

The ITI Statistical Machine Translation System

J. Tomás*, J.M. Vilar†, F. Casacuberta‡

*Departamento de Comunicaciones,
Institut Tecnològic d'Informàtica,
Universitat Politècnica de València,
46071 Valencia, Spain.
jtomas@dcom.upv.es

†Dpt. de Llenguatges i Sistemes Informàtics,
Universitat Jaume I,
12071 Castelló de la Plana, Spain
jvilar@lsi.uji.es

‡Dpt. de Sistemes Informàtics i Computació,
Institut Tecnològic d'Informàtica,
Universitat Politècnica de València,
46071 Valencia, Spain.
fcn@dsic.upv.es

Abstract

One of the translation systems currently under development at ITI uses a phrase based approach: translations are produced considering sequences of words as the elementary unit. We discuss here the implications of monotonicity and non-monotonicity assumptions and their relation with the search algorithms. Also different criteria for the selection of segments and training of parameters are discussed and experimentally tested.

1. Introduction

Early approaches to *Statistical machine translation* (SMT) used the assumption that the translation process can be modeled as a composition of complementary models: some models explain *single-word* alignments through *statistical dictionaries*; some other models deal 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 relying on a word-by-word translation.

Different ways to relax this assumption range from using statistical context-dependent dictionaries as in (Berger et al., 1996b; García-Varea and Casacuberta, 2005) to modeling the alignment of groups of adjacent words in the source sentence with groups of adjacent target words. This last approach is known as the *template-based* (TB) approach (Och and Ney, 2004). A template establishes the alignment (possibly through reordering) between two sequences of word classes (which can also be automatically learned from a bilingual corpus). However, the lexical model inside the templates is still based on word-to-word correspondences (Och et al., 1999; Och and Ney, 2000; Och and Ney, 2004).

A simple alternative to these models has been introduced in recent works: the *phrase-based* (PB) approach (Tomás and Casacuberta, 2001; Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003; Tomás et al., 2005). These methods model 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 dic-

tionaries of single-word pairs are substituted by statistical dictionaries of *bilingual phrases* or *bilingual segments*.

The main problem with the PB approach is the selection of appropriate phrases. Given the large number of possible segment pairs, an exhaustive search is infeasible, thus heuristics are necessary. Most of the methods from the literature are based on some “symmetrisation” of word alignments (Och, 2002; Koehn et al., 2003; Zens and Ney, 2004). An alternative approach is to use *recursive alignments* (Nevado et al., 2004; Vilar and Vidal, 2005).

We describe a system based on statistical PB models. Two versions are considered: using monotonicity in the models and the search or allowing for a restricted degree of non-monotonicity.

In the following section, the statistical framework for machine translation is reviewed and the PB models are introduced. The translation engines that use these models are based on search algorithms, which are presented in section 3. Section 4 describes the corpora used and the experiments performed. Finally, a discussion and the conclusions can be found in section 5.

2. Statistical Machine Translation

The goal of a machine translation system is to translate a given source language sentence s into a target sentence t . Following (Brown et al., 1993), a statistical approach can be used. Assume that the probability distribution $\Pr(t|s)$ is known. Then the best translation is the one that maximizes that probability, ie

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

Usually, this equation is rewritten as

$$\hat{t} = \underset{t}{\operatorname{argmax}} \Pr(t|s) = \underset{t}{\operatorname{argmax}} \Pr(t) \Pr(s|t). \quad (2)$$

Equation 2 presents two basic problems to be solved:

- The construction of models for the output language and for the translation process. The model for the output language ($\Pr(\mathbf{t})$) is used to ensure that the translation produced are grammatically correct. The translation model ($\Pr(\mathbf{s}|\mathbf{t})$) ensures that \hat{t} is indeed a good translation of \mathbf{t} .
- The search process. Solving the argmax is by no means trivial, specially if good response times are required.

Each of those problems requires some compromises. In the case of the language model, it is common to use an n -gram approach. For the translation model, different simplifications are used, as explained below. Finally, the search process itself has to be tuned in order to obtain adequate response times, which usually implies suboptimal results.

2.1. Phrase-based models

Statistical translation models typically assume that the input and the output sentences can be divided in smaller units. These units are related to each other by means of an alignment and they are translated independently. When the units are words, we get the conventional word-based translation models. These include the well known IBM models (Brown et al., 1993), the HMM based models (Ney et al., 2000) or the template models (Och and Ney, 2004) (under this model, the lexicon is still word-based and the alignments are restricted by the available templates). Phrase-base models, on the other hand, divide the sentence in segments each composed of a series of words. The translation probabilities now relate a sequence of words in a source sentence (\tilde{s}) with another sequence of words in the target sentence (\tilde{t}). The simplest formulation with such models is based on monotone models (Tomás and Casacuberta, 2001). In this model, the source sentence \mathbf{s} is segmented into K phrases (\tilde{s}_1^K) and the target sentence \mathbf{t} into other K phrases (\tilde{t}_1^K). A uniform probability distribution over all possible segmentation is assumed. The monotonicity assumption implies that the target phrase in position k is produced by the source phrase in the same position k . This can be expressed as

$$\Pr(\mathbf{s}|\mathbf{t}) \propto \sum_{K, \tilde{t}_1^K, \tilde{s}_1^K} \prod_{k=1}^K p(\tilde{s}_k|\tilde{t}_k). \quad (3)$$

The distribution $p(\tilde{s}|\tilde{t})$ can be interpreted as a dictionary that returns the probability of translating the phrase \tilde{t} into the phrase \tilde{s} . As a phrase can be a single word, a conventional word to word statistical dictionary can be considered as part of the model.

If monotonicity is not admissible, a hidden variable α can be introduced. This represents the fact that the target phrase in position k is produced by the source phrase in position α_k . Symbolically:

$$\Pr(\mathbf{s}|\mathbf{t}) \propto \sum_{K, \tilde{t}_1^K, \tilde{s}_1^K, \alpha_1^K} p(\alpha_1^K) \prod_{k=1}^K p(\tilde{s}_k|\tilde{t}_{\alpha_k}) \quad (4)$$

This model allows efficient search algorithms, details are explained in section 3.

2.2. Log-linear model combination

The above approach (modeling directly the distributions following Equation 2) has two problems: the difficulty of coming up with good models using a generative approach and the difficulty of introducing other sources of knowledge in the process. Those problems can be solved using a log-linear combination of models. In the experiments, we have adopted the following log-linear model combination in the monotone search for a given segmentation of (\mathbf{s}, \mathbf{t}) into K segments $\sigma = (\tilde{s}_1^K; \tilde{t}_1^K)$:

$$\begin{aligned} p_{\text{PB}}(\mathbf{s}_1^J, \mathbf{t}_1^I; \sigma) = & \sum_{i=1}^I \left[c_1 + \lambda_1 \cdot \log p(\mathbf{t}_i|\mathbf{t}_{i-2}^{i-1}) \right. \\ & + \lambda_2 \cdot \log p(T_i|T_{i-4}^{i-1}) \\ & + \lambda_3 \cdot \log \sum_{j=1}^J p(\mathbf{t}_i|\mathbf{s}_j) \\ & \left. + \lambda_4 \cdot \log \sum_{j=1}^J p(\mathbf{s}_j|\mathbf{t}_i) \right] + \\ & \sum_{k=1}^K \left[c_2 + \lambda_5 \cdot \log p_{\mathbf{s} \rightarrow \mathbf{t}}(\tilde{\mathbf{t}}_k|\tilde{\mathbf{s}}_k) \right. \\ & \left. + \lambda_6 \cdot \log p_{\mathbf{t} \rightarrow \mathbf{s}}(\tilde{\mathbf{s}}_k|\tilde{\mathbf{t}}_k) \right]. \end{aligned} \quad (5)$$

This integrates the following knowledge sources:

- Language models for the target language. There are two models, a conventional trigram model: $p(\mathbf{t}_i|\mathbf{t}_{i-2}^{i-1})$ and a five-gram class model: $p(T_i|T_{i-4}^{i-1})$. As explained above, the aim of these models is that the resulting sentence is correct in the target language. Word classes are obtained using the software *mkcls* (Och et al., 1999).
- Simple translation models (like IBM model 1) both direct ($p(\mathbf{t}_i|\mathbf{s}_j)$) and inverse ($p(\mathbf{s}_j|\mathbf{t}_i)$). These models act as “smoothers” for the translation probabilities.
- Direct and inverse phrase based translation models: $p_{\mathbf{s} \rightarrow \mathbf{t}}(\tilde{\mathbf{t}}_k|\tilde{\mathbf{s}}_k)$ and $p_{\mathbf{t} \rightarrow \mathbf{s}}(\tilde{\mathbf{s}}_k|\tilde{\mathbf{t}}_k)$. These are the most complex models and should capture the main bulk of the work.

Each of the sources is controlled by a weight (a scaling factor), the λ_i , and two penalties c_1 and c_2 are included to control the values of I and K .

2.3. Learning phrase-based alignment models

There are different approaches to the parameter estimation of the parameters in the previous equations. Details of the estimation of monotone and no-monotone phrase-based models can be found in (Tomás et al., 2005). Some of these techniques correspond to a direct learning of the parameters from a sentence-aligned corpus using a maximum likelihood approach (Tomás and Casacuberta, 2001; Marcu and Wong, 2002). Other techniques are heuristics based on previous computation of word alignments in the training corpus (Zens et al., 2002; Koehn et al., 2003).

Word alignments are the basis for the most widely used methods of finding bilingual segments. 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 et al., 1999; Och, 2002) deals with this problem in two steps. In the first step, *symmetrized alignments* are computed from the alignments obtained in a translation direction ($s \rightarrow t$) and the alignments obtained in the opposite translation direction ($t \rightarrow s$). Different combinations of these two types of alignments were proposed in (Och and Ney, 2003) (*intersection*, *union* and *refined*). From these symmetrized alignments, the bilingual segments are built following different criteria in the second step (Och and Ney, 2003). These criteria consider that a segment from a source sentence and a segment from a target sentence give way to a bilingual segment if all the words in the source segment are aligned (according to the symmetrized alignments) with one word in the target segment and vice versa. Adjacent or internal (source or target) words that are not aligned with any (target or source) word can also be added to the bilingual segment.

In this work, an alternative strategy is proposed. It consists also in two steps but different from the steps proposed in (Och, 2002). In the first step, two separate PB models (bilingual segments and the corresponding probabilities) were built, one model from word-alignments in one direction ($s \rightarrow t$) and another model from word-alignments in the opposite direction ($t \rightarrow s$). The bilingual segments are obtained following a similar procedure as in the second step of the method proposed in (Och, 2002). In the second step of our strategy, these two models are combined using log-linear interpolation.

3. Search

Given a source sentence s_1^J , the aim of the search in MT is to obtain a target sentence $\hat{t}_1^{\hat{J}}$ that maximizes Equation 5:

$$\hat{t}_1^{\hat{J}} = \operatorname{argmax}_{I, t_1^I, \sigma} p_{PB}(s_1^J, t_1^I; \sigma) \quad (6)$$

The search algorithm is a crucial part in statistical machine translation. Its performance directly affects the quality and efficiency of translation. In this section, we describe two search algorithms which are based on multi-stack-decoding (Berger et al., 1996a) for the monotone and for a non-monotone version of Equation 3 (Tomás and Casacuberta, 2004).

The most common statistical decoder algorithms use the concept of partial translation hypothesis to perform the search. In a partial hypothesis, some of the source words have been used to generate a target prefix. Each hypothesis is scored according to the translation and language model. In our implementation for the monotone model, (Tomás and Casacuberta, 2001) we define a hypothesis as the triple $(J', t_1^{J'}, g)$, where J' is the length of the current source prefix (ie, that prefix is $s_1^{J'}$), $t_1^{J'}$ is its translation and g is the score of that translation computed from Equation 5.

3.1. Monotone search

The translation procedure can be described as follows. The system maintains a large set of hypotheses, each of them

with its translation score. The set is divided in lists so that each hypothesis in the list covers the same number of source words. Within each list the hypotheses are sorted according to the translation score. The algorithm consists in an iterative process. In each iteration, the system extracts from each list the best scored hypothesis and extends it. The extension consists in selecting one or more untranslated source words and attaching one or more target words to the current output prefix. The extension of a hypothesis can generate hundreds of new hypotheses. The process is iterated *Max-iter* times. Thus, at most *Max-iter* hypotheses are extended from each list. The output of the search is the final hypothesis with the highest score and with no untranslated source words.

3.2. Non-monotone search

If a non-monotone model is used, the search can be made allowing for *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 $\langle \text{nul} \rangle$ in the target hypothesis. The meaning of this token is that, in a future expansion, the token $\langle \text{nul} \rangle$ must be replaced by a sequence of words. In our implementation, we allow only one token $\langle \text{nul} \rangle$. Therefore, we can distinguish between two classes of hypotheses. A hypothesis is closed if it does not contain the token $\langle \text{nul} \rangle$, and it is open if it contains this token.

In the process of extending a partial hypothesis, those bilingual phrase-pairs (\tilde{s}, \tilde{t}) in which \tilde{s} matches the source segment after the last translated word are considered. If the hypothesis to be extended is closed (it has no $\langle \text{nul} \rangle$ token), two new hypotheses are created adding \tilde{t} and $\langle \text{nul} \rangle \tilde{t}$, respectively, to the target prefix. On the other hand, if the hypothesis is open, four new hypotheses are created: one closes the hypothesis by replacing the token $\langle \text{nul} \rangle$ by \tilde{t} ; and three new open hypotheses are obtained putting \tilde{t} to the left or to the right of $\langle \text{nul} \rangle$ and at the end of the target prefix. 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. A decoding example using this algorithm is shown in Figure 1.

The restriction to at most one $\langle \text{nul} \rangle$ token implies that the practical costs of the monotone and non-monotone search algorithms are very similar. In practice, the parameter *Max-iter* can be used to increase the speed of the translation.

The language model causes another problem. In an open hypothesis we cannot calculate the language model contribution of the right part of the prefix after the $\langle \text{nul} \rangle$ 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 $\langle \text{nul} \rangle$ 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.

There are some proposals (Vogel et al., 2003; Koehn, 2004; Och and Ney, 2004) that try to solve this problem by select-

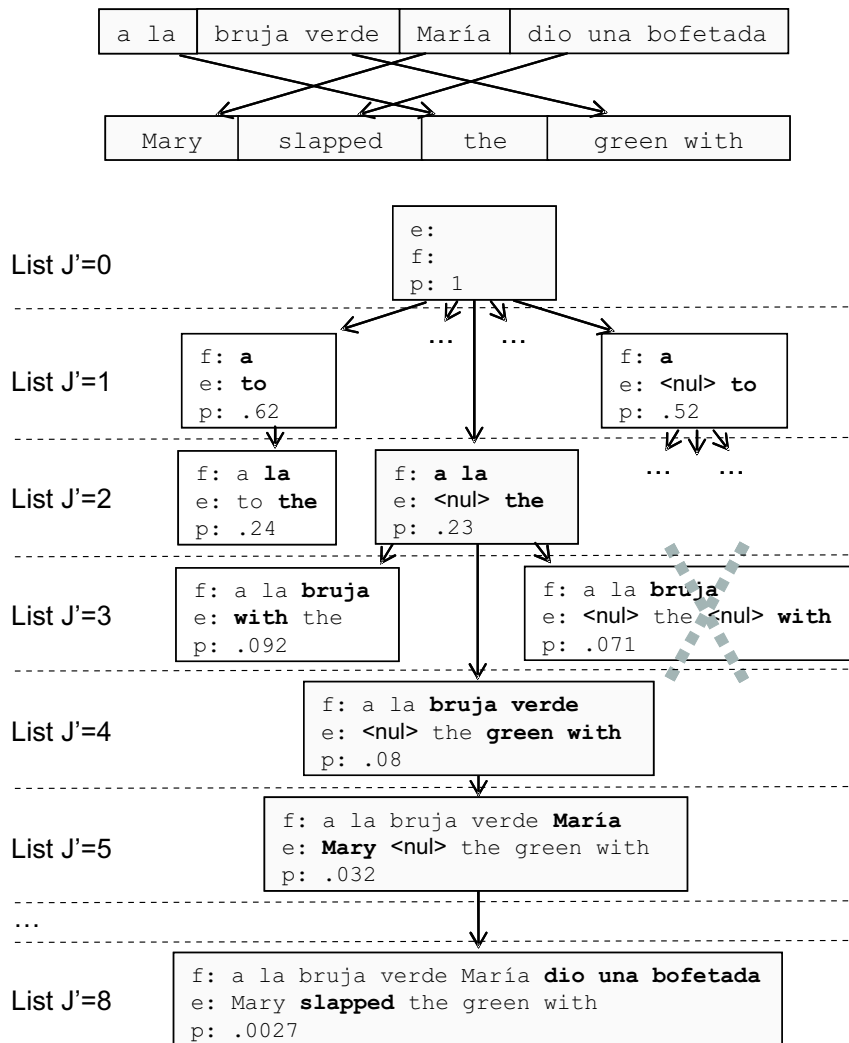


Figure 1: Decoding Example using TWR algorithm. The Spanish sentence “A la bruja verde Mario dio un fofetada” is translate into the English sentence “Mary slapped the green witch”. Partial hypotheses are stored in sorted list with numbers from 0 to 8. In each hypothesis the first J’ words of the source sentence have been translated.

ing source words in different positions (SWR) and generating the target words left to right. In this approach, a partial hypothesis is a triple (C, t_1^J, g) , where C is the coverage set (the source positions that have been translated). Hypotheses with the same number of elements in C are stored in the same list. As in TWR approach J such lists are needed.

Compared with TWR, this approach has different limitations. First, each partial hypothesis needs a coverage set, which increases spatial costs. Second, hypotheses with different C can be stored in the same list. In order to properly compare hypothesis that cover different parts of the source sentence an estimation (usually heuristic) of the contribution to the score of the parts that are not yet covered can be introduced (Wang and Waibel, 1997). Third, in order to reduce the computational cost of the algorithm re-ordering restrictions must be introduced. For example, (Berger et al., 1996a) proposes than only the first l yet uncovered word positions can be translated. Fourth, many different paths lead to hypothesis with the same coverage set and target

segment. In order to reduce the search space hypothesis recombination is used. In our approach this phenomenon is less frequent, and hypothesis recombination is not essential. On the other hand, the main limitations of the TWR approach are: First, in an open hypothesis a estimation of the language model is used. Second, only one token $\langle \text{nul} \rangle$ is allowed. Then, we do not cover all possible reorderings.

4. Experimental results

The models and procedures introduced in this work were assessed through a series of experiments with different corpora. One of these corpora, the assessment measures used and the results are described in this section.

4.1. Corpora

The corpora used for the experiments were extracted from the Bulletin of the European Union (from English to Spanish, French, and German; and from Spanish, French, and German to English) (Khadivi and Goutte, 2003). 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 1.

4.2. Assessment

In all the experiments reported in this paper, the translations of the source test sentences produced by the translation systems were compared with target test references and two measures 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 100.0 (best score) (Papineni et al., 2002).

4.3. Results

Several experiments were carried out to assess the approach presented. The default parameters in the following experiments were: maximum phrase length of 8 words; the parameter estimation was the relative frequencies using the direct model (without symmetrization or lexicon model); and the search was monotone (Max-iter=10).

The scaling factors in Equation 5 can be estimated by optimizing the value of a training criterion over a development corpus (Och, 2003). In our case, the optimization consisted in minimizing the difference between the translation word error rate and the BLEU scores. The optimization was carried out using the downhill simplex algorithm (Nedler and Mead, 1965). Table 2 shows the scaling factors used in the experiments.

parameter	model	scaling factor
c_1		1.2
c_2		0.2
λ_1	$p(\mathbf{t}_i \mathbf{t}_{i-2}^{i-1})$	1
λ_2	$p(T_i T_{i-4}^{i-1})$	2
λ_3	$p(\mathbf{t}_i \mathbf{s}_j)$	4
λ_4	$p(\mathbf{s}_j \mathbf{t}_i)$	5
λ_5	$p_{s \rightarrow t}(\tilde{\mathbf{t}}_k \tilde{\mathbf{s}}_k)$	13
λ_6	$p_{t \rightarrow s}(\tilde{\mathbf{t}}_k \tilde{\mathbf{s}}_k)$	4

Table 2: Scaling factors used in the experiments (English-French).

Two different search algorithms have been presented in section 3.: a monotone search algorithm and a non-monotone search algorithm. A complete comparison between these monotone and non-monotone models is presented in Table 3. In this experiment, *Max-iter* has been set to 64 for non-monotone search. Similar results has been observed using a SWR algorithm for non-monotone search (Tomás and Casacuberta, 2004).

The results obtained in these experiments are comparable. This is an interesting result, since the monotone models are simpler than non-monotone ones.

Languages	Monotone		Non-monotone	
	WER	BLEU	WER	BLEU
English Spanish	46.7	42.1	46.7	42.3
English French	45.2	42.8	45.1	42.8
English German	57.4	30.2	57.3	30.3

Table 3: Effect of different procedures for searching on the WER (%) and BLEU (%) for the EU corpus.

The maximum length of a phrase can be restricted in order to limit the number of parameters of the models. The influence of upper bounds in the segment lengths can be seen in Table 4.

The use of long segments to significantly improves the results obtained. As expected, too, the use of long segments leads to a huge number of parameters.

Additional models are combined with PB model in Equation 5. Table 5 shows the effect of these new models in the system performance.

description	model	WER	BLEU
PB direct model	$p_{s \rightarrow t}(\tilde{\mathbf{t}}_k \tilde{\mathbf{s}}_k)$	46.9	41.0
+ language model	$+ p(\mathbf{t}_i \mathbf{t}_{i-2}^{i-1})$	45.8	42.1
+ word/phrase penalty	$+ c_1, c_2$	45.5	42.4
+ class language model	$+ p(T_i T_{i-4}^{i-1})$	45.2	42.8
+ lexicon model ($t \rightarrow s$)	$+ p(\mathbf{s}_j \mathbf{t}_i)$	44.6	44.7
+ lexicon model ($s \rightarrow t$)	$+ p(\mathbf{t}_i \mathbf{s}_j)$	44.5	44.6
+ symmetrization	$+ p_{t \rightarrow s}(\tilde{\mathbf{t}}_k \tilde{\mathbf{s}}_k)$	44.2	44.9

Table 5: Effect of combining different models on the WER (%) and BLEU (%) for the EU corpus (English-French).

Lexicon models are the most helpful additional features. Probably these models smooth the probabilities of the bilingual phrases that are poorly estimated; for example, when a phrase appears one or two times in the training set. Using an additional PB model, trained from the word aligned corpus opposite direction ($p_{t \rightarrow s}(\tilde{\mathbf{t}}_k | \tilde{\mathbf{s}}_k)$), is also interesting for improve the results.

5. Conclusions

Phrase-based models constitute the state-of-the-art in 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. 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.

Following these ideas a Phrase-based SMT system was developed in the ITI. The main innovations of this system is a new method for symmetrization and a new decoder algorithm for non-monotone search.

The phrase-based models (estimation and search) have been tested through experiments on a difficult corpus and differ-

		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

Table 1: The “EU” corpus from English to Spanish, German, and French. Trigram models were used to compute the test-set perplexity.

max.phr.len.	English Spanish			English French			English German		
	WER	BLEU	param.	WER	BLEU	param.	WER	BLEU	param.
2	55.5	26.5	1.0M	54.6	27.2	1.0M	62.9	22.5	1.3M
4	49.3	36.4	4.3M	47.6	38.1	4.1M	58.5	28.1	5.0M
6	47.2	40.6	7.9M	46.2	41.7	7.4M	57.8	29.8	9.4M
8	46.7	42.1	11.3M	45.2	42.8	10.5M	57.4	30.2	13.8M
10	46.6	42.2	12.7M	44.9	43.0	12.9M	-	-	-

Table 4: Effect of maximum phrase length on the WER (%), BLEU (%) and number of parameters for the EU corpus.

ent languages. The main conclusions that can be drawn from these experiments are the following: 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. The use of long bilingual phrases can achieve less translation errors. The use of additional models like lexical models or symmetrized models improve significantly the results.

A 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. Monotone phrase-based models are closely related to stochastic finite-state transducers, and this could help in the design of more efficient search algorithms.

6. Acknowledgements

This work has been partially supported by the Spanish project TIC2003-08681-C02-02 and by the IST Programme of the European Union IST-2001-32091.

7. References

- A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, A. S. Kehler, and R. L. Mercer. 1996a. Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981, Apr.
- Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996b. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March.
- P. F. Brown, 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.
- I. García-Varea and F. Casacuberta. 2005. Learn context-dependent lexicon models for statistical machine translation. *Machine Learning*, 59:1–24.
- S. Khadivi 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.
- Philipp Koehn, 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)*, Edmonton, Alberta.
- P. Koehn. 2004. 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.
- Daniel Marcu and William Wong. 2002. 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*.
- J.A. Nedler and R. Mead. 1965. A simplex method for function minimization. *Computer Journal*, 7:308–313.
- F. Nevado, F. Casacuberta, and J. Landa. 2004. Translation memories enrichment by statistical bilingual segmentation. In *Proceedings of the IV International Conference on Language Resources and Evaluation - LREC2004*, volume 1, pages 335–338, Lisbon, May. ELRA.
- H. Ney, 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.
- F.J. Och and H. Ney. 2000. Improved statistical align-

- ment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 440–447, Hongkong, China, October.
- F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–450.
- F.J. Och, 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.
- F.J. Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, October.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.
- Kishore Papineni, 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.
- C. Tillmann and H. Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, March.
- J. Tomás and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Proceedings of the Machine Translation Summit VIII*, pages 357–361, Santiago de Compostela.
- J. Tomás and F. Casacuberta. 2004. Statistical machine translation decoding using target word reordering. In *Structural, Syntactic, and Statistical Pattern Recognition*, volume 3138 of *Lecture Notes in Computer Science*, pages 734–743. Springer-Verlag.
- J. Tomás, 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. Springer-Verlag, Estoril (Portugal), June.
- TransType-2. 2001. TT2. TransType2 - computer assisted translation. Project technical annex.
- J. M. Vilar and E. Vidal. 2005. A recursive statistical translation model. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, USA.
- S. Vogel, 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.
- Y.-Y. Wang and A. Waibel. 1997. Decoding algorithm in statistical machine translation. In *Proceedings of the 35th. Annual Meeting of the Association on Computational Linguistics*, Madrid, Spain.
- R. Zens 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.
- R. Zens, 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 AI, KI 2002.*, volume 2479 of *LNAI*, pages 18–32. Springer Verlag, September.