Symmetric Word Alignments for Statistical Machine Translation

Evgeny Matusov and **Richard Zens** and **Hermann Ney**
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{matusov,zens,ney}@cs.rwth-aachen.de

Abstract

In this paper, we address the word alignment problem for statistical machine translation. We aim at creating a symmetric word alignment allowing for reliable one-to-many *and* many-to-one word relationships. We perform the iterative alignment training in the source-to-target and the target-to-source direction with the well-known IBM and HMM alignment models. Using these models, we robustly estimate the local costs of aligning a source word and a target word in each sentence pair. Then, we use efficient graph algorithms to determine the symmetric alignment with minimal total costs (i. e. maximal alignment probability). We evaluate the automatic alignments created in this way on the German–English Verbmobil task and the French–English Canadian Hansards task. We show statistically significant improvements of the alignment quality compared to the best results reported so far. On the Verbmobil task, we achieve an improvement of more than 1% absolute over the baseline error rate of 4.7%.

1 Introduction

Word-aligned bilingual corpora provide important knowledge for many natural language processing tasks, such as the extraction of bilingual word or phrase lexica (Melamed, 2000; Och and Ney, 2000). The solutions of these problems depend heavily on the quality of the word alignment (Och and Ney, 2000). Word alignment models were first introduced in statistical machine translation (Brown et al., 1993). An alignment describes a mapping from source sentence words to target sentence words.

Using the IBM translation models IBM-1 to IBM-5 (Brown et al., 1993), as well as the Hidden-Markov alignment model (Vogel et al., 1996), we can produce alignments of good quality. However, all these models constrain the alignments so that a source word can be aligned to at most one target word. This constraint is useful to reduce the computational complexity of the model training, but makes it hard to align phrases in the target language (English) such as 'the day after tomorrow' to one word in the source language (German) 'übermorgen'. We will present a word alignment algorithm which avoids this constraint and produces symmetric word alignments. This algorithm considers the alignment problem as a task of finding the edge cover with minimal costs in a bipartite graph. The parameters of the IBM models and HMM, in particular the state occupation probabilities, will be used to determine the costs of aligning a specific source word to a target word.

We will evaluate the suggested alignment methods on the German–English Verbmobil task and the French–English Canadian Hansards task. We will show statistically significant improvements compared to state-of-the-art results in (Och and Ney, 2003).

2 Statistical Word Alignment Models

In this section, we will give an overview of the commonly used statistical word alignment techniques. They are based on the source-channel approach to statistical machine translation (Brown et al., 1993). We are given a source language sentence $f_1^J := f_1 ... f_j ... f_J$ which has to be translated into a target language sentence $e_1^I := e_1 ... e_i ... e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\begin{aligned} \hat{e}_1^I &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \end{aligned}$$

This decomposition into two knowledge sources allows for an independent modeling of

target language model $Pr(e_1^I)$ and translation model $Pr(f_1^J|e_1^I)$. Into the translation model, the word alignment A is introduced as a hidden variable:

$$Pr(f_1^J|e_1^I) = \sum_A Pr(f_1^J, A|e_1^I)$$

Usually, the alignment is restricted in the sense that each source word is aligned to at most one target word, i.e. $A = a_1^J$. The alignment may contain the connection $a_j = 0$ with the 'empty' word e_0 to account for source sentence words that are not aligned to any target word at all. A detailed description of the popular translation/alignment models IBM-1 to IBM-5 (Brown et al., 1993), as well as the Hidden-Markov alignment model (HMM) (Vogel et al., 1996) can be found in (Och and Ney, 2003). Model 6 is a loglinear combination of the IBM-4, IBM-1, and the HMM alignment models.

A *Viterbi alignment* $\hat{A}$ of a specific model is an alignment for which the following equation holds:

$$\hat{A} = \underset{A}{\operatorname{argmax}} \left\{ Pr(f_1^J, A|e_1^I) \right\}.$$

3 State Occupation Probabilities

The training of all alignment models is done using the EM-algorithm. In the E-step, the counts for each sentence pair (f_1^J, e_1^I) are calculated. Here, we present this calculation on the example of the HMM. For its lexicon parameters, the marginal probability of a target word e_i to occur at the target sentence position i as the translation of the source word f_j at the source sentence position j is estimated with the following sum:

$$p_j(i, f_1^J|e_1^I) = \sum_{a_1^J : a_j = i} Pr(f_1^J, a_1^J|e_1^I)$$

This value represents the likelihood of aligning f_j to e_i via every possible alignment $A = a_1^J$ that includes the alignment connection $a_j = i$. By normalizing over the target sentence positions, we arrive at the *state occupation probability*:

$$p_j(i|f_1^J, e_1^I) = \frac{p_j(i, f_1^J|e_1^I)}{\sum_{i'=1}^{I} p_j(i', f_1^J|e_1^I)}$$

In the M-step of the EM training, the state occupation probabilities are aggregated for all words in the source and target vocabularies by taking the sum over all training sentence pairs. After proper renormalization the lexicon probabilities $p(f|e)$ are determined.

Similarly, the training can be performed in the inverse (target-to-source) direction, yielding the state occupation probabilities $p_i(j|e_1^I, f_1^J)$.

The negated logarithms of the state occupation probabilities

$$w(i, j; f_1^J, e_1^I) := -\log p_j(i|f_1^J, e_1^I) \quad (1)$$

can be viewed as *costs* of aligning the source word f_j with the target word e_i. Thus, the word alignment task can be formulated as the task of finding a mapping between the source and the target words, so that each source and each target position is covered and the total costs of the alignment are minimal.

Using state occupation probabilities for word alignment modeling results in a number of advantages. First of all, in calculation of these probabilities with the models IBM-1, IBM-2 and HMM the EM-algorithm is performed exact, i.e. the summation over all alignments is efficiently performed in the E-step. For the HMM this is done using the Baum-Welch algorithm (Baum, 1972). So far, an efficient algorithm to compute the sum over all alignments in the fertility models IBM-3 to IBM-5 is not known. Therefore, this sum is approximated using a subset of promising alignments (Och and Ney, 2000). In both cases, the resulting estimates are more precise than the ones obtained by the maximum approximation, i. e. by considering only the Viterbi alignment.

Instead of using the state occupation probabilities from only one training direction as costs (Equation 1), we can interpolate the state occupation probabilities from the source-to-target and the target-to-source training for each pair (i,j) of positions in a sentence pair (f_1^J, e_1^I). This will improve the estimation of the local alignment costs. Having such symmetrized costs, we can employ the graph alignment algorithms (cf. Section 4) to produce reliable alignment connections which include many-to-one and one-to-many alignment relationships. The presence of both relationship types characterizes a symmetric alignment that can potentially improve the translation results (Figure 1 shows an example of a symmetric alignment).

Another important advantage is the efficiency of the graph algorithms used to deter-

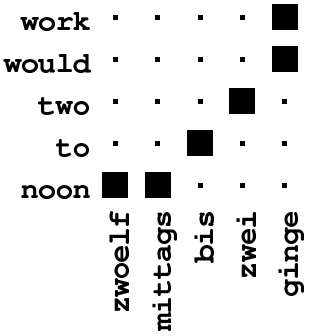

Figure 1: Example of a symmetric alignment with one-to-many and many-to-one connections (Verbmobil task, spontaneous speech).

mine the final symmetric alignment. They will be discussed in Section 4.

4 Alignment Algorithms

In this section, we describe the alignment extraction algorithms. We assume that for each sentence pair (f_1^J, e_1^I) we are given a cost matrix C.[1] The elements of this matrix c_{ij} are the local costs that result from aligning source word f_j to target word e_i. For a given alignment $A \subseteq I \times J$, we define the costs of this alignment $c(A)$ as the sum of the local costs of all aligned word pairs:

$$c(A) = \sum_{(i,j)\in A} c_{ij} \quad (2)$$

Now, our task is to find the alignment with the minimum costs. Obviously, the empty alignment has always costs of zero and would be optimal. To avoid this, we introduce additional constraints. The first constraint is source sentence coverage. Thus each source word has to be aligned to at least one target word or alternatively to the empty word. The second constraint is target sentence coverage. Similar to the source sentence coverage thus each target word is aligned to at least one source word or the empty word.

Enforcing only the source sentence coverage, the minimum cost alignment is a mapping from source positions j to target positions a_j, including zero for the empty word. Each target position a_j can be computed as:

$$a_j = \operatorname*{argmin}_i \{c_{ij}\}$$

This means, in each column we choose the row with the minimum costs. This method resembles the common IBM models in the sense that the IBM models are also a mapping from source positions to target positions. Therefore, this method is comparable to the IBM models for the source-to-target direction. Similarly, if we enforce only the target sentence coverage, the minimum cost alignment is a mapping from target positions i to source positions b_i. Here, we have to choose in each row the column with the minimum costs. The complexity of these algorithms is in $O(I \cdot J)$.

The algorithms for determining such a non-symmetric alignment are rather simple. A more interesting case arises, if we enforce both constraints, i.e. each source word as well as each target word has to be aligned at least once. Even in this case, we can find the global optimum in polynomial time.

The task is to find a symmetric alignment A, for which the costs $c(A)$ are minimal (Equation 2). This task is equivalent to finding a *minimum-weight edge cover* (MWEC) in a complete bipartite graph[2]. The two node sets of this bipartite graph correspond to the source sentence positions and the target sentence positions, respectively. The costs of an edge are the elements of the cost matrix C.

To solve the minimum-weight edge cover problem, we reduce it to the maximum-weight bipartite matching problem. As described in (Keijsper and Pendavingh, 1998), this reduction is linear in the graph size. For the maximum-weight bipartite matching problem, well-known algorithm exist, e.g. the Hungarian method. The complexity of this algorithm is in $O((I+J) \cdot I \cdot J)$. We will call the solution of the minimum-weight edge cover problem with the Hungarian method "the MWEC algorithm". In contrary, we will refer to the algorithm enforcing either source sentence coverage or target sentence coverage as the *one-sided* minimum-weight edge cover algorithm (o-MWEC).

The cost matrix of a sentence pair (f_1^J, e_1^I) can be computed as a weighted linear interpolation of various cost types h_m:

$$c_{ij} = \sum_{m=1}^{M} \lambda_m \cdot h_m(i,j)$$

In our experiments, we will use the negated logarithm of the state occupation probabilities as described in Section 3. To obtain a more symmetric estimate of the costs, we will interpolate both the source-to-target direction and

[1] For notational convenience, we omit the dependency on the sentence pair (f_1^J, e_1^I) in this section.

[2] An edge cover of G is a set of edges E' such that each node of G is incident to at least one edge in E'.

the target-to-source direction (thus the state occupation probabilities are interpolated log-linearly). Because the alignments determined in the source-to-target training may substantially differ in quality from those produced in the target-to-source training, we will use an interpolation weight α:

$$c_{ij} = \alpha \cdot w(i, j; f_1^J, e_1^I) + (1 - \alpha) \cdot w(j, i; e_1^I, f_1^J) \quad (3)$$

Additional feature functions can be included to compute c_{ij}; for example, one could make use of a bilingual word or phrase dictionary.

To apply the methods described in this section, we made two assumptions: first, the costs of an alignment can be computed as the sum of *local* costs. Second, the features have to be *static* in the sense that we have to fix the costs before aligning any word. Therefore, we cannot apply dynamic features such as the IBM-4 distortion model in a straightforward way. One way to overcome these restrictions lies in using the state occupation probabilities; e.g. for IBM-4, they contain the distortion model to some extent.

5 Results

5.1 Evaluation Criterion

We use the same evaluation criterion as described in (Och and Ney, 2000). We compare the generated word alignment to a reference alignment produced by human experts. The annotation scheme explicitly takes the ambiguity of the word alignment into account. There are two different kinds of alignments: sure alignments (S) which are used for unambiguous alignments and possible alignments (P) which are used for alignments that might or might not exist. The P relation is used especially to align words within idiomatic expressions and free translations. It is guaranteed that the sure alignments are a subset of the possible alignments ($S \subseteq P$). The obtained reference alignment may contain many-to-one and one-to-many relationships.

The quality of an alignment A is computed as appropriately redefined precision and recall measures. Additionally, we use the alignment error rate (AER), which is derived from the well-known F-measure.

$$\text{recall} = \frac{|A \cap S|}{|S|}, \quad \text{precision} = \frac{|A \cap P|}{|A|}$$

$$\text{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

Table 1: Verbmobil task: corpus statistics.

Source/Target:		German	English
Train	Sentences	34 446	
	Words	329 625	343 076
	Vocabulary	5 936	3 505
	Singletons	2 600	1 305
Dictionary	Entries	4 404	
Test	Sentences	354	
	Words	3 233	3 109
S reference relations		2 559	
P reference relations		4 596	

Table 2: Canadian Hansards: corpus statistics.

Source/Target:		French	English
Train	Sentences	128K	
	Words	2.12M	1.93M
	Vocabulary	37 542	29 414
	Singletons	12 986	9 572
Dictionary	Entries	28 701	
Test	Sentences	500	
	Words	8 749	7 946
S reference relations		4 443	
P reference relations		19 779	

With these definitions a recall error can only occur if a S(ure) alignment is not found and a precision error can only occur if a found alignment is not even P(ossible).

5.2 Experimental Setup

We evaluated the presented lexicon symmetrization methods on the Verbmobil and the Canadian Hansards task. The German–English Verbmobil task (Wahlster, 2000) is a speech translation task in the domain of appointment scheduling, travel planning and hotel reservation. The French–English Canadian Hansards task consists of the debates in the Canadian Parliament.

The corpus statistics are shown in Table 1 and Table 2. The number of running words and the vocabularies are based on full-form words including punctuation marks. As in (Och and Ney, 2003), the first 100 sentences of the test corpus are used as a development corpus to optimize model parameters that are not trained via the EM algorithm, e.g. the interpolation weights. The remaining part of the test corpus is used to evaluate the models.

We use the same training schemes (model sequences) as presented in (Och and Ney, 2003): $1^5H^5 3^3 4^3 6^3$ for the Verbmobil Task , i.e. 5 iteration of IBM-1, 5 iterations of the

HMM, 3 iteration of IBM-3, etc.; for the Canadian Hansards task, we use $1^5H^{10}3^34^36^3$. We refer to these schemes as the *Model 6 schemes*. For comparison, we also perform less sophisticated trainings, to which we refer as the *HMM schemes* (1^5H^{10} and 1^5H^5, respectively), as well as the *IBM Model 4 schemes* ($1^5H^{10}3^34^3$ and $1^5H^53^34^3$).

In all training schemes we use a conventional dictionary (possibly containing phrases) as additional training material. Because we use the same training and testing conditions as (Och and Ney, 2003), we will refer to the results presented in that article as the baseline results.

5.3 Non-symmetric Alignments

In the first experiments, we use the state occupation probabilities from only one translation direction to determine the word alignment. This allows for a fair comparison with the Viterbi alignment computed as the result of the training procedure. In the source-to-target translation direction, we cannot estimate the probability for the target words with fertility zero and choose to set it to 0. In this case, the minimum weight edge cover problem is solved by the one-sided MWEC algorithm. Like the Viterbi alignments, the alignments produced by this algorithm satisfy the constraint that multiple source (target) words can only be aligned to one target (source) word.

Tables 3 and 4 show the performance of the one-sided MWEC algorithm in comparison with the experiment reported by (Och and Ney, 2003). We report not only the final alignment error rates, but also the intermediate results for the HMM and IBM-4 training schemes.

For IBM-3 to IBM-5, the Viterbi alignment and a set of promising alignments are used to determine the state occupation probabilities. Consequently, we observe similar alignment quality when comparing the Viterbi and the one-sided MWEC alignments.

We also evaluated the alignment quality after applying alignment generalization methods, i.e. we combine the alignment of both translation directions. Experimentally, the best generalization heuristic for the Canadian Hansards task is the intersection of the source-to-target and the target-to-source alignments. For the Verbmobil task, the refined method of (Och and Ney, 2003) is used. Again, we observed similar alignment error rates when merging either the Viterbi alignments or the o-MWEC alignments.

Table 3: AER [%] for non-symmetric alignment methods and for various models (HMM, IBM-4, Model 6) on the Canadian Hansards task.

Alignment method	HMM	IBM4	M6
Baseline T→S	14.1	12.9	11.9
S→T	14.4	12.8	11.7
intersection	8.4	6.9	7.8
o-MWEC T→S	14.0	13.1	11.9
S→T	14.3	13.0	11.7
intersection	8.2	7.1	7.8

Table 4: AER [%] for non-symmetric alignment methods and for various models (HMM, IBM-4, Model 6) on the Verbmobil task.

Alignment method	HMM	IBM4	M6
Baseline T→S	7.6	4.8	4.6
S→T	12.1	9.3	8.8
refined	7.1	4.7	4.7
o-MWEC T→S	7.3	4.8	4.5
S→T	12.0	9.3	8.5
refined	6.7	4.6	4.6

5.4 Symmetric Alignments

The heuristically generalized Viterbi alignments presented in the previous section can potentially avoid the alignment constraints[3]. However, the choice of the optimal generalization heuristic may depend on a particular language pair and may require extensive manual optimization. In contrast, the symmetric MWEC algorithm is a systematic and theoretically well-founded approach to the task of producing a symmetric alignment.

In the experiments with the symmetric MWEC algorithm, the optimal interpolation parameter α (see Equation 3) for the Verbmobil corpus was empirically determined as 0.8. This shows that the model parameters can be estimated more reliably in the direction from German to English. In the inverse English-to-German alignment training, the mappings of many English words to one German word are not allowed by the modeling constraints, although such alignment mappings are significantly more frequent than mappings of many German words to one English word.

The experimentally best interpolation parameter for the Canadian Hansards corpus was $\alpha = 0.5$. Thus the model parameters estimated in the translation direction from French to English are as reliable as the ones estimated

[3]Consequently, we will use them as baseline for the experiments with symmetric alignments.

in the direction from English to French.

Lines 2a and 2b of Table 5 show the performance of the MWEC algorithm. The alignment error rates are slightly lower if the HMM or the full Model 6 training scheme is used to train the state occupation probabilities on the Canadian Hansards task. On the Verbmobil task, the improvement is more significant, yielding an alignment error rate of 4.1%.

Columns 4 and 5 of Table 5 contain the results of the experiments, in which the costs c_{ij} were determined as the loglinear interpolation of state occupation probabilities obtained from the HMM training scheme with those from IBM-4 (column 4) or from Model 6 (column 5). We set the interpolation parameters for the two translation directions proportional to the optimal values determined in the previous experiments. On the Verbmobil task, we obtain a further improvement of 19% relative over the baseline result reported in (Och and Ney, 2003), reaching an AER as low as 3.8%.

The improvements of the alignment quality on the Canadian Hansards task are less significant. The manual reference alignments for this task contain many possible connections and only a few sure connections (cf. Table 2). Thus automatic alignments consisting of only a few reliable alignment points are favored. Because the differences in the number of words and word order between French and English are not as dramatic as e.g. between German and English, the probability of the empty word alignment is not very high. Therefore, plenty of alignment points are produced by the MWEC algorithm, resulting in a high recall and low precision. To increase the precision, we replaced the empty word connection costs (previously trained as state occupation probabiliities using the EM algorithm) by the global, word- and position-independent costs depending only on one of the involved languages. The alignment error rates for these experiments are given in lines 3a and 3b of Table 5. The global empty word probability for the Canadian Hansards task was empirically set to 0.45 for French and for English, and, for the Verbmobil task, to 0.6 for German and 0.1 for English. On the Canadian Hansards task, we achieved further significant reduction of the AER. In particular, we reached an AER of 6.6% by performing only the HMM training. In this case the effectiveness of the MWEC algorithm is combined with the efficiency of the HMM training, resulting in a fast and robust alignment training procedure.

We also tested the more simple one-sided MWEC algorithm. In contrast to the experiments presented in Section 5.3, we used the loglinear interpolated state occupation probabilities (given by the Equation 3) as costs. Thus, although the algorithm is not able to produce a symmetric alignment, it operates with symmetrized costs. In addition, we used a combination heuristic to obtain a symmetric alignment. The results of these experiments are presented in Table 5, lines 4-6 a/b.

The performance of the one-sided MWEC algorithm turned out to be quite robust on both tasks. However, the o-MWEC alignments are not symmetric and the achieved low AER depends heavily on the differences between the involved languages, which may favor many-to-one alignments in one translation direction only. That is why on the Verbmobil task, when determining the mininum weight in each row for the translation direction from English to German, the alignment quality deteriorates, because the algorithm cannot produce alignments which map several English words to one German word (line 5b of Table 5).

Applying the generalization heuristics (line 6a/b of Table 5), we achieve an AER of 6.0% on the Canadian Hansards task when interpolating the state occupation probabilities trained with the HMM and with the IBM-4 schemes. On the Verbmobil task, the interpolation of the HMM and the Model 6 schemes yields the best result of 3.7% AER. In the latter experiment, we reached 97.3% precision and 95.2% recall.

6 Related Work

A description of the IBM models for statistical machine translation can be found in (Brown et al., 1993). The HMM-based alignment model was introduced in (Vogel et al., 1996). An overview of these models is given in (Och and Ney, 2003). That article also introduces the Model 6; additionally, state-of-the-art results are presented for the Verbmobil task and the Canadian Hansards task for various configurations. Therefore, we chose them as baseline. Additional linguistic knowledge sources such as dependeny trees or parse trees were used in (Cherry and Lin, 2003; Gildea, 2003). Bilingual bracketing methods were used to produce a word alignment in (Wu, 1997). (Melamed, 2000) uses an alignment model that enforces one-to-one alignments for nonempty words.

Table 5: AER[%] for different alignment symmetrization methods and for various alignment models on the Canadian Hansards and the Verbmobil tasks (MWEC: minimum weight edge cover, EW: empty word).

	Symmetrization Method	HMM	IBM4	M6	HMM + IBM4	HMM + M6
Canadian Hansards	1a. Baseline (intersection)	8.4	6.9	7.8	–	–
	2a. MWEC	7.9	9.3	7.5	8.2	7.4
	3a. MWEC (global EW costs)	6.6	7.4	6.9	6.4	6.4
	4a. o-MWEC T→S	7.3	7.9	7.4	6.7	7.0
	5a. S→T	7.7	7.6	7.2	6.9	6.9
	6a. S↔T (intersection)	7.2	6.6	7.6	6.0	7.1

	Symmetrization Method	HMM	IBM4	M6	HMM + IBM4	HMM + M6
Verbmobil	1b. Baseline (refined)	7.1	4.7	4.7	–	–
	2b. MWEC	6.4	4.4	4.1	4.3	3.8
	3b. MWEC (global EW costs)	5.8	5.8	6.6	6.0	6.7
	4b. o-MWEC T→S	6.8	4.4	4.1	4.5	3.7
	5b. S→T	9.3	7.2	6.8	7.5	6.9
	6b. S↔T (refined)	6.7	4.3	4.1	4.6	3.7

7 Conclusions

In this paper, we addressed the task of automatically generating symmetric word alignments for statistical machine translation. We exploited the state occupation probabilties derived from the IBM and HMM translation models. We used the negated logarithms of these probabilities as local alignment costs and reduced the word alignment problem to finding an edge cover with minimal costs in a bipartite graph. We presented efficient algorithms for the solution of this problem. We evaluated the performance of these algorithms by comparing the alignment quality to manual reference alignments. We showed that interpolating the alignment costs of the source-to-target and the target-to-source translation directions can result in a significant improvement of the alignment quality.

In the future, we plan to integrate the graph algorithms into the iterative training procedure. Investigating the usefulness of additional feature functions might be interesting as well.

Acknowledgment

This work has been partially funded by the EU project TransType 2, IST-2001-32091.

References

L. E. Baum. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.

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):263–311, June.

C. Cherry and D. Lin. 2003. A probability model to improve word alignment. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, Sapporo, Japan, July.

D. Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87, Sapporo, Japan, July.

J. Keijsper and R. Pendavingh. 1998. An efficient algorithm for minimum-weight bibranching. *Journal of Combinatorial Theory Series B*, 73(2):130–145, July.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

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

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

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

W. Wahlster, editor. 2000. *Verbmobil: Foundations of speech-to-speech translations*. Springer Verlag, Berlin, Germany, July.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, September.