Integration of Speech to Computer-Assisted Translation Using Finite-State Automata

Shahram Khadivi **Richard Zens** **Hermann Ney**
Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{khadivi,zens,ney}@cs.rwth-aachen.de

Abstract

State-of-the-art computer-assisted translation engines are based on a statistical prediction engine, which interactively provides completions to what a human translator types. The integration of human speech into a computer-assisted system is also a challenging area and is the aim of this paper. So far, only a few methods for integrating statistical machine translation (MT) models with automatic speech recognition (ASR) models have been studied. They were mainly based on N-best rescoring approach. N-best rescoring is not an appropriate search method for building a real-time prediction engine. In this paper, we study the incorporation of MT models and ASR models using finite-state automata. We also propose some transducers based on MT models for rescoring the ASR word graphs.

1 Introduction

A desired feature of computer-assisted translation (CAT) systems is the integration of the human speech into the system, as skilled human translators are faster at dictating than typing the translations (Brown et al., 1994). Additionally, incorporation of a statistical prediction engine, i.e. a statistical interactive machine translation system, to the CAT system is another useful feature. A statistical prediction engine provides the completions to what a human translator types (Foster et al., 1997; Och et al., 2003). Then, one possible procedure for skilled human translators is to provide the oral translation of a given source text and then to post-edit the recognized text. In the post-editing step, a prediction engine helps to decrease the amount of human interaction (Och et al., 2003).

In a CAT system with integrated speech, two sources of information are available to recognize the speech input: the target language speech and the given source language text. The target language speech is a human-produced translation of the source language text. Statistical machine translation (MT) models are employed to take into account the source text for increasing the accuracy of automatic speech recognition (ASR) models.

Related Work

The idea of incorporating ASR and MT models was independently initiated by two groups: researchers at IBM (Brown et al., 1994), and researchers involved in the TransTalk project (Dymetman et al., 1994; Brousseau et al., 1995). In (Brown et al., 1994), the authors proposed a method to integrate the IBM translation model 2 (Brown et al., 1993) with an ASR system. The main idea was to design a language model (LM) to combine the trigram language model probability with the translation probability for each target word. They reported a perplexity reduction, but no recognition results. In the TransTalk project, the authors improved the ASR performance by rescoring the ASR N-best lists with a translation model. They also introduced the idea of a dynamic vocabulary for a speech recognition system where translation models were generated for each source language sentence. The better performing of the two is the N-best rescoring.

Recently, (Khadivi et al., 2005) and (Paulik et al., 2005a; Paulik et al., 2005b) have studied the integration of ASR and MT models. The first work showed a detailed analysis of the effect of different MT models on rescoring the ASR N-best lists. The other two works considered two parallel N-best lists, generated by MT and ASR systems,

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 467–474,
Sydney, July 2006.

respectively. They showed improvement in the ASR N-best rescoring when some proposed features are extracted from the MT N-best list. The main concept among all features was to generate different kinds of language models from the MT N-best list.

All of the above methods are based on an N-best rescoring approach. In this paper, we study different methods for integrating MT models to ASR word graphs instead of N-best list. We consider ASR word graphs as finite-state automata (FSA), then the integration of MT models to ASR word graphs can benefit from FSA algorithms. The ASR word graphs are a compact representation of possible recognition hypotheses. Thus, the integration of MT models to ASR word graphs can be considered as an N-best rescoring but with very large value for N. Another advantage of working with ASR word graphs is the capability to pass on the word graphs for further processing. For instance, the resulting word graph can be used in the prediction engine of a CAT system (Och et al., 2003).

The remaining part is structured as follows: in Section 2, a general model for an automatic text dictation system in the computer-assisted translation framework will be described. In Section 3, the details of the machine translation system and the speech recognition system along with the language model will be explained. In Section 4, different methods for integrating MT models into ASR models will be described, and also the experimental results will be shown in the same section.

2 Speech-Enabled CAT Models

In a speech-enabled computer-assisted translation system, 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 $e_1^I = e_1 \dots e_i \dots e_I$, and an acoustic signal $x_1^T = x_1 \dots x_t \dots x_T$, which is the spoken target language sentence. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\text{argmax}} \{Pr(e_1^I|f_1^J, x_1^T)\} \quad (1)$$

$$\cong \underset{I,e_1^I}{\text{argmax}} \{Pr(e_1^I) Pr(f_1^J|e_1^I) Pr(x_1^T|e_1^I)\} \quad (2)$$

Eq. 1 is decomposed into Eq. 2 by assuming conditional independency between x_1^T and f_1^J. The decomposition into three knowledge sources allows for an independent modeling of the target language model $Pr(e_1^I)$, the translation model $Pr(f_1^J|e_1^I)$ and the acoustic model $Pr(x_1^T|e_1^I)$.

Another approach for modeling the posterior probability $Pr(e_1^I|f_1^J, x_1^T)$ is direct modeling using a log-linear model. The decision rule is given by:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\text{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, x_1^T) \right\} \quad (3)$$

Each of the terms $h_m(e_1^I, f_1^J, x_1^T)$ denotes one of the various models which are involved in the recognition procedure. Each individual model is weighted by its scaling factor λ_m. As there is no direct dependence between f_1^J and x_1^T, the $h_m(e_1^I, f_1^J, x_1^T)$ is in one of these two forms: $h_m(e_1^I, x_1^T)$ and $h_m(e_1^I, f_1^J)$. Due to the argmax operator which denotes the search, no renormalization is considered in Eq. 3. This approach has been suggested by (Papineni et al., 1997; Papineni et al., 1998) for a natural language understanding task, by (Beyerlein, 1998) for an ASR task, and by (Och and Ney, 2002) for an MT task. This approach is a generalization of Eq. 2. The direct modeling has the advantage that additional models can be easily integrated into the overall system. The model scaling factors λ_1^M are trained on a development corpus according to the final recognition quality measured by the word error rate (WER)(Och, 2003).

Search

The search in the MT and the ASR systems is already very complex, therefore a fully integrated search to combine ASR and MT models will considerably increase the complexity. To reduce the complexity of the search, we perform two independent searches with the MT and the ASR systems, the search result of each system will be represented as a large word graph. We consider MT and ASR word graphs as FSA. Then, we are able to use FSA algorithms to integrate MT and ASR word graphs. The FSA implementation of the search allows us to use standard optimized algorithms, e.g. available from an open source toolkit (Kanthak and Ney, 2004).

The recognition process is performed in two steps. First, the baseline ASR system generates a word graph in the FSA format for a given utterance x_1^T. Second, the translation models rescore each word graph based on the corresponding source language sentence. For each utterance, the decision about the best sentence is made according to the recognition and the translation models.

3 Baseline Components

In this section, we briefly describe the basic system components, namely the MT and the ASR systems.

3.1 Machine Translation System

We make use of the RWTH phrase-based statistical machine translation system for the English to German automatic translation. The system includes the following models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: German to English and English to German. Additionally, a word penalty and a phrase penalty are included. The reordering model of the baseline system is distance-based, i.e. it assigns costs based on the distance from the end position of a phrase to the start position of the next phrase. More details about the baseline system can be found in (Zens and Ney, 2004; Zens et al., 2005).

3.2 Automatic Speech Recognition System

The acoustic model of the ASR system is trained on the VerbMobil II corpus (Sixtus et al., 2000). The corpus consists of German large-vocabulary conversational speech: 36k training sentences (61.5h) from 857 speakers. The test corpus is created from the German part of the bilingual English-German XEROX corpus (Khadivi et al., 2005): 1562 sentences including 18k running words (2.6h) from 10 speakers. The test corpus contains 114 out-of-vocabulary (OOV) words. The remaining part of the XEROX corpus is used to train a back off trigram language model using the SRI language modeling toolkit (Stolcke, 2002). The LM perplexity of the speech recognition test corpus is about 83. The acoustic model of the ASR system can be characterized as follows:

- recognition vocabulary of 16716 words;
- 3-state-HMM topology with skip;
- 2500 decision tree based generalized within-word triphone states including noise plus one state for silence;
- 237k gender independent Gaussian densities with global pooled diagonal covariance;
- 16 MFCC features;
- 33 acoustic features after applying LDA;
- LDA is fed with 11 subsequent MFCC vectors;
- maximum likelihood training using Viterbi approximation.

Table 1: Statistics of the machine translation corpus.

		English	German
Train:	Sentences	47 619	
	Running Words	528 779	467 633
	Vocabulary	9 816	16 716
	Singletons	2 302	6 064
Dev:	Sentences	700	
	Running Words	8 823	8 050
	Unknown words	56	108
Eval:	Sentences	862	
	Running Words	11 019	10 094
	Unknown words	58	100

The test corpus recognition word error rate is 20.4%. Compared to the previous system (Khadivi et al., 2005), which has a WER of 21.2%, we obtain a 3.8% relative improvement in WER. This improvement is due to a better and complete optimization of the overall ASR system.

4 Integration Approaches

In this section, we will introduce several approaches to integrate the MT models with the ASR models. To present the content of this section in a more reader-friendly way, we will first explain the task and corpus statistics, then we will present the results of N-best rescoring. Afterwards, we will describe the new methods for integrating the MT models with the ASR models. In each sub-section, we will also present the recognition results.

4.1 Task

The translation models are trained on the part of the English-German XEROX corpus which was not used in the speech recognition test corpus. We divide the speech recognition test corpus into two parts, the first 700 utterances as the development corpus and the rest as the evaluation corpus. The development corpus is used to optimize the scaling factors of different models (explained in Section 2). The statistics of the corpus are depicted in Table 1. The German part of the training corpus is also used to train the language model.

4.2 N-best Rescoring

To rescore the N-best lists, we use the method of (Khadivi et al., 2005). But the results shown here are different from that work due to a better optimization of the overall ASR system, using a

Table 2: Recognition WER [%] using N-best rescoring method.

Models		Dev	Eval
MT		47.1	50.5
ASR		19.3	21.3
ASR+MT	IBM-1	17.8	19.0
	HMM	18.2	19.2
	IBM-3	17.1	18.4
	IBM-4	17.1	18.3
	IBM-5	16.6	18.2
	Phrase -based	18.8	20.3

better MT system, and generating a larger N-best list from the ASR word graphs. We rescore the ASR N-best lists with the standard HMM (Vogel et al., 1996) and IBM (Brown et al., 1993) MT models. The development and evaluation sets N-best lists sizes are sufficiently large to achieve almost the best possible results, on average 1738 hypotheses per each source sentence are extracted from the ASR word graphs.

The recognition results are summarized in Table 2. In this table, the translation results of the MT system are shown first, which are obtained using the phrase-based approach. Then the recognition results of the ASR system are shown. Afterwards, the results of combined speech recognition and translation models are presented.

For each translation model, the N-best lists are rescored based on the translation probability $p(e_1^I|f_1^J)$ of that model and the probabilities of speech recognition and language models. In the last row of Table 2, the N-best lists are rescored based on the full machine translation system explained in Section 3.1.

The best possible hypothesis achievable from the N-best list has the WER (oracle WER) of 11.2% and 12.4% for development and test sets, respectively.

4.3 Direct Integration

At the first glance, an obvious method to combine the ASR and MT systems is the integration at the level of word graphs. This means the ASR system generates a large word graph for the input target language speech, and the MT system also generates a large word graph for the source language text. Both MT and ASR word graphs are in the target language. These two word graphs can be considered as two FSA, then using FSA theory, we can integrate two word graphs by applying the composition algorithm.

We conducted a set of experiments to integrate the ASR and MT systems using this method. We obtain a WER of 19.0% and 20.9% for development and evaluation sets, respectively. The results are comparable to N-best rescoring results for the phrase-based model which is presented in Table 2. The achieved improvements over the ASR baseline are statistically significant at the 99% level (Bisani and Ney, 2004). However, the results are not promising compared to the results of the rescoring method presented in Table 2 for HMM and IBM translation models. A detailed analysis revealed that only 31.8% and 26.7% of sentences in the development and evaluation sets have identical paths in both FSA, respectively. In other words, the search algorithm was not able to find any identical paths in two given FSA for the remaining sentences. Thus, the two FSA are very different from each other. One explanation for the failure of this method is the large difference between the WERs of two systems, as shown in Table 2 the WER for the MT system is more than twice as high as for the ASR system.

4.4 Integrated Search

In Section 4.3, two separate word graphs are generated using the MT and the ASR systems. Another explanation for the failure of the direct integration method is the independent search to generate the word graphs. The search in the MT and the ASR systems is already very complex, therefore a full integrated search to combine ASR and MT models will considerably increase the complexity.

However, it is possible to reduce this problem by integrating the ASR word graphs into the generation process of the MT word graphs. This means, the ASR word graph is used in addition to the usual language model. This kind of integration forces the MT system to generate identical paths to those in the ASR word graph. Using this approach, the number of identical paths in MT and ASR word graphs are increased to 39.7% and 34.4% of the sentences in development and evaluation sets, respectively. The WER of the integrated system are 19.0% and 20.7% for development and evaluation sets.

4.5 Lexicon-Based Transducer

The idea of a dynamic vocabulary, restricting and weighting the word lexicon of the ASR was first

introduced in (Brousseau et al., 1995). The idea was also seen later in (Paulik et al., 2005b), they extract the words of the MT N-best list to restrict the vocabulary of the ASR system. But they both reported a negative effect from this method on the recognition accuracy. Here, we extend the dynamic vocabulary idea by weighting the ASR vocabulary based on the source language text and the translation models. We use the lexicon model of the HMM and the IBM MT models. Based on these lexicon models, we assign to each possible target word e the probability $Pr(e|f_1^J)$. One way to compute this probability is inspired by IBM Model 1:

$$Pr(e|f_1^J) = \frac{1}{J+1}\sum_{j=0}^{J} p(e|f_j)$$

We can design a simple transducer (or more precisely an acceptor) using probability in Eq. 4 to efficiently rescore all paths (hypotheses) in the word graph with IBM Model 1:

$$\begin{aligned} P_{\text{IBM-1}}(e_1^I|f_1^J) &= \frac{1}{(J+1)^I}\prod_{i=1}^{I}\sum_{j=0}^{J} p(e_i|f_j) \\ &= \prod_{i=1}^{I}\frac{1}{(J+1)}\cdot p(e_i|f_1^J) \end{aligned}$$

The transducer is formed by one node and a number of self loops for each target language word. In each arc of this transducer, the input label is target word e and the weight is $-\log\frac{1}{J+1}\cdot p(e|f_1^J)$.

We conducted experiments using the proposed transducer. We built different transducers with the lexicons of HMM and IBM translation models. In Table 3, the recognition results of the rescored word graphs are shown. The results are very promising compared to the N-best list rescoring, especially as the designed transducer is very simple. Similar to the results for the N-best rescoring approach, these experiments also show the benefit of using HMM and IBM Models to rescore the ASR word graphs.

Due to its simplicity, this model can be easily integrated into the ASR search. It is a sentence specific unigram LM.

4.6 Phrase-Based Transducer

The phrase-based translation model is the main component of our translation system. The pairs of source and corresponding target phrases are extracted from the word-aligned bilingual training corpus (Zens and Ney, 2004). In this section, we design a transducer to rescore the ASR word graph using the phrase-based model of the MT system. For each source language sentence, we extract all possible phrases from the word-aligned training corpus. Using the target part of these phrases we build a transducer similar to the lexicon-based transducer. But instead of a target word on each arc, we have the target part of a phrase. The weight of each arc is the negative logarithm of the phrase translation probability.

Table 3: Recognition WER [%] using lexicon-based transducer to rescore ASR word graphs.

Models		Dev	Eval
ASR		19.3	21.3
ASR+MT	IBM-1	17.5	19.0
	HMM	17.8	19.2
	IBM-3	17.7	18.8
	IBM-4	17.8	18.8
	IBM-5	17.6	18.9

This transducer is a good approximation of non-monotone phrase-based-lexicon score. Using the designed transducer it is possible that some parts of the source texts are not covered or covered more than once. Then, this model can be compared to the IBM-3 and IBM-4 models, as they also have the same characteristic in covering the source words. The above assumption is not critical for rescoring the ASR word graphs, as we are confident that the word order is correct in the ASR output. In addition, we assume low probability for the existence of phrase pairs that have the same target phrase but different source phrases within a particular source language sentence.

Using the phrase-based transducer to rescore the ASR word graph results in WER of 18.8% and 20.2% for development and evaluation sets, respectively. The improvements are statistically significant at the 99% level compared to the ASR system. The results are very similar to the results obtained using N-best rescoring method. But the transducer implementation is much simpler because it does not consider the word-based lexicon, the word penalty, the phrase penalty, and the reordering models, it just makes use of phrase translation model. The designed transducer is much faster in rescoring the word graph than the MT system in rescoring the N-best list. The average speed to rescore the ASR word graphs with this transducer is 49.4 words/sec (source language

text words), while the average speed to translate the source language text using the MT system is 8.3 words/sec. The average speed for rescoring the N-best list is even slower and it depends on the size of N-best list.

A surprising result of the experiments as has also been observed in (Khadivi et al., 2005), is that the phrase-based model, which performs the best in MT, has the least contribution in improving the recognition results. The phrase-based model uses more context in the source language to generate better translations by means of better word selection and better word order. In a CAT system, the ASR system has much better recognition quality than MT system, and the word order of the ASR output is correct. On the other hand, the ASR recognition errors are usually single word errors and they are independent from the context. Therefore, the task of the MT models in a CAT system is to enhance the confidence of the recognized words based on the source language text, and it seems that the single word based MT models are more suitable than phrase-based model in this task.

4.7 Fertility-Based Transducer

In (Brown et al., 1993), three alignment models are described that include fertility models, these are IBM Models 3, 4, and 5. The fertility-based alignment models have a more complicated structure than the simple IBM Model 1. The fertility model estimates the probability distribution for aligning multiple source words to a single target word. The fertility model provides the probabilities $p(\phi|e)$ for aligning a target word e to ϕ source words. In this section, we propose a method for rescoring ASR word graphs based on the lexicon and fertility models.

In (Knight and Al-Onaizan, 1998), some transducers are described to build a finite-state based translation system. We use the same transducers for rescoring ASR word graphs. Here, we have three transducers: lexicon, null-emitter, and fertility. The lexicon transducer is formed by one node and a number of self loops for each target language word, similar to IBM Model 1 transducer in Section 4.5. On each arc of the lexicon transducer, there is a lexicon entry: the input label is a target word e, the output label is a source word f, and the weight is $-\log p(f|e)$. The null-emitter transducer, as its name states, emits the null word with a pre-defined probability after each input word. The fertility transducer is also a simple transducer to map zero or several instances of a source word to one instance of the source word.

The ASR word graphs are composed successively with the lexicon, null-emitter, fertility transducers and finally with the source language sentence. In the resulting transducer, the input labels of the best path represent the best hypothesis.

The mathematical description of the proposed method is as follows. We can decompose Eq. 1 using Bayes' decision rule:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\text{argmax}}\{Pr(e_1^I|f_1^J, x_1^T)\} \quad (4)$$

$$\cong \underset{I,e_1^I}{\text{argmax}}\{Pr(f_1^J)Pr(e_1^I|f_1^J)Pr(x_1^T|e_1^I)\} \quad (5)$$

In Eq. 5, the term $Pr(x_1^T|e_1^I)$ is the acoustic model and can be represented with the ASR word graph[1], the term $Pr(e_1^I|f_1^J)$ is the translation model of the target language text to the source language text. The translation model can be represented by lexicon, fertility, and null-emitter transducers. Finally, the term $Pr(f_1^J)$ is a very simple language model, it is the source language sentence.

The source language model in Eq. 5 can be formed into the acceptor form in two different ways:

1. a linear acceptor, i.e. a sequence of nodes with one incoming arc and one outgoing arc, the words of source language text are placed consecutively in the arcs of the acceptor,

2. an acceptor containing possible permutations. To limit the permutations, we used an approach as in (Kanthak et al., 2005).

Each of these two acceptors results in different constraints for the generation of the hypotheses. The first acceptor restricts the system to generate exactly the same source language sentence, while the second acceptor forces the system to generate the hypotheses that are a reordered variant of the source language sentence. The experiments conducted do not show any significant difference in the recognition results among the two source language acceptors, except that the second acceptor is much slower than the first acceptor. Therefore, we use the first model in our experiments. Table 4 shows the results of rescoring the ASR word graphs using the fertility-based transducers.

[1]Actually, the ASR word graph is obtained by using $Pr(x_1^T|e_1^I)$ and $Pr(e_1^I)$ models. However, It does not cause any problem in the modeling, especially when we make use of the direct modeling, Eq. 3

Table 4: Recognition WER [%] using fertility-based transducer to rescore ASR word graphs.

Models		Dev	Eval
ASR		19.3	21.3
ASR+MT	IBM-3	17.4	18.6
	IBM-4	17.4	18.5
	IBM-5	17.6	18.7

As Table 4 shows, we get almost the same or slightly better results when compared to the lexicon-based transducers.

Another interesting point about Eq. 5 is its similarity to speech translation (translation from target spoken language to source language text). Then, we can describe a speech-enabled CAT system as similar to a speech translation system, except that we aim to get the best ASR output (the best path in the ASR word graph) rather than the best translation. This is because the best translation, which is the source language sentence, is already given.

5 Conclusion

We have studied different approaches to integrate MT with ASR models, mainly using finite-state automata. We have proposed three types of transducers to rescore the ASR word graphs: lexicon-based, phrase-based and fertility-based transducers. All improvements of the combined models are statistically significant at the 99% level with respect to the baseline system, i.e. ASR only.

In general, N-best rescoring is a simplification of word graph rescoring. As the size of N-best list is increased, the results obtained by N-best list rescoring approach the results of the word graph rescoring. But we should consider that the statement is correct when we use exactly the same model and the same implementation to rescore the N-best list and word graph. Figure 1 shows the effect of the N-best list size on the recognition WER of the evaluation set. As we expected, the recognition results of N-best rescoring improve as N becomes larger, until the point that the recognition result converges to its optimum value. As shown in Figure 1, we should not expect that word graph rescoring methods outperform the N-best rescoring method, when the size of N-best lists are large enough. In Table 2, the recognition results are calculated using a large enough size for N-best lists, a maximum of 5,000 per sentence, which results in the average of 1738 hypotheses

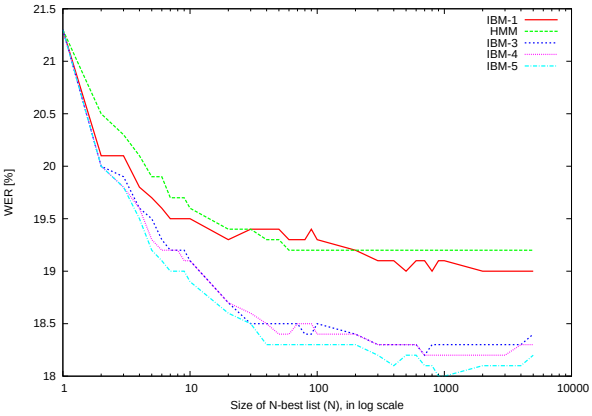

Figure 1: The N-best rescoring results for different N-best sizes on the evaluation set.

per sentence. An advantage of the word graph rescoring is the confidence of achieving the best possible results based on a given rescoring model.

The word graph rescoring methods presented in this paper improve the baseline ASR system with statistical significance. The results are competitive with the best results of N-best rescoring. For the simple models like IBM-1, the transducer-based integration generates similar or better results than N-best rescoring approach. For the more complex translation models, IBM-3 to IBM-5, the N-best rescoring produces better results than the transducer-based approach, especially for IBM-5. The main reason is due to exact estimation of IBM-5 model scores on the N-best list, while the transducer-based implementation of IBM-3 to IBM-5 is not exact and simplified. However, we observe that the fertility-based transducer which can be considered as a simplified version of IBM-3 to IBM-5 models can still obtain good results, especially if we compare the results on the evaluation set.

Acknowledgement

This work has been funded by the European Union under the RTD project TransType2 (IST 2001 32091) and the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org).

References

P. Beyerlein. 1998. Discriminative model combination. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 481 – 484, Seattle, WA, May.

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluationx. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 409–412, Montreal, Canada, May.

J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon. 1995. French speech recognition in an automatic dictation system for translators: the transtalk project. In *Proceedings of Eurospeech*, pages 193–196, Madrid, Spain.

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.

P. F. Brown, S. F. Chen, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer. 1994. Automatic speech recognition in machine-aided translation. *Computer Speech and Language*, 8(3):177–187, July.

M. Dymetman, J. Brousseau, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon. 1994. Towards an automatic dictation system for translators: the TransTalk project. In *Proceedings of ICSLP-94*, pages 193–196, Yokohama, Japan.

G. Foster, P. Isabelle, and P. Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194.

S. Kanthak and H. Ney. 2004. FSA: An efficient and flexible C++ toolkit for finite state automata using on-demand computation. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–517, Barcelona, Spain, July.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, Ann Arbor, Michigan, June.

S. Khadivi, A. Zolnay, and H. Ney. 2005. Automatic text dictation in computer-assisted translation. In *Interspeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, pages 2265–2268, Portugal, Lisbon.

K. Knight and Y. Al-Onaizan. 1998. Translation with finite-state devices. In D. Farwell, L. Gerber, and E. H. Hovy, editors, *AMTA*, volume 1529 of *Lecture Notes in Computer Science*, pages 421–437. Springer Verlag.

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.

F. J. Och, R. Zens, and H. Ney. 2003. Efficient search for interactive statistical machine translation. In *EACL03: 10th Conf. of the Europ. Chapter of the Association for Computational Linguistics*, pages 387–393, Budapest, Hungary, April.

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.

K. A. Papineni, S. Roukos, and R. T. Ward. 1997. Feature-based language understanding. In *EUROSPEECH*, pages 1435–1438, Rhodes, Greece, September.

K. A. Papineni, S. Roukos, and R. T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 189–192, Seattle, WA, May.

M. Paulik, S. Stüker, C. Fügen, , T. Schultz, T. Schaaf, and A. Waibel. 2005a. Speech translation enhanced automatic speech recognition. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 121–126, Puerto Rico, San Juan.

M. Paulik, C. Fügen, S. Stüker, T. Schultz, T. Schaaf, and A. Waibel. 2005b. Document driven machine translation enhanced ASR. In *Interspeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, pages 2261–2264, Portugal, Lisbon.

A. Sixtus, S. Molau, S.Kanthak, R. Schlüter, and H. Ney. 2000. Recent improvements of the RWTH large vocabulary speech recognition system on spontaneous speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1671 – 1674, Istanbul, Turkey, June.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

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.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 257–264, Boston, MA, May.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.