INTEGRATING SPEECH RECOGNITION AND MACHINE TRANSLATION: WHERE DO WE STAND?

Evgeny Matusov, Stephan Kanthak, Hermann Ney

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

ABSTRACT

This paper describes state-of-the-art interfaces between speech recognition and machine translation. We modify two different machine translation systems to effectively process dense speech recognition lattices. In addition, we describe how to fully integrate speech translation with machine translation based on weighted finite-state transducers. With a thorough set of experiments, we show that both the acoustic model scores and the source language model positively and significantly affect the translation quality. We have found consistent improvements on three different corpora compared with translations of single best recognition results.

1. INTRODUCTION

Over the last decade it has been demonstrated by many publications and research projects that automatic speech recognition (ASR) and machine translation (MT) can be coupled in order to directly translate spoken utterances into another language. Whereas the most simple speech translation systems translate single best recognizer output, a few attempt to benefit from considering multiple recognition hypotheses for an utterance. Such attempts can be classified by the type of input that the systems use. A simple extension to translating only the single best ASR output is translations of the N-best ASR hypotheses. Recently, moderate improvements with this approach were reported by e. g. [3] and [4]. A more tighter coupling of ASR and MT is reached when word lattices are translated; the lattices can also be converted to confusion networks. In the past, some improvement of translation quality was achieved by using lattices with small densities [12]. Finally, a fully integrated approach where the whole search space of ASR and MT is integrated can be pursued. In the past, this approach was successful only on very small tasks [13].

When coupling speech recognition and machine translation, the recognition model scores and the translation model scores can be combined to improve translation performance. A theoretical basis for the score combination was given in [9]. One can differentiate between joint probability speech translation systems and conditional probability systems. In both types of systems, the ASR acoustic and language model scores can be combined with the translation features. The recognition features can either be included directly in the search, or in a post-processing step by rescoring word lattices or N-best lists.

This paper is organized as follows. Based on the presentation of [9], Section 2 reviews the Bayes' decision rule for speech translation. Starting from there, in Section 3 we show how ASR word lattices can be translated and review the basics of our two speech translation systems: the joint-probability system and the phrase-based system that employs log-linear modeling. Section 4 explains the functionality of the fully integrated speech translation system. In Section 5 we present significant improvements in quality of translation when we utilize recognition features in translation and optimize the model scaling factors.

2. BAYES' DECISION RULE FOR SPEECH TRANSLATION

In speech translation, we try to find the 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 [9]). 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.

As already stated in [9], the conditional probability term $Pr(f_1^J|e_1^I)$ and $P(e_1^I)$ can be rewritten when using a *joint* probability translation model:

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

This simplifies coupling the systems since the joint probability translation model can be used instead of the usual language model in speech recognition (see Section 4).

It should be noted that in comparison to integrated speech translation which uses the decision rule from above, speech translation based on word lattices uses the additional approximations that word boundary times are fixed and that many word sequences may never be contained in the word lattice due to the word-pair or word-triple approximation.

1-4244-0469-X/06/$20.00 ©2006 IEEE

3. SPEECH TRANSLATION SYSTEMS AT RWTH

3.1. WFST-Based Joint Probability System

The joint probability MT system (referred to as FSA, for a more detailed description see also [6]) is implemented with weighted finite-state transducers (WFSTs). First, the training corpus is transformed as shown in Figure 1, based on a word alignment. Then a statistical m-gram model is trained on the bilingual corpus. This language model is represented as a finite-state transducer Tr which is the final translation model. Searching for the best target sentence is done in the composition of the input represented as a WFST and the translation transducer Tr.

Coupling the FSA system with ASR is fairly simple since the output of the recognizer represented as WFST can be used directly as input to the machine translation search. For the FSA-based speech translation system the only features used are the acoustic probability from the input word lattice and the translation model probability. The source language model scores are not included, since the joint m-gram translation probability contains dependencies on the predecessor source words and thus serves as a source language model.

```
vorrei|i'd_like del|some gelato|ice_cream
per|ε favore|please
```

Fig. 1. Example of a transformed sentence pair.

3.2. Phrase-Based System

The phrase-based translation system (referred to as PBT) follows a direct modeling approach. Probability distributions are represented as features in a log-linear model. In particular, the translation model probability $Pr(f_1^J|e_1^I)$ is decomposed into several probabilities. The main feature is the phrasal translation lexicon. It is supplemented by single word based lexicon probabilities. Lexica from both translation directions are used. In addition, we include the target language model, as well as the word and phrase penalty features to avoid too short/long translations.

Each feature is scaled by a separate exponent. The scaling factors are optimized in a minimum error training framework [10] with the Downhill Simplex algorithm iteratively, 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.

For speech translation we additionally include the acoustic model probabilities $Pr(x_1^T|f_1^J)$ of the hypotheses in the ASR word lattices and probabilities of the source language model as features. Details are given in [7]. When searching for the best translation, the system has to optimize over alternative recognition word sequences f_1^J (as given by the input word lattice), over all possible monotone segmentations of a given recognized sequence into source language phrases, and over all possible translations of these phrases.

The utilization of multiple features and the direct optimization for an objective error measure is the main advantage of this system in comparison to the FSA system. However, it is paid by a less efficient search, which makes heavy pruning unavoidable.

3.3. Reordering

Appropriate reordering of words/phrases in translation is very important for good performance of MT systems, since there are significant differences in typical word order between most languages (see also [6]). In case of ASR word lattice input, the reordering in search is a complex problem. Here, we present two basic solutions.

In the FSA system, the search is monotone. However, we reorder words in each sentence in the target training corpus based on the initial word alignment such that the resulting alignment becomes monotone. Obviously, resulting translations will have the word order of the source sentence. To fix the wrong word order, we use a similar idea to that described in [2]. Given the best translation hypothesis we first permute its words and then compose the resulting permutation automaton with an n-gram target language model in order to select the word order with the highest probability. The computational complexity can be reduced by using constraint permutation automata.

In a recent modification of the PBT system, limited word reordering is possible. While traversing the input lattice, a matched source phrase can be skipped and processed later. This type of reordering helped to improve translation quality, see Section 5.

4. FULL INTEGRATION OF ASR AND MT

As the PBT system is more complicated to integrate with speech recognition search, we only use the FSA system for the fully integrated speech translation. We start by representing the static ASR search network as a composition of multiple WFSTs (see also [8]), namely H for the HMM topology, C for the context-dependency (CART), L for the lexicon, and G for the language model. As the transducer cascade $H \circ C \circ L$ already represents the conditional probability term $Pr(s_1^T|f_1^J)$ for a given HMM state sequence s_1^T, we only need to replace the source language model G by the translation model Tr to get the final optimized (det = determinized) speech translation search network ST :

$$ST = \det(H \circ \det(C^{-1})^{-1} \circ \det(L \circ Tr))$$

The problems faced in the optimized composition are:

- Tr is ambiguous on the input side. This can be solved by adding disambiguation symbols to the input side of Tr as described in [8] for the lexicon.
- unknown words, i.e. source words contained in the lexicon, but not in the input language of Tr, must be passed to the output language of the translation model Tr. This is performed by preprocessing Tr appropriately.

5. EXPERIMENTS

5.1. Corpus Statistics

The speech translation experiments were carried out on three different tasks. Experiments for all tasks were based on bilingual sentence-aligned training corpora. Corpus statistics for these tasks are given in Table 1.

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. Speech translation experiments were also performed on a smaller Chinese-English BTEC corpus [1] in the framework of the IWSLT 2005 evaluation campaign [14]. 16 reference translations of the correct transcriptions for the BTEC test corpora were available.

The Italian-English Eutrans II FUB task contains sentences from the domain of hotel help-desk requests. It is significantly smaller than the BTEC task and has evolved from one of the first European-funded speech translation projects. Only a single reference translation is available for the test corpus on this task.

Table 1. Corpus statistics of the speech translation tasks BTEC and Eutrans II.

		BTEC				Eutrans II FUB	
		Italian	English	Chinese	English	Italian	English
Train	Sentences	66 107		20 000		3 257	
	Running Words	410 275	427 402	176 199	189 927	47 681	57 663
	Vocabulary	15 983	10 971	8 687	6 870	2 453	1 695
	Singletons	6 386	3 974	4 006	2 888	975	519
Test	Sentences	253		506		300	
	Running Words	1 459	1 510	3 918	3 909	5 305	6 419
	Out-Of-Vocabulary rate [%]	2.5	0.9	2.3	1.8	2.3	1.3
	ASR WER [%]	21.4	-	42.0	-	23.7	-

5.2. Evaluation Criteria

For the automatic evaluation, we used word error rate (WER), position-independent word error rate (PER), and the BLEU score [11]. The BLEU score measures accuracy, i. e. larger scores are better. On all tasks, training and evaluation were performed using the corpus and references in lowercase and without punctuation marks.

5.3. Translation of Word Lattices

We compare the performance of the transducer-based joint probability system and of the phrase-based system on the BTEC Italian-English task. We consider three translation conditions: translating single best recognition output, translating ASR word lattices without the acoustic model scores, and including the acoustic model scores in the ASR word lattice in the global decision process.

For the FSA system, a 4-gram translation model was estimated on the bilingual representation of the training corpus for this task. A 4-gram target language model was used in search for the PBT system as well as to score permutations of the final hypotheses from the FSA system. In order to include the source language model feature in the PBT system, we extended each word lattice by the scores of a trigram language model.

The objective error measures for the two systems on the BTEC Italian-English task are summarized in Table 2. We observe that exploring the word lattice topology in translation already results in some improvement in the translation quality. However, the improvements are more significant when we combine recognition model features with the translation model features. In the case of the FSA system, as mentioned in Section 3, we interpolate the acoustic model score and the translation model score. It is important to optimize the scaling factor for the translation model score. On this task, the scaling factor is 45 and is higher than the usual LM scaling factor in speech recognition.

When using the PBT system, we include both the acoustic model and the source language model score. The language model score is used to model the context dependency for the source language which had been captured only within the source phrases of the phrasal lexicon. The scaling factors for the recognition features only or for translation and recognition features simultaneously were optimized in the log-linear model on a development set for the word error rate. Table 2 shows the improvements in translation quality on the test set when using optimized scaling factors.

Table 2 also shows that the PBT system not only performs better in terms of absolute error measures, but also is able to achieve a larger relative improvement (8% vs. 5.4% in WER) with the integrated approach of word lattice translation based on log-linear modeling.

Table 2. Translation results [%] on the BTEC Italian-English task. Comparison of the log-linear 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

5.4. Importance of Word Reordering

As mentioned in Section 3.3, both of the described speech translation systems can be improved by allowing limited reordering. In the case of the FSA system, target sentences were reordered in training, but the lattice was processed monotonically. After translation, the resulting single best hypotheses were permuted under the IBM reordering constraints with a window size of 3 and scored with a target language model. This has further reduced the number of translation errors on the BTEC Italian-English task, as shown in Table 3.

Postponing the translation of a matched phrase and thus allowing limited reordering in the search also helps to improve the performance of the PBT system. However, this improvement is significant only when translating from a language with the word order largely different from English, e. g. Chinese. Local reorderings which are typical for the Italian-English translations are already captured in the bilingual phrasal lexicon of the system. Table 4 presents the translation results on the Chinese-English BTEC task. Performing the limited reordering clearly results in better translation quality for both ASR single best output and word lattice translation.

Table 3. Effect of target reordering in training and after translation for word lattice translation on the BTEC Italian-English task (FSA system, results in [%]).

Reordering:	WER	PER	BLEU
none	31.6	27.6	54.3
target	30.6	26.0	55.4

Table 4. Effect of phrase reordering in search on the BTEC Chinese-English task (PBT system, results in [%]).

Reordering:	Translation of:	WER	PER	BLEU
none	single best	62.1	52.7	31.1
	lattice	58.3	48.1	34.1
skip	single best	61.3	51.7	33.1
	lattice	57.7	47.2	35.1

Table 5. Translation results [%] on the Eutrans II FUB Italian-English task. The last line contains results when directly coupling the speech recognition and machine translation systems by using a single optimized finite-state network.

Input:	WER	PER	BLEU
correct text	29.1	22.1	58.8
single best	37.4	29.1	51.3
word lattice	38.2	29.5	50.2
+ ac. scores	36.6	28.1	52.4
integrated	36.3	28.0	52.6

Table 6. Comparison of speech recognition and speech translation search characteristics for the Eutrans II FUB Italian-English task (AMD Athlon64 2.0GHz; RTF: real-time factor).

system	# active states	RTF
ASR	1 872	0.35
ST	14 379	1.26

5.5. Fully Integrated Speech Translation

The experiments for fully integrated speech translation were performed on the Eutrans II FUB corpus. For better comparison we generated lattices with different densities. The lattice error rate, i.e. the minimum WER among all paths through the lattice, was 9.1% on average for the largest lattice density of 2098. We optimized the system with respect to both the lattice density and the translation model scaling factor λ simultaneously. In contrast to the results presented in [12], the WER consistently drops with larger lattices and shows a clear minimum for $\lambda = 90$ (for comparison: the optimal language model scaling factor for the ASR system is 16). Results of all error measures for the optimal settings are given in Table 5. The target word reordering in training and after translation was performed as described in Section 3.3. Different from [6], we consistently use a trigram language model to generate lattices and a trigram translation model here. The last line of the table shows that the fully integrated system performs better than the system using large lattices which is another proof that the error rate does not rise with larger lattice densities. Note that although the speech recognition system has a slightly worse WER on this task compared to [5], we obtain a much better speech translation WER.

Table 6 compares the search space of the network described in Section 4, using either the usual language model G, or the translation model Tr. In both cases, pruning thresholds were adjusted to be minimal and to not produce search errors. Both static pre-compiled search networks were about the same size with the speech translation network being slightly bigger. Speech translation had about 7.7 times more active state hypotheses and was slower than speech recognition by a factor of 3.6. This can partly be contributed to the high ambiguity of the translation model as the same input sentence may have many different translations. On the Eutrans II FUB task, we observed an average of 2.9 target phrases per source word in the bi-language.

6. CONCLUSIONS

In this paper, we gave a short overview of the current research on coupling speech recognition and machine translation. We presented two state-of-the-art speech translation systems which consistently perform better when translating ASR word lattices with acoustic and/or language model scores, or even in a fully integrated speech translation architecture. These improvements are significant and were achieved on several tasks. However, on (large vocabulary) tasks with good ASR performance, the MT performance is yet to be generally improved to avoid translating word sequences which contain recognition errors. Also, the key to success of speech translation is a closer cooperation of the ASR and MT researchers who have to agree on common standards for e. g. the lattice structure, definition of vocabulary, segmentation, and other practical interface issues.

7. REFERENCES

[1] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. IWSLT*, pp. 1–12, Kyoto, Japan, September.

[2] S. Bangalore and G. Riccardi, “Finite-State Models for Lexical Reordering in Spoken Language Translation”, Proc. Int. Conf. on Spoken Language Processing, vol. 4, pp. 422–425, Beijing, China, 2000.

[3] A. Bozarov, Y. Sagisaka, R. Zhang, G. Kikui. “Improved Speech Recognition Word Lattice Translation by Confidence Measure”. In Proc. Interspeech 2005, pp. 3197–3200, Lisbon, Portugal, 2005.

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

[5] F. Casacuberta, D. Llorens, C. Martínez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Picó, A. Sanchis, E. Vidal, and J. M. Vilar, “Speech-to-speech Translation Based on Finite-State Transducers”, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 613–616, Salt Lake City, UH, 2001.

[6] E. Matusov, S. Kanthak, and H. Ney, “On the Integration of Speech Recognition and Statistical Machine Translation”, In Proc. Interspeech 2005, pp. 3177–3180, Lisbon, Portugal, 2005.

[7] E. Matusov and H. Ney, “Phrase-based Translation of Speech Recognizer Word Lattices Using Loglinear Model Combination”, To appear in Proc. Int. Workshop on Automatic Speech Recognition and Understanding, Cancun, Mexico, 2005.

[8] M. Mohri, F. C. N. Pereira and M. Riley, “Weighted Finite-State Transducers in Speech Recognition”, Proc. ISCA Workshop, ASR2000, Paris, France, 2000.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney “The RWTH Phrase-based Statistical Machine Translation System”, to appear in Proc. Int. Workshop on Spoken Language Translation (IWSLT), Pittsburgh, USA, 2005.