DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN
UNIVERSIDAD POLITÉCNICA DE VALENCIA

P.O. Box: 22012 E-46071 Valencia (SPAIN)

Informe Técnico / Technical Report

Ref. No.: DSIC-II/18/07 **Pages:** 43
Title: A pattern recognition approach to machine translation: monotone and non-monotone phrase-based statistical models
Author(s): Jesús Tomás, Francisco Casacuberta
Date: 20/9/2007
Keywords: phrase-based models, statistical methods, machine translation

Vº Bº

Leader of research Group

Author(s)

A pattern recognition approach to machine translation: monotone and non-monotone phrase-based statistical models

Jesús Tomás*
Universidad Politécnica de Valencia

Francisco Casacuberta†
Universidad Politécnica de Valencia

Statistical machine translation has proven to be an interesting framework for automatically building machine translation systems from available parallel corpora. Most statistical machine translation approaches are based on single-word translation models and do not take contextual information into account for translation.

The models in the phrase-based approach define 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. A simplified version of these models is introduced to allow for simple search strategies that have been applied with success in the translation between some pairs of languages. These models are based on a monotonicity assumption of alignments between source and target segments. Different methods to select adequate bilingual segments and to train the parameters of these models are presented and discussed in this article. A simplified decoder has also been developed for these models.

This 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.

1. Introduction.

The development of a classical *machine translation* (MT) system requires great human effort. *Statistical machine translation* (SMT) has proven to be an interesting framework for (quasi) automatically building MT systems if adequate parallel corpora are available (Brown et al., 1990).

The most common approach to SMT is based on two types of statistical models: A *(statistical) target language model* and a *translation model*. The most widely used target language model is the well-known (smoothed) n-gram model (Jelinek, 1998), which represents the statistical dependency

* Escuela Politécnica Superior de Gandia, Instituto de Tecnología Informática, 46730 Gandia, Spain. E-mail: jtomas@upv.es

† Departamento de Sistemas Informáticos y Computación, Instituto de Tecnología Informática, 46071 Valencia, Spain. E-mail: fcn@iti.upv.es

of sequences of n target words. The translation model is currently composed of families of complementary models that attempt to model *single-word* alignments through *statistical dictionaries* and models for dealing with the relation between positions in the source and in the target sentence (Brown et al., 1993; Ney et al., 2000; Och and Ney, 2003). In this case, the basic assumption is that each source word is generated by only one target word. This assumption does not correspond to the nature of natural language; in some cases, it is necessary to know the context of the word to be translated and, in other cases, it is convenient to translate whole word sequences instead of a word-by-word translation.

One way to upgrade this simple assumption is the use of statistical context-dependent dictionaries as in (Berger, Della Pietra, and Della Pietra, 1996; García-Varea and Casacuberta, 2005). Another way of overcoming the above-mentioned restriction of single-word models is known as the *template-based* (TB) approach (Och and Ney, 2004). In this approach, an entire group of adjacent words in the source sentence may be aligned with an entire group of adjacent target words. As a result, the word context has a greater influence and the changes in word order from source to target language can be learned explicitly. The alignment of word groups is carried out through *templates*. A template establishes the alignment (possibly through reordering) between two sequences of word classes (these classes are learned automatically using a bilingual corpus). However, the lexical model, which is inside the templates, continues to be based on word-to-word correspondences (Och, Tillmann, and Ney, 1999; Och and Ney, 2000; Och and Ney, 2004). A finite-state implementation of these models was presented in (Kumar, Deng, and Byrne, 2006).

A simple alternative to these models has been introduced in recent works: the *phrase-based* [1] (PB) approaches (Tomás and Casacuberta, 2001; Marcu and Wong, 2002; Zens, Och, and Ney, 2002; Zens and Ney, 2004; Tomás, Lloret, and Casacuberta, 2005). These methods allow for learning the probability that a sequence of contiguous words (*source segment*) in a source sentence is a translation of another sequence of contiguous words (*target segment*) in the target sentence. In this case, the statistical dictionaries of single-word pairs are substituted by statistical dictionaries of *bilingual phrases* or *bilingual segments*.

One shortcoming of the PB alignment models is the generalization capability, since only sequences of segments that have been seen in the training corpus are accepted. One possible solution is to combine this approach with templates (Tomás and Casacuberta, 2003). Another problem with the PB approach is the selection of appropriate phrases. Most of the

[1] Although the term "phrase" has a more restricted meaning, we will use it in this article as a possible word sequence.

methods used to generate bilingual segments are based on some "symmetrization" of word alignments (Och, 2002; Koehn, Och, and Marcu, 2003; Zens and Ney, 2004). There are other proposals that attempt to obtain phrases without having to build initial word-to-word alignments (Marcu and Wong, 2002; Zhang, Vogel, and Waibel, 2003; Tomás and Casacuberta, 2001; Watanabe, Sumita, and Okuno, 2003). Chunking is an interesting proposal for selecting linguistic-based word segments (Watanabe, Imamura, and Sumita, 2002; Koehn and Knight, 2002). Another possible way to select phrases is to use linguistic parsers (Koehn, Och, and Marcu, 2003; Tomás and Casacuberta, 2003; Zens and Ney, 2004). Phrases can also be obtained from *head-transducer translation models*, which are collections of weighted finite-state transducers that are organized in a way similar to recursive transition networks (Alshawi and Douglas, 1998). Other alternatives are based on *recursive alignments* (Nevado, Casacuberta, and Landa, 2004).

Other interesting models are the *stochastic finite-state transducers* (Casacuberta and Vidal, 2004) that are very appropriate for some speech-to-speech translation tasks (Casacuberta et al., 2004). These models are closely related to some restricted PB models.

SMT systems are far from being perfect. However, these SMT systems (using PB models in particular) can also be used in *computer-assisted translation* (CAT) to increase the productivity of the (human) translation process. The idea is to use a text-to-text translation system to produce portions of target text that can be accepted or amended by a human translator using text or speech. These user-validated portions are then used by the text-to-text translation system to produce further, hopefully improved, suggestions (Och, Zens, and Ney, 2003; Civera et al., 2004). PB models can also be used for CAT (Bender et al., 2005).

Statistical PB models are used in this article. The *monotone PB* models were initially introduced by the authors in (Tomás and Casacuberta, 2001) and are widely developed in this work. Recently, this approach has been adopted by another research group (Crego et al., 2005). These models seem to be adequate for pairs of related languages (such as the Romanic languages). These models have also proven to be adequate for some specific tasks where the languages were not Romanic languages. In addition to the development of these models, a new approach to obtain bilingual segments is proposed; new adaptations of the multi-stack decoding have also been also developed for monotone and non-monotone models and exhaustive experiments to prove the adequacy of the proposed techniques for some tasks are presented. The results obtained are comparable or better than the ones obtained using other statistical and non-statistical machine translation systems.

In the following section, the statistical framework for machine translation is reviewed and the monotone PB models are introduced. These models can be estimated from corpora of training pairs by using the algorithms in section 3. The translation engines that use these models are based on search algorithms, which are presented in section 4. Section 5 describes the corpora used and the experiments preformed. The experimental results are presented in section 6 and the method for computing the confidence of the results is discussed in an appendix. Finally, the conclusions can be found in section 7.

2. Statistical framework for machine translation.

The MT problem can be statistically stated as follows. Given a sentence $\mathbf{s}$ from a source language (based on a source vocabulary Σ), search for a target-language sentence $\hat{\mathbf{t}}$ (based on a target vocabulary Δ) which maximizes the posterior probability[2]:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}|\mathbf{s}) \ . \tag{1}$$

It is commonly accepted that a convenient way to deal with this equation is to transform it by using Bayes' theorem (Brown et al., 1990):

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s}|\mathbf{t}) \ , \tag{2}$$

where $\Pr(\mathbf{t})$ is estimated by a *target language model* q(typically a smoothed n-gram (Jelinek, 1998)) , which gives high probability to well-formed target sentences, and where $\Pr(\mathbf{s}|\mathbf{t})$ accounts for source-target word(-position) relations and is based on *stochastic dictionaries* and *alignment models* (Brown et al., 1993; Ney et al., 2000).

The translation models introduced by Brown et al. (Brown et al., 1993) to deal with $\Pr(\mathbf{s}|\mathbf{t})$ in equation 2 are based on the concept of alignment between the components of a pair $(\mathbf{s}, \mathbf{t})$ (*statistical alignment models*). Formally, if the number of words in $\mathbf{s}$ and in $\mathbf{t}$ are J and I, respectively, an *alignment* is a function $\mathbf{a} : \{1, ..., J\} \rightarrow \{0, ..., I\}$[3]. The particular case $\mathbf{a}_j = 0$ means that the position j in $\mathbf{s}$ is not aligned with any position in $\mathbf{t}$.

By introducing the alignment function $\mathbf{a}$ in $\Pr(\mathbf{s}|\mathbf{t})$,

$$\Pr(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} \Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) \ . \tag{3}$$

[2]For simplicity, the true distribution $\Pr(X = x)$ and $\Pr(X = x|Y = y)$ are denoted as $\Pr(x)$ and $\Pr(x|y)$. The model distributions are denoted as $P()$. The model parameters are denoted by $p()$.

[3]The image of j by $\mathbf{a}$ will be denoted as $\mathbf{a}_j$.

2.1 Single-word alignment models.
Different models have been proposed in (Brown et al., 1993; Ney et al., 2000) for dealing with $\Pr(\mathbf{s}, \mathbf{a}|\mathbf{t})$ in equation 3: Zero-order models such as the so-called *model 1* (M1), *model 2* (M2), and *model 3* (M3) (Brown et al., 1993); and first-order models such as *model 4* (M4), *model 5* (M5) (Brown et al., 1993), *t*he hidden Markov model (HMM) (Ney et al., 2000), and *model 6* (M6) (Och and Ney, 2003).

In all of these approaches, there are two types of models: *statistical dictionaries*, and models that take into account the probabilistic relationship between source and target positions (simple *alignment models* or *fertility distribution* plus *distortion distributions*) (Brown et al., 1993; Och and Ney, 2003).

Issues about learning (estimation) of such models and searching (decoding) can be found in (Brown et al., 1993; Berger et al., 1996; Tillmann et al., 1997; García-Varea, Casacuberta, and Ney, 1998; Knight, 1999; Germann, 2003; Och and Ney, 2003).

In models M3, M4, and M5, there is an implicit idea of (source) word groups that are aligned to a target word, *cept* (Brown et al., 1993). Another extension for dealing with multiple target words, *multicept*, was discussed in (Brown et al., 1993) and in (Goutte, Yamada, and Gaussier, 2004). In these cases, the words cannot be contiguous, and these models are based on word-to-word alignments.

2.2 Phrase-based alignment models.
To formalize the concept of bilingual phrase, only non-null segments of contiguous words are considered. It is also assumed that the number of source segments is equal to the number of target segments. Let J be the number of words in $\mathbf{s}$ [4], and let K be the number of bilingual segments (source / target segments). In this case, $\Pr(\mathbf{s}|\mathbf{t})$ in equation 2 can be rewritten as [5]:

$$\Pr(\mathbf{s}|\mathbf{t}) = \Pr(J|\mathbf{t}) \cdot \sum_{K} \Pr(K|\mathbf{t}, J) \cdot \Pr(\mathbf{s}_1^J|\mathbf{t}, J, K) \; . \tag{4}$$

Let I be the number of target words in $\mathbf{t}$. The segmentation of the target sentence is introduced as a function μ:

$$\mu : \{1, \ldots, K\} \rightarrow \{1, \ldots, I\} : \mu_k > \mu_{k-1}, \;\; 1 < k \leq K \;\; \& \;\; \mu_K = I,$$

[4] It is not strictly necessary to introduce this variable now, but it allows for a more clear presentation of the models.

[5] We will use the notation $\mathbf{s}_j^{j'}$ as a sequence of words in $\mathbf{s}$ from j to j'; this sequence is empty if $j' < j$.

On the other hand, the segmentation of the source sentence can be introduced as a function γ:

$$\gamma : \{1, \ldots, K\} \rightarrow \{1, \ldots, J\} : \gamma_k > \gamma_{k-1}, \;\; 1 < k \leq K \;\; \& \;\; \gamma_K = J,$$

Then,

$$\Pr(\mathbf{s}|\mathbf{t}) = \Pr(J|\mathbf{t}) \cdot \sum_{K} \sum_{\mu_1^K} \sum_{\gamma_1^K} \Pr(K|\mathbf{t}, J) \cdot \Pr(\mu_1^K, \gamma_1^K|\mathbf{t}, J, K) \cdot \Pr(\mathbf{s}|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K) \; . \tag{5}$$

The following example from Spanish to English will be used to illustrate the models to be proposed: Let *"Por favor , pmdame un taxi"* be a source sentence $\mathbf{s}$, and *"Could you ask for a taxi , please ?"* be the corresponding target sentence $\mathbf{t}$.

In this example, let K be 3; a possible segmentation of the source sentence into three segments could be $\gamma_1 = 3$, $\gamma_2 = 4$ and $\gamma_3 = 6$. Thus, the first source segment corresponds to *"por favor ,"*; the second source segment corresponds to *"pídame"*; and the third source segment corresponds to *"un taxi"*. On the other hand, a possible segmentation of the target sentence into three segments could be $\mu_1 = 4$, $\mu_2 = 6$, and $\mu_3 = 9$, Thus, the first target segment corresponds to *"could you ask for"*; the second target segment corresponds to *"a taxi"*; and the last target segment corresponds to *", please ?"*.

In translation, the correspondence between source and target segments can introduce a reordering of segments through a permutation:

$$\alpha : \{1, \ldots, K\} \rightarrow \{1, \ldots, K\} : \alpha_k = \alpha_{k'} \;\; \text{iff} \;\; k = k' \; .$$

Introducing the permutation in the last term of equation 5,

$$\Pr(\mathbf{s}|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K) = \sum_{\alpha_1^K} \Pr(\alpha_1^K|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K) \cdot \Pr(\mathbf{s}|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K, \alpha_1^K) \; . \tag{6}$$

In the above example, a possible reordering of the segments could be $\alpha_1 = 3$, $\alpha_2 = 1$ and $\alpha_3 = 2$, i.e. *"por favor ,"* is aligned with *", please ?"*; *"pmdame"* is aligned with *"could you ask for"*; *"un taxi"* is aligned with *"a taxi"*.

The two last terms in equation 6 can be factorized in different ways, one of which leads to the following [6]:

[6] To simplify the notation, we extend γ and μ for any $i, j \leq 0$ to $\gamma_j = 0$ and $ß_i = 0$, respectively.

$$\Pr(\mathbf{s}|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K) = \sum_{\alpha_1^K} \prod_{k=1}^{K} \Big\{ \Pr(\alpha_k|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K, \alpha_1^{k-1}) \cdot$$

$$\Pr(\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k}, \gamma_k|\mathbf{t}, J, K, \mu_1^K, \gamma_1^{k-1}, \alpha_1^k, \mathbf{s}_1^{\gamma_1}, \ldots, \mathbf{s}_{\gamma_{k-2}+1}^{\gamma_{k-1}}) \Big\} \ . \quad (7)$$

Finally, from equation 7 and equation 5,

$$\Pr(\mathbf{s}|\mathbf{t}) = \Pr(J|\mathbf{t}) \cdot \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \Pr(K|\mathbf{t}, J) \cdot \Pr(\mu_1^K, \gamma_1^K|\mathbf{t}, J, K)$$
$$\cdot \sum_{\alpha_1^K} \prod_{k=1}^{K} \Big\{ \Pr(\alpha_k|\mathbf{t}, J, K, \mu_1^K, \gamma_1^K, \alpha_1^{k-1}) \cdot$$
$$\Pr(\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k}, \gamma_k|\mathbf{t}, J, K, \mu_1^K, \gamma_1^{k-1}, \alpha_1^k, \mathbf{s}_1^{\gamma_1}, \ldots, \mathbf{s}_{\gamma_{k-2}+1}^{\gamma_{k-1}}) \Big\} \ . \quad (8)$$

Other possibilities can include a factorization of $\Pr(\mu_1^K, \gamma_1^K|\mathbf{t}, J, K)$ in order to produce more complex models.

2.2.1 Monotone phrase-based alignment models. Different approaches can be adopted for equation 8. For the simplest one, it can be assumed that all segmentations have the same probability:

$$\Pr(J|\mathbf{t}) \cdot \Pr(K|\mathbf{t}, J) \cdot \Pr(\mu_1^K, \gamma_1^K|\mathbf{t}, J, K) \approx p_1 \ . \quad (9)$$

Another simple assumption is that each source phrase depends only on the target phrase that has been aligned:

$$\Pr(\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k}, \gamma_k|\mathbf{t}, J, K, \mu_1^K, \gamma_1^{k-1}, \alpha_1^k, \mathbf{s}_1^{\gamma_{k-1}}) \approx p(\widetilde{s}_k|\widetilde{t}_{\alpha_k}) \ , \quad (10)$$

where $\widetilde{s}_k$ corresponds to the value $\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k}$ of a random variable on Σ^* [7] and $\widetilde{t}_{\alpha_k}$ corresponds to the value $\mathbf{t}_{\mu_{k'-1}+1}^{\mu_{k'}}$ of a random variable on Δ^* aligned with $\widetilde{s}_k$ ($k' = \alpha_k$).

If monotonicity is assumed, $\alpha_k = k$, then,

$$\Pr(\mathbf{s}|\mathbf{t}) \approx P(\mathbf{s}|\mathbf{t}) = p_1 \cdot \sum_K \sum_{\widetilde{s}_1^K} \sum_{\widetilde{t}_1^K} \prod_{k=1}^{K} p(\widetilde{s}_k|\widetilde{t}_k) \ . \quad (11)$$

¿From equation 11, $\widetilde{s}_1 \ldots \widetilde{s}_K = \mathbf{s}$ and $\widetilde{t}_1 \ldots \widetilde{t}_K = \mathbf{t}$. The parameter p_1 is not relevant for translation and will be omitted. The only parameters of this

[7] By Σ^* and Δ^*, we will denote the sets of finite-length sentences that can be obtained from the vocabularies Σ and Δ, respectively.

model are of the form $p(\widetilde{s}|\widetilde{t})$ for a pair of a generic source segment $\widetilde{s}$ (from Σ^*) and a generic target segment $\widetilde{t}$ (from Δ^*). $(\widetilde{s},\widetilde{t})$ defines a *bilingual segment*. These parameters estimate that the probability of translation of a given segment $\widetilde{t}$ is the translation of a segment $\widetilde{s}$.

In practice, the sums in equation 11 are approximated by a maximization:

$$P(\mathbf{s}|\mathbf{t}) \approx \max_K \max_{\widetilde{s}_1^K} \max_{\widetilde{t}_1^K} \prod_{k=1}^{K} p(\widetilde{s}_k|\widetilde{t}_k) \; . \tag{12}$$

2.2.2 Non-monotone phrase-based models. In this case, reordering of segments is permitted and the probability of an alignment α_k in equation 7 can depend on the last alignment α_{k-1} (first-order alignment):

$$\Pr(\alpha_1^K|\mathbf{t}, J, K, \mu_1^K) \approx \prod_{k=1}^{K} p(\alpha_k|\alpha_{k-1}) \; . \tag{13}$$

Following the notation used with monotone PB models,

$$\Pr(\mathbf{s}|\mathbf{t}) \approx P(\mathbf{s}|\mathbf{t}) = \sum_K \sum_{\widetilde{s}_1^K} \sum_{\widetilde{t}_1^K} \sum_{\alpha_1^K} \prod_{k=1}^{K} p(\alpha_k|\alpha_{k-1}) \cdot p(\widetilde{s}_k|\widetilde{t}_{\alpha_k}) \; . \tag{14}$$

For the distortion model -$p(\alpha_k|\alpha_{k-1})$-, it is assumed that an alignment depends only on the distance of the two segments (Och and Ney, 2000):

$$p(\alpha_k|\alpha_{k-1}) = p_0^{|\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}}|} \; , \tag{15}$$

where p_0 is a parameter to be adjusted.

As for equation 12, the maximization can be used as an approximation to the sums:

$$P(\mathbf{s}|\mathbf{t}) \approx \max_K \max_{\widetilde{s}_1^K} \max_{\widetilde{t}_1^K} \max_{\alpha_1^K} \prod_{k=1}^{K} p_0^{|\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}}|} \cdot p(\widetilde{s}_k|\widetilde{t}_{\alpha_k}) \; . \tag{16}$$

2.3 Log-linear models.
In practice, the following equation is used for SMT instead of equation 2:

$$\begin{aligned} \hat{\mathbf{t}} &= \operatorname*{argmax}_{\mathbf{t}} \frac{\lambda_1 \cdot \log P(\mathbf{t}) + \lambda_2 \cdot \log P(\mathbf{s}|\mathbf{t})}{\sum_{\mathbf{t}'} \lambda_1 \cdot \log P(\mathbf{t}') + \lambda_2 \cdot \log P(\mathbf{s}|\mathbf{t}')} \\ &= \operatorname*{argmax}_{\mathbf{t}} \{\lambda_1 \cdot \log P(\mathbf{t}) + \lambda_2 \cdot \log P(\mathbf{s}|\mathbf{t})\} \end{aligned} \tag{17}$$

This equation is a particular case of a log-linear model for $\Pr(\mathbf{t}|\mathbf{s})$ (Och and Ney, 2004):

$$\begin{aligned}\hat{\mathbf{t}} &= \underset{\mathbf{t}}{\operatorname{argmax}} \frac{\sum_{i=1}^{N} \lambda_i \cdot \log f_i(\mathbf{t},\mathbf{s})}{\sum_{\mathbf{t}'} \sum_{i=1}^{N} \lambda_i \cdot \log f_i(\mathbf{t}',\mathbf{s})} \\ &= \underset{\mathbf{t}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \lambda_i \cdot \log f_i(\mathbf{t},\mathbf{s}) \right\} . \end{aligned} \tag{18}$$

With $f_1(\mathbf{t},\mathbf{s}) = P(\mathbf{t})$, $f_2(\mathbf{t},\mathbf{s}) = P(\mathbf{s}|\mathbf{t})$ and $N = 2$, equation 18 becomes 17. Another possibility is $f_2(\mathbf{t},\mathbf{s}) = P(\mathbf{t}|\mathbf{s})$, where $P(\mathbf{t}|\mathbf{s})$ can be built in a similar way as in the previous section for $P(\mathbf{s}|\mathbf{t})$ (Tomás and Casacuberta, 2002; Och, 2002):

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \{\lambda_1 \cdot \log P(\mathbf{t}) + \lambda_2 \cdot \log P(\mathbf{t}|\mathbf{s})\} . \tag{19}$$

Other heuristic assumptions can be adopted, for example:

$$\begin{aligned}\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \{&\lambda_1 \cdot \log P(\mathbf{t}) + \lambda_2 \cdot \log P(\mathbf{t}|\mathbf{s}) + \\ &\lambda_3 \cdot \log I + \lambda_4 \cdot \log P(\mathcal{C}(\mathbf{t})) + \lambda_5 \cdot \log K\} , \end{aligned} \tag{20}$$

where $\mathcal{C}(\mathbf{t})$ is the sequence of word categories in the target sentence $\mathbf{t}$, I is the number of target words in $\mathbf{t}$, and K is the number of segments. If n-grams are adopted as a target language model, greater values for n can be used for $P(\mathcal{C}(\mathbf{t}))$ than for $P(\mathbf{t})$ for the same training data due to the well-known estimation problems. On the other hand, the use of I allows for the control of the number of target words to be produced. Finally, the use of $P(\mathbf{t}|\mathbf{s})$ has allowed us to obtain better results than the use of $P(\mathbf{s}|\mathbf{t})$ in many MT experiments (Och and Ney, 2004). Obviously, a combination of both should also be explored. The heuristic assumptions that have been adopted in this article will be presented in section 5.3.

3. Learning phrase-based alignment models.

There are different approaches to the parameter estimation with the PB models from equations 11 and 14. The following subsections are devoted to the estimation of monotone PB models. For non-monotone models, the procedures are similar and the parameter that controls the phrase reordering p_0 is adjusted using a validation set.

The first approach for the parameter estimation of equation 11 corresponds to a direct learning of the parameters from a sentence-aligned corpus using a maximum likelihood approach. The second one is a heuristic that tries to use word alignments.

3.1 Training with a sentence-aligned corpus.

Given a sentence-aligned corpus $\mathcal{T}$, composed of pairs of sentences $(\mathbf{s}, \mathbf{t})$, the maximum likelihood criterium attempts to estimate the parameters $p(\widetilde{s}|\widetilde{t})$ for all bilingual segments $(\widetilde{s}, \widetilde{t})$ that maximize (Tomás and Casacuberta, 2001; Marcu and Wong, 2002):

$$\prod_{(\mathbf{s},\mathbf{t})\in\mathcal{T}} P(\mathbf{s}|\mathbf{t}), \tag{21}$$

using equation 11 or equation 14, subject to the constraints that hold for each target segment $\widetilde{t}$:

$$\sum_{\widetilde{s}} p(\widetilde{s}|\widetilde{t}) = 1\ . \tag{22}$$

By applying an EM procedure (Moon, 1996), the corresponding re-estimation formula for the monotone model is:

$$\widehat{p}\,(\widetilde{s}|\widetilde{t}) = \Gamma_{\widetilde{t}}^{-1} \cdot \sum_{(\mathbf{s},\mathbf{t})\in\mathcal{T}} \sum_{K} \sum_{\widetilde{s}_1^K} \sum_{\widetilde{t}_1^K} \left(\prod_{k=1}^{K} p(\widetilde{s}_k|\widetilde{t}_k) \cdot \sum_{l=1}^{K} \delta(\widetilde{s} = \widetilde{s}_k) \cdot \delta(\widetilde{t} = \widetilde{t}_k) \right), \tag{23}$$

where $\Gamma_{\widetilde{t}}$ is a normalization factor, and $\delta(a) = 1$ if $a = true$ and $\delta(a) = 0$ id $a = false$.

The computation of equation 23 can be performed using a *forward-backward* algorithm similar to the one proposed in (Casacuberta, 1995; Picó and Casacuberta, 2001) for stochastic finite-state transducers.

3.2 Training with a word-aligned corpus.

The parameters of the model can also be obtained using a word-aligned corpus (Zens, Och, and Ney, 2002; Koehn, Och, and Marcu, 2003; Tomás, Lloret, and Casacuberta, 2005). These alignments can be obtained from the statistical models of section 2.1 using the available public software GIZA++ (Och and Ney, 2003).

Training of PB models from a word-aligned corpus consists of building of a set of bilingual segments and the estimation of the corresponding model parameters.

The most widely used approaches to define the set of bilingual segments are based on word alignments. However, the word-alignment models usually adopted do not permit the alignment of one source word to many target words (Brown et al., 1993). The strategy proposed in (Och, Tillmann, and Ney, 1999; Och, 2002) tried to solve this problem in two steps. In the first step, *symmetrized alignments* are computed from a combination of the alignments obtained in a translation direction (s→t) and alignments obtained from the opposite direction (t→s). Different methods of combination were proposed: *intersection*, *union* and *refined* (Och

and Ney, 2003). The bilingual segments are built from these symmetrized alignments, following different criteria in the second step.

Another alternative strategy also consists of two steps but is different from the steps in (Och, 2002). In the first step, two sets of bilingual segments were obtained: one set from word-alignments in one direction (s→t) and another set from word-alignments in the opposite direction (t→s). In the second step, these two sets of bilingual segments are combined to generate the definitive set of bilingual segments (*symmetrization*). This strategy is developed in this section.

Extracting bilingual segments. In the first step of the proposed strategy (and in the second step of the approach proposed in (Och and Ney, 2003)), different criteria can be used to define the set of bilingual segments $BilPhr$ in the sentence pair $(\mathbf{s}, \mathbf{t})$ with a word alignment $\mathbf{a}$. Figure 1 shows three different criteria.

The first criterium for selecting bilingual phrases (segments) is strict and it is defined as:

$$BilPhrStr(\mathbf{s},\mathbf{t},\mathbf{a}) = \left\{ (\mathbf{s}_{j_1}^{j_2}, \mathbf{t}_{i_1}^{i_2}) : \begin{array}{c} \forall j : j_1 \le j \le j_2; \exists i : i_1 \le i \le i_2 : \mathbf{a}_j = i \\ \forall i : i_1 \le i \le i_2; \exists j : j_1 \le j \le j_2 : \mathbf{a}_j = i \end{array} \right\}. \tag{24}$$

This criterion considers $(\mathbf{s}_{j_1}^{j_2}, \mathbf{t}_{i_1}^{i_2})$ as a bilingual phrase if all the words in $\mathbf{s}_{j_1}^{j_2}$ are aligned with a word in $\mathbf{t}_{i_1}^{i_2}$, and vice versa (Tomás and Casacuberta, 2003).

The second criterium can be considered as an extension of the previous one:

$$BilPhrExt(\mathbf{s},\mathbf{t},\mathbf{a}) = \left\{ (\mathbf{s}_{j_1}^{j_2}, \mathbf{t}_{i_1}^{i_2}) : \begin{array}{c} \forall j : j_1 \le j \le j_2; (i_1 \le \mathbf{a}_j \le i_2) \vee (\mathbf{a}_j = 0) \\ \forall j\text{: } (j < j_1) \vee (j_2 < j)\text{: } (\mathbf{a}_j < i_1) \vee (i_2 < \mathbf{a}_j)) \end{array} \right\}, \tag{25}$$

but it allows some words in $\mathbf{s}_{j_1}^{j_2}$ or in $q\mathbf{t}_{i_1}^{i_2}$ to be unaligned (Zens, Och, and Ney, 2002; Tomás and Casacuberta, 2003).

The third criterium forces the bilingual phrases (segments) to be extracted in a monotone way (Casacuberta and Vidal, 2004):

$$BilPhrMon(\mathbf{s},\mathbf{t},\mathbf{a}) = \left\{ (\mathbf{s}_{j_1}^{j_2}, \mathbf{t}_{i_1}^{i_2}) : \begin{array}{c} \forall j : j_1 \le j \le j_2; (i_1 \le \mathbf{a}_j \le i_2) \vee (\mathbf{a}_j = 0) \\ \forall j : j < j_1; \mathbf{a}_j < i_1 \quad \forall j : j > j_2; \mathbf{a}_j > i_2 \end{array} \right\}. \tag{26}$$

In other words, it does not permit the extraction of a bilingual phrase if there is a word at the left of the source sequence that has been aligned to a word at the right of the target sequence (or vice versa).

Symmetrization. The second step in the approach developed here is devoted to obtaining the final set of bilingual segments. This set is a combination of the models obtained from the word alignments in one direction

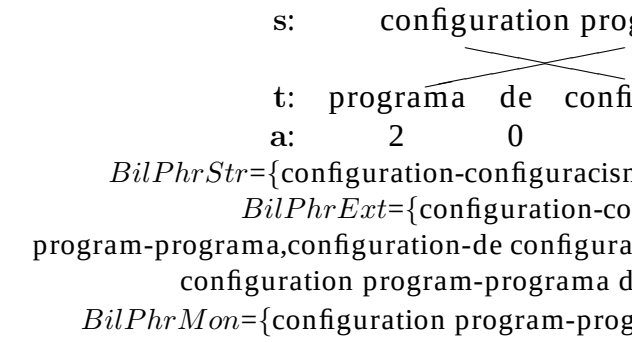

Figure 1

Example of the extraction of a set of bilingual multiword sequences from a word-aligned sentence using three different criteria. In the example, s_1 has been aligned with t_2 ($\mathbf{a}_1 = 2$), s_2 has not been aligned with ($\mathbf{a}_2 = 0$), and s_3 has been aligned with t_1 ($\mathbf{a}_3 = 1$).

and the models obtained from the word alignments in the opposite direction. We present two new alternatives in order to combine these two sets of bilingual segments. The first one is simple: The final set is obtained as a simple addition of the two previous sets (*addition*).

The second approach consists of building two separate PB models, one from the alignments in one direction (to compute $P_{s\rightarrow t}(\mathbf{t}|\mathbf{s})$) and the other from the alignments in the opposite direction (to compute $P_{t\rightarrow s}(\mathbf{t} \mid \mathbf{s})$). Then, both models are combined using a log-linear interpolation (*interpolation*):

$$P(\mathbf{t}|\mathbf{s}) = P_{s\rightarrow t}(\mathbf{t}|\mathbf{s})^{1-\lambda} \cdot P_{t\rightarrow s}(\mathbf{t}|\mathbf{s})^{\lambda} \tag{27}$$

Estimating the parameters of the models. The estimation of the parameters of the model can be done via relative frequencies, for each pair of segments $(\mathbf{s}, \mathbf{t})$:

$$p(\widetilde{s}|\widetilde{t}) = \frac{N(\widetilde{s}, \widetilde{t})}{N(\widetilde{t})}, \tag{28}$$

where $N(\widetilde{t})$ denotes the number of times that phrase $\widetilde{t}$ has appeared in the training set, and $N(\widetilde{s}, \widetilde{t})$ is the number of times that the bilingual segment $(\widetilde{s}, \widetilde{t})$ has appeared in the training set.

A refined way of estimating the parameters can be carried out by combining the re-estimation formulae in equation 23 and the bilingual phrases $BilPhr$ introduced in this section. In this case, the bilingual phrases that are involved in equation 23 must be consistent with word-alignments computed with the training pairs. The corresponding re-

estimation formula is:

$$\widehat{p}\,(\widetilde{s}|\widetilde{t}) \;=\; \lambda_{\widetilde{t}}^{-1} \cdot \sum_{(\mathbf{s},\mathbf{t})\in\mathcal{T}} \sum_{K} \sum_{\widetilde{s}_1^K} \sum_{\widetilde{t}_1^K} \left(\prod_{k=1}^{K} p(\widetilde{s}_k|\widetilde{t}_k) \cdot \sum_{l=1}^{K} \delta(\widetilde{s},\widetilde{s}_k) \cdot \delta(\widetilde{t},\widetilde{t}_k) \cdot \delta((\widetilde{s},\widetilde{t}) \in BilPhr(\mathcal{T})) \right), \qquad (29)$$

where $BilPhr(\mathcal{T})$ is the set of bilingual phrases obtained from $\mathcal{T}$ by applying $BilPhrStr$, $BilPhrExt$ or $BilPhrMon$.

3.3 Model smoothing

The unseen events (bilingual phrases) in a training set are modelled through a simple (smoothing) method: When a segment in a source sentence is not found in the repertory of bilingual segments, shorter segments are looked for and, in the extreme case, these short segments can be single words. The edit distance (substitutions, insertions, and deletions) can also be used as another smoothing technique.

In some cases the phrase translation probabilities are poorly estimated, as for example, when a phrase appears only one or two times in the training set. A lexical model can be used as an additional feature to smooth the PB models. Using these lexical models in equation 20:

$$\begin{aligned} \hat{\mathbf{t}} \;=\; \underset{\mathbf{t}}{\operatorname{argmax}} \{ & \lambda_1 \cdot \log P(\mathbf{t}) + \lambda_2 \cdot \log P(\mathbf{t}|\mathbf{s}) + \\ & \lambda_3 \cdot \log I + \lambda_4 \cdot \log P(\mathcal{C}(\mathbf{t})) + \lambda_5 \cdot \log K + \\ & \lambda_6 \cdot \log P_L(\mathbf{t}|\mathbf{s}) + \lambda_7 \cdot \log P_L(\mathbf{s}|\mathbf{t}) \}, \end{aligned} \qquad (30)$$

where $P_L(\mathbf{t}|\mathbf{s})$ and $P_L(\mathbf{s}|\mathbf{t})$ are the probabilities computed by the lexical models.

Model M1 of section 2.1 can be used to computed the probability P_L (Och et al., 2004). In this case, only a *stochastic dictionary* is required. As an alternative, P_L can be computed from *lexical weights* (LW) that are estimated from the word alignments in the bilingual segments (Koehn, Och, and Marcu, 2003).

4. Search algorithms for machine translation.

The goal of searching in machine translation is to compute a target sentence $\hat{\mathbf{t}}$ that maximizes the probability of some models for $\Pr(\mathbf{s}|\mathbf{t})$. The search problem with statistical alignment models is a computationally difficult problem (Knight, 1999). In practice, only approximations can be computed. Different search strategies have been proposed to define

the way in which the search space is explored. A best-first strategy is used in the A^* algorithms in (Wang and Waibel, 1997; Wang and Waibel, 1998; Och, Ueffing, and Ney, 2001). A breadth-first strategy is used in the *dynamic-programming* algorithms proposed in (García-Varea, Casacuberta, and Ney, 1998; Tillmann and Ney, 2003) and in a *beam-search* algorithm proposed in (Och and Ney, 2004; Koehn, 2004a). Some type of combination of depth-first and best-first strategies is used in the so-called *multi-stack-decoding* algorithm (Berger et al., 1996; Ortiz, García-Varea, and Casacuberta, 2003). Finally, a *greedy* strategy can also be used (Germann, 2003).

The basis of most of these algorithms is to generate partial hypotheses about the target sentence in an incremental way. Each of these hypotheses is composed of a prefix of the target sentence, a subset of source positions that have been aligned with the positions of the prefix of the target sentence, and a score. New hypotheses can be generated from a previous hypothesis by adding a (some) target word(s) to the prefix of the target sentence, which is the translation of a(some) source word(s) that has(have) not yet been translated.

For PB models, similar search strategies can be proposed, for example, a *beam-search* algorithm in (Zens, Och, and Ney, 2002; Marcu and Wong, 2002; Koehn, 2004a) which is basically a breadth-first strategy. *Pharaoh* is a beam-search decoder that is freely available for researchers (Koehn, 2004a)

The search algorithm for PB models adopted in this work is based on a simplified best-first strategy where the hypothesis are stored in different data structures (sorted lists) depending on which words in the source sentence have been translated. This strategy is similar to a *multi-stack decoding scheme*. This procedure allows us to force the expansion of hypotheses with a different degree of completion. In each iteration, the algorithm explores all lists and extends the best hypothesis found in each list. For an optimal search, the algorithm must continue until there are no more partial hypotheses to be expanded. However, in practice we stop the search after several iterations (*Max-iter*). The approximations of Eqs. 12 or 16 are used.

In the following subsections, three alternative algorithms are proposed: the first algorithm for the monotone models of section 2.2.1, and the last two algorithms for the non-monotone models of section 2.2.2.

4.1 Monotone search algorithm.

In the particular case of monotone models, a hypothesis in the *monotone search* (MS) algorithm consists of a prefix of the source sentence, a prefix of the target sentence (which is the translation of the prefix of the

source sentence), and the score given for the models for this partial translation. More formally, given a source sentence $\mathbf{s}$, a hypothesis is a tuple $(\mathbf{s}_1^j, \mathbf{t}_1^i, S = P(\mathbf{t}_1^i) \cdot P(\mathbf{s}_1^j | \mathbf{t}_1^i))$, where $P(\mathbf{t}_1^i)$ is calculated according to a target language model (n-gram), and $P(\mathbf{s}_1^j | \mathbf{t}_1^i)$ is calculated according to equation 12.

The algorithm requires a different list for each length of the source prefixes. In these lists, the hypotheses that have been generated by the algorithm are stored according to the length of the source prefix.

The initialization consists of building a hypothesis for the empty target and empty source prefixes with a score of 1.0. The algorithm selects the best hypothesis from each list and generates a new hypothesis for each (theoretically) possible bilingual phrase whose source segment matches the words following the source prefix of the selected hypothesis.

After each iteration, the best final hypothesis can be used to define a pruning criterion to remove those partial hypotheses that cannot improve the best final hypothesis.

The computational cost of this algorithm is proportional to: the number of words in the source sentence to be translated (J); the maximum number of source words in a bilingual segment; the maximum number of bilingual segments that have the same source segment and the maximum number of iterations (*Max-iter*). The costs of the basic list operations are not considered. In practice, the parameter *Max-iter* has been used to increase the speed of the translation.

4.2 A search algorithm based on source-word reordering.

The procedures for non-monotone models can be quite similar to the MSA, but the search algorithm is based on a *source word reordering* (SWR). In this case, a hypothesis consists of a prefix of the target sentence ($\mathbf{t}_1^i$), a coverage set of source positions ($\mathcal{C}$), and a score (S). There is one list for each possible set of source positions whose words have already been translated. The possible number of lists can be very high; consequently, the lists are only created when they are required.

In each iteration, as in the previous algorithm, the best hypothesis from each available list is selected to generate new extended hypothesis. The target element of each new hypotheses is the concatenation of the target element of the selected hypotheses and the target element of each possible bilingual phrase. However, the source phrase must match consecutive free source positions; i.e., positions that are not in the set of visited source positions of the selected hypothesis.

The computational cost of this non-monotone search is clearly higher than the monotone search of the previous section. In the present case, the algorithm cost is proportional to: The number of words in the source

sentence to be translated (J); the maximum number of source words in a bilingual segment; the maximum number of bilingual segments that have the same source segment; the number of sets of source positions (2^J) and the maximum number of iterations (*Max-iter*). And in the previous section, the costs of the basic list operations are not considered. The main difference of this cost and the cost of monotone search is in the use of sets of source positions that introduces an exponential complexity. In practice, this exponential cost requires the use of heuristic strategies to prune the search space. In our implementations, the time complexity has been controlled by the parameter *Max-iter* and by the use of the search constraints proposed in (Berger et al., 1996).

4.3 A search algorithm based on target-word reordering.

The previous algorithm for non-monotone models presents a high computational cost, given that the maximum number of possible lists can be very high (2^J). There are some proposals (Vogel et al., 2003; Koehn, 2004a; Och and Ney, 2004) that try to solve this problem by storing the hypothesis with the same number of elements in $\mathcal{C}$ in the same list. The main problem with such proposals is the comparison of hypotheses that cover different parts of the source sentence. To solve this problem, an estimation (heuristic function) of the contribution to the score of the parts that are not yet covered can be introduced (Wang and Waibel, 1997).

We present a new approach, where the source sentence is translated left to right, and we introduce a possible *target-word reordering* (TWR) (Tomás and Casacuberta, 2004). Here, we define a hypothesis in the same way as in the monotone algorithm, and each hypothesis is also stored in a separate list according to the source-length prefix. In contrast to the monotone case, we can introduce the special token $\langle nul \rangle$ in the target hypothesis. The meaning of this token is that, in a future expansion, the token $\langle nul \rangle$ must be replaced by a sequence of words.

A hypothesis could include an arbitrary number of $\langle nul \rangle$ tokens; however, in our implementation, we allow only one. Therefore, we can distinguish between two classes of hypotheses. A hypothesis is closed if it does not contain the token $\langle nul \rangle$, and it is open if it contains this token.

In the process of generating a new hypothesis from a selected one, a new bilingual phrase ($\widetilde{s}$,$\widetilde{t}$) is considered. $\widetilde{s}$ must match the source segment that follows the last source word that has been translated. If the hypothesis to be extended is closed, two new hypotheses are created: by concatenation of $\widetilde{t}$ and of $\langle nul \rangle \widetilde{t}$ to the the target prefix of the old hypothesis, respectively.

On the other hand, if the hypothesis is open, four new hypotheses are created: One for closing the open hypothesis by replacing the token

actual $\langle nul \rangle$ by $\widetilde{t}$; and three new open hypotheses. These last three hypotheses are built by putting $\widetilde{t}$ at the left or right of $\langle nul \rangle$, or by putting $\widetilde{t}$ at right of the target prefix (Note that, in an open hypothesis $\langle nul \rangle \widetilde{t}$ can be in any position of the target part of the hypothesis). This algorithm uses a different approach to approximate the distortion probability because, when we open a hypothesis, the final position of the target part of the hypothesis cannot be known. Thus, we have a different parameter distortion for each type of extension: If the hypothesis is closed, we use the probability p_o to open it, and $1 - p_o$ to keep it closed. If the hypothesis is open, we use the probabilities p_c to close it. $(1 - p_c)/3$ is used for the other three extension types.

Another problem to be solved is the evaluation of the language model in an open hypothesis since we cannot calculate the language model contribution of the right part of the prefix after the $\langle nul \rangle$ token (we do not know which words will replace this special token). To solve this problem, we compute an estimation of the language model contribution. It consists of assigning the probability of its unigram to the word at the right of $\langle nul \rangle$ times the probability of the bigram for the next word, etc. When a hypothesis is closed, this estimation is replaced by the true language model contribution.

Taking into account that the number of lists is upper bounded by the length of the source sentence and the possible local reordering is also bounded by a constant, the computational cost of this algorithm is of the same order as the monotone search algorithm.

5. Experimental framework.

The models introduced in section 2 and the procedures presented in sections 3 and 4 were assessed through a series of experiments with different corpora. These corpora and the assessment measures used are described in this section and the results are presented in section 6.

5.1 Corpora.

Four different corpora were used:

- *"El Periódico"*: a bilingual newspaper (from Spanish to Catalan) (Tomás et al., 2001).
- *XRCE*: Xerox printer manuals (from English to Spanish, French and German and from Spanish, French, and German to English) (Khadivi and Goutte, 2003).
- *EU*: Bulletin of the European Union (from English to Spanish, French, and German; and from Spanish, French, and German to

English) (Khadivi and Goutte, 2003).

- *Hansards*: Proceedings of the Canadian Parliament (from French to English) (Brown et al., 1990).

Table 1
The "El Periódico" corpus from Spanish to Catalan. There was no overlap between training and test sentences. Trigram models were used to compute the test-set perplexity. (mK means $m \times 1,000$ and mM means $m \times 1,000,000$).

		Spanish	Catalan
Train	Sentence pairs	644K	
	Running words	7.1M	7.4M
	Vocabulary	129K	128K
Test	Sentences	240	
	Running words	4.3K	4,4K
	Test perplexity	199	204

The first corpus was obtained from the electronic newspaper "El Periódico de Cataluña" (Tomás et al., 2001). The training set corresponds to daily information from the newspaper over 10 months. The test set was acquired from different sources: part from the same newspaper (which was not in the training set); part from a technical manual; part from "Diario Oficial de la Generalidad Valenciana", the legislative bulletin of the local government, etc. The main characteristics of this corpus are presented in Table 1. This corpus allowed the comparison of the proposed approach with other knowledge-based approaches.

The second corpus was the XRCE corpus that was obtained from different user manuals for Xerox printers in the framework of the TT2 project (TransType-2, 2001). The main characteristics of this corpus are presented in Table 2.

The third corpus was the EU corpus (Khadivi and Goutte, 2003). The data source is the Bulletin of the European Union, which is published in the eleven official languages of the European Union. This corpus is publicly available on the Internet. This corpus was acquired and processed in the framework of the TT2 project (TransType-2, 2001). The features of this corpus are presented in Table 3.

The last corpus was the Proceedings of the Canadian Parliament (Brown et al., 1990). We used the aligned corpus provided in (Germann, 2001). More specifically, we used the sub-corpus *Senate Debates training*

Table 2
The "XRCE" corpus from English to Spanish, German, and French, and vice-versa. There was no overlap between the training and test sentences, and the test set did not contain out-of-vocabulary words with respect to any of the training sets. Trigram models were used to compute the test-set perplexity. (mK means $m \times 1,000$)

		English Spanish		English German		English French	
Train	Sentence pairs	56K		49K		53K	
	Running words	665K	753K	633K	696K	587K	534K
	Vocabulary	8K	11K	8K	10K	8K	19K
Test	Sentence pairs	1,125		984		996	
	Running words	8K	10K	11K	12K	12K	12K
	Test perplexity	48	33	51	87	73	52

Table 3
The "EU" corpus from English to Spanish, German, and French, and vice-versa . Trigram models were used to compute the test-set perplexity. (mK means $m \times 1,000$ and mM means $m \times 1,000,000$)

		English Spanish		English German		English French	
Train	Sentence pairs	214K		223K		215K	
	Running words	5.9M	6.6M	6.5M	6.1M	6.0M	6.6M
	Vocabulary	84K	97K	87K	152K	85K	91K
Test	Sentence pairs	800		800		800	
	Running words	22K	25K	22K	21K	22K	24K
	Test perplexity	96	72	95	153	97	71

for learning the models, and the first 1000 sentences of the *testing 1* for evaluation. The main characteristics of this corpus are presented in Table 4,

5.2 Assessment.

In all the experiments reported in this paper, the translations of the source test sentences produced by the translation systems were compared with

Table 4
The "Hansards" corpus from French to English. Trigram models were used to compute the test-set perplexity. (mK means $m \times 1,000$ and mM means $m \times 1,000,000$)

		English	French
Train	Sentence pairs	182K	
	Running words	3.0M	3.2M
	Vocabulary	39K	53K
Test	Sentences	1,000	
	Running words	19K	21K
	Test perplexity	71	-

target test references and some values were computed:

- *Word error rate* (WER): The minimum number of substitution, insertion, and deletion operations needed to convert the word string hypothesized by the translation system into a given single reference word string (Och and Ney, 2003; Tillmann and Ney, 2003).
- *BiLingual Evaluation Understudy* (BLEU): it is based on the n-grams of the hypothesized translation that occur in the reference translations. The BLEU metric ranges from 0.0 (worst score) to 1.0 (best score) (Papineni et al., 2002).

An important question arises when these assessment values are computed from a given test data: to determine whether the observed differences between two methods are statistically significant or just caused by chance. Paired bootstrap can be used for this purpose (Bisani and Ney, 2004; Koehn, 2004b; Zhang and Vogel, 2004). More details can be found in appendix A.

In some of the experiments in this paper, we highlight the statistical significance as follows. First, we select a system as reference (in the tables, the corresponding results are presented in boldface), and we compare it with the rest of the systems in the same task. A result labeled with a "▼" ("▽") means that the system is worse than the reference with a confidence of 95% (90%). A $''-''$ means for no significant differences. The use of "▲" ("△") is similar to "▼" ("▽") but, in this case, the system is better than the reference.

5.3 Optimization of Model Scaling Factors.

In the experiments, we used the approach presented in section 2.3 that produced the best results in practice. This approach consists of a log-linear combination of the language model, the translation, and several feature functions (target length penalty, phrase penalty). Each feature function needs a scaling factor parameter in order to adjust the importance of the different features. Equation 20 is used in the experiments. The search for the most adequate scaling factors was based on the criterium of maximizing the error rate of a developed corpus which was independent from the test corpus (Och, 2003). This maximization process was carried out by using the downhill simplex algorithm (Bender et al., 2004). Some scaling factors are presented in Table 5. In this process, we used a developed set which was different from the test set [8] and we minimize the combined score obtained from WER minus BLEU.

Table 5
Some scaling factors used in the experiments for each corpus.

	λ_1 $P(\mathbf{t})$	λ_2 $P(\mathbf{t}\|\mathbf{s})$	λ_3 I	λ_4 $P(\mathcal{C}(\mathbf{t}))$	λ_5 K
"El Periódico"	1	13	0	0	2
XRCE (English-Spanish)	1	12	4	0	-2
EU (English-Spanish)	1	13	-3	2	-2
Hansards	1	10	5	0	0

6. Results.

Different sets of experiments were carried out to study different aspects of PB modeling for machine translation. In the first set of experiments, different estimation procedures of the model parameters were compared. The second set of experiments were carried out to study the effect of the maximum length of bilingual segments on the translation performance. The third set of experiments was designed to study the different ways to generate bilingual segments. In the fourth set of experiments, the monotone and non-monotone models were compared. The fifth set of experiments was devoted to studying the effect of the size of the training set on the system performance. In the following experiments, a smoothing

[8] In the XRCE task, the developed set was the same as the test set.

technique was explored. In the last experiments, the proposed approach was compared with other machine translation systems.

The aspects that are not studied in each specific experiment are fixed to a standard configuration: Maximum phrase length was 3, 16, 8 and 6 for the "El Periódico", XRCE, EU, and "Hansards", respectively; the size corpus was the one in the corresponding feature tables (tables 1, 2, 3 and 4); the parameter estimation was the relative frequencies using the direct model; the extraction criterion was $BilPhrExt$ without symmetrization and the search was monotone (*Max-iter*=32). Equation 20 was used in the experiments, with the exception of those in subsection 6.6, where equation 30 was used.

6.1 Estimating the parameters of the models.

In section 3, three different parameter-estimation procedures were described. The first one was based on corpora aligned at the sentence level (equation 23), while the other procedures were based on corpora aligned at the word level (equation 28 and equation 29). In addition, the search was carried out using equation 20 (*direct approach* (Tomás and Casacuberta, 2002) (Och, 2002)) or the alternative version that consisted of substituting $P(\mathbf{t}|\mathbf{s})$ by $P(\mathbf{s}|\mathbf{t})$ in equation 20 (*source-channel approach* (Brown et al., 1990)). The results obtained using the source-channel approach and the direct approach on the "EL Periódico" and the XRCE corpora are presented in Table 6.

As Table 6 indicates, the direct approach obtains better results than the source-channel approach. This behavior has been observed in other corpora but not in other (non machine-translation) applications (Ney, Popović, and Sündermann, 2004). A possible reason for this is that the quality of the estimated models is not very high. Another possible reason is that the use of heuristics to speed up the search obtain, suboptimal solutions.

Table 6 shows that the parameter estimation with a word-aligned corpus using relative frequencies obtains the best results. One possible reason might be that the EM training presented allows only a monotone alignment; and in some cases, this assumption does not correspond to the nature of the corpus.

The direct approach with models trained with relative frequencies are used in the rest of this section.

6.2 Maximum length of segments.

In practice, the number of possible bilingual phrases can be very large, which can introduce computational problems. To overcome these problems, the phrases cannot be of arbitrary lengths. The introduction of

Table 6
Effects of the parameter estimation procedures and the approaches used (source-channel or direct) on the WER (%) and BLEU (%) in one of the XRCE tasks (English to Spanish) and in the "El Periódico" task (Spanish to Catalan). "SA" refers to equation 23, "WA, freq." refers to 28, and "WA, EM alg." refers to equation 29. The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

Approach	Parameter estimation	XRCE (English-Spanish)		"El Perisdico" (Spanish-Catalan)	
		WER	BLEU	WER	BLEU
Source-channel	SA	29.0$^{\blacktriangledown}$	60.8$^{\blacktriangledown}$	13.5$^{\blacktriangledown}$	75.1$^{\blacktriangledown}$
	WA, freq.	26.6$^{\blacktriangledown}$	63.7$^{\blacktriangledown}$	10.8^{-}	79.2^{-}
	WA, EM alg.	28.3$^{\blacktriangledown}$	61.6$^{\blacktriangledown}$	12.9$^{\blacktriangledown}$	76.3$^{\blacktriangledown}$
Direct model	SA	27.9$^{\blacktriangledown}$	61.4$^{\blacktriangledown}$	13.4$^{\blacktriangledown}$	75.3$^{\blacktriangledown}$
	WA, freq.	**24.7**	**64.9**	**10.7**	**79.2**
	WA, EM alg.	26.8$^{\blacktriangledown}$	62.2$^{\blacktriangledown}$	12.9$^{\blacktriangledown}$	76.4$^{\blacktriangledown}$

upper bounds in the segment lengths was studied in some experiments whose results are presented in Figures 2, 3 and 4 for the (English-Spanish) XRCE, (English-French) EU and the (French-English) "Hansards" corpora, respectively.

The use of long segments in XRCE and EU tasks allow us to significantly improve the results obtained. This behavior is not so evident in the "Hansards" task. In contrast, the use of long segments leads to a huge number of parameters.

6.3 Different methods for building segments.

In section 3.2, two new methods (addition and interpolation) were proposed within a new strategy (first select the two sets of bilingual segments and then combine them) to obtain an adequate set of symmetrized bilingual segments. In this section, we compare them (using $BilPhrExt$ as the criterium for the selection of bilingual segments) with two methods that use alignments in only one direction (s→t or t→s), as well as with the three methods described in (Och and Ney, 2003) (intersection, union and refined[9]) within the other strategy (combining the models obtained in both directions) to obtain an adequate set of symmetrized bilingual seg-

[9]These bilingual segments have been obtained using the toolkit presented in (Ortiz, García-Varea, and Casacuberta, 2005).

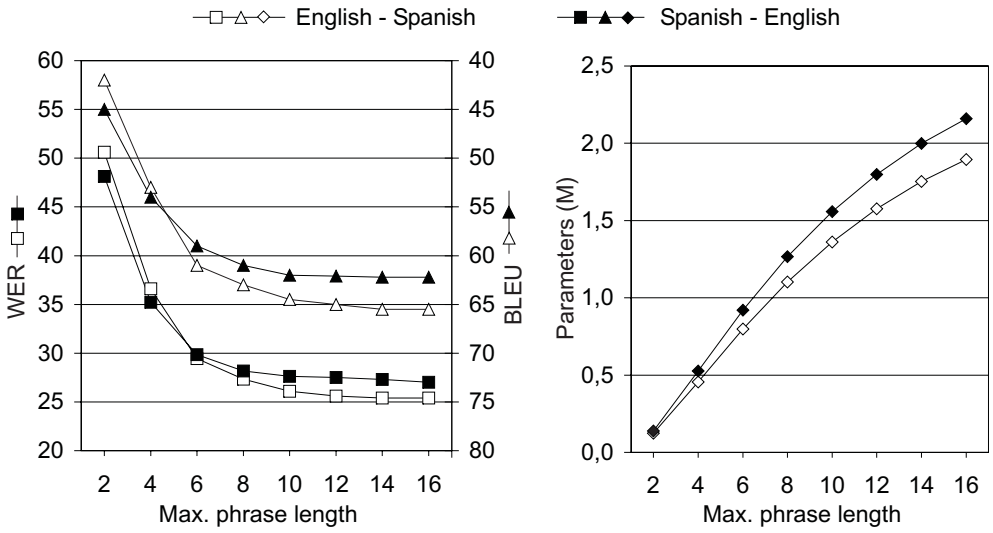

Figure 2
Effect of the size of the segment length on the WER (%), BLEU (%), and the number of parameters ($\times 1,000,000$) for one of the XRCE tasks (English-Spanish).

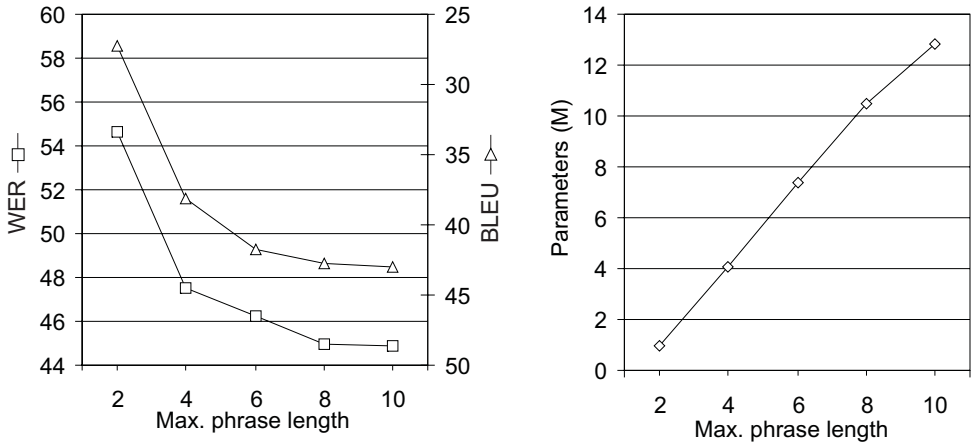

Figure 3
Effect of the size of the segment length on the WER (%), BLEU (%), and the number of parameters ($\times 1,000,000$) for one of the EU tasks (English-French).

ments. The experimental comparison of all of these methods is presented in Table 7.

The differences among the different approaches are not so important,

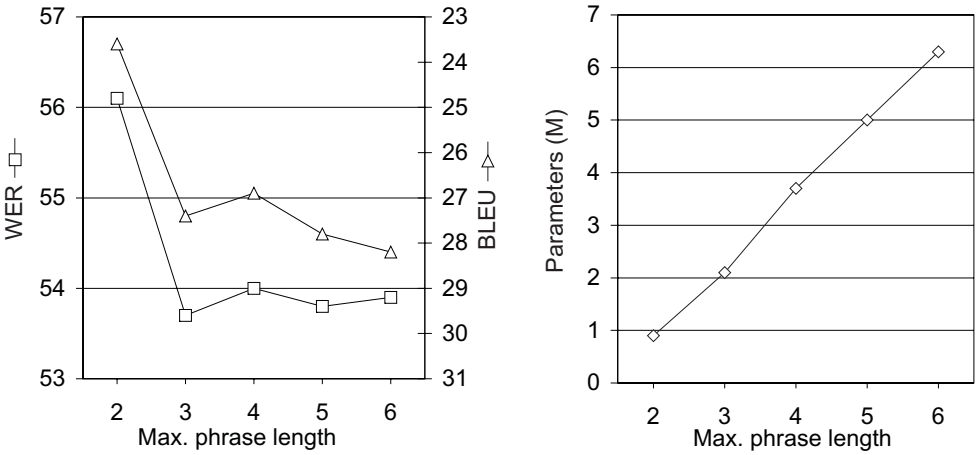

Figure 4
Effect of the size of the segment length on the WER (%), BLEU (%), and the number of parameters ($\times 1,000,000$) for the "Hansards" task (French-English).

Table 7
Effect of the symmetrization methods (using $BilPhrExt$) in WER (%), BLEU (%) and number of model parameters (#Par) on one of the XRCE (English to Spanish) and "Hansards" corpora (French to English). The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

Symmetrization method	XRCE (English-Spanish)			"Hansards" (French-English)		
	WER	BLEU	#Par	WER	BLEU	#Par
s→t	24.7$^{\blacktriangledown}$	64.9$^{\blacktriangledown}$	1.7M	54.0$^{\blacktriangle}$	28.2$^{\blacktriangledown}$	6.3M
t→s	25.5$^{\blacktriangledown}$	53.8$^{\blacktriangledown}$	2.0M	55.9$^{\blacktriangledown}$	29.6$^{\blacktriangledown}$	5.6M
Intersection	24.8$^{\blacktriangledown}$	64.8$^{\blacktriangledown}$	1.3M	55.6$^{\blacktriangledown}$	30.9^{-}	2.6M
Union	25.7$^{\blacktriangledown}$	63.3$^{\blacktriangledown}$	2.0M	56.1$^{\blacktriangledown}$	27.0$^{\blacktriangledown}$	7.9M
Refined	24.8$^{\blacktriangledown}$	64.9$^{\blacktriangledown}$	1.6M	54.5$^{\triangle}$	30.2^{-}	5.1M
Addition	24.3^{-}	65.3^{-}	2.6M	54.1$^{\blacktriangle}$	29.1$^{\blacktriangledown}$	9.4M
Interpolation	**24.1**	**65.8**	3.7M	**54.9**	**31.0**	11.9M

but the symmetrization by interpolation seems to present the best behavior. The number of parameters in each method corresponds to the size of the set of bilingual phrases. In the addition approach, this number is the cardinal of the intersection of the sets obtained in each direction. In the interpolation approach, this number is the sum of the sizes of both

sets. The interpolation parameter (λ) in equation 27 has been adjusted in a way similar to the scaling parameters in section 5.3. The value of this parameter was set to 0.8.

Table 8
Effect of the extraction criterion of the set of bilingual phrases in WER (%) and BLEU (%) on the XRCE tasks. Results are obtained using *addition* as the symmetrization method. The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

XRCE languages	$BilPhrStr$ WER	$BilPhrStr$ BLEU	$BilPhrExt$ WER	$BilPhrExt$ BLEU	$BilPhrMon$ WER	$BilPhrMon$ BLEU
English-Spanish	26.0$^{\blacktriangledown}$	62.8$^{\blacktriangledown}$	**24.3**	**65.3**	25.4$^{\blacktriangledown}$	64.3$^{\blacktriangledown}$
Spanish-English	26.4^{-}	60.0$^{\blacktriangledown}$	**26.2**	**62.1**	27.6$^{\blacktriangledown}$	58.5$^{\blacktriangledown}$
English-French	52.2$^{\blacktriangle}$	34.4^{-}	**52.8**	**34.3**	52.9^{-}	33.9^{-}
French-English	50.8$^{\blacktriangle}$	33.2$^{\triangledown}$	**51.4**	**33.9**	52.8$^{\blacktriangledown}$	31.6$^{\blacktriangledown}$
English-German	63.9$^{\triangle}$	21.5$^{\blacktriangledown}$	**64.4**	**23.3**	66.9$^{\blacktriangledown}$	21.7$^{\blacktriangledown}$
German-English	53.2$^{\triangledown}$	28.2$^{\blacktriangledown}$	**53.8**	**31.1**	56.2$^{\blacktriangledown}$	28.4$^{\blacktriangledown}$

The criteria for extracting bilingual segments (section 3.2) were also explored on the XRCE corpus and the results are presented in Table 8.

¿From these results, $BilPhrExt$ presents the best behavior in most of the cases. $BilPhrExt$ represents a good trade-off between $BilPhrStr$ and $BilPhrMon$.

6.4 Monotone vs. non-monotone search.

Three different search algorithms have been presented in section 4: A monotone search algorithm in section 4.1 and two non-monotone search algorithms in sections 4.2 and 4.3. Figure 5 compares the translation speed and WER/BLEU for several values of *Max-iter* (1,2,4,..,64) for one pair of languages in the XRCE and in the EU tasks. Threshold pruning was not used in these experiments. Values of *Max-iter* greater than 16/32 do not improve significantly the results. Source-word reordering and target-word reordering algorithms present a similar behavior; however, the second one is faster and requires less memory for lists [10].

A complete comparison between these monotone and non-monotone models are presented in Table 9. For non-monotone models we have used a source-word reordering algorithm, with *Max-iter*=64.

The results obtained in these experiments are comparable. This is an

[10] The experiments were performed on a Pentium 4 (3.2 GHz) with 1.5 G of RAM memory.

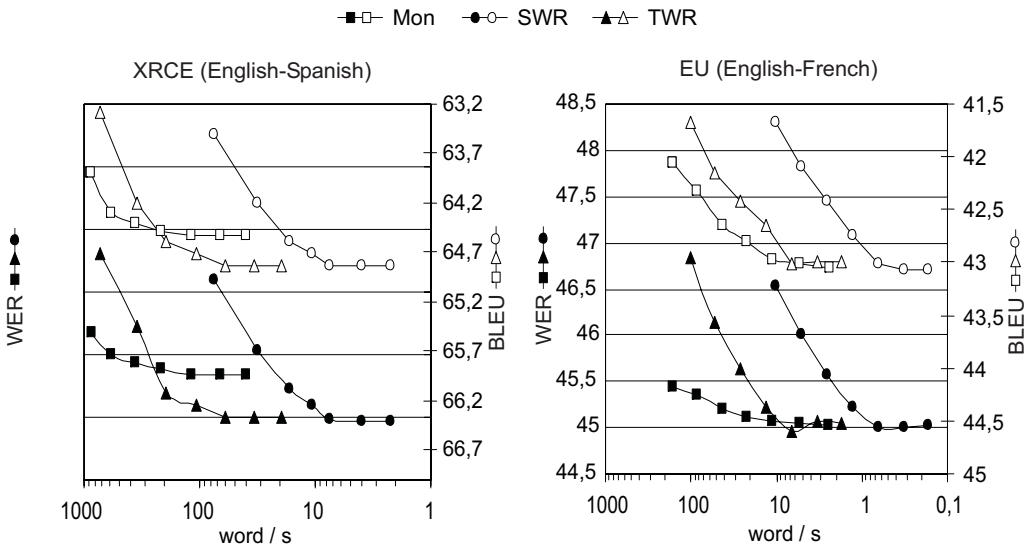

Figure 5
WER (%), BLEU (%) and translation speed as words by second (word / s) obtained in the three search algorithm on XRCE task (English to Spanish) and on EU task (English to French). Each curve is composed by 7 values that correspond left-to-right to the results obtained for several values of *Max-iter* (1, 2, 4, 8, 16, 32 and 64). "MS" stands for monotone search (section 4.1), "SWR" for non-monotone search based on source-word reordering (section 4.2) and "TWR" for non-monotone search based on target-word reordering (section 4.3).

interesting result, since the non-monotone search presents a higher qcomputational cost than the monotone search (sections 4.1 and 4.2).

Further experiments were also carried out using the beam-search decoder *Pharaoh* (Koehn, 2004a). In this case, the performance of the translations was also comparable to the ones obtained with the search algorithms proposed in this article. However, Pharaoh required less computational time than our decoder, but considerably more memory. This memory problem was important for the EU task.

6.5 Effect of the training set on the system performance.

Further experiments were carried out on the "El Periódico" and the Hansards corpora to study the effect of the amount of training data on the quality of the PB models. The results with training sets from $5,000$ to $640,000$ pairs for the "El Periódico" corpus, and from $5,000$ to the whole training corpus for the Hansards corpus are presented in Figures 6 and 7.

Figure 6shows that when the training set is large enough (from $320,000$

Table 9
Effect of different procedures for searching on the WER (%) and BLEU (%) for the XRCE, EU and "Hansards" tasks. The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

Corpus	Languages	Monotone		Non-monotone	
		WER	BLEU	WER	BLEU
XRCE	English-Spanish	**24.7**	**64.9**	24.4^{-}	65.5^{-}
	Spanish-English	**26.6**	**61.1**	26.8^{-}	60.7^{-}
	English-French	**53.7**	**32.9**	53.5^{-}	33.6^{-}
	French-English	**52.2**	**32.2**	52.0^{-}	$32.9^{\triangle}$
	English-German	**66.1**	**21.9**	66.3^{-}	$22.1^{\blacktriangledown}$
	German-English	**54.1**	**31.0**	54.4^{-}	30.8^{-}
EU	English-Spanish	**46.7**	**42.1**	46.7^{-}	42.3^{-}
	Spanish-English	**46.7**	**41.8**	46.6^{-}	42.1^{-}
	English-French	**45.0**	**42.8**	45.1^{-}	42.8^{-}
	French-English	**42.1**	**46.5**	42.2^{-}	46.6^{-}
	English-German	**60.2**	**27.1**	60.1^{-}	27.2^{-}
	German-English	**53.6**	**34.2**	53.6^{-}	34.4^{-}
Hansards	French-English	**53.9**	**28.2**	53.9^{-}	28.4^{-}

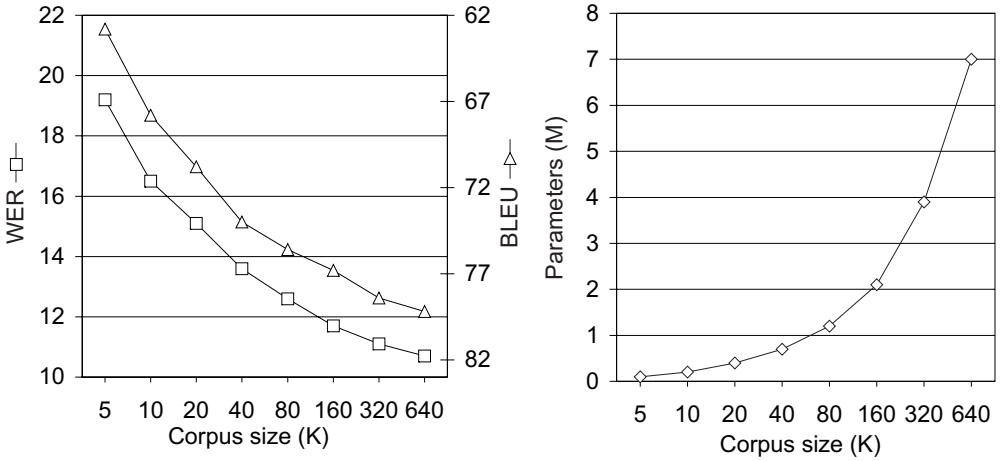

Figure 6
Effect of the size of the training data on the WER (%), BLEU (%), and the number of model parameters ($\times 1,000,000$) in the "El Periódico" task (Spanish to Catalan).

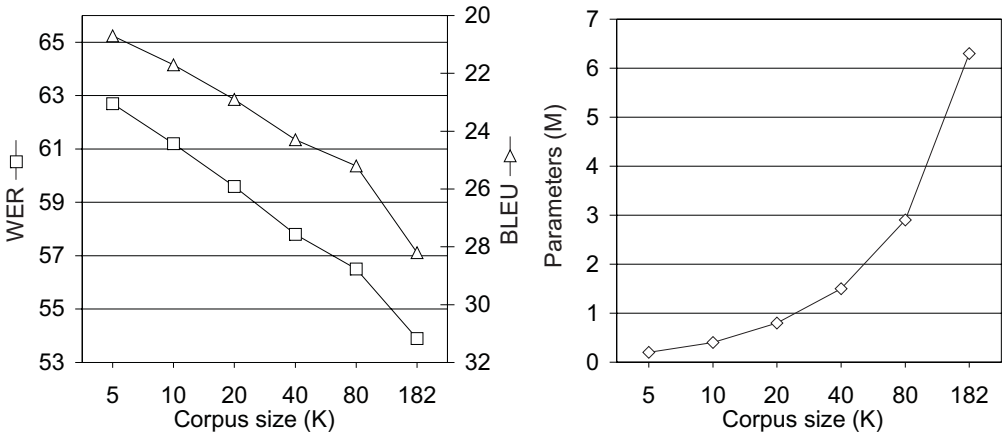

Figure 7
Effect of the size of the training data on the WER (%), BLEU (%), and the number of model parameters ($\times 1,000,000$) in the "Hansards" task (French to English).

to $640,000$ pairs), the improvement of performance is slight (3.6% of WER), but the size of the models increases dramatically (95%). A similar behavior was observed in the Hansard corpus.

6.6 Using lexical models as additional features.

Some experiments were carried out to study the influence of the combination of lexical models (subsection 3.3) and PB models. In this section equation 30 is used. The weights λ_6 and λ_7 are estimated in a way similar to the one in section 5.3. Some results are presented in Table 10 using both approaches

The introduction of M1 as lexical models significantly improved the results obtained with the baseline for the XRCE corpus and only in one case for the Hansards corpus. However, the lexical weights do not appear to be what causes the improvements.

6.7 Comparison with other machine translation systems.

A final series of experiments was carried out to compare the MT system based on PB models with other available machine translation systems. The corpus used was the "El Periódico" (Spanish to Catalan). One of these systems was **Salt**, a knowledge-based machine translation system supported by the Government of the Generalitat Valenciana (http://www.cultgva.es). The second system was the translation system of the **Instituto Cervantes**, developed by AutomaticTrans

Table 10
Effect of the introduction of several types of lexical models as a additional feature in WER (%) and BLEU (%) on of the XRCE tasks (English to Spanish), Hansards corpora (French to English) and EU corpora (English to French). As lexical models, by *M1* it is denoted the use of model M1 and by *LW* the use of lexical weights. The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

Lexical models	XRCE English-Spanish		"Hansards" French-English		EU English-French	
	WER	BLEU	WER	BLEU	WER	BLEU
Baseline	**24.7**	**64.9**	**53.9**	**28.2**	**45.2**	**42.8**
+ *M1* $P_L(\mathbf{t}\|\mathbf{s})$ ($\lambda_6 = 0$)	24.1$^{\blacktriangle}$	65.4$^{\blacktriangle}$	53.5$^{\blacktriangle}$	28.3^{-}	45.1^{-}	42.7^{-}
+ *M1* $P_L(\mathbf{s}\|\mathbf{t})$ ($\lambda_5 = 0$)	24.4^{-}	65.9$^{\blacktriangle}$	54.2^{-}	29.7$^{\blacktriangle}$	44.6$^{\blacktriangle}$	44.7$^{\blacktriangle}$
+ *M1* $P_L(\mathbf{t}\|\mathbf{s})\&P_L(\mathbf{t}\|\mathbf{s})$	23.9$^{\blacktriangle}$	66.3$^{\blacktriangle}$	53.9^{-}	29.8$^{\blacktriangle}$	44.5$^{\blacktriangle}$	44.6$^{\blacktriangle}$
+ *LW* $P_L(\mathbf{t}\|\mathbf{s})$	24.6^{-}	65.0^{-}	53.7^{-}	28.5^{-}	45.1^{-}	42.8^{-}
+ *LW* $P_L(\mathbf{s}\|\mathbf{t})$	24.5^{-}	65.1^{-}	53.6$^{\triangle}$	28.6$^{\triangle}$	45.0^{-}	42.9^{-}

(http://oesi.cervantes.es/traduccionAutomatica.html). Another system was **Incyta**, a knowledge-based commercial system (http://www.incyta.com). Finally, the fourth system was **InterNOSTRUM**, a hybrid knowledge-based and finite-state translation system (http://www.internostrum.com) (Canals-Marote et al., 2001). The results are presented in Table 11.

Table 11 shows that the PB system presents comparable results with other knowledge-based systems; however, these systems require great amounts of human effort over many years. The proposed system only requires a few months. On the other hand, the performance-speed ratio of PB systems is better than the other approaches.

In the TransType2 project (TransType-2, 2001), several statistical translation models were tested using the XRCE and EU corpora. In (Zens and Ney, 2004), some results were presented using *alignment templates* (Och, 2002) for the XRCE corpus. With this approach, the WER was 28.9% in the Spanish-English direction. In the same task, PB models obtained a WER of 26.2%.

Finally, the results obtained with the EU corpus are presented in Table 12. In this case, the results are compared with the AT technique (Bender et al., 2005). Statistical significance is not reported since the translation outputs using the AT technique were obtained in the TT2 project and they are not available. For some pairs of languages, the results are similar, but for the other pairs, PB models achieved better results.

Table 11
Results obtained with different machine translation systems on the "El Periódico" task (Spanish to Catalan) in WER (%) and BLEU (%). The numbers of translated words per second are also reported. The values in boldface and the symbols $^{\blacktriangledown}$, $^{\triangledown}$, $^{-}$, $^{\triangle}$ and $^{\blacktriangle}$ are used to represent the confidence of the results (section 5.2).

MT system	WER	BLEU	Translation speed (words/sec.)
Inst. Cervantes	$9.3^{\blacktriangle}$	$82.5^{\blacktriangle}$	160
Salt	9.5^{-}	82.2^{-}	12
Phrase-based	**10.2**	**81.1**	510
Incyta	10.6^{-}	81.0^{-}	Not available
InterNOSTRUM	$11.3^{\blacktriangledown}$	$79.2^{\blacktriangledown}$	5000

Table 12
Comparative results of PB models with AT models for the EU corpus (Bender et al., 2005) in WER (%) and BLEU (%). The PB models were obtained using monotone search, without symmetrization and one additional language model based on word categories.

EU Languages	AT		PB	
	WER	BLEU	WER	BLEU
English-Spanish	46.9	41.5	46.7	42.1
Spanish-English	48.3	41.2	46.7	41.8
English-French	45.1	42.1	45.0	42.8
French-English	44.0	44.6	42.1	46.5
English-German	61.1	32.4	60.2	27.1
German-English	56.2	33.8	53.6	34.2

7. Discussion and conclusions.

Phrase-based models constitute a very promising approach for statistical machine translation. In these models, the translation unit is the word sequence or segment ("phrase"), and the relationship between a source segment and a target segment is formalized through monotone and non-monotone segment alignments. The main parameters in these models are

the probabilities of a dictionary composed of bilingual phrases or segments. One of the merits of such models is their ability to take into account the context in translation. In addition, this phrase-based approach is very simple (especially the one based on monotone alignments), and the search is very fast. This method can obtain good translation results for certain tasks such as restricted-domain tasks or translations between Romanic languages. For an unrestricted task in Spanish-Catalan translation, results similar to those obtained from some rule-based commercial systems have been obtained using the method presented here.

The *monotone phrase-based models* constitute one of the main contributions of this article. To perform the translation of a given source sentence with these monotone models, a specific efficient search algorithm has been developed. The search with this algorithm allows for a translation speed of several hundred words per second.

Another contribution for the search with non-monotone models has also been proposed in this article. In this algorithm, the source sentence is translated left-to-right, and the target sentence is reordered. This algorithm presents a lower computational cost than others like the beam-search presented in (Och and Ney, 2004; Koehn, 2004a), and it does not require the definition of a heuristic function.

An important drawback of this method is that it does not have generalization capability in word reordering; however, this does not seem important in the corpora that were used in the experiments. Similar conclusions were drawn (Zens and Ney, 2004) with *alignment templates.*

Different procedures for statistical estimation of the parameters of the phrase-based models have been proposed elsewhere. Another contribution of this article is constituted by two new procedures. The first one (Tomás and Casacuberta, 2001) is based on a corpus aligned at the sentence level. The second procedure is based on a corpus aligned at the word level and consists of: first, obtaining bilingual segments in both translation directions; second, estimating the corresponding probabilistic parameters; and third, combining both models by addition or by addition or interpolation.

The phrase-based models (estimation and search) have been tested through many experiments using different corpora and languages. The main conclusions that can be drawn from these experiments are the following: the use of long bilingual phrases can achieve less translation errors; the best training procedure is based on relative frequencies; monotone and non-monotone searches obtain similar translation results; monotone search requires less computational requirements than non-monotone search. However, the distortion model used in the non-monotone search is quite simple, and more complex models can be explored in the future.

Another minor conclusion is that the direct approach produces better results on the corpora used than the source-channel.

In the future, other methods to extract bilingual phrases should be explored. Another way to deal with the low generalization capability of the proposed models can be the combination of the phrase-based approach and the alignment-template approach. More robust smoothing methods should also be explored. Monotone phrase-based models are closely related to stochastic finite-state transducers, and this could help in the design of more efficient search algorithms.

A. Statistical significance of the results

To determine whether the observed differences between the results (WER or BLEU) of the two methods were statistically significant, paired bootstrap was used (Bisani and Ney, 2004; Koehn, 2004b; Zhang and Vogel, 2004).

Given an initial test T_0, a thousand new artificial tests, T_i, were created by resampling with replacement T_0. Then, we obtained *delta-WER* as:

$$\delta_i = \text{WER}_A(T_i) - \text{WER}_B(T_i),$$

where $\text{WER}_s(T_i)$ was the WER obtained by the system s ($s \in \{A, B\}$) using the artificial test T_i. We used δ_i to estimate the two-tailed confidence interval and the *probability of improvements* (poi), i.e. the probability of $\delta_i < 0$. The poi was estimated by counting the times that $\text{WER}_A(T_i) < \text{WER}_B(T_i)$ (Bisani and Ney, 2004). The same process was performed using for BLEU measure to obtain *delta-BLEU*.

The poi as a function of the different *delta-WER* and *delta-BLEU* obtained in some of the experiments in section 6 are represented in table 8.

Figure 8 shows that if WER difference is greater than 0.7, the poi is always greater than 97.5%, so we can state that this difference is statistically significant (using the standard criterion of 95% confidence interval). Note that in some case, a minor difference (e.g. 0.5) can be also significant. In the experiments in Table 11, systems of different origin are compared. In this case, we need a BLEU difference greater than 1.6 to guarantee the statistical significance. In the rest of experiments, the same system is compared with slight modifications. Here, the results are highly correlated, and a BLEU difference of 0.8 guarantees the significance.

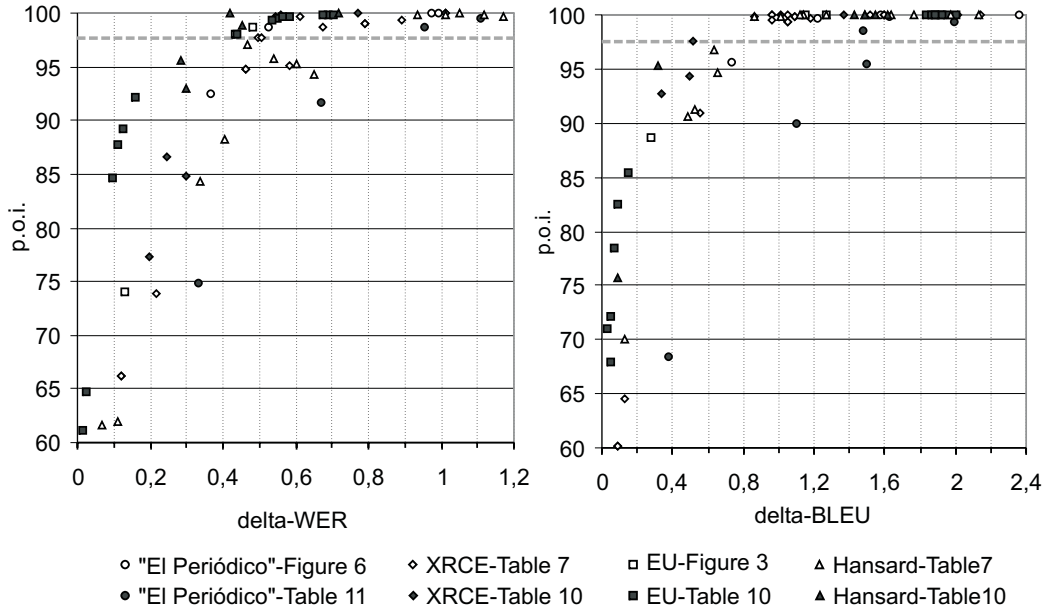

Figure 8
The delta-WER and delta-BLEU versus probability of improvement for some experiments in section 6: For XRCE, “Hansards” and “El Periódico” tasks.

Acknowledgments

TIC2003-08681-C02-02 the IST Programme of the European Union under grant IST-2001-32091. The authors wish to thank the anonymous reviewers for their criticisms and suggestions.

References

Alshawi, H. and S. Douglas. 1998. Learning phrase-based head transduction models for translation of spoken utterances. In *Proceedings of the International Conference on Speech and Language Processing 98 (ICSLP-98)*, pages 2767–2770, Sydney, Australia, October.

Bender, O., F. Casacuberta, C. Goutte, S. Hasan, S. Khadivi, K. Macherey, E. Matusov, A. Sanchis, H. Ney, M. Popovic, J. Tomás, N. Ueffing, D. Vilar, and R. Zens. 2005. Combined models, extended alignment models and morphosyntax incorporated model for long-range dependencies, me models and use of wordnet (deliverable d6.3). Technical report, TransType2(IST-2001-32091), RWTH Aachen and ITI.

Bender, O., R. Zens, E. Matusov, and H. Ney. 2004. Alignment templates: the rwth smt system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 79–84, Kyoto, Japan, September.

Berger, A. L., P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. 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, Apr.

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

Bisani, M. and H. Ney. 2004. Bootstrap estimates for confidence intervals in asr performance evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412, Montreal, Canada, May.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–310.

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M.I.. Guardiola-Savall, A. Iturraspe-Bellver,

S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P. Pérez-Antón, and M.L. Forcada. 2001. The Spanish-Catalan machine translation system interNOSTRUM. In *Proceedings of the Machine Translation Summit VIII*, pages 73–76, Santiago de Compostela, Spain, September.

Casacuberta, F. 1995. Probabilistic estimation of stochastic regular syntax-directed translation schemes. In A. Calvo and R. Medina, editors, *Proceedings of the VI Spanish Symposium on Pattern Recognition and Image Analysis*, pages 201–207, Córdoba, España.

Casacuberta, F., H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47.

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

Civera, Jorge, Juan M. Vilar, Elsa Cubel, Antonio L. Lagarda, Sergio Barrachina, Enrique Vidal, Francisco Casacuberta, David Picó, and Jorge González. 2004. From machine translation to computer assisted translation using finite-state models. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP04)*, pages 349–356, Barcelona, July.

Crego, J.M., M.R. Costa-jussà, J.B. Mariño, and J.A.R. Fonollosa. 2005. Ngram-based versus phrase-based statistical machine translation. In *Proceedings of the International Workshop Spoken Language Translation (IWSLT)*, pages 177–184, Pittsburgh, PA, USA, October.

García-Varea, I. and F. Casacuberta. 2005. Learn context-dependent lexicon models for statistical machine translation. *Machine Learning*, 59:1–24.

García-Varea, I., F. Casacuberta, and H. Ney. 1998. An iterative, dp-based search algorithm for statistical machine translation. In *Proceeding of the International Conference on Spoken Language Processing (ICSLP'98)*, pages 1235–1238, Australia, November.

Germann, U. 2001. Aligned Hansards of the 36th Parliament of Canada. Release 2001-1a.

Germann, U. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the Human Language Technology and North*

American Association for Computational Linguistics Conference (HLT/NAACL), pages 1–8, Edmonton, Canada, May.

Goutte, Cyril, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 502–509, Barcelona, Spain, July.

Jelinek, F. 1998. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts.

Khadivi, S. and C. Goutte. 2003. Tools for corpus alignment and evaluation of the alignments (deliverable d4.9). Technical report, TransType2(IST-2001-32091), RWTH Aachen and Xerox Co.

Knight, K. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

Koehn, P. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the The 6th Conference of the Association for Machine Translation in the Americas (AMTA04)*, volume 3265 of *Lecture Notes in Artificial Intelligence*, pages 115–124, Georgetown University, Washington DC, USA, September-October. Springer.

Koehn, Philipp. 2004b. Statistical significance tests for machine translation evaluation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain, July.

Koehn, Philipp and Kevin Knight. 2002. ChunkMT: Statistical machine translation with richer linguistic knowledge. Draft, Unpublished.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta, May-June.

Kumar, S., Y. Deng, and W. Byrne. 2006. A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering*. In press.

Marcu, Daniel and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora EMNLP-02*, pages 133–139, Philadelphia, PA, USA, July.

Moon, T.K. 1996. The expectation-maximization algorithm. *Signal*

Processing Magazine, 13(6):47 – 60, Nov.

Nevado, F., F. Casacuberta, and J. Landa. 2004. Translation memories enrichment by statistical bilingual segmentation. In *Proceeedings of the IV International Conference on Language Resources and Evaluation - LREC2004*, volume 1, pages 335–338, Lisbon, May. ELRA.

Ney, H., S. Nießen, F. Och, H. Sawaf, C. Tillmann, and S. Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Transactions on Speech and Audio Processing*, 8(1):24–36.

Ney, H., M. Popović, and D. Sündermann. 2004. Error measures and bayes decision rules revisited with applications to pos tagging. In *Proceedings of the Conference on Empirical methods in Natural Language Processing (EMNLP)*, pages 270–276, Barcelona, Spain, July.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Och, F. J., N. Ueffing, and H. Ney. 2001. An efficient a* search algorithm for statistical machine translation. In *Proceedings of the Data-Driven Machine Translation Workshop*, pages 55–62, Toulouse, France, July.

Och, F.J. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, October.

Och, F.J., D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A.x Fraser, S. Kumar, L.n Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, pages 161–168, Boston, USA, May.

Och, F.J. and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 440–447, Hongkong, China, October.

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

Och, F.J. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics,*, 30(4):417–450.

Och, F.J., C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, pages 20–28, University of Maryland, College Park, MD, USA, June.

Och, F.J., R. Zens, and H. Ney. 2003. Efficient search for interactive statistical machine translation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 387–393, Budapest, Hungary, April.

Ortiz, D., I. García-Varea, and F. Casacuberta. 2003. An empirical comparison of stack-decoding algorithms for statistical machine translation. In *Pattern Recongnition and Image Analysis, First Iberia Congerence*, volume 2652 of *Lecture Notes in Computer Science*, pages 654–663, Puerto de Andratx, Mallorca, June. Springer-Verlag.

Ortiz, D., I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*. AMTA, Phuket, Thailand, September, pages 141–148.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July.

Picó, D. and F. Casacuberta. 2001. Some statistical-estimation methods for stochastic finite-state transducers. *Machine Learning*, 44:121–141.

Tillmann, C. and H. Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, March.

Tillmann, C., S. Vogel, H. Ney, and A. Zubiaga. 1997. A dp-based search using monotone alignments in statistical translation. In *Procededings of the 35th Annual Conference of the Association for Computational Linguistics*, pages 289–296, Madrid, Spain, July.

Tomás, J. and F. Casacuberta. 2002. Binary feature classification for word disambigaution in statistical machine translation. In *Proceedings of the 2nd International Workshop on Pattern Recognition and Information Systems*, pages 213–224, Alicante, Spain, April.

Tomás, J. and F. Casacuberta. 2003. Combining phrase-based and template-based alignment models in statistical translation. In *Pattern Recongnition and Image Analysis, First Iberia Congerence*,

volume 2652 of *Lecture Notes in Computer Science*, pages 1020–1031, Puerto de Andratx, Mallorca, June. Springer-Verlag.

Tomás, J. and F. Casacuberta. 2004. Statistical machine translation decoding using target word reordering. In *Structural, Syntactic, and Statistical Pattern Recongnition*, volume 3138 of *Lecture Notes in Computer Science*, pages 734–743. Springer-Verlag, August.

Tomás, J., J. Lloret, and F. Casacuberta. 2005. Phrase-based alignment models for statistical machine translation. In *Iberian Conference on Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 605–613, Estoril (Portugal), June. Springer-Verlag.

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

Tomás, J., J. M. de Val, F. Fabregat, F. Casacuberta, D. Picó, A. Sanchis, and E. Vidal. 2001. Automatic development of Spanish-Catalan corpora for machine translation. In *Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies*, pages 175–179, Jaén, September.

TransType-2. 2001. TT2. TransType2 - computer assisted translation. Project technical annex.

Vogel, S., Y. Zhang, F. Huang, A. Venugopal, B. Zhao, A. Tribble, M. Eck, and A. Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of the Machine Translation Summit IX*, pages 110–117, September.

Wang, Y.-Y. and A. Waibel. 1997. Decoding algorithm in statistical machine translation. In *Proceedings of the 35th. Annual Meeting of the Association on Computational Linguistics*, pages 366–372, Madrid, Spain, July.

Wang, Y.-Y. and A. Waibel. 1998. Fast decoding for statistical machine translation. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1357–1363, Sydney, Australia, November.

Watanabe, T., K. Imamura, and E. Sumita. 2002. Statistical machine translation based on hierarchical phrase alignment. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 188–198, Keihanna, Japan, March.

Watanabe, T., E. Sumita, and H. G. Okuno. 2003. Chunk-based statistical translation. In Erhard

Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 303–310, July.

Zens, R. and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 257–264, Boston, MA, May.

Zens, R., F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In G. Lakemeyer M. Jarke, J. Koehler, editor, *Advances in artificial intelligence. 25. Annual German Conference on Artificial Intelligence*, volume 2479 of *Lecture Notes on Artificial Intelligence*, pages 18–32, Aachen, Germany, September. Springer Verlag.

Zhang, Y. and S. Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 294–301, Baltimore, MD USA, October.

Zhang, Ying, Stepham Vogel, and Alex Waibel. 2003. Integrated phrase segmentation and alignment algorithm for statistical machine translation. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 567– 573, Beijing, China, October.