Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 93-100 recibido 29-04-2005; aceptado 01-06-2005

Bilingual phrases for statistical machine translation*

I. García-Varea†, F. Nevado‡, D. Ortiz‡, J. Tomás⋆, F. Casacuberta‡

† Dpto. de Informática
Univ. de Castilla-La Mancha
ismael.garcia@uclm.es

‡ Dpto. de Sist. Inf. y Comp.
Univ. Politécnica de Valencia
{fcn|dortiz|fnevado}@iti.upv.es

⋆Dpto. de Comunicaciones
Univ. Politécnica de Valencia
jtomas@dcom.upv.es

Abstract: The statistical framework has proved to be very successful in machine translation. The main reason for this success is the existence of powerful techniques that allow to build machine translation systems automatically from available parallel corpora. Most of statistical machine translation approaches are based on single-word translation models, which do not take bilingual contextual information into account. The translation model in the phrase-based approach defines correspondences between sequences of contiguous source words (source segments) and sequences of contiguous target words (target segments) instead of only correspondences between single source words and single target words. That is, statistical phrase-based translation models make use of explicit bilingual contextual information. Different methods for the selection of adequate bilingual word sequences and for training the parameters of these models are reviewed in this paper. Improved techniques for the selection and training model parameters are also introduced. The phrase-based approach has been assessed in different tasks using different corpora and the results obtained are comparable or better than the ones obtained using other statistical and non-statistical machine translation systems.
Keywords: Statistical machine translation, Phrase-based translation models, Bilingual segmentation

1 Introduction

The interest for the statistical approach to machine translation (SMT) has greatly increased due to the successful results obtained for typical restricted-domain translation tasks.

The translation process can be formulated from a statistical point of view as follows: A source language string $f_1^J = f_1 \dots f_J$ is to be translated into a target language string $e_1^I = e_1 \dots e_I$. Every target string is regarded as a possible translation for the source language string with maximum a-posteriori probability $Pr(f_1^J|e_1^I)$. According to Bayes' decision rule, the target string $\hat{e}_1^I$ that maximizes the product of both the target language model $Pr(e_1^I)$ and the string translation model $Pr(f_1^J|e_1^I)$ must be chosen. The equation that models this process is:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \qquad (1)$$

Different *translation models* (TMs) have been proposed depending on how the relation between the source and the target languages is structured. This relation is summarized using the concept of *alignment*; that is, how the words of a pair of sentences are aligned to each other.

Classical statistical translation models can be classified as *single-word based* (SWB) alignment models. Models of this kind assume than a source word is generated by only one target word (Brown et al., 1993)(Ney et al., 2000). This assumption does not correspond to the nature of natural language; in some cases, we need to know a multiword sequence in order to obtain a correct translation. In essence these SWB aligment models lack of useful bilingual contextual information. Previous work to deal with bilingual contextual information have used maximum-entropy models (García-Varea and Casacuberta, 2005; Berger, Della Pietra, and Della Pietra, 1996), but the estimation and the definition of a search algorithm for these kind of models is difficult and high costly.

Recent works present an alternative to these models, the *phrase-based* (PB) approach (Tomás and Casacuberta, 2001; Zens, Och, and Ney, 2002; Marcu and Wong, 2002).

* This work has been partially supported by the Spanish project TIC2003-08681-C02, the *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031, the *Generalitat Valenciana*, and the project AMETRA (INTEK-CN03AD02)

These methods explicitly learn the probability of a segment in a source sentence being translated to another segment of words in the target sentence. These bigger units allow us to represent bilingual contextual information in an explicit and easy way.

The organization of the paper is as follows. First, we review the PB translation model, the estimation of these models and search algorithms. Also, experimental results on different translation tasks are presented. Then, the obtaining of PB translation units (bilingual segments) is reviewed using three different methods. Experiments assessing the quality of the bilingual segments obtained are also presented. Finally, we give some conclusions and some lines of future work.

2 Phrase-based translation

Different models that deal with structures or phrases instead of single words have been proposed: syntax translation models are described in (Yamada and Knight, 2001) , alignment templates are used in (Och, 2002), and the alignment template approach is reframed into the so-called *phrase based translation* (PBT) in (Marcu and Wong, 2002; Zens, Och, and Ney, 2002; Koehn, Och, and Marcu, 2003; Tomás and Casacuberta, 2001).

PBT can be explained from a generative point of view as follows (Zens, Och, and Ney, 2002):

1. The source sentence f_1^J is segmented into K phrases ($\tilde{f}_1^K$).
2. Each source phrase $\tilde{f}_k$ is translated into a target phrase $\tilde{e}$.
3. Finally the target phrases are reordered in order to compose the target sentence $\tilde{e}_1^K = e_1^I$.

2.1 Phrase-based models

In PBT, it is assumed that the relations between source and target phrases/segments can be explained by means of the hidden variable $\tilde{\mathbf{a}} = \tilde{a}_1^K$, which contains all the decisions made during the generative process. Additionally, a one-to-one phrase alignment is used, i.e, one source phrase is translated by exactly one target phrase. Assuming that the alignment between phrases is modeled using a bigram model, this process can be formulated as follows:

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{K,\tilde{f}_1^K,\tilde{e}_1^K,\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}, \tilde{f}_1^K|\tilde{e}_1^K) \qquad (2) \\ &= \sum_{K,\tilde{f}_1^K,\tilde{e}_1^K,\tilde{\mathbf{a}}} \prod_{k=1}^{K} p(\tilde{a}_k|\tilde{a}_{k-1}) p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) \end{aligned}$$

where $\tilde{a}_k$ denotes the index of the target phrase $\tilde{e}$ that is aligned with the k-th source phrase $\tilde{f}_k$ (Tomás and Casacuberta, 2001).

Different additional assumptions can be made from equation (2), as for example, in (Zens, Och, and Ney, 2002; Tomás and Casacuberta, 2001), and following the maximum approximation, equation (2) can be rewritten as:

$$Pr(f_1^J|e_1^I) = \alpha(e_1^I) \max_{K,\tilde{f}_1^K,\tilde{e}_1^K} \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k) \quad (3)$$

where it is assumed that all segmentations have the same probability $\alpha(e_1^I)$, and only monotone translation are allowed. This results in a very efficient search. Also, the summation over every possible alignment can be removed according to the monotonicity restriction.

2.2 Model estimation

Different methods has been proposed for learning the bilingual phrase translation probabilities ($p(\tilde{f}|\tilde{e})$), used in equations (2) and (3), from a parallel training corpus.

In (Koehn, Och, and Marcu, 2003) three different methods for learning bilingual phrase translations probabilities are described:

1. From word-based alignments.
2. From syntactic phrases (see (Yamada and Knight, 2001) for more details).
3. From sentence-based alignments, using the EM algorithm for training (Marcu and Wong, 2002; Tomás and Casacuberta, 2001).

Here, we focus on the first method, in which a set of bilingual phrases ($\mathcal{BP}$) must be previously extracted from a bilingual, word-aligned training corpus. The extraction process is driven by an additional constraint: the bilingual phrase must be consistent with its corresponding word alignment set A (a set of pairs (i, j) of (source,target) position indices)

as shown in equation (4) (which is the same given in (Och, 2002) for the alignment template approach).

$$\mathcal{BP}(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n}) : \forall (i', j') \in A : j \leq j' \leq j+m \iff i \leq i' \leq i+n\} \quad (4)$$

Typically, in order to obtain better bilingual phrases, different word alignment sets are combined. The common combination procedure consists of estimating SWB models in both directions and performing different operations with the resulting alignment sets. The most common operations are union and intersection of the single-word alignment sets, and also the refined *symmetrization method* proposed in (Och, 2002).

Additionally, in (Venugopal, Vogel, and Waibel, 2003), two methods of phrase extractions are proposed (based on source n-grams and HMM alignments respectively). They improve a translation lexicon, instead of defining a phrase-based model, which is also used within a word-based decoder. In the same line, a method to produce phrase-based alignments from word-based alignments is proposed in (Lambert and Castell., 2004).

Once the phrase pairs are collected, the phrase translation probability parameters are estimated via the relative frequency as follows:

$$p(\tilde{f}|\tilde{e}) = \frac{\text{count}(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} \text{count}(\tilde{f}, \tilde{e})} \quad (5)$$

2.3 Phrase-based search

The aim of the search in MT is solve the maximization of equation (1) in order to obtain the target sentence e_1^I that maximizes the product probabilities $Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$. The search algorithm is a crucial part in statistical machine translation. Its performance directly affects the quality and efficiency of translation. In this section, we describe two search algorithms which are based on multi-stack-decoding (Berger et al., 1996) for the monotone and for the non-monotone model.

The most common statistical decoder algorithms use the concept of partial translation hypothesis to perform the search (Berger et al., 1996). In a partial hypothesis, some of the source words have been used to generate a target prefix. Each partial hypothesis is scored according to the translation and language model. In our implementation for the monotone model, we define a hypothesis search as the triple $(J', e_1^{I'}, g)$, where J' is the length of the source prefix we are translating to the hypothesis (that is $f_1^{J'}$). The sequence of I' words, $e_1^{I'}$, is the target prefix that has been generated. And g is the score of the partial hypothesis ($g = Pr(e_1^{I'})Pr(f_1^{J'}|e_1^{I'})$).

The translation procedure can be described as: The system maintains a large set of hypotheses, each of which has a corresponding translation score. This set starts with an initial empty hypothesis. Each hypothesis is stored in a different stack, according to the source words that have been considered in the hypothesis (J'). The algorithm consists of an iterative process. In each iteration, the system selects the best scored partial hypothesis to extend in each stack. The extension consists in selecting one (or more) untranslated words in the source and selecting one (or more) target words that are attached to the existing output prefix. In the new hypothesis, the source words are marked as translated (increasing J') and the probability cost of the hypothesis is updated. The extension of a partial hypothesis can generate hundreds of new partial hypotheses. The process continues several times or until there are no more sentences to extend. The final hypothesis with the highest score and with no untranslated source words is the output of the search. This algorithm has a good performance, it can translate more than a hundred words per second in the experiments we carried out for this work.

We propose extending the search to allow for non-monotone translation. In this extension, several reorderings in the target sequence of phrases are scored with a corresponding probability. We define a hypothesis search as the triple $(w, e_1^{I'}, g)$, where $w \subseteq \{1, 2, \ldots, J\}$ is the coverage set that defines which positions of source words have been translated. For a better comparison of hypotheses, Berger et al. (1996) proposes storing each hypothesis in different stacks according to their value of w. The number of possible stacks can be very high (2^J); thus, the stacks are created on demand. The translation procedure is similar to the previous one: In each iteration, the system selects the best scored partial hypothesis to extend in each created stack and extends it. In order to speed up the search, beam search and rest-cost estimation (of partial hypothesis) tech-

		English	Spanish
Train	Sentences	55,761	
	Running words	665k	753k
	Vocabulary	7,957	11,051
Test	Sentences	1,125	
	Running words	8,370	10,106

Table 1: XRCE corpus statistics. (k≡ $\times 1,000$)

	spa→eng	eng→spa
Monotone search	24.3	26.2
Non-monotone search	24.1	26.2

Table 2: Effect of the type of search on WER (in %) using the 1,215 test sentences of the XRCE corpus.

niques has been used.

2.4 Translation experiments

In order to evaluate the performance of these approaches, we carried out several experiments using several corpora. We selected trigram models for the target language model. As an evaluation criterium we use *word error rate* (WER), the minimum number of substitution, insertion and deletion operations needed to convert the hypothesized translation by the MT into a given reference translation (Och, 2002).

2.4.1 XRCE corpus

The XRCE corpus was compiled using some Xerox technical manuals published in several languages. This is a reduced-domain task that has been defined in the TransType2 project (TT2, 2002). Table 1 presents some statistical information about this corpus after the pre-processing phase.

In the formal description of the model, we do not limit the number of words in a phrase. However, in a practical implementation, we limit the maximum number of words in a phrase to 16. Table 2 compares translation results, by means of WER, for this task for the monotone and non-monotone decoders, and for the Spanish to English (spa→eng) and English to Spanish translation directions (eng→spa). For this task the quality of the translation results was very similar for the monotone and non-monotone decoders.

		English	Spanish
Train	Sentences	1,246,789	
	Running words	34,631k	35,838k
	Vocabulary	44,568	85,568
Test	Sentences	500	
	Running words	11,359	12,095

Table 3: EPPS corpus statistics.

maximum phrase length	WER(%)	num. bilingual phrases
4	61.7	14M
6	60.9	18M
8	59.2	23M

Table 4: Effect of maximum numbers of words in a phrase on WER (in %) using the EPPS corpus. (M≡ $\times 1,000,000$)

2.4.2 EPPS corpus

The EPPS corpus was compiled from European Parliament Plenary Sessions, that are available at http://www.europarl.eu.int. It is composed by the proceedings of the European Parliament from year 1996 to 2004. This task has been defined in the TCStar project (www.tc-star.org). Table 3 presents some statistical information about this corpus after the pre-processing phase.

Table 4 reports some results, for the EPPS corpus, using phrases of different length. The translation quality results (WER) were obtained by using a non-monotone decoder. The second column of the table shows the number of bilingual phrases extracted from the training corpus in order to train the model parameters. As can be seen in Table 4 the larger the size of the maximum phrase length used the better results obtained. On the other hand, the number of bilingual phrases was substantially increased when larger sizes of length phrases were used. The translation results show that this task is a very difficult task, as the corpus statistics suggested (Table 3).

2.4.3 El Periódico corpus

The El Periódico corpus is obtained from the electronic publication of the newspaper *El Periódico de Catalunya* (http://www.elperiodico.es). This general information newspaper is published daily in a bilingual edition. The domain of this corpus corresponds to the language used in a

		Spanish	Catalan
Train	Sentences	644,961	
	Running words	7,180k	7,435k
	Vocabulary	129k	128k
Test	Sentences	120	
	Running words	2,179	2,211

Table 5: El Periódico corpus statistics.

Translator	WER(%)
Salt	9.9
Statistical PBT	10.7
Incyta	10.9
Internostrum	11.9

Table 6: Comparative evaluation for several Spanish–Catalan translators.

journalistic context, including sections such as editorials, politics, sports, TV programming, etc. Table 5 presents some statistical information about this corpus.

In order to carry out the evaluation, we compared the results obtained with our statistical translator with three Spanish–Catalan commercial systems: Salt (www.cultgva.es), Incyta (www.incyta.com), and Internostrum (www.internostrum.com). The test sentences were taken from different media: a newspaper, a technical manual, legal text, etc. The references used to compute the WER were also taken from the Catalan version of the same documents.

In this task we use a maximum phrase length of 3 words, obtaining a total of 7 millions of bilingual phrases for training the model parameters. The translation results for this task are shown in Table 6, where the non-monotone version of decoder was used. As it can be seen in Table 6, the statistical phrase-based translator obtains an intermediate position, but with very low differences with respect to the commercial rule-based systems. The best results were obtained by the Salt system. The main advantage of our phrase-based statistical translation system with respect to the commercial systems used here is that it is automatically built, in contrast to the expert knowledge that these rule-based systems need to use in order to be performed.

3 Bilingual Segmentation

The purpose of bilingual segmentation is to obtain translation units at a subsentence level. We redefine the formal definition of the bilingual segmentation (or simply *bisegmentation*) concept in (Simard and Plamondon, 1998) as follows:

Let $f_1^J = \{f_1, f_2, \ldots, f_J\}$ be a source sentence and $e_1^I = \{e_1, e_2, \ldots, e_I\}$ the corresponding target sentence in a bilingual corpus. A segmentation S of f_1^J and e_1^I is defined as a set of ordered pairs included in $\mathcal{P}(f_1^J) \times \mathcal{P}(e_1^I)$, where $\mathcal{P}(f_1^J)$ and $\mathcal{P}(e_1^I)$ are the set of all subsets of consecutive sequences of words, of f_1^J and e_1^I, respectively. Each of the ordered pairs of the segmentation is a *bisegment*.

A bilingual segmentation or *bisegmentation* of length K of a sentence pair (f_1^J, e_1^I) is defined as a triple $(\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$, where $\tilde{a}_1^K$ is a specific one-to-one mapping between the K segments/phrases of both sentences.

3.1 Using phrase-based models to obtain bilingual segmentations

Phrase-based models can be used in order to perform bilingual segmentation. For that purpose, first we obtain a phrase-based dictionary from a word-level aligned parallel corpus, by using the bilingual phrase extraction method described in section 2.2.

Then, given a pair of sentences (f_1^J, e_1^I) and a word alignment between them, we have to obtain the best bisegmentation in K bisegments ($1 \leq K \leq \min(J, I)$), and implicitly the best phrase-alignment $\tilde{a}_1^K$ (or Viterbi phrase-alignment) between them.

The probability of a bilingual segmentation of length K is computed as:

$$p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K) = \prod_{k=1}^{K} p(\tilde{f}_{\tilde{a}_k} | \tilde{e}_k) \qquad (6)$$

Basically, the algorithm, which will be referred as *SPBalign* algorithm, works as follows: Given a sentence pair (f_1^J, e_1^I) and an alignment set $A(f_1^J, e_1^I)$:

1. For every possible $K \in \{1 \cdots \min(J, I)\}$
   (a) Extract all possible bilingual segmentations of size K according to the restrictions of $A(f_1^J, e_1^I)$.
   (b) Compute and store the probability $p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K)$ of these bisegmentations.

2. Return the bilingual segmentation $(\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$ of highest probability.

3.2 Other bisegmentation techniques

There exist other bisegmentation techniques that are described in the literature. In the following sections we will briefly introduce the *GIATI-based bisegmentation* and the *Recursive bilingual segmentation* techniques which will be compared with the one presented above.

3.2.1 GIATI-based bisegmentation

The GIATI technique is an automatic method to infer statistical finite-state transducers described in (Casacuberta, 2000), which can also be used for obtaining bisegmentations.

This technique carries out a labelling of the words of the source sentence with the words of the output sentence from a word alignment between both sentences.

This kind of labelling can produce a bisegmentation if we consider that the bisegments are composed of the source words and their corresponding labels of target words. The method labels every source word with its connected target words except when a reordering is done in the alignment. In this case, the method groups all the necessary source and target words in order to consider the reordering inside the bisegment. This system will be referred as *GIATIalign*.

3.2.2 Recursive bilingual segmentation

Basically, a recursive alignment is an alignment between phrases of a source sentence and phrases of a target sentence. The bisegmentations can be obtained as a byproduct from the recursive alignments as it is described in (Nevado, Casacuberta, and Landa, 2004).

A recursive alignment represents the translation relations between two sentences, but it also includes information about the possible reorderings needed in order to generate the target sentence from the source sentence. This system will be referred as *RECalign*.

3.3 Bilingual segmentation experiments

The bisegmentations obtained with the three presented techniques are compared with a reference bisegmentation computed manually by experts. In order to evaluate them we used the three bilingual segmentation error rates recall, precision, and F-measure described in (Simard and Plamondon, 1998).

		English	Spanish
Train	Sentences	10,000	
	Runing words	99,292	97,131
	Vocabulary	513	686
Test	Sentences	40	
	Running words	491	487

Table 7: EuTrans-I corpus statistics.

We carried out different experiments according to the type of word alignment that was used to bisegment the test corpus. That is, the source-to-target word alignment (E-S) for English-to-Spanish, and, additionally, three different combinations of both alignments were used: the intersection ($\cap$), the union ($\cup$) , and the refined (R) symmetrization methods that were mentioned in section 2.2.

3.3.1 Corpus description

For the experiments, we have used the EuTrans-I corpus, which is a Spanish–English bilingual corpus whose domain is a subset of the Tourist task (Amengual et al., 2000). From this corpus, 10,000 different sentence pairs were selected for training purposes. We also selected a test corpus, not included in the training corpus, consisting on 40 randomly selected pairs of sentences. The 40-sentence test corpus was bilingually segmented by human experts. Table 7 shows the characteristics of the training and test sets we have used for this corpus.

3.3.2 Bisegmentation quality results

The bisegmentation results for the Spanish–English are presented in Table 8. The four different types of word alignment are used with every technique. For every experiment the recall, precision and F-measure are presented. The F-measure is the harmonic mean of precision and recall, so it give us a compromise between the coverage and exactness of the automatic obtained bilingual segmentation. For every technique, the best result is highlighted in bold.

The union of word alignments obtains the better results. That is exactly what we expected, because the alignment union re-

Technique	recall	precision	F-measure
RECalign+(S-E)	39.67	87.11	54.51
RECalign+(∩)	36.60	87.42	51.60
RECalign+(∪)	52.96	79.01	**63.41**
RECalign+(R)	48.86	80.67	60.85
GIATIalign+(S-E)	39.91	85.92	54.50
GIATIalign+(∩)	36.22	80.26	49.91
GIATIalign+(∪)	39.99	85.52	**54.50**
GIATIalign+(*R*)	37.35	84.68	51.84
SPBalign+(S-E)-5	68.09	68.47	68.28
SPBalign+(∩)	67.21	67.38	67.29
SPBalign+(∪)	72.58	65.49	**68.85**
SPBalign+(R)	66.27	65.84	66.06

Table 8: Bisegmentation results for EuTrans-I task for the Spanish–English translation direction.

marks the alignments between words which are translations of each other, which finally results in a better bisegmentation (and implicitly the alignment at segment level).

In general, a similar accuracy is obtained in both translation directions. In all cases the *SPBalign* technique proposed here, outperforms the *RECalign* and *GIATIalign* techniques and obtains more balanced values of precision and recall.

4 Concluding remarks

Phrase-based models can gather an important part of bilingual contextual information for translation and they can be built from training bilingual corpora. Related with these models, the bilingual segmentation of bilingual corpora is an important challenge in machine translation.

Some approaches to phrase-based translation models have been introduced in this paper. They are based on monotone and non-monotone alignments. The monotone approach is very simple and the search can be performed in reduced time. This method can obtain good translation results in certain tasks such as some reduced-domain tasks or between Romance languages. For an unrestricted task, such as Spanish–Catalan translation, better or comparable results than some rule-based commercial systems have been obtained. Note that these models require a drastically lower human effort than conventional rule-based translation systems.

In contrast, phrase-based models have a low capability of generalization and are not able to deal adequately with unseen events. Due to this fact, in the future we plan to study the combination of phrase-based models with single-word models by means of interpolation or using the maximum-entropy formalism, that allows us to integrate different knowledge sources.

According to the state of the art of the phrase-based approach to statistical machine translation there is still quite room for improvement. For example, for phrase-based translation models, we plan to study the following things:

- To include more dependencies into the models, by relaxing the assumptions that are currently taken into account. In that way we plan to learn a phrase-alignment ($p(\tilde{a}_k|\tilde{a}_1^{k-1})$) model and a phrase-length ($p(K|e_1^I)$) model.

- To carry out an exact maximum-likelihood estimation of the phrase translation parameters by using relative frequency and the EM algorithm training methods from single-word based statistical alignment models.

With respect to the bilingual segmentation techniques we have in mind to explore new symmetrization methods to combine alignments at word level in order to obtain better results, as for example the union of a list of n-best word alignments. In the same direction, we think that it could be useful to use weighted word-alignments in order to pay more attention to those alignment relations that are really relevant for phrase alignments.

References

Amengual, J.C., J.M. Benedí, F. Casacuberta, M.A. Castaño, A. Castellanos, V.M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J.M. Vilar. 2000. The EuTrans-I speech translation system. *Machine Translation*, 1.

Berger, Adam L., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, A. S. Kehler, and R. L. Mercer. 1996. Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981, April.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.

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

Casacuberta, F. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In A.L. Oliveira, editor, *Grammatical Inference: Algorithms and Applications*, volume 1891 of *Lecture Notes in Computer Science*. Springer-Verlag, Lisbon, Portugal, September, pages 1–14.

García-Varea, Ismael and Francisco Casacuberta. 2005. Maximum entropy modeling: A suitable framework to learn context-dependent lexicon models for statistical machine translation. *Machine Learning*, 59:1–24.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL*, Edmonton, Canada, May.

Lambert, Patrik and Núria Castell. 2004. Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proc. of the Fourth Int. Conf. on LREC*, Lisbon, Portugal.

Marcu, Daniel and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the EMNLP Conference*, pages 1408–1414, Philadelphia, USA, July.

Nevado, F. F. Casacuberta, and J. Landa. 2004. Translation memories enrichment by statistical bilingual segmentation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

Ney, Hermann, Sonja Nießen, Franz J. Och, Hassan Sawaf, Christoph Tillmann, and Stephan Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Trans. on Speech and Audio Processing*, 8(1):24–36, January.

Och, Franz Joseph. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, Computer Science Department, RWTH Aachen, Germany, October.

Simard, M. and P. Plamondon. 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80.

Tomás, J. and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Procs. of the Machine Translation Summit VIII*, pages 357–361, Santiago de Compostela, Spain.

TT2. 2002. Transtype2-computer-assisted translation (TT2). Technical report. Information Society Technologies (IST) Programme. IST-2001-32091.

Venugopal, Ashish, Stephan Vogel, and Alex Waibel. 2003. Effective phrase translation extraction from alignment models. In *Proc. of the 41th Annual Meeting of ACL*, pages 319–326, Sapporo, Japan, July.

Yamada, Kenji and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proc. of the 39th Annual Meeting of ACL*, pages 523–530, Toulouse, France, July.

Zens, R., F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI*, volume 2479 of *Lecture Notes in Computer Science*. Springer Verlag, September, pages 18–32.