Statistical Machine Translation of Spontaneous Speech with Scarce Resources

Evgeny Matusov, Maja Popović, Richard Zens,
Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University, Aachen, Germany.
{matusov,popovic,zens,ney}@cs.rwth-aachen.de

Abstract

This paper deals with the task of statistical machine translation of spontaneous speech using a limited amount of training data. We propose a method for selecting relevant additional training data from other sources that may come from other domains. We present two ways to solve the data sparseness problem by including morphological information into the EM training of word alignments. We show that the use of part-of-speech information for harmonizing word order between source and target sentences yields significant improvements in the BLEU score.

1. Introduction

When developing a system to automatically translate spontaneous speech, we regard the following aspects as important:

- Usually, only a limited corpus of bilingual sentence pairs is available for training.
- Rule-based transfer machine translation methods are hardly applicable, since the utterances are often spontaneous and colloquial and may not represent well-formed sentences. Furthermore, in case of automatically recognized speech, the input sentence may contain recognition errors which may completely destroy the sentence structure.
- The training data sparsely covers only a limited vocabulary and a very limited number of possible cases of non-monotonous translations.

In this paper, we present some methods for mitigating these problems. We follow a statistical approach to machine translation in which we estimate translation model parameters from a training corpus of bilingual sentence pairs. In section 2 we will briefly describe the source-channel approach to the statistical word alignment model and the alignment template system for machine translation.

There are different aspects of the data sparseness problem. First, it may be that it is difficult to obtain enough bilingual sentence-aligned training data for a specific language; this is not within the scope of this work. Another problem is obtaining additional bilingual data for specific domain or genre like medical texts or travel conversations. In Section 3 we will describe a method for extending the training corpus with relevant bilingual data from larger (more general) corpora.

The next problem we may face is the limited coverage of the vocabulary, when many words appear only once in the training corpus. This is especially true for highly inflected languages. Section 4 will present a possibility to use morphological information like word base forms to improve automatic word alignments for such languages. Some research in this direction has been performed in [1]; they proposed hierarchical lexicon models containing base forms and part-of-speech tags for the translation from German into English. In our work, we will use such lexicon models directly in the alignment training, whereas [1] created the models from the final (Viterbi) alignment obtained after the standard training procedure.

Finally, when only a small bilingual corpus is available for training, not enough examples of word or phrase reordering are learned, and non-monotonous translations can not be produced or have a very poor quality. In Section 5 we will propose a method for re-ordering the source sentences in training and in testing using part-of-speech (POS) tags. We will try to reduce the differences in word order between the source and the corresponding target sentence. This strategy monotonizes the translation process. As a result, good estimates of model parameters become possible even with scarce training data.

We will present experimental results in which we apply all of these methods to two machine translation tasks. These are the Nespole! [2] German-English corpus and the German-English Verbmobil corpus. We achieve substantial improvements with some of the presented techniques.

2. Statistical Machine Translation

2.1. Word Alignment

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we choose the

sentence with the highest probability:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I|f_1^J) \right\} \quad (1)$$

$$= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J|e_1^I) \right\} \quad (2)$$

The decomposition into two knowledge sources in Equation 2 is known as the source-channel approach to statistical machine translation [3]. It allows an independent modeling of the target language model $Pr(e_1^I)$ and translation model $Pr(f_1^J|e_1^I)$. The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence.

The word alignment A is introduced into the translation model as a hidden variable:

$$Pr(f_1^J|e_1^I) = \sum_A Pr(f_1^J, A|e_1^I) \quad (3)$$

Usually, restricted alignments are used in the sense that each source word is aligned to at most one target word. Thus, an alignment A is a mapping from source sentence positions to target sentence positions $A = a_1...a_j...a_J$, $(a_j \in \{0, \ldots, I\})$. The alignment a_1^J may contain alignments $a_j = 0$ with the 'empty' word e_0 to account for source sentence words that are not aligned to any target word at all. A detailed comparison of the commonly used translation models IBM-1 to IBM-5 [4], as well as the Hidden-Markov alignment model (HMM) [5] can be found in [6]. All these models include parameters $p(f|e)$ for the single-word based lexicon. They differ in the alignment model. All of the model parameters are trained iteratively with the EM-Algorithm.

2.2. Translation: Alignment Template Approach

The argmax operation in Eq. 2 denotes the search problem, i.e. the generation of the output sentence in the target language. We have to maximize over all possible target language sentences.

For the search, we choose an alternative to the classical source-channel approach and model the posterior probability $Pr(e_1^I|f_1^J)$ directly. Using a log-linear model [7], we obtain:

$$Pr(e_1^I|f_1^J) = Z(f_1^J) \cdot \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right)$$

Here, $Z(f_1^J)$ denotes the appropriate normalization constant, h_m are the feature functions and λ_m are the corresponding scaling factors. We thus arrive at the decision rule:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

This approach has the advantage that additional models or feature functions can be easily integrated into the overall system. The model scaling factors λ_1^M are trained according to the maximum entropy principle, e.g. using the Generalized Iterative Scaling (GIS) algorithm. Alternatively, one can train them with respect to the final translation quality measured by some error criterion [8].

We follow the alignment template translation approach of [9], where a phrase tranlation model is used as one of the main features. The key elements of this translation approach are the *alignment templates*. These are pairs of source and target language phrases together with an alignment within the phrases. The phrases are extracted from the automatically estimated word alignments. The alignment templates are build at the level of word classes, which improves their generalization capability.

Besides the alignment template translation model probabilities, we use additional feature functions. These are the word translation model and two language models: a word-based trigram model and a class-based five-gram model. Furthermore, we use two heuristics, namely the word penalty and alignment template penalty feature functions. To model the alignment template reorderings, we use a feature function that penalizes reorderings linear in the jump width.

We use a dynamic programming beam search algorithm to generate the translation hypothesis with maximum probability. This search algorithm allows for arbitrary reorderings at the level of alignment templates. Within the alignment templates, the word order is learned in training and kept fix during the search process.

This is only a brief description of the alignment template approach. For further details, see [9, 7].

3. Acquiring Additional Training Data

When only a small corpus of sentence pairs is available for training of the statistical translation models, it may be reasonable to include additional bilingual training data from other sources. Since this additional data may come from another domain and substantially differ from the original training corpus, a method for selecting relevant sentences is desirable.

In our experiments, we use a relevance measure of n-*gram coverage*. To this end, we compute the set C of n-grams occurring in the source part of the initial training corpus ($n = 1, 2, 3, 4$). Then, for each candidate source sentence in the additional corpus, we compute a score based on the occurrence of the n-grams from C in that sentence. The score is defined as the geometric mean of n-gram precisions and is therefore similar to the BLEU score used in machine translation evaluation [10]. Such score provides a quantitative measure of how "out-of-domain" or "in-domain" the additional training data may be. We add only those sentence pairs to the initial training corpus, for which this score is sufficiently high.

4. Morphological Information for Word Alignments

4.1. Lexicon Smoothing

Existing statistical translation systems usually treat different derivations of the same base form as they were independent of each other. In our approach, the dependencies between such derivations are taken into account during the EM training of the statistical alignment models.

Typically, the statistical lexicon model $p(f|e)$ is based only on the full forms of the words. For highly inflected languages like German this might cause problems because the coverage of the lexicon might be low. In particular, the coverage problem arises in alignment trainings with small amount of data.

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 well known from language modeling [11]. It is absolute discounting with interpolation:

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

Here, $\overline{e}$ 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 $\beta(f, \overline{e})$ is the normalized backing-off distribution.

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

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

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 $\beta(f|\overline{e})$ is estimated using relative frequencies:

$$\beta(f|\overline{e}) = \frac{N(f,\overline{e})}{\sum_{f'} N(f',\overline{e})} \quad (4)$$

Here, $N(f, \overline{e})$ denotes the count of the event that the source language word f and the target language base form $\overline{e}$ 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 $\overline{e}$.

4.2. Hierarchical Lexicon Counts

Another way to exploit morphological information for creating automatic word alignment is to make use of the hierarchical representation of the statistical lexicon model as proposed in [1]. A constraint grammar parser GERCG (http://www.lingsoft.fi) for lexical analysis and morphological and syntactic disambiguation for German language is used to obtain morpho-syntactic information. The performance of the tool is quite robust even when parsing spontaneous utterances. For each German word, this tool provides its base form and the sequence of morpho-syntactic tags, and this information is then added to the original corpus. For example, the German word "gehe" (go), a verb in the indicative mood and present tense which is derived from the base form "gehen" is annotated as "gehe#gehen-V-IND-PRES#gehen". Conventional statistical translation models cannot handle the fact that for example the German words "gehe" and "gehst" are derivatives of the same base form "gehen" and both can be translated into the same English word "go", whereas the hierarchical representation makes it possible to take such interdependencies into account.

The EM training with hierarchical lexicon counts is performed as follows:
In the E-step the following types of counts are collected:

- full form counts:

$$N(f,e) = \sum_s \sum_A p(A|f_{1s}^{J_s}, e_{1s}^{I_s}) \cdot \sum_{i,j} \delta(f, f_{js}) \delta(e, e_{is})$$

  where f is the full form of the word, e.g. "gehe";

- base form+tag counts:

$$N(\tilde{f},e) = \sum_s \sum_A p(A|f_{1s}^{J_s}, e_{1s}^{I_s}) \cdot \sum_{i,j} \delta(\tilde{f}, \tilde{f}_{js}) \delta(e, e_{is})$$

  where $\tilde{f}$ represents the base form of the word f with sequence of corresponding tags, e.g. "gehen-V-IND-PRES";

- base form counts:

$$N(\overline{f},e) = \sum_s \sum_A p(A|f_{1s}^{J_s}, e_{1s}^{I_s}) \cdot \sum_{i,j} \delta(\overline{f}, \overline{f}_{js}) \delta(e, e_{is})$$

  where $\overline{f}$ is the base form of the word f, e.g. "gehen".

For each full form, refined hierarchical counts are obtained in the following way:

$$N_{hier}(f,e) \quad = \quad N(f,e) + N(\tilde{f},e) + N(\overline{f},e)$$

The M-step is then performed using hierarchical counts:

$$p(f|e) \quad = \quad \frac{N_{hier}(f,e)}{\sum\limits_{f'} N_{hier}(f',e)}$$

5. Part-of-Speech Information for Source Sentence Reordering

The training data sparseness may not allow for reliable word alignment training, especially for language pairs with significantly different word order. In such cases, it is always of benefit to have monotonous alignments. This is not always possible due to word order differences. These differences can be reduced through initial re-ordering of the source training sentences.

We propose to perform such re-ordering based on the POS information for the source words and information about the typical sentence structure of the target language. Ideally, a syntactical parse tree is more useful for this purpose. However, when dealing with spontaneous speech, the standard parse algorithms trained on well-formed sentences perform very poorly, and there are not enough annotated data to train a statistical parser for spontaneous utterances. Keeping this in mind, we decided to use a statistical POS tagger developed by [12] to annotate the source (German) sentences. POS information exhibits less context dependency than syntactic information (subject, predicate, object tags, etc.) and thus the POS tags can be relied upon even in case of spontaneous speech.

The goal of source sentence re-ordering was not to exactly match the structure of the target sentence, but rather to reduce the distance between words which are translations of each other so that the following can be achieved:

- A more robust word alignment training, which results in extraction of more and better bilingual phrases for the Alignment Template system (for instance, non-contiguous phrases not extracted in regular training due to model limitations become contiguous after the re-ordering and can be effectively learned).
- If the input test corpus is re-ordered, translations of higher quality and especially better fluency can be produced. This was only partially possible in the usual hypotheses generation because of the re-ordering constraints. Such constraints allow e. g. only "jumps" of maximum 4 words during search (IBM-style constraints [4]).

The re-ordering is done with context-dependent rules which are specific to the involved language pair and are based on the information about typical syntactic structures. For German-English translation, for instance, we derived the following rules[1]:

1. put verb prefixes standing on the end of a (sub)sentence directly after the first preceding verb:

```
  Ich fahre um 9 Uhr vom Bahnhof ab
* Ich fahre ab um 9 Uhr vom Bahnhof
```

2. put compound verb parts (infinitives, participles) standing at the end of an affirmative (sub)sentence directly after the first (auxiliary) verb in the (sub)sentence:

```
  Ich kann Ihnen noch heute meine
  Nummer geben
* Ich kann geben Ihnen noch heute
* meine Nummer
  Sie haben mir gestern nicht die
  richtige Abfahrtzeit genannt
* Sie haben genannt gestern nicht
* die richtige Abfahrtzeit
```

3. put the verb(s) standing at the end of a (sub)sentence (subordinate clause) directly after the first noun/pronoun in the (sub)sentence:

```
  ... weil ich erst dann Ihnen
  meine Nummer geben kann
* ... weil ich kann geben erst
* dann Ihnen meine Nummer
```

4. change word order in interrogative sentences.

To avoid many erroneous reorderings, we applied these rules to subsentences, which were defined as words between commas (if comma is not between two nouns) and some subordinating conjunctions.

More thorough linguistic considerations would have probably produced better rules, but even with these heuristics we achieved a significant improvement in the translation quality (see Section 6). The fluency of the resulting translation was improved dramatically since the source sentence now approximately had the word order of the target sentence.

The derived rules are especially effectively applicable to corpora with relatively short sentences, where usually there is only one verb group and a few nouns/pronouns, one of which is the subject. More often than not this seems to be true for spontaneous utterances.

[1] * in the examples marks the re-ordered sentences.

Table 1: *Corpus statistics of the Verbmobil task.*

		German	English
Train	Sentences	34K	
	Words	329 625	343 076
	Vocabulary	5 936	3 505
	Singletons	2 600	1 305
Dictionary	Entries	4 404	
Alignment	Sentences	354	
test corpus	Words	3 233	3 109

6. Experimental Results

6.1. Improving Word Alignment

The presented methods of dealing with training data sparsity by using the word morphology were expected to improve the automatically trained word alignment. We evaluated the word alignment quality on the Verbmobil task. The German–English Verbmobil task [13] is a speech translation task in the domain of appointment scheduling, travel planning and hotel reservation. The corpus statistics are shown in Table 1. The number of running words and the vocabularies are based on full-form words including punctuation marks. As in [6], the first 100 sentences of the alignment test corpus are used as a development corpus to optimize model parameters that are not trained via the EM algorithm, e.g. the smoothing parameters.

We use the same evaluation criterion as described in [14]. The generated word alignment is compared to a reference alignment which is produced by human experts. The obtained reference alignment may contain many-to-one and one-to-many relationships and includes sure (S) and possible (P) alignment points. The quality of an alignment A is computed as appropriately redefined precision and recall measures. We also 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).

In word alignment training we use the same training scheme (model sequence) as presented in [6]: $1^5 H^5 3^3 4^3 6^5$, i.e. 5 iteration of IBM-1, 5 iterations of the HMM, 3 iteration of IBM-3, etc.. We include a conventional dictionary (possibly containing phrases) as additional training material. Since we use the same training and testing conditions as [6], we will refer to the results presented in that article as the baseline results.

In Table 2, we present the results for the lexicon smoothing as described in Section 4.1 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. 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.

Table 2: *Effect of smoothing the lexicon probabilities on the AER[%] for the Verbmobil task (34k sentences; 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
34k 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
8k Base	92.5	95.4	6.2	88.7	88.3	11.5
smooth	93.2	94.9	6.0	89.9	87.8	11.1

We also tested lexicon smoothing using only a part of the Verbmobil training corpus of only 8000 sentence pairs. The reduction in the AER was also achieved, however a minor one. With the reduction of the corpus size, the number of different words which have the same baseform is quite limited so that the resulting smoothing distribution is similar to (and is estimated as poorly as) the original full form distribution. In case of an extremly small training corpus substraction of probability mass through lexicon smoothing may even be harmful to parameter estimation. This is not a problem, however, for the hierarchical lexicon estimation, where the count collection is additive over all morphological hierarchy levels.

Table 3 shows the alignment error rate (AER) when training with a very small Verbmobil corpus of only 500 sentences, as well as with the full Verbmobil corpus. The hierarchical lexicon estimation is used as described in Section 4.2. Results are presented for the Viterbi alignments from both translation directions (German→English and English→German) as well as for combination of those two alignments. The alignment combination is performed using the refined heuristic described in [6]. The results show a consistent decrease in AER for all training schemes, especially for the small training corpus. It can be also seen that greater improvements are yielded for the simpler models.

6.2. Improving Translation Quality

In this section we present the translation experiments performed in the framework of the PF-STAR project using the German-English Nespole! [2] corpus of manually transcribed telephone inquiries concerning travel information and hotel reservations, as well as the experiments on the Verbmobil task.

Table 4 summarizes the statistics of the available Ne-

[2]The base forms were determined using the GERCG constraint grammar parser.

corpus size = 0.5k				
Training	Model	$G \rightarrow E$	$E \rightarrow G$	combined
1^4H^5	hmm	18.8	24.0	16.9
	+hier	16.9	21.5	14.8
$1^4H^53^34^3$	ibm4	16.9	21.5	16.2
	+hier	15.8	20.7	14.9
$1^4H^53^34^36^5$	ibm6	16.7	21.1	15.9
	+hier	15.6	20.9	14.8

corpus size = 34k				
Training	Model	$G \rightarrow E$	$E \rightarrow G$	combined
1^4H^5	hmm	8.9	14.9	7.9
	+hier	8.4	13.7	7.3
$1^4H^53^34^3$	ibm4	6.3	10.9	6.0
	+hier	6.1	10.8	5.7
$1^4H^53^34^36^5$	ibm6	5.7	9.9	5.5
	+hier	5.5	9.7	5.0

Table 3: *AER [%] for Verbmobil corpus for the baseline system (name of the model) and the system using hierarchical method (+hier).*

Table 4: *Nespole! German-English PF-STAR training corpus statistics.*

	German	English
Sentence pairs	3046	
Running words	14437	14743
Vocabulary	1452	1118
Singletons	734	472
Extension through n-gram coverage		
Sentence pairs	15835	
Running words	201907	207515
Vocabulary	17361	12367
Singletons	10423	4583

Table 5: *Nespole! German PF-STAR test corpus statistics.*

	Development	Test
Sentence pairs	300	106
Running words	1437	933
OOV-Rate	0.84 %	0.96 %

spole! training corpus. Originally, the training corpus contained only 3046 sentence pairs. Using the technique described in Section 3, we extended this training corpus by about 12000 sentence pairs of the in-domain training data from the Verbmobil and Zeres [15] bilingual corpora. The domain of both of these bilingual corpora is traveling; however, Verbmobil corpus mostly consists of manually transcribed business meetings arrangements, whereas Zeres corpus is a collection of leaflets describing some hotels and thus does not represent spontaneous speech. We have produced translations with the lowest error rates by using all of the sentences from Verbmobil and Zeres corpora which have at least one unigram in common with the original Nespole! corpus.

For estimation of the model scaling factors and testing we used the development and test corpora, respectively (see Table 5; the percentage of the words not seen in training, or *out-of-vocabulary (OOV)* words, is also given in this table).

In a preprocessing step, we performed splitting of compound german words in training and in test corpora, using a slightly modified algorithm of [16]. This resulted in a reduction in the number of German singletons. We also added a conventional dictionary to the training corpus. It consisted of the conventional dictionary for the Verbmobil task, as well as verbatim translations of named entities (Italian city and hotel names mentioned in the Nespole! corpus).

The model training was performed as described in Section 2 and included model scaling factor optimization on an N-best list of the development corpus with maximum entropy. The objective error rates (as computed with respect to one reference translation) are presented in Table 6. As baseline we consider the configuration in which we use only the original training corpus of 3046 sentences.

The acquisition of additional relevant training data helped to reduce the error rates quite significantly. In contrast, this had *not* been the case when we had used the whole Verbmobil and Zeres training corpora together with the Nespole! corpus to train our system, since this large corpus had contained a significant amount of out-of-domain text.

In the next experiment, we reordered the source (German) sentences both in training and testing corpora using POS information as described in Section 5. The application of the re-ordering rules resulted in modifications in 19.7 % of German sentences in the extended training corpus, 8 % of sentences in the development corpus, and 13.2 % of sentences in the test corpus.

We then re-trained the whole translation system using the modified source sentences. We achieved further reduction of the word error rate and a substantial increase in the BLEU score (Table 6). The position-independent word error rate remained almost unchanged. This indicates that, first and foremost, the fluency of the automatic translations was improved. It is interesting to note that the improvement in fluency was observed not only in translations of re-ordered German test sentences, but also in translations of other sentences. This suggests that the system is capable of generalization with respect to correct word order.

We performed the same experiment on the Verbmobil task. We purposely chose to train our system only on 8000 sentence pairs from the original 34K Verbmobil training corpus in order to intensify the problem of data sparsity. We used standard Verbmobil development and test corpora, the statistics for which is given in Table 7. The application of the POS-based re-ordering rules resulted in modifications in 17.0 % of German sentences in the 8K training corpus, 24.3 % of sentences in the development corpus, and 23.9 %

Table 6: *Translation results German to English for the Nespole! (PF-STAR) task (error rates in %).*

	WER	PER	BLEU
Baseline	60.7	47.4	0.212
+ in-domain corpus	56.1	45.2	0.238
+ sentence reordering (German)	53.7	45.5	0.270

Table 7: *Verbmobil German test corpus statistics (OOV-rate with respect to 8K training corpus).*

	Development	Test
Sentence pairs	276	251
Running words	3159	2628
OOV-Rate	3.3 %	4.0 %

of sentences in the test corpus. When calculating the objective translation error measures for the test translation hypotheses, we used multiple reference translations.

The results of the Verbmobil 8K experiment are given in Table 8. Compared to the baseline, where no re-ordering of German source sentence was performed, we again achieve quite significant improvements in WER and BLEU, i.e. especially in translation fluency.

7. Conclusions

In this paper we addressed the task of statistical machine translation of spontaneous utterances using a limited amount of training data. We described a consistent way of selecting additional in-domain training data from foreign sources. We presented two ways to solve the data sparsity problem by including morphological information into the training of word alignments. On a task of highly spontaneous speech we showed that by using part-of-speech information for initial re-ordering of the input sentences we can achieve significant improvements in the translation fluency. So far this re-ordering is based on heuristic, but effective rules that aim at monotonizing the translation process.

In our future research, we plan to develop a method of integrating the POS-based reordering into the search process. We also plan to perform experiments on the automatically transcribed speech and address the problem of reducing the Out-Of-Vocabulary rate in the test corpus using syntax and word morphology.

8. Acknowledgements

This work has been partially funded by the European Commission under the projects LC-Star, IST-2001-32216, and PF-Star, IST-2001-37599. We would like to thank Marcello Federico (ITC IRST) for providing us with the Nespole! German-English training data.

Table 8: *Verbmobil translation results German to English using a small training corpus of 8K sentence pairs (error rates in %).*

	WER	PER	BLEU
Baseline	56.3	38.2	0.241
+ reordering (German)	52.3	37.9	0.261

9. References

[1] S. Nießen and H. Ney. 2001. Toward hierarchical models for statistical machine translation of inflected languages. In *Data-Driven Machine Translation Workshop*, pages 47–54, Toulouse, France, July.

[2] NEgotiating through SPOken Language in e-commerce. Project homepage, 2000. http://nespole.itc.it/.

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

[4] 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.

[5] 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.

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

[7] F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

[8] F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

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

[10] K. A. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, September.

[11] 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.

[12] D. Sündermann and H. Ney. 2003. Synther – a new m-gram POS tagger. In *Proc. NLP-KE-2003, International Conference on Natural Language Processing and Knowledge Engineering*, pages 628–633, Beijing, China, October.

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

[14] 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.

[15] J. C. Amengual, J. M. Benedí, A. Castaño, A. Marzal, F. Prat, E. Vidal, J. M. Vilar, C. Delogu, A. di Carlo, H. Ney, and S. Vogel. 1996. *Example-Based Understanding and Translation Systems (EuTrans): Final Report, Part I.* Deliverable of ESPRIT project No. 20268, October.

[16] M. Larson, D. Willet, J. Köhler, and G. Rigoll. 2000. Compound Splitting and Lexical Unit Recombination for Improved Performance of a Speech Recognition System for German Parliamentary Speeches. In *Proc. 6th Int. Conf. On Spoken Language Processing (ICSLP)*, pages 945–948, Bejing, China, October.