sentence).

See [9] for further details.

3.2. Pruning Hypotheses

The search graph structure is thought to perform very accurate comparisons (only hypotheses covering the same source words are compared) in order to allow for very high pruning levels. Despite of this, the number of lists when allowing for reordering grows exponentially (an upper bound is 2^J, where J is the number of words of the source sentence) and forces the search to be further pruned out for efficiency reasons.

Only the best N hypotheses are kept on each list (histogram pruning, b), with best scores within a margin, given the best score in the list (threshold pruning, t). Not just the lists, but the groups are pruned out, following the same pruning strategies (B and T). To score a list, the cost of its best scored hypothesis is used.

3.3. Reordering capabilities

When allowing for reordering, the pruning strategies are not enough to reduce the combinatory explosion without an important lost in translation performance. With this purpose, two reordering strategies are used:

- A distortion limit (m). A source word (phrase or tuple) is only allowed to be reordered if it does not exceed a distortion limit, measured in words.
- A reorderings limit (j). Any translation path is only allowed to perform j reordering jumps.

The use of the reordering strategies suppose a necessary trade-off between quality and efficiency.

4. Comparison

4.1. Evaluation Framework

Experiments have been carried out using two databases: the EPPS database (Spanish-English) and the BTEC [20] database (Chinese-English).

The BTEC is a small corpus translation task, used in the IWSLT'04 spoken language campaign[1]. Table 1 shows the main statistics of the used data, namely number of sentences, words, vocabulary, and mean sentence lengths for each language.

The EPPS data set corresponds to the parliamentary session transcriptions of the European Parliament and is currently available at the Parliament's website (`http://www.euro parl.eu.int/`). In the case of the results presented here, we have used the version of the EPPS data that was made available by RWTH Aachen University through the TC-STAR consortium[2]. The training data used included session transcriptions from April 1996 until

[1] www.slt.atr.jp/IWSLT2004

[2] 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/`

BTEC	*Chinese*	*English*
Training Sentences	20 k	20 k
Words	182.9 k	188.9 k
Vocabulary	8.1 k	7.6 k
Development Sentences	506	506
Words	3.5 k	3.7 k
Vocabulary	870	874
Test Sentences	500	500
Words	3.7 k	3.8 k
Vocabulary	893	906

Table 1: *BTEC Corpus: Training, Development and Test data sets. The Development data set and the Test data set have 16 references, (k stands for thousands)*

EPPS	*Spanish*	*English*
Training Sentences	1.2 M	1.2 M
Words	34.8 M	33.4 M
Vocabulary	169 k	105 k
Development Sentences	504	504
Words	15.4 k	15.3 k
Vocabulary	2.8 k	2.3 k
Test Sentences	840	840
Words	22.7 k	20.3 k
Vocabulary	4 k	4.3 k

Table 2: *EuroParl Corpus: Basic statistics for the considered training, The Development data set and the Test data set have 2 references, (M and k stands for millions and thousands, respectively)*

September 2004, the development data used included parliamentary session transcriptions from October 21st until October 28th, 2004, and the test data from November 15th until November 18th, 2004.

Table 2, presents some basic statistics of training, development and test data sets for each considered language: English (en) and Spanish (es). More specifically, the statistics presented in table 2 are, the total number of sentences, the total number of words and the vocabulary size (or total number of distinct words).

4.2. Units

We used GIZA++ to perform the word alignment of the whole training corpus, and refined the links by the union of both alignment directions.

In the phrase-based model, we extract phrases up to length 4 and, in addition, those phrases up to length 7 which could not be generated by smaller phrases. This lengths are applied to the BTEC corpus. In the case of the EPPS task, we extract phrases up to length 3, without any extension.

The regular tuples extraction method was used in the monotone configuration of the ngram-model, while the unfolded extraction method was used in the reordering configuration.

	vocabulary	total	intersection
phrases	124.4 k	281.8 k	-
tuples	31.2 k	110.6 k	22.8 k
tuples'	298.6 k	-	62.3 k

Table 3: *For each set (rows), vocabulary, number of units extracted from the corpus and intersection with the phrases vocabulary set are shown (for the Chinese to English translation task). The unfolded tuples are used to build the tuple sets.*

Figure 3 shows how tuples and phrases vocabulary sets are related. In addition, an extended tuples vocabulary set (*tuples'*) is shown, which is built by concatenation of tuples. Consecutive tuples of each training sentence are concatenated building a new set of bilingual units. Pruned tuples in the sentence sequence are not taken into account to build the extended set.

This extended set approaches the tuples set to the phrases set. Also it allows us to show how many phrases units can be reached with the tuples units. In table 3 are given the vocabulary sizes of these sets for the BTEC corpus using the unfolding method to extract tuples. In principle, all tuples should be included as phrases. However, there are longer tuples that have been pruned out as phrases. There are also some tuples extracted from word-to-null alignments (39 word-to-null tuples).

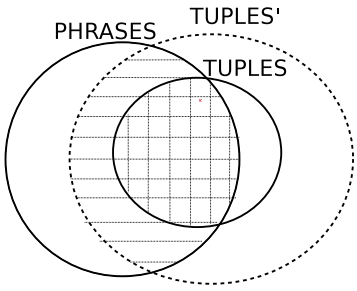


Figure 3: *Phrases and tuples vocabulary sets.*

Table 4 shows the number of Ngrams used by the decoder to translate the test file. For the phrase-based system, only 1grams (phrases) are used. The difference in number of loaded units implies a substantial impact in efficiency (in terms of computing time and memory size).

4.3. Experiments

In this section we introduce the experiments that have been carried out in order to evaluate both approaches.

All the computation time and memory size results are approximated. The experiments were performed on a Pentium

System	1gr	2gr	3gr	4gr
PB zh2en	59,610	-	-	-
NB zh2en	8,999	23,335	3,429	1,999
PB es2en	7,017,894	-	-	-
NB es2en	335,299	1,426,582	767,827	-

Table 4: *Number of Ngrams (translation model) loaded by the decoder to translate the test file. PB and NB stands for phrase-based and ngram-based, the first two rows correspond to the Chinese to English task, while the last two rows are related to the Spanish to English task.*

System	1gr	2gr	3gr	4gr
PB zh2en	2,518	-	-	-
NB zh2en	1,653	1,241	284	89
PB es2en	15,619	-	-	-
NB es2en	2,988	8,490	9,333	-

Table 5: *Number of Ngrams used by the decoder when translating the test file. PB and NB stands for phrase-based and ngram-based,the first two rows correspond to the Chinese to English task, while the last two rows are related to the Spanish to English task.*

IV (Xeon 3.06GHz), with 4Gb of RAM memory.

All the experiments reported in this paper have been performed setting the order of the target language model to $N = 3$. The order of the bilingual language model used for the BTEC task was $N = 4$, for the EPPS task was $N = 3$. When applying non-monotone decoding, the reordering constraints were set to $m = 5$ and $j = 3$ (in both Ngram-based and phrase-based approaches). Regarding the pruning adjustments, b and t are set to 10 units for the BTEC task and to 50 for the EPPS task, when applying reordering the B and T pruning values are also set to 10.

The evaluation in the BTEC task has been carried out using references and translations in lowercase and without punctuation marks. We applied the SIMPLEX algorithm to optimize the model weights (on the development set) [21]. Results in the test set with 16 references are reported.

Table 5 shows the number of 1-grams, 2-grams, 3-grams and 4-grams used when translating the test file using the best configuration of each system (allowing for reordering).

The experiments in table 6 correspond to the Chinese to English translation task under the phrase-based SMT system. Results corresponding to the same translation task, under the ngram-based SMT system, are shown in table 7.

The Spanish to English translation task results under the phrase-based SMT system, are shown in table 8. Results corresponding to the same translation task, under the ngram-based SMT system, are shown in table 9. The regular tuples extraction method was used in all cases as the translation was always performed under monotone conditions.

As can be seen, very similar results are achieved by both systems, when translating with the baseline and extended configurations. Thresholds for confidence margins are ± 1.6 and ± 0.6 (respectively for the Chinese-to-English and Spanish-to-English tasks given the number of words in the test sets for the mWER measure). In both cases the additional models (either the IBM1 lexicon model or the posterior probability model) seem to be used by the corresponding systems as a way to refine the translation probabilities. Examples of these situations are the overestimation problem introduced in previous sections for the phrase-based approach, and the apparition of bad tuples following incorrect word-to-word alignments.

5. Conclusions

In this paper we have performed a comparison of two state-of-the-art statistical machine translation approaches, which only differ in the modeling of the translation context. The comparison was made as fair as possible, in terms of using the same training/development/test corpora, word-to-word alignment, decoder and additional shared models (ibm1, word penalty, target LM and reordering model).

The comparison has been performed on two different translation tasks (in terms of reordering needs and related to the corpus size). Similar accuracy results in all tasks are reached for the baseline configurations. When upgrading the systems with additional features, slight differences appear. Although improvements added by each feature depends on the task and system, similar performances are reached in the best system's configurations. Under reordering conditions, the ngram-based system seems to take advantage of the unfolding method applied in training, outperforming the phrase-based system. However, last results obtained for the IWSLT'05 show an opposite behaviour of both systems, see [22] and [23].

We can conclude that both approaches have a similar performance in terms of translation quality. The slight differences seen in the experiments are related to how the systems take advantage of each feature model and to the current system's implementation. In terms of the memory size and computation time, the ngram-based system has obtained consistently better results. This indicates how even though using a smaller vocabulary of bilingual units, it has been more efficiently built and managed. The last characteristic becomes of great importance when working with large databases.

6. Acknowledgments

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 Spanish government, under grant TIC-2002-04447-C02 (Aliado Project), Universitat Politècnica de Catalunya and the TALP Research Center under UPC-RECERCA and TALP-UPC-RECERCA grants.

Phrase-based	mWER	BLEU	TIME (sec)	SIZE (Mb)
Baseline	50.02	36.32	23	2.4
Baseline + P(e|f)	49.57	37.02	28	2.8
Baseline + P(e|f) + IBM1 + Reord.	48.60	39.65	438	3.2

Table 6: *Results for the Chinese to English translation task using the phrase-based translation model and different features. The baseline uses translation model, language model, word penalty and phrase penalty. The IBM1 is used in both directions. The last row shows the best system and it includes reordering.*

Ngram-based	mWER	BLEU	TIME (sec)	SIZE (Mb)
Baseline	49.68	35,41	17	1.2
Baseline + IBM1	48.42	35.75	21	1.4
Baseline + IBM1 + Reord.	45.30	41.66	225	1.6

Table 7: *Results for the Chinese to English translation task using the ngram-based translation model and different features. The baseline configuration uses translation model, language model and word penalty. The IBM1 is used in both directions. The last row shows the best system and it includes reordering*

Phrase-based	mWER	BLEU	TIME (sec)	SIZE (Mb)
Baseline	39.35	48.84	900	1,180
Baseline + P(e|f) + IBM1	35.10	54.19	1084	1,640

Table 8: *Results for the Spanish to English translation task using the phrase translation model and different features. The baseline uses translation model, language model, word penalty and phrase penalty. The IBM1 model is used in both directions.*

Ngram-based	mWER	BLEU	TIME (sec)	SIZE (Mb)
Baseline	39.61	48.49	641	580
Baseline + IBM1	34.86	54.38	801	600

Table 9: *Results for the Spanish to English translation task using the phrase translation model and different features. The IBM1 is used in both directions. The baseline uses translation model, language model and word penalty. The IBM1 model is used in both directions.*

7. References

[1] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] R. Zens, F. Och, and H. Ney, "Phrase-based statistical machine translation," in *KI - 2002: Advances in artificial intelligence*, M. Jarke, J. Koehler, and G. Lakemeyer, Eds. Springer Verlag, September 2002, vol. LNAI 2479, pp. 18–32.

[3] K. Yamada and K. Knight, "A syntax-based statistical translation model," *39th Annual Meeting of the Association for Computational Linguistics*, pp. 523–530, July 2001.

[4] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pp. 133–139, July 2002.

[5] A. de Gispert and J. Mariño, "Análisis de las relaciones cruzadas en el alineado estadístico para la traducción automática," *II Jornadas en Tecnología del Habla*, December 2002.

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

[7] J. M. Crego, J. Mariño, and A. Gispert, "Reordered search and tuple unfolding for ngram-based smt," *Proc. of the MT Summit X*, September 2005.

[8] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney, "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*, pp. 167–174, June 2005.

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

[10] F. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," *40th Annual Meeting of the Association for Computational Linguistics*, pp. 295–302, July 2002.

[11] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–72, March 1996.

[12] R. Zens, F. Och, and H. Ney, "Improvements in phrase-based statistical machine translation," *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pp. 257–264, May 2004.

[13] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, December 2004.

[14] M. R. Costa-jussà and J. Fonollosa, "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*, June 2005.

[15] C. Callison-Burch, C. Bannard, and J. Scroeder, "Scaling phrase-based statistical machine translation to larger corpora and longer phrases," *ACL05*, June 2005.

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

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

[18] F. 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, "A smorgasbord of features for statistical machine translation," *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pp. 161–168, May 2004.

[19] F. Och, "Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html," RWTH Aachen University, Tech. Rep., 2003.

[20] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual curpus for speech translation of travel conversations in the real world," *LREC 2002*, pp. 147–152, May 2002.

[21] J. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[22] J. M. Crego, J. Mariño, and A. Gispert, "Talp: The upc tuple-based smt system," *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, October 2005.

[23] M. R. Costa-jussà and J. Fonollosa, "Tuning a phrase-based statistical translation system for the iwslt 2005 chinese to english and arabic to english tasks," *IWSLT05*, October 2005.