# Improved Word Alignment Using a Symmetric Lexicon Model

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

## Abstract

Word-aligned bilingual corpora are an important knowledge source for many tasks in natural language processing. We improve the well-known IBM alignment models, as well as the Hidden-Markov alignment model using a symmetric lexicon model. This symmetrization takes not only the standard translation direction from source to target into account, but also the inverse translation direction from target to source. We present a theoretically sound derivation of these techniques. In addition to the symmetrization, we introduce a smoothed lexicon model. The standard lexicon model is based on full-form words only. We propose a lexicon smoothing method that takes the word base forms explicitly into account. Therefore, it is especially useful for highly inflected languages such as German. We evaluate these methods 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 system reported so far. For the Canadian Hansards task, we achieve an improvement of more than 30% relative.

## 1 Introduction

Word-aligned bilingual corpora are an important knowledge source for many tasks in natural language processing. Obvious applications are the extraction of bilingual word or phrase lexica (Melamed, 2000; Och and Ney, 2000). These applications 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). The alignment describes the 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. In (Och and Ney, 2003), it is shown that the statistical approach performs very well compared to alternative approaches, e.g. based on the Dice coefficient or the competitive linking algorithm (Melamed, 2000).

A central component of the statistical translation models is the lexicon. It models the word translation probabilities. The standard training procedure of the statistical models uses the EM algorithm. Typically, the models are trained for one translation direction only. Here, we will perform a simultaneous training of both translation directions, source-to-target and target-to-source. After each iteration of the EM algorithm, we combine the two lexica to a symmetric lexicon. This symmetric lexicon is then used in the next iteration of the EM algorithm for both translation directions. We will propose and justify linear and loglinear interpolation methods.

Statistical methods often suffer from the data sparseness problem. In our case, many words in the bilingual sentence-aligned texts are singletons, i.e. they occur only once. This is especially true for the highly inflected languages such as German. It is hard to obtain reliable estimations of the translation probabilities for these rarely occurring words. To overcome this problem (at least partially), we will smooth the lexicon probabilities of the full-form words using a probability distribution that is estimated using the word base forms. Thus, we exploit that multiple full-form words share the same base form and have similar meanings and translations.

We will evaluate these 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). On the Canadian

Hansards task, the symmetrization methods will result in an improvement of more than 30% relative.

## 2 Statistical Word Alignment Models

In this section, we will give a short description of the commonly used statistical word alignment models. These alignment models stem from 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:

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

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, we use restricted alignments in the sense that each source word is aligned to at most one target word, i.e. $A = a_1^J$. A detailed description of the popular translation 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). All these models include parameters $p(f|e)$ for the single-word based lexicon. They differ in the alignment model.

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

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

We measure the quality of an alignment model using the quality of the Viterbi alignment compared to a manually produced reference alignment.

In Section 3, we will apply the lexicon symmetrization methods to the models described previously. Therefore, we will now sketch the standard training procedure for the lexicon model. The EM algorithm is used to train the free lexicon parameters $p(f|e)$.

In the E-step, the lexical counts for each sentence pair $(f_1^J, e_1^I)$ are calculated and then summed over all sentence pairs in the training corpus:

$$N(f,e) = \sum_{(f_1^J, e_1^I)} \sum_{a_1^J} p(a_1^J | f_1^J, e_1^I) \sum_{i,j} \delta(f, f_j) \delta(e, e_i)$$

In the M-step the lexicon probabilities are:

$$p(f|e) = \frac{N(f,e)}{\sum_{\tilde{f}} N(\tilde{f}, e)}$$

## 3 Symmetrized Lexicon Model

During the standard training procedure, the lexicon parameters $p(f|e)$ and $p(e|f)$ were estimated independent of each other in strictly separate trainings. In this section, we present two symmetrization methods for the lexicon model. As a starting point, we use the *joint* lexicon probability $p(f, e)$ and determine the conditional probabilities for the source-to-target direction $p(f|e)$ and the target-to-source direction $p(e|f)$ as the corresponding marginal distribution:

$$p(f|e) = \frac{p(f,e)}{\sum_{\tilde{f}} p(\tilde{f}, e)} \quad (1)$$

$$p(e|f) = \frac{p(f,e)}{\sum_{\tilde{e}} p(f, \tilde{e})} \quad (2)$$

The nonsymmetric auxiliary $Q$-functions for reestimating the lexicon probabilities during the EM algorithm can be represented as follows. Here, $N_{ST}(f,e)$ and $N_{TS}(f,e)$ denote the lexicon counts for the source-to-target ($ST$) direction and the target-to-source ($TS$) direction, respectively.

$$Q_{ST}(\{p(f|e)\}) = \sum_{f,e} N_{ST}(f,e) \cdot \log \frac{p(f,e)}{\sum_{\tilde{f}} p(\tilde{f}, e)}$$

$$Q_{TS}(\{p(e|f)\}) = \sum_{f,e} N_{TS}(f,e) \cdot \log \frac{p(f,e)}{\sum_{\tilde{e}} p(f, \tilde{e})}$$

### 3.1 Linear Interpolation

To estimate the joint probability using the EM algorithm, we define the auxiliary $Q$-function as a linear interpolation of the $Q$-functions for the source-to-target and the target-to-source direction:

$$
\begin{aligned}
Q_\alpha(\{p(f,e)\}) &= \alpha \cdot Q_{ST}(\{p(f|e)\}) \\
&\quad +(1-\alpha) \cdot Q_{TS}(\{p(e|f)\}) \\
&= \alpha \cdot \sum_{f,e} N_{ST}(f,e) \cdot \log p(f,e) \\
&\quad +(1-\alpha) \cdot \sum_{f,e} N_{TS}(f,e) \cdot \log p(f,e) \\
&\quad -\alpha \cdot \sum_{e} N_{ST}(e) \cdot \log \sum_{\tilde{f}} p(\tilde{f},e) \\
&\quad -(1-\alpha) \cdot \sum_{f} N_{TS}(f) \cdot \log \sum_{\tilde{e}} p(f,\tilde{e})
\end{aligned}
$$

The unigram counts $N(e)$ and $N(f)$ are determined, for each of the two translation directions, by taking a sum of $N(f,e)$ over $f$ and over $e$, respectively. We define the *combined lexicon count* $N_\alpha(f,e)$:

$$N_\alpha(f,e) := \alpha \cdot N_{ST}(f,e) + (1-\alpha) \cdot N_{TS}(f,e)$$

Now, we derive the symmetrized $Q$-function over $p(f,e)$ for a certain word pair $(f,e)$. Then, we set this derivative to zero to determine the reestimation formula for $p(f,e)$ and obtain the following equation:

$$\frac{N_\alpha(f,e)}{p(f,e)} = \alpha \cdot \frac{N_{ST}(e)}{\sum_{\tilde{f}} p(\tilde{f},e)} + (1-\alpha) \cdot \frac{N_{TS}(f)}{\sum_{\tilde{e}} p(f,\tilde{e})}$$

We do not know a closed form solution for this equation. As an approximation, we use the following term:

$$\hat{p}(f,e) = \frac{N_\alpha(f,e)}{\sum_{\tilde{f},\tilde{e}} N_\alpha(\tilde{f},\tilde{e})}$$

This estimate is an exact solution, if the unigram counts for $f$ and $e$ are independent of the translation direction, i.e. $N_{ST}(f) = N_{TS}(f)$ and $N_{ST}(e) = N_{TS}(e)$. We make this approximation and thus we interpolate the lexicon counts linear after each iteration of the EM algorithm. Then, we normalize these counts (according to Equations 1 and 2) to determine the lexicon probabilities for each of the two translation directions.

### 3.2 Loglinear Interpolation

We will show in Section 5 that the linear interpolation results in significant improvements over the nonsymmetric system. Motivated by these experiments, we investigated also the loglinear interpolation of the lexicon counts of the two translation directions. The combined lexicon count $N_\alpha(f,e)$ is now defined as:

$$N_\alpha(f,e) = N_{ST}(f,e)^\alpha \cdot N_{TS}(f,e)^{1-\alpha}$$

The normalization is done in the same way as for the linear interpolation. The linear interpolation resembles more a union of the two lexica whereas the loglinear interpolation is more similar to an intersection of both lexica. Thus for the linear interpolation, a word pair $(f,e)$ obtains a large combined count, if the count in at least one direction is large. For the loglinear interpolation, the combined count is large only if both lexicon counts are large.

In the experiments, we will use the interpolation weight $\alpha = 0.5$ for both the linear and the loglinear interpolation, i.e. both translation directions are weighted equally.

### 3.3 Evidence Trimming

Initially, the lexicon contains all word pairs that cooccur in the bilingual training corpus. The majority of these word pairs are not translations of each other. Therefore, we would like to remove those lexicon entries. Evidence trimming is one way to do this. The *evidence* of a word pair $(f,e)$ is the estimated count $N(f,e)$. Now, we discard a word pair if its evidence is below a certain threshold $\tau$.[1] In the case of the symmetric lexicon, we can further refine this method. For estimating the lexicon in the source-to-target direction $\hat{p}(f|e)$, the idea is to keep all entries from this direction and to boost the entries that have a high evidence in the target-to-source direction $N_{TS}(f,e)$. We obtain the following formula:

$$\bar{N}_{ST}(f,e) = \begin{cases} \alpha N_{ST}(f,e) + (1-\alpha) N_{TS}(f,e) \\ \qquad \text{if } N_{ST}(f,e) > \tau \\ 0 \text{ else} \end{cases}$$

The count $\bar{N}_{ST}(f,e)$ is now used to estimate the source-to-target lexicon $\hat{p}(f|e)$. With this method, we do not keep entries in the source-to-target lexicon $\hat{p}(f|e)$ if their evidence is low, even if their evidence in the target-to-source

[1] Actually, there is always implicit evidence trimming caused by the limited machine precision.

direction $N_{TS}(f,e)$ is high. For the target-to-source direction, we apply this method in a similar way.

## 4 Lexicon Smoothing

The lexicon model described so far is based on full-form words. For highly inflected languages such as German this might cause problems, because many full-form words occur only a few times in the training corpus. Compared to English, the token/type ratio for German is usually much lower (e.g. Verbmobil: English 99.4, German 56.3). The information that multiple full-form words share the same base form is not used in the lexicon model. To take this information into account, we smooth the lexicon model with a backing-off lexicon that is based on word base forms. The smoothing method we apply is absolute discounting with interpolation:

$$p(f|e) = \frac{\max\{N(f,e)-d,0\}}{N(e)} + \alpha(e)\cdot\beta(f,\bar{e})$$

This method is well known from language modeling (Ney et al., 1997). Here, $\bar{e}$ denotes the generalization, i.e. the base form, of the word $e$. The nonnegative value $d$ is the discounting parameter, $\alpha(e)$ is a normalization constant and $\beta(f,\bar{e})$ is the normalized backing-off distribution.

The formula for $\alpha(e)$ is:

$$\begin{aligned}\alpha(e) &= \frac{1}{N(e)}\left(\sum_{f:N(f,e)>d} d + \sum_{f:N(f,e)\le d} N(f,e)\right)\\ &= \frac{1}{N(e)}\sum_{f}\min\{d,N(f,e)\}\end{aligned}$$

This formula is a generalization of the one typically used in publications on language modeling. This generalization is necessary, because the lexicon counts may be fractional whereas in language modeling typically integer counts are used. Additionally, we want to allow for discounting values $d$ greater than one. The backing-off distribution $\beta(f,\bar{e})$ is estimated using relative frequencies:

$$\beta(f,\bar{e}) \;=\; \frac{N(f,\bar{e})}{\sum\limits_{\tilde{f}} N(\tilde{f},\bar{e})}$$

Here, $N(f,\bar{e})$ denotes the count of the event that the source language word $f$ and the target language base form $\bar{e}$ occur together. These counts are computed by summing the lexicon counts $N(f,e)$ over all full-form words $e$ which share the same base form $\bar{e}$.

## 5 Results

### 5.1 Evaluation Criteria

We use the same evaluation criterion as described in (Och and Ney, 2000). The generated word alignment is compared to a reference alignment which is 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 alignments that are unambiguous 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, free translations, and missing function words. 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.

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

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

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

Table 1: Verbmobil: Corpus statistics.

| | | German | English |
|---|---|---|---|
| Train | Sentences | 34K | |
| | Words | 329 625 | 343 076 |
| | Vocabulary | 5 936 | 3 505 |
| | Singletons | 2 600 | 1 305 |
| Test | Sentences | 354 | |
| | Words | 3 233 | 3 109 |

Table 2: Canadian Hansards: Corpus statistics.

| | | French | English |
|---|---|---|---|
| Train | Sentences | 128K | |
| | Words | 2.12M | 1.93M |
| | Vocabulary | 37 542 | 29 414 |
| | Singletons | 12 986 | 9 572 |
| Test | Sentences | 500 | |
| | Words | 8 749 | 7 946 |

(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 discounting parameter for lexicon smoothing. 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). As 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. In (Och and Ney, 2003), the alignment quality of statistical models is compared to alternative approaches, e.g. using the Dice coefficient or the competitive linking algorithm. The statistical approach showed the best performance and therefore we report only the results for the statistical systems.

### 5.3 Lexicon Symmetrization

In Table 3 and Table 4, we present the following experiments performed for both the Verbmobil and the Canadian Hansards task:

- **Base:** the system taken from (Och and Ney, 2003) that we use as baseline system.
- **Lin.:** symmetrized lexicon using a linear interpolation of the lexicon counts after each training iteration as described in Section 3.1.
- **Log.:** symmetrized lexicon using a log-linear interpolation of the lexicon counts after each training iteration as described in Section 3.2.

Table 3: Comparison of alignment performance for the Verbmobil task (S→T: source-to-target direction, T→S: target-to-source direction; all numbers in percent).

| | S→T | | | T→S | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | AER | Pre. | Rec. | AER |
| Base | 93.5 | 95.3 | 5.7 | 91.4 | 88.7 | 9.9 |
| Lin. | 96.0 | 95.4 | 4.3 | 93.7 | 89.6 | 8.2 |
| Log. | 93.6 | 95.6 | 5.5 | 94.5 | 89.4 | 7.9 |

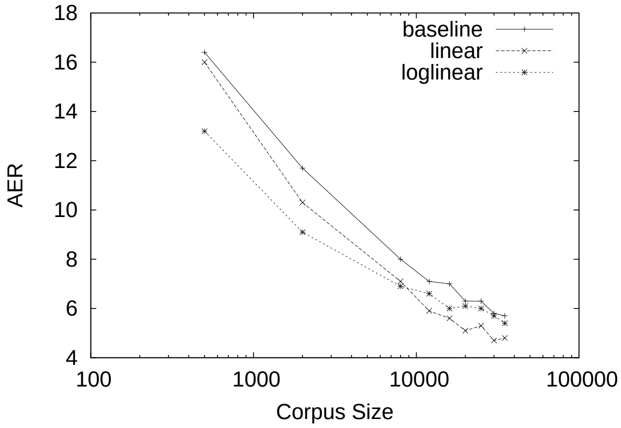

Figure 1: AER[%] of different alignment methods as a function of the training corpus size for the Verbmobil task (source-to-target direction).

In Table 3, we compare both interpolation variants for the Verbmobil task to (Och and Ney, 2003). We observe notable improvements in the alignment error rate using the linear interpolation. For the translation direction from German to English (S→T), an improvement of about 25% relative is achieved from an alignment error rate of 5.7% for the baseline system to 4.3% using the linear interpolation. Performing the loglinear interpolation, we observe a substantial reduction of the alignment error rate as well. The two symmetrization methods improve both precision *and* recall of the resulting Viterbi alignment in both translation directions for the Verbmobil task. The improvements with the linear interpolation is for both translation directions statistically significant at the 99% level. For the loglinear interpolation, the target-to-source translation direction is statistically significant at the 99% level. The statistical significance test were done using boostrap resampling.

We also performed experiments on subcorpora of different sizes. For the Verbmobil task, the results are illustrated in Figure 1.

Table 4: Comparison of alignment performance for the Canadian Hansards task (S→T: source-to-target direction, T→S: target-to-source direction; all numbers in percent).

| | S→T | | | T→S | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | AER | Pre. | Rec. | AER |
| Base | 85.4 | 90.6 | 12.6 | 85.6 | 90.9 | 12.4 |
| Lin. | 89.3 | 91.4 | 9.9 | 89.0 | 92.0 | 9.8 |
| Log. | 91.0 | 92.0 | 8.6 | 91.2 | 92.1 | 8.4 |

We observe that both symmetrization variants result in improvements for all corpus sizes. With increasing training corpus size the performance of the linear interpolation becomes superior to the performance of the loglinear interpolation.

In Table 4, we compare the symmetrization methods with the baseline system for the Canadian Hansards task. Here, the loglinear interpolation performs best. We achieve a relative improvement over the baseline of more than 30% for both translation directions. For instance, the alignment error rate for the translation direction from French to English (S→T) improves from 12.6% for the baseline system to 8.6% for the symmetrized system with loglinear interpolation. Again, the two symmetrization methods improve both precision *and* recall of the Viterbi alignment.

For the Canadian Hansards task, all the improvements of the alignment error rate are statistically significant at the 99% level.

### 5.4 Generalized Alignments

In (Och and Ney, 2003) generalized alignments are used, thus the final Viterbi alignments of both translation directions are combined using some heuristic. Experimentally, the best heuristic for the Canadian Hansards task is the intersection. For the Verbmobil task, the refined method of (Och and Ney, 2003) is used. The results are summarized in Table 5. We see that both the linear and the loglinear lexicon symmetrization methods yield an improvement with respect to the alignment error rate. For the Verbmobil task, the improvement with the loglinear interpolation is statistically significant at the 99% level. For the Canadian Hansards task, both lexicon symmetrization methods result in statistically significant improvements at the 95% level. Additionally, we observe that precision and recall are more balanced for the symmetrized lexicon variants, especially for the Canadian Hansards task.

Table 6: Effect of smoothing the lexicon probabilities on the alignment performance for the Verbmobil task (S→T: source-to-target direction, smooth English; T→S: target-to-source direction, smooth German; all numbers in percent).

| | S→T | | | T→S | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | AER | Pre. | Rec. | AER |
| Base | 93.5 | 95.3 | 5.7 | 91.4 | 88.7 | 9.9 |
| smooth | 94.8 | 94.8 | 5.2 | 93.4 | 88.2 | 9.1 |

### 5.5 Lexicon Smoothing

In Table 6, we present the results for the lexicon smoothing as described in Section 4 on the Verbmobil corpus[2]. As expected, a notable improvement in the AER is reached if the lexicon smoothing is performed for German (i.e. for the target-to-source direction), because many full-form words with the same base form are present in this language. These improvements are statistically significant at the 95% level.

## 6 Related Work

The popular IBM models for statistical machine translation are described in (Brown et al., 1993). The HMM-based alignment model was introduced in (Vogel et al., 1996). A good overview of these models is given in (Och and Ney, 2003). In that article Model 6 is introduced as the loglinear interpolation of the other models. 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. Compared to our work, these publications kept the training of the two translation directions strictly separate whereas we integrate both directions into one symmetrized training. Additional linguistic knowledge sources such as dependency trees or parse trees were used in (Cherry and Lin, 2003) and (Gildea, 2003). In (Cherry and Lin, 2003) a probability model $Pr(a_1^J|f_1^J, e_1^I)$ is used, which is symmetric per definition. 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. In

[2] The base forms were determined using LingSoft tools.

Table 5: Effect of different lexicon symmetrization methods on alignment performance for the generalized alignments for the Verbmobil task and the Canadian Hansards task.

| task: | Verbmobil | | | Canadian Hansards | | |
|---|---|---|---|---|---|---|
| | Precision[%] | Recall[%] | AER[%] | Precision[%] | Recall[%] | AER[%] |
| Base | 93.3 | 96.0 | 5.5 | 96.6 | 86.0 | 8.2 |
| Lin. | 96.1 | 94.0 | 4.9 | 95.2 | 88.5 | 7.7 |
| Loglin. | 95.2 | 95.3 | 4.7 | 93.6 | 90.8 | 7.5 |

(Toutanova et al., 2002), extensions to the HMM-based alignment model are presented.

## 7 Conclusions

We have addressed the task of automatically generating word alignments for bilingual corpora. This problem is of great importance for many tasks in natural language processing, especially in the field of machine translation.

We have presented lexicon symmetrization methods for statistical alignment models that are trained using the EM algorithm, in particular the five IBM models, the HMM and Model 6. We have evaluated these methods on the Verbmobil task and the Canadian Hansards task and compared our results to the state-of-the-art system of (Och and Ney, 2003). We have shown that both the linear and the loglinear interpolation of lexicon counts after each iteration of the EM algorithm result in statistically significant improvements of the alignment quality. For the Canadian Hansards task, the AER improved by about 30% relative; for the Verbmobil task the improvement was about 25% relative.

Additionally, we have described lexicon smoothing using the word base forms. Especially for highly inflected languages such as German, this smoothing resulted in statistically significant improvements.

In the future, we plan to optimize the interpolation weights to balance the two translation directions. We will also investigate the possibility of generating directly an unconstrained alignment based on the symmetrized lexicon probabilities.

## Acknowledgment

This work has been partially funded by the EU project LC-Star, IST-2001-32216.

## References

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.

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

H. Ney, S. Martin, and F. Wessel. 1997. Statistical language modeling using leaving-one-out. In S. Young and G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech Processing*, pages 174–207. Kluwer.

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.

K. Toutanova, H. T. Ilhan, and C. D. Manning. 2002. Extensions to hmm-based statistical word alignment models. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, pages 87–94, Philadelphia, PA, July.

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.