HMM-Based Word Alignment in Statistical Translation

Stephan Vogel **Hermann Ney** **Christoph Tillmann**
Lehrstuhl für Informatik V, RWTH Aachen
D-52056 Aachen, Germany
{vogel,ney,tillmann}@informatik.rwth-aachen.de

Abstract

In this paper, we describe a new model for word alignment in statistical translation and present experimental results. The idea of the model is to make the alignment probabilities dependent on the differences in the alignment positions rather than on the absolute positions. To achieve this goal, the approach uses a first-order Hidden Markov model (HMM) for the word alignment problem as they are used successfully in speech recognition for the time alignment problem. The difference to the time alignment HMM is that there is no monotony constraint for the possible word orderings. We describe the details of the model and test the model on several bilingual corpora.

1 Introduction

In this paper, we address the problem of word alignments for a bilingual corpus. In the recent years, there have been a number of papers considering this or similar problems: (Brown et al., 1990), (Dagan et al., 1993), (Kay et al., 1993), (Fung et al., 1993).

In our approach, we use a first-order Hidden Markov model (HMM) (Jelinek, 1976), which is similar, but not identical to those used in speech recognition. The key component of this approach is to make the alignment probabilities dependent not on the absolute position of the word alignment, but on its relative position; i.e. we consider the differences in the index of the word positions rather than the index itself.

The organization of the paper is as follows. After reviewing the statistical approach to machine translation, we first describe the conventional model (mixture model). We then present our first-order HMM approach in full detail. Finally we present some experimental results and compare our model with the conventional model.

2 Review: Translation Model

The goal is the translation of a text given in some language F into a target language E. For convenience, we choose for the following exposition as language pair French and English, i.e. we are given a French string $f_1^J = f_1 ... f_j ... f_J$, which is to be translated into an English string $e_1^I = e_1 ... e_i ... e_I$. Among all possible English strings, we will choose the one with the highest probability which is given by Bayes' decision rule:

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \{Pr(e_1^I|f_1^J)\} \\ &= \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \end{aligned}$$

$Pr(e_1^I)$ is the language model of the target language, whereas $Pr(f_1^J|e_1^I)$ is the string translation model. The argmax operation denotes the search problem. In this paper, we address the problem of introducing structures into the probabilistic dependencies in order to model the string translation probability $Pr(f_1^J|e_1^I)$.

3 Alignment Models

A key issue in modeling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the English sentence and the words of the French sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $[f_1^J; e_1^I]$. We further constrain this model by assigning each French word to *exactly one* English word. Models describing these types of dependencies are referred to as *alignment models.*

In this section, we describe two models for word alignment in detail:

- a mixture-based alignment model, which was introduced in (Brown et al., 1990);
- an HMM-based alignment model.

In this paper, we address the question of how to define specific models for the alignment probabilities. The notational convention will be as follows. We use the symbol $Pr(.)$ to denote general

probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(.)$.

3.1 Alignment with Mixture Distribution

Here, we describe the mixture-based alignment model in a formulation which is different from the original formulation in (Brown et al., 1990). We will use this model as reference for the HMM-based alignments to be presented later.

The model is based on a decomposition of the joint probability for f_1^J into a product over the probabilities for each word f_j:

$$Pr(f_1^J|e_1^I) = p(J|I) \cdot \prod_{j=1}^{J} p(f_j|e_1^I) \quad ,$$

where, for normalization reasons, the sentence length probability $p(J|I)$ has been included. The next step now is to assume a sort of pairwise interaction between the French word f_j and *each* English word $e_i, i = 1, ...I$. These dependencies are captured in the form of a mixture distribution:

$$\begin{aligned} p(f_j|e_1^I) &= \sum_{i=1}^{I} p(i, f_j|e_1^I) \\ &= \sum_{i=1}^{I} p(i|j, I) \cdot p(f_j|e_i) \end{aligned}$$

Putting everything together, we have the following mixture-based model:

$$Pr(f_1^J|e_1^I) = p(J|I) \cdot \prod_{j=1}^{J} \sum_{i=1}^{I} [p(i|j, I) \cdot p(f_j|e_i)] \quad (1)$$

with the following ingredients:

- sentence length probability: $p(J|I)$;
- mixture alignment probability: $p(i|j, I)$;
- translation probability: $p(f|e)$.

Assuming a uniform alignment probability

$$p(i|j, I) = \frac{1}{I}$$

we arrive at the first model proposed by (Brown et al., 1990). This model will be referred to as IBM1 model.

To train the translation probabilities $p(f|e)$, we use a bilingual corpus consisting of sentence pairs $\left[f_1^{J_s}; e_1^{I_s}\right], s = 1, ..., S$. Using the maximum likelihood criterion, we obtain the following iterative equation (Brown et al., 1990):

$$\hat{p}(f|e) = \frac{A(f,e)}{\sum_{f'} A(f',e)} \qquad \text{with}$$

$$A(f,e) = \sum_s \left(\frac{p(f|e)}{\sum_{i=1}^{I_s} p(f|e_{is})} \sum_{j=1}^{J_s} \delta(f, f_{js}) \sum_{i=1}^{I_s} \delta(e, e_{is}) \right)$$

For uniform alignment probabilities, it can be shown (Brown et al., 1990), that there is only one optimum and therefore the EM algorithm (Baum, 1972) *always* finds the global optimum.

For mixture alignment model with nonuniform alignment probabilities (subsequently referred to as IBM2 model), there are too many alignment parameters $p(i|j, I)$ to be estimated for small corpora. Therefore, a specific model for the alignment probabilities is used:

$$p(i|j, I) = \frac{r(i - j\frac{I}{J})}{\sum_{i'=1}^{I} r(i' - j\frac{I}{J})} \quad . \qquad (2)$$

This model assumes that the position distance relative to the diagonal line of the (j, i) plane is the dominating factor (see Fig. 1). To train this model, we use the maximum likelihood criterion in the so-called maximum approximation, i.e. the likelihood criterion covers only the most likely alignment rather than the set of all alignments:

$$Pr(f_1^J|e_1^I) \cong \prod_{j=1}^{J} \max_i [p(i|j, I) \cdot p(f_j|e_i)] \qquad (3)$$

In training, this criterion amounts to a sequence of iterations, each of which consists of two steps:

- *position alignment:* Given the model parameters, determine the most likely position alignment.
- *parameter estimation:* Given the position alignment, i.e. going along the alignment paths for all sentence pairs, perform maximum likelihood estimation of the model parameters; for model-free distributions, these estimates result in relative frequencies.

Due to the nature of the mixture model, there is no interaction between adjacent word positions. Therefore, the optimal position i for each position j can be determined independently of the neighbouring positions. Thus the resulting training procedure is straightforward.

3.2 Alignment with HMM

We now propose an HMM-based alignment model. The motivation is that typically we have a strong localization effect in aligning the words in parallel texts (for language pairs from Indoeuropean languages): the words are not distributed arbitrarily over the sentence positions, but tend to form clusters. Fig. 1 illustrates this effect for the language pair *German - English*.

Each word of the German sentence is assigned to a word of the English sentence. The alignments have a strong tendency to preserve the local neighborhood when going from the one language to the other language. In many cases, although not always, there is an even stronger restriction: the difference in the position index is smaller than 3.

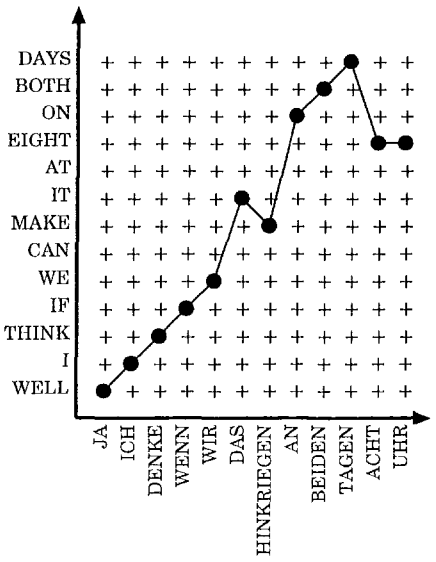

Figure 1: Word alignment for a German - English sentence pair.

To describe these word-by-word alignments, we introduce the mapping $j \rightarrow a_j$, which assigns a word f_j in position j to a word e_i in position $i = a_j$. The concept of these alignments is similar to the ones introduced by (Brown et al., 1990), but we will use another type of dependence in the probability distributions. Looking at such alignments produced by a human expert, it is evident that the mathematical model should try to capture the strong dependence of a_j on the previous alignment. Therefore the probability of alignment a_j for position j should have a dependence on the previous alignment a_{j-1}:

$$p(a_j|a_{j-1}, I) \quad ,$$

where we have included the conditioning on the total length I of the English sentence for normalization reasons. A similar approach has been chosen by (Dagan et al., 1993). Thus the problem formulation is similar to that of the time alignment problem in speech recognition, where the so-called Hidden Markov models have been successfully used for a long time (Jelinek, 1976). Using the same basic principles, we can rewrite the probability by introducing the 'hidden' alignments $a_1^J := a_1 ... a_j ... a_J$ for a sentence pair $[f_1^J; e_1^I]$:

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \\ &= \sum_{a_1^J} \prod_{j=1}^{J} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \end{aligned}$$

So far there has been no basic restriction of the approach. We now assume a first-order dependence on the alignments a_j only:

$$\begin{aligned} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) &= p(f_j, a_j|a_{j-1}, e_1^I) \\ &= p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j}) \end{aligned}$$

where, in addition, we have assumed that the translation probability depends only on a_j and not on a_{j-1}. Putting everything together, we have the following HMM-based model:

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})] \quad (4)$$

with the following ingredients:

- HMM alignment probability: $p(i|i', I)$ or $p(a_j|a_{j-1}, I)$;
- translation probability: $p(f|e)$.

In addition, we assume that the HMM alignment probabilities $p(i|i', I)$ depend only on the jump width $(i - i')$. Using a set of non-negative parameters $\{s(i - i')\}$, we can write the HMM alignment probabilities in the form:

$$p(i|i', I) = \frac{s(i - i')}{\sum_{l=1}^{I} s(l - i')} \quad . \quad (5)$$

This form ensures that for each word position i', $i' = 1, ..., I$, the HMM alignment probabilities satisfy the normalization constraint.

Note the similarity between Equations (2) and (5). The mixture model can be interpreted as a zeroth-order model in contrast to the first-order HMM model.

As with the IBM2 model, we use again the maximum approximation:

$$Pr(f_1^J|e_1^I) \cong \max_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})] \quad (6)$$

In this case, the task of finding the optimal alignment is more involved than in the case of the mixture model (IBM2). Therefore, we have to resort to dynamic programming for which we have the following typical recursion formula:

$$Q(i, j) = p(f_j|e_i) \max_{i'=1,..,I} [p(i|i', I) \cdot Q(i', j - 1)]$$

Here, $Q(i, j)$ is a sort of partial probability as in time alignment for speech recognition (Jelinek, 1976).

4 Experimental Results

4.1 The Task and the Corpus

The models were tested on several tasks:

- the Avalanche Bulletins published by the Swiss Federal Institute for Snow and Avalanche Research (SFISAR) in Davos, Switzerland and made available by the European Corpus Initiative (ECI/MCI, 1994);
- the Verbmobil Corpus consisting of spontaneously spoken dialogs in the domain of appointment scheduling (Wahlster, 1993);

- the EuTrans Corpus which contains typical phrases from the tourists and travel domain. (EuTrans, 1996).

Table 1 gives the details on the size of the corpora and their vocabulary. It should be noted that in all these three cases the ratio of vocabulary size and number of running words is not very favorable.

Table 1: Corpora

Corpus	Language	Words	Voc. Size
Avalanche	French	62849	1993
	German	44805	2265
EuTrans	Spanish	13768	2008
	English	15888	1630
Verbmobil	German	150279	4017
	English	154727	2443

For several years between 83 and 92, the Avalanche Bulletins are available for both German and French. The following is a typical sentence pair from the corpus:

Bei zuerst recht hohen, später tieferen Temperaturen sind von Samstag bis Dienstag morgen auf der Alpennordseite und am Alpenhauptkamm oberhalb 2000 m 60 bis 80 cm Neuschnee gefallen.
Par des températures d' abord élevées, puis plus basses, 60 à 80 cm de neige sont tombs de samedi à mardi matin sur le versant nord et la crête des Alpes au-dessus de 2000 m.

An example from the Verbmobil corpus is given in Figure 1.

4.2 Training and Results

Each of the three corpora were used to train both alignment models, the mixture-based alignment model in Eq.(1) and the HMM-based alignment model in Eq.(4). Here, we will consider the experimental tests on the Avalanche corpus in more detail. The training procedure consisted of the following steps:

- Initialization training: IBM1 model trained for 10 iterations of the EM algorithm.
- Refinement training: The translation probabilities from the initialization training were used to initialize both the IBM2 model and the HMM-based alignment model
  - IBM2 Model: 5 iterations using the maximum approximation (Eq.(3))
  - HMM Model: 5 iterations using the maximum approximation (Eq.(6))

The resulting perplexity (inverse geometric average of the likelihoods) for the different models are given in the Tables 2 and 3 for the Avalanche corpus. In addition to the total perplexity, which is the global optimization criterion, the tables also show the perplexities of the translation probabilities and of the alignment probabilities. The last line in Table 2 gives the perplexity measures when applying the maximum approximation and computing the perplexity in this approximation. These values are equal to the ones after initializing the IBM2 and HMM models, as they should be. From Table 3, we can see that the mixture alignment gives slightly better perplexity values for the translation probabilities, whereas the HMM model produces a smaller perplexity for the alignment probabilities. In the calculation of the perplexities, the sentence length probability was not included.

Table 2: IBM1: Translation, alignment and total perplexity as a function of the iteration.

Iteration	Translation	Alignment	Total
0	99.36	20.07	1994.00
1	3.72	20.07	74.57
2	2.67	20.07	53.62
...	...	...	...
9	1.87	20.07	37.55
10	1.86	20.07	37.36
Max.	3.88	20.07	77.95

Table 3: Translation, alignment and total perplexity as a function of the iteration for the IBM2 (A) and the HMM model (B)

	Iter.	Translation	Alignment	Total
A	0	3.88	20.07	77.95
	1	3.17	10.82	34.27
	2	3.25	10.15	33.03
	3	3.22	10.10	32.48
	4	3.20	10.06	32.18
	5	3.18	10.05	32.00
B	0	3.88	20.07	77.95
	1	3.37	7.99	26.98
	2	3.46	6.17	21.36
	3	3.47	5.90	20.48
	4	3.46	5.85	20.24
	5	3.45	5.84	20.18

Another interesting question is whether the HMM alignment model helps in finding good and sharply focussed word-to-word correspondences. As an example, Table 4 gives a comparison of the translation probabilities $p(f|e)$ between the mixture and the HMM alignment model for the German word *Alpensüdhang*. The counts of the words are given in brackets. There is virtually no difference between the translation tables for the two models (IBM2 and HMM). But in general, the HMM model seems to give slightly better results in the cases of German compound words like *Alpensüdhang - versant sud des Alpes* which require function words in the translation.

Table 4: Alpensüdhang.

IBM1	Alpes (684)	0.171
	des (1968)	0.035
	le (1419)	0.039
	sud (416)	0.427
	sur (769)	0.040
	versant (431)	0.284
IBM2	Alpes (684)	0.276
	sud (416)	0.371
	versant (431)	0.356
HMM	Alpes (684)	0.284
	des (1968)	0.028
	sud (416)	0.354
	versant (431)	0.333

This is a result of the smoother position alignments produced by the HMM model. A pronounced example is given in Figure 2. The problem of the absolute position alignment can be demonstrated at the positions (a) and (c): both *Schneebrettgefahr* und *Schneeverfrachtungen* have a high probability on *neige*. The IBM2 models chooses the position near the diagonal, as this is the one with the higher probability. Again, *Schneebrettgefahr* generates *de* which explains the wrong alignment near the diagonal in (c).

However, this strength of the HMM model can also be a weakness as in the case of *est developpe – ist ... entstanden* (see (b) in Figure 2. The required two large jumps are correctly found by the mixture model, but not by the HMM model. These cases suggest an extention to the HMM model. In general, there are only a small number of big jumps in the position alignments in a given sentence pair. Therefore a model could be useful that distinguishes between local and big jumps.

The models have also been tested on the Verbmobil Translation Corpus as well as on a small Corpus used in the EuTrans project. The sentences in the EuTrans corpus are in general short phrases with simple grammatical structures. However, the training corpus is very small and the produced alignments are generally of poor quality. There is no marked difference for the two alignment models.

Table 5: Perplexity results for (a) EuTrans and (b) Verbmobil Corpus.

	Model	Iter.	Transl.	Align.	Total
a	IBM1	10	2.610	6.233	16.267
	IBM2	5	2.443	4.003	9.781
	HMM	5	2.461	3.934	9.686
b	IBM1	10	4.373	10.674	46.672
	IBM2	5	4.696	6.538	30.706
	HMM	5	4.859	5.452	26.495

The Verbmobil Corpus consists of spontaneously spoken dialogs in the domain of appointment scheduling. The assumption that every word in the source language is aligned to a word in the target language breaks down for many sentence pairs, resulting in poor alignment. This in turn affects the quality of the translation probabilities.

Several extensions to the current HMM based model could be used to tackle these problems:

- The results presented here did not use the concept of the empty word. For the HMM-based model this, however, requires a second-order rather than a first-order model.
- We could allow for multi-word phrases in both languages.
- In addition to the absolute or relative alignment positions, the alignment probabilities can be assumed to depend on part of speech tags or on the words themselves. (confer model 4 in (Brown et al., 1990)).

5 Conclusion

In this paper, we have presented an HMM-based approach for modelling word alignments in parallel texts. The characteristic feature of this approach is to make the alignment probabilities explicitly dependent on the alignment position of the previous word. We have tested the model successfully on real data. The HMM-based approach produces translation probabilities comparable to the mixture alignment model. When looking at the position alignments those generated by the HMM model are in general much smoother. This could be especially helpful for languages such as German, where compound words are matched to several words in the source language. On the other hand, large jumps due to different word orderings in the two languages are successfully modeled. We are presently studying and testing a multilevel HMM model that allows only a small number of large jumps. The ultimate test of the different alignment and translation models can only be carried out in the framework of a fully operational translation system.

6 Acknowledgement

This research was partly supported by the German Federal Ministery of Education, Science, Research and Technology under the Contract Number 01 IV 601 A (Verbmobil) and under the Esprit Research Project 20268 (EuTrans).

References

L. E. Baum. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1-8.

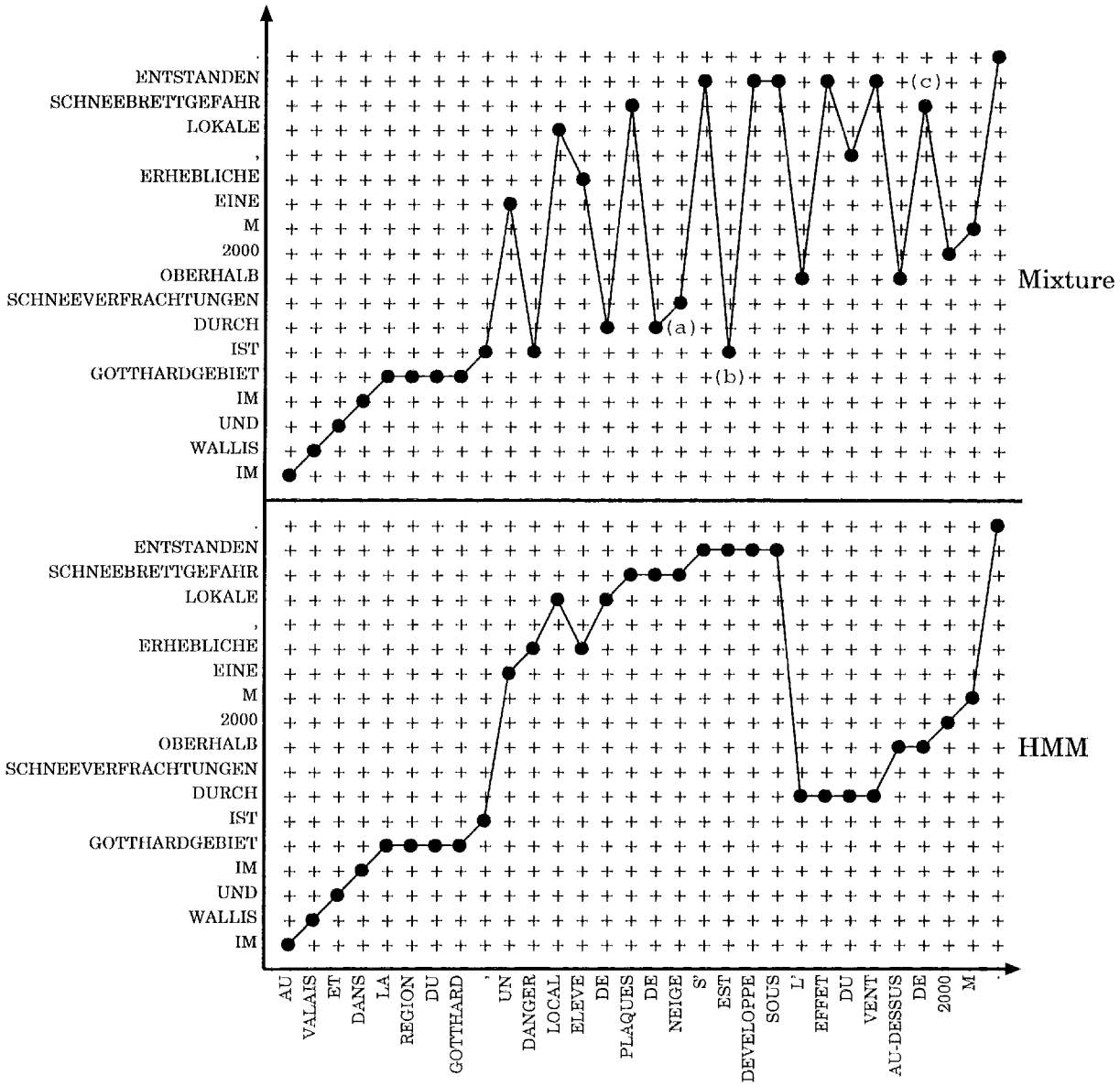

Figure 2: Alignments generated by the IBM2 and the HMM model.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.

Ido Dagan, Ken Church, and William A. Gale. 1993. Robust Bilingual Word Alignment for Machine Aided Translation *Proceedings of the Workshop on Very Large Corpora*, Columbus, Ohio, 1-8

ECI/MCI: The European Corpus Initiative Multilingual Corpus I. 1994. Association for Computational Linguistics.

EuTrans. The Definition of a MT Task. Technical Report, EuTrans Project. 1996(Forthcomming). Depto. de Sistemas Informaticos y Computacion (DSIC), Universidad Politecnica de Valencia.

Pascale Fung, and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. *Proceedings of COLING 94*, 1096-1102, Kyoto, Japan.

Frederik Jelinek. 1976. Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, Vol. 64, 532-556, April 1976.

Martin Kay, and Martin Röscheisen. 1993. Text-Translation Alignment. *Computational Linguistics*, 19(1):121-142

Wolfgang Wahlster. 1993. Verbmobil: Translation of Face-to-Face Dialogs. *Proceedings of the MT Summit IV*, 127-135, Kobe, Japan.