PHRASE-BASED TRANSLATION OF SPEECH RECOGNIZER WORD LATTICES USING LOGLINEAR MODEL COMBINATION

Evgeny Matusov, Hermann Ney, Ralph Schlüter

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
52056 Aachen, Germany
{matusov,ney,schlueter}@informatik.rwth-aachen.de

ABSTRACT

This paper presents a phrase-based speech translation system that combines phrasal lexicon, language, and acoustic model features in a loglinear model. Automatic speech recognition and machine translation are coupled by using large word lattices as the input for translation. For the first time, all features are directly integrated into the decoding process. The feature weights are iteratively optimized for an objective error measure. We prove that acoustic recognition scores of the recognized words in the lattices together with a source language model score positively and significantly affect the translation quality. We show the advantage of using loglinear model combination for a robust optimization of scaling factors. We report consistent improvements compared with translations of single best recognition output on an Italian-to-English translation task. First encouraging results were also obtained on a large vocabulary task of translating European parliamentary speeches.

1. INTRODUCTION

It has been shown in the past that automatic speech recognition (ASR) and machine translation (MT) can be coupled in order to directly translate spoken utterances into another language. Here we present a framework of phrase-based translation of ASR word lattices, in which ASR and MT models can be effectively combined to achieve improvement of translation quality.

Various approaches to speech translation have been proposed and investigated in the past. [8] presents an integrated speech translation system for tasks from the Eutrans project. However, the experimental results were inconsistent as the integrated speech translation performed much worse than the serial approach on real-world data. [4] presented only the theory of integrated speech translation, but lacked experimental results. More recently, [7] concluded that improvements from tighter coupling may only be observed when ASR lattices are sparse, i.e. when there are only few hypothesized words per spoken word in the lattice. This would mean that a fully integrated speech translation would not work at all. [3] presented a joint probability approach to speech translation based on weighted finite-state transducers (WFSTs). They were able to show consistent and significant improvements in translation quality on three different tasks, using lattices of high density with acoustic model scores. The translation system of [3] produced translation hypotheses with a single score; it was stressed that the optimization of a scaling factor for either the translation or the acoustic model score is crucial for good performance. In the approach presented in [11], loglinear models with multiple features were used for speech translation. In contrast to our work, the scaling factors for the feature functions were optimized in a rescoring procedure on a translation word graph that was generated using a single word based translation system. Here, we directly integrate all models, including phrase based and single word based lexica and recognition features in the decoding process. Also, in [11] the authors extract ASR N-best lists of only 100 hypotheses per utterance to approximate the coupling of speech recognition and translation; here we explore a much tighter integration using dense lattices. Finally, [1] used loglinear model combination to translate N-best lists, but was not able to achieve improvements with confusion networks as computed from lattices.

This paper is organized as follows. Based on the presentation of [4], Section 2 reviews the Bayes' decision rule for speech translation. Starting from there, we present our phrase-based statistical translation approach in the framework of loglinear modeling and minimum error training in Section 3. Here, we translate ASR word lattices and benefit from acoustic and source language model scores. Section 4 touches on some of the practical problems that arise when we translate word lattices. In Section 5 we present significant improvements in quality of translation from Italian to English when we use the acoustic and source language model scores together with the translation model features and optimize the model scaling factors. We then describe the initial promising experiments with lattice translation on a large vocabulary English-to-Spanish task.

2. BAYES' DECISION RULE FOR SPEECH TRANSLATION

In speech translation, we are looking for a target language sentence e_1^I which is the translation of a speech utterance represented by acoustic vectors x_1^T. In order to minimize the number of sentence errors, we maximize the posterior probability of the target language translation given the speech signal (see [4]). The source words f_1^J are introduced as a hidden variable:

$$\begin{aligned}
\hat{e}_1^{\hat{I}} &= \operatorname*{argmax}_{I,e_1^I} Pr(e_1^I|x_1^T) \\
&= \operatorname*{argmax}_{I,e_1^I} \{Pr(e_1^I) \cdot Pr(x_1^T|e_1^I)\} \\
&= \operatorname*{argmax}_{I,e_1^I} \{Pr(e_1^I) \cdot \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J, e_1^I)\} \\
&\cong \operatorname*{argmax}_{I,e_1^I} \{Pr(e_1^I) \cdot \max_{f_1^J} \{Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J)\}\}
\end{aligned}$$

Note that we made the natural assumption that the speech signal does not depend on the target sentence and approximated the sum over all possible source language transcriptions by the maximum. $Pr(x_1^T|f_1^J)$ may be a standard acoustic model, and $Pr(e_1^I)$ is the target language model. For the translation model $Pr(f_1^J|e_1^I)$, we introduce an *alignment* between a source and target sentence as the hidden variable $\mathcal{A}$:

$$Pr(f_1^J|e_1^I) = \sum_{\mathcal{A}} Pr(\mathcal{A}|e_1^I) \cdot Pr(f_1^J|e_1^I, \mathcal{A})$$

The hidden alignment $\mathcal{A}$ represents all possible interpretations of source words by target words. It is used to estimate phrase based and single word based lexicon models.

The derived decision rule does not include a source language model probability $Pr(f_1^J)$. To take into account the requirement for the "well-formedness" of the source sentence f_1^J, the translation model has to include context dependency on the previous source words [4]. In a phrase-based translation model, this dependency is present within a phrase. We can approximate this dependency for the whole sentence by including a source language model in the log-linear modeling framework that is presented in the next section.

3. LOGLINEAR MODEL COMBINATION

In practice, we follow a direct translation approach. Here, probability distributions are represented as features in a log-linear model. In particular, the translation model probability is decomposed into several probabilities. We estimate phrase-based lexicon models using statistical word alignments as described in [10]. The probabilities of phrasal translations are supplemented by single word based model probabilities. Following a unified speech translation approach, we also include acoustic model probabilities $Pr(x_1^T|f_1^J)$ of the hypotheses in the ASR word lattices as a feature. As it was mentioned, probabilities of a source language model can also be included.

For a hypothesized recognized source sentence f_1^J and a hypothesized translation e_1^I, let $k \rightarrow (j_k, i_k), k = 1, \ldots, K$ be a *monotone* segmentation of the sentence pair into K bilingual phrases (without empty phrases and overlap). Our phrase-based approach to integrated statistical speech translation is then expressed by the following equation:

$$\begin{aligned}
\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I,e_1^I} \Bigg\{ \prod_{i=1}^{I} \Big[\mathbf{c_1}^{\lambda_1} \cdot \mathbf{lm}^{\lambda_2}(i) \Big] \cdot \max_{J,f_1^J} \Bigg(\prod_{j=1}^{J} \mathbf{asr}(j) \cdot \\
\max_{K,(j_k,i_k)} \prod_{k=1}^{K} \Big[\mathbf{c_2}^{\lambda_3} \cdot \mathbf{phr\_std}^{\lambda_4}(k) \cdot \mathbf{phr\_inv}^{\lambda_5}(k) \cdot \\
\cdot \mathbf{swb\_std}^{\lambda_6}(k) \cdot \mathbf{swb\_inv}^{\lambda_7}(k) \Big] \Bigg) \Bigg\}
\end{aligned}$$

Here, we optimize over alternative recognition word sequences f_1^J, over all possible monotone segmentations of a given recognized sequence into source language phrases, and over all possible translations of these phrases. The following models contribute to the global decision criterion:

- $\mathbf{lm}(i) = p(e_i \,|\, e_{i-m+1}^{i-1})$ is the m-gram target language model.
- For a given phrasal segmentation (j_k, i_k), let $(f_{j_{k-1}+1}^{j_k}, e_{i_{k-1}+1}^{i_k})$, $k = 1, \ldots, K$ be the resulting bilingual pairs, each consisting of a source phrase $(f_{j_{k-1}+1}^{j_k})$ and one of its possible target phrase translations $(e_{i_{k-1}+1}^{i_k})$. Then the phrasal lexicon models in both translation directions are given by:

  $$\begin{aligned}
  \mathbf{phr\_std}(k) &= p(f_{j_{k-1}+1}^{j_k} \,|\, e_{i_{k-1}+1}^{i_k}) \\
  \mathbf{phr\_inv}(k) &= p(e_{i_{k-1}+1}^{i_k} \,|\, f_{j_{k-1}+1}^{j_k})
  \end{aligned}$$

  The phrase translation probabilities are computed as a loglinear interpolation of the relative frequencies and the IBM Model 1 probability.
- For a given phrasal segmentation, we also use the single word based lexicon models:

  $$\begin{aligned}
  \mathbf{swb\_std}(k) &= \prod_{j=j_{k-1}+1}^{j_k} p(f_j \,|\, e_{i_{k-1}+1}^{i_k}) \\
  \mathbf{swb\_inv}(k) &= \prod_{i=i_{k-1}+1}^{i_k} p(e_i \,|\, f_{j_{k-1}+1}^{j_k})
  \end{aligned}$$

The probability $p(f_j \mid e_{i_{k-1}+1}^{i_k})$ is defined as the product of the single word based lexicon probabilities $p(f_j \mid e_i)$ over all words e_i within the target phrase, and $p(e_i \mid f_{j_{k-1}+1}^{j_k})$ is the inverse model of the same type.

- $\mathbf{c_1}$ and $\mathbf{c_2}$ are word and phrase penalty features, respectively.

- the recognition model feature $\mathbf{asr}(j)$ is represented by:

$$\mathbf{asr}(j) = p^{\mu_1}(f_j \mid f_{j-m+1}^{j-1}) \cdot p^{\mu_2}(x_j|f_j)$$

The probability $p(x_j|f_j)$ is the acoustic probability of a word hypothesis f_j in the ASR word lattice which covers the portion x_j of the acoustic vectors [4]. The probability $p(f_j \mid f_{j-m+1}^{j-1})$ is the m-gram source language model probability. Both probabilities can be scaled with an exponent; then, the source language model feature and the acoustic feature will get the scaling factors μ_1 and μ_2, respectively.

The translation model features and the target language model are scaled with a set of exponents $\lambda = \{\lambda_1, \ldots \lambda_7\}$. Their values can be optimized together with the scaling factors for the recognition model features μ_1 and μ_2 in a single loglinear model simultaneously. The scaling factors are optimized in a minimum error training framework [5] iteratively with the Downhill Simplex algorithm, by performing 100 to 200 translations of a development set. The criterion for optimization is an objective machine translation error measure like word error rate or BLEU score. Negative logarithms of the probabilities are used.

4. PRACTICAL ASPECTS OF LATTICE TRANSLATION

4.1. Generation of Word Lattices

The speech recognition systems used here produce word lattices where arcs are labeled with start and end time, the recognized entity (word, noise, hesitation, silence), the negative log probability of acoustic vectors between start and end time given the entity. In a first step we mapped all entities that were not spoken words onto the empty arc label ε. As the time information is not used in our approach, we removed it from the lattices and compressed the structure by applying ε-removal, determinization, and minimization. For all of these operations, we employ the finite-state transducer toolkit of [2] which efficiently implemented them on-demand. This step significantly reduced runtime without changing the results.

4.2. Phrase Extraction

Even if we limit the maximum phrase length (e. g. to 12 words), the number of different phrase pairs which can be extracted from a bilingual training corpus is very large. However, for efficiency of translation, candidate phrase pairs have to be kept in main memory. To overcome this problem, for off-line experiments, only phrase pairs in which the source phrase appears in the input test corpus are extracted. In case of ASR word lattice input, we reduce the memory requirements with the following approach. The lattice for each test utterance is traversed, and only phrases which match (sub)sequences of arcs in the lattice are extracted. Thus, only phrases which can be used in translation will be loaded. In an alternative approach, a phrase pair can be extracted only if each word in the candidate source phrase is present in the lattice vocabulary. This vocabulary is smaller than the ASR system vocabulary because it includes only the different words which actually appear in the lattice. This approach has the advantage that the lattices do not have to be searched before translation.

4.3. Pruning

A phrase-based translation system that can take a word lattice of high density as input has an enormous search space so that pruning is necessary. In our system, we apply coverage pruning (here, hypotheses which cover the same set of source words are compared) and histogram pruning. These pruning methods are based on the total costs of a hypothesis. The absolute value of these costs depend on the scaling factors of the individual models. Since scaling factor values of substantially different magnitude are tested during the optimization, it can happen that too many or too few hypotheses will be pruned. To avoid this, we normalize the scaling factors in each iteration of the optimization procedure so that they sum up to 1.

It may also be necessary to prune the input word lattices. In our experiments, we have separate scaling factors for the acoustic and the source language model features. For this purpose, the arcs of the word lattices are only labeled with acoustic model scores. Then, during the translation process, the arc weights in the recognition lattice are extended by the scores of a source language model. Only after this operation the resulting automaton is pruned using a relatively large beam (pruning of the lattices based on acoustic scores only would have resulted in suboptimal performance). Again, the scaling factors which will be “tried” in the optimization process have to be considered when choosing the pruning threshold.

Table 1. Corpus statistics of the BTEC translation task.

		Italian	English
Train:	Sentences	66 107	
	Running Words	410 275	427 402
	Vocabulary	15 983	10 971
	Singletons	6 386	3 974
Dev:	Sentences	253	
	Running Words	1 472	1 510
	Out-Of-Vocabulary rate [%]	3.1	0.8
	ASR WER [%]	23.3	-
	avg. lattice density	49	-
	ASR graph error rate [%]	15.6	-
Test:	Sentences	253	
	Running Words	1 459	1 513
	Out-Of-Vocabulary rate [%]	2.5	0.8
	ASR WER [%]	21.4	-
	avg. lattice density	59	-
	ASR graph error rate [%]	15.4	-

Table 2. Corpus statistics of the EPPS translation task.

		English	Spanish
Train:	Sentences	1 652 174	
	Running Words	31 148 131	32 554 806
	Vocabulary	80 125	124 192
	Singletons	27 631	41 148
Dev	Sentences	500	
	Running Words	6899	6446
	Out-Of-Vocabulary rate [%]	0.2	0.1
	ASR WER [%]	14.5	-
	avg. lattice density	8	-
	ASR graph error rate [%]	6.3	-
Test	Sentences	792	
	Running Words	19 306	19 047
	Out-Of-Vocabulary rate [%]	1.6	–
	ASR WER [%]	14.6	-
	avg. lattice density	17	-
	ASR graph error rate [%]	8.7	-

5. EXPERIMENTAL RESULTS

5.1. Corpus Statistics

The speech translation experiments were carried out on two different tasks. Experiments for both tasks were based on bilingual sentence-aligned corpora.

The Italian-English *Basic Travel Expression Corpus* (BTEC) task contains tourism-related sentences usually found in phrase books for tourists going abroad. We were kindly provided with this corpus by ITC-irst. Corpus statistics for this task are given in Table 1. Word lattices of a 506 sentence test corpus have also been provided. The corpus was divided in two equal parts, one of which was used as a development set to tune model scaling factors. The lattice density in Table 1 is defined as the number of arcs in a lattice divided by the segment reference length, averaged over all segments. It is measured after determinization and minimization of the original lattices. The ASR graph error rate is the minimum WER among all paths through the lattice. For the evaluation, 16 reference translations of the correct transcriptions were made available.

We also tested our system on a large vocabulary task, namely machine translation of parliamentary speeches given in the European Parliament Plenary Sessions (EPPS). The training corpus for this task has been collected in the framework of the European research project TC-STAR. It contains over 30 million words of bilingual Spanish-English data. In March 2005, an open MT evaluation has been conducted in the project. The phrase-based SMT system presented here had shown the best translation performance [9], especially on the conditions of translating verbatim spoken text and single best recognizer output. Here, we present experimental results for translation from English to Spanish. We use the same test corpus as in the 2005 TC-STAR evaluation, for which two reference translations were made available. However, we use the RWTH ASR output (single best and word lattices) instead of the official evaluation data for the ASR condition. The corpus statistics for this task are given in Table 2. The translation vocabulary sizes are as large as 125 thousand words. The vocabulary used for English speech recognition is smaller – about 50 thousand words. The held-out development corpus was selected to have a similar ASR word error rate to the test corpus.

5.2. Evaluation Criteria

For the automatic evaluation, we used word error rate (WER), position-independent word error rate (PER), and the BLEU score [6]. The BLEU score measures accuracy, i. e. larger scores are better. The error rates and scores were computed with respect to multiple reference translations. On both tasks, training and evaluation were performed using the corpus and references in lowercase and without punctuation marks.

5.3. BTEC Italian-English Task

On the BTEC task, we estimated and used in search a 4-gram target language model. To include the source language model feature, in some experiments we extended each word lattice by the scores of a trigram language model and applied moderate beam pruning to the resulting automaton, as described in Section 4.3.

The experimental results for the BTEC development corpus are given in Table 3. In this table, the results are

Table 3. Translation results for the BTEC Italian-to-English task (development set). Here, λ denotes the set of scaling factors for translation model features and the target language model.

optimal λ found on:	Input/ transcription:	WER [%]	PER [%]	BLEU [%]
correct text	correct text	23.7	21.3	64.3
	single best	32.2	29.0	54.7
single best	single best	31.0	28.0	56.0
	word lattice	31.0	28.3	55.3
	+ ac. score	30.3	27.5	56.6
	+ LM score	30.4	27.7	56.2
	ac. + LM scores	29.8	27.0	57.5
lattice	opt all factors	29.2	26.3	58.7

Table 4. Translation results [%] on the BTEC test set. Comparison of the loglinear model approach (PBT) with the WFST-based joint probability approach (FSA).

System:	Input:	WER	PER	BLEU
PBT	single best	32.4	27.2	55.4
	word lattice	31.9	28.0	54.7
	ac. + LM scores	30.6	26.6	56.2
	opt all factors	29.8	25.8	57.7
FSA	single best	33.4	29.1	52.7
	lattice + ac. scores	31.6	27.6	54.3

Table 5. Examples of improvements with the integrated speech translation approach (BTEC test set, Italian-to-English translation).

Translation of	
single best	*I'm very sick lost*
lattice	*I feel much better now*
reference	*I feel much better now*
single best	*when should I take it ma'am*
lattice	*when should I take it sir*
reference	*when should I bring it sir*

grouped according to the type of optimization that was done in the loglinear model. In the first group of experiments, an optimal set of translation parameters λ was determined on the correct text by minimizing the word error rate. We then use these parameters to translate single best recognizer output and observe that the WER of the correct text translation is lower than the WER of the single best ASR translation by about 26 % relative.

However, we can also optimize the translation model scaling factors on the single best recognizer output, taking the parameters λ for initialization. Thus we obtained the optimal parameter settings λ' which were used to produce the second group of results in Table 3. Here, when the recognition features are used, the translation parameters λ' are kept fixed, and only the scaling factors for the acoustic model score (and, if applicable, for the source LM score) are optimized. In this way, we reduce the time complexity that is necessary to optimize all factors iteratively with the Downhill Simplex algorithm. Overfitting of parameters may also be avoided.

The translation quality for the single best recognition input improves with the new parameter set λ'. The translation model seemed to adapt to recognition errors by changing the scaling factors and giving more weight to the target language model and the inverse phrase based and single word based lexicons. However, later we observed no similar improvement on the test set for the single best input. We attribute this to the fact that recognition errors may be utterance-specific.

In the next experiment we take the word lattices with multiple hypotheses of the recognized utterances as input, but first do not use acoustic scores i. e. exploit only the lattice topology. We observe no improvement in translation quality. Thus, in some cases the system gets confused by hypotheses which are easy to translate but have little in common with the spoken words. Including the acoustic scores of the labels in the input lattice, and optimizing the scaling factor μ_2, we can achieve a significant reduction of the error rate and improvement of the BLEU score (Table 3). A similar improvement is achieved when we use only the source language model score and optimize the scaling factor μ_1. If we use both recognition scores and optimize μ_1 and μ_2, we are able to combine the positive effects of the two models and further improve the translation quality.

The best results can be achieved by optimizing the translation model scaling factors λ and the recognition model scaling factors μ simultaneously. With this settings, the relative difference in WER to the translation of correct text is reduced from 26% to less than 19%.

The results on the test set are given in Table 4. Here, the observations are similar to the development set; the best translation quality can be achieved by using both the acoustic and the source language model score, as well as optimizing all of the involved parameters of the loglinear model. Examples of translation quality improvements are given in Table 5. Table 4 also compares the system performance of the loglinear model (denoted with PBT) with the joint probability based WFST system of [3] (denoted with FSA). In that system, the translation model includes context dependency for the source words, so that it is used instead of the source language model; thus, the improvement is reached by using word lattices with acoustic scores. Our system not only performs better in terms of absolute error measures, but also is able to achieve a larger relative improvement (8% vs.

Table 6. Translation results [%] on the EPPS English-to-Spanish task.

Corpus	Input	WER	PER	BLEU
Dev	correct text	42.9	34.2	46.3
	single best	47.8	39.0	39.9
	lattice + ASR scores	46.6	38.8	40.2
Test	correct text	45.1	34.6	43.0
	single best	51.0	40.1	37.4
	lattice + ASR scores	51.3	40.6	36.8

5.4% in WER) with the integrated approach of word lattice translation based on loglinear modeling.

5.4. Experiments on the EPPS Task

On the EPPS task, we used a trigram language model in decoding. To avoid the enormous computational complexity on this large vocabulary task, we applied heavy pruning. Optimizing all of the scaling factors λ, the source language model factor μ_1, and the acoustic model factor μ_2 simultaneously on the development set for the WER or BLEU score, we were able to improve all translation measures. The best results could be achieved when optimizing for the BLEU score (see Table 6). However, we were not able to improve translation quality on the test set with the optimized parameters. We believe that the reasons for this may be search errors (resulting from heavy pruning) and overfitting. More thorough experiments are under way for this task.

6. CONCLUSIONS

In this paper, we used ASR word lattices as input for a statistical translation system. Coupling of speech recognition and machine translation was done using loglinear model combination. By using word lattices with acoustic model scores instead of single best recognition results, and also by including source language model scores we were able to avoid the negative effect of recognition errors and consistently improved translation quality. We optimized the scaling factors for the translation and recognition features in the minimum error training framework, and integrated all the features in the global search process, without using an approximation of N-best translation hypotheses. To the best of our knowledge, these are the first experiments, in which a state-of-the-art, phrase-based machine translation system with multiple features was applied to word lattice translation, and significant improvements were gained. On the large vocabulary task of MT for the European parliamentary speeches, the work is in progress, but first results are encouraging.

7. ACKNOWLEDGEMENT

This work was in part funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738). We would like to thank our colleagues R. Zens, S. Kanthak, C. Gollan, and J. Lööf who have contributed to this work.

8. REFERENCES

[1] N. Bertoldi, "Statistical Models and Search Algorithms for Machine Translation", PhD thesis, Università degli Studi di Trento, Italy, February 2005.

[2] S. Kanthak and H. Ney, "FSA: An Efficient and Flexible C++ Toolkit for Finite State Automata using On-demand Computation", Proc. 42nd Annual Meeting of the ACL, pp. 510 – 517, Barcelona, Spain, 2004.

[3] E. Matusov, S. Kanthak, and H. Ney, "On the Integration of Speech Recognition and Statistical Machine Translation", To appear in Proc. Interspeech 2005, Lisbon, Portugal, 2005.

[4] H. Ney, "Speech Translation: Coupling of Recognition and Translation", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1149–1152, Phoenix, AZ, 1999.

[5] F.J. Och, "Minimum Error Rate Training in Statistical Machine Translation", In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 160–167, Sapporo, Japan, July 2003.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proc. 40th Annual Meeting of the ACL, Philadelphia, PA, pp. 311–318, 2002.

[7] S. Saleem, S.-C. Jou, S. Vogel, and T. Schultz, "Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems", Proc. Int. Conf. on Spoken Language Processing, pp. 41–44, Jeju Island, Korea, 2004.

[8] E. Vidal, "Finite-State Speech-to-Speech Translation", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 111–114, Munich, Germany, 1997.

[9] D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney, "Statistical Machine Translation of European Parliamentary Speeches", To appear in Proc. of the Machine Translation Summit X, Phuket, Thailand, September 2005.

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

[11] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, W. K. Lo, "A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation", In Proc. of The 20th International Conference on Computational Linguistics (CoLing'04), Geneve, Switzerland, 2004.