SPEECH TRANSLATION ENHANCED ASR FOR EUROPEAN PARLIAMENT SPEECHES - ON THE INFLUENCE OF ASR PERFORMANCE ON SPEECH TRANSLATION

Sebastian Stüker[1], *Matthias Paulik*[1,2], *Muntsin Kolss*[1], *Christian Fügen*[1], *and Alex Waibel*[1,2]

[1]Institut für Theoretische Informatik, Universiät Karlsruhe (TH), Karlsruhe, Germany
[2]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA
{stueker|paulik|kolss|waibel}@ira.uka.de

ABSTRACT

In this paper we describe our work in coupling automatic speech recognition (ASR) and machine translation (MT) in a speech translation enhanced automatic speech recognition (STE-ASR) framework for transcribing and translating European Parliament speeches. We demonstrate the influence of the quality of the ASR component on the MT performance, by comparing a series of WERs with the corresponding automatic translation scores. By porting an STE-ASR framework to the task at hand, we show how the word errors for transcribing English and Spanish speeches can be lowered by 3.0% and 4.8% relative, respectively.

***Index Terms*—** Speech Recognition, Machine Translation, European Parliamentary Plenary Sessions, TC-STAR, STE-ASR

1. INTRODUCTION

For many years automatic speech recognition (ASR) and machine translation (MT) evolved independently from each other. Speech-To-Speech Translation (SST) is one field that brings together these two separate sciences now. Projects, such as the European Union sponsored, integrated project TC-STAR, have set out to improve and build closely integrated SST systems.

Naturally, errors committed by the ASR components lead to additional errors in the Machine Translation component on top of the ones that would be observed on error free transcriptions of speech. Therefore, one of the major research directions is still the improvement of the individual recognition and translation components. But the field of closely integrating machine translation and speech recognition are special interest as well. One way to do this, is the use of a speech translation enhanced ASR setup, as we have described in previous work [1].

Section 2 introduces the TC-STAR project and the task of translating European Parliament Plenary Sessions. Section 3 then describes our systems for this task that were used to conduct the experiments reported in this paper. In section 4 we then report on our findings in the relation between the WER of our ASR systems and the quality of our MT systems using three automatic translation quality scores. Finally, in section 5.1, we describe our work in applying a speech translation enhanced ASR setup to the speeches in order to improve the automatic transcriptions of them.

2. TC-STAR AND EPPS

The experiments described in this paper were performed in the context of the European Integrated Project *Technologies and Corpora for Speech-to-Speech-Translation (TC-STAR)* (http://www.tc-star.org) which is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation.

TC-STAR currently focuses on the three languages English, Spanish, and Chinese. The tasks on which recognition, translation, and synthesis are performed are Broadcast News for the translation direction Chinese to English, and speeches given in the European Parliament for the directions English to Spanish and vice versa.

2.1. European Parliamentary Speeches Corpus

The parliament of the European Union operates in 21 official languages. The fact that all official transactions, whether in verbal or written form, within the European Parliament have to be made available in all official languages, makes the parliament an ideal environment for SST research, since it produces a wealth of often even parallel audio and text material.

The *European Parliamentary Speeches (EPPS) task* within TC-STAR focuses on transcribing speeches given in the European Parliament in English and Spanish, and translating them into the other language. In order to adequately address this task a number of language resources was created within TC-STAR [2]. For Automatic Speech Recognition the data for this task has been recorded from the European Union's TV Information service *Europe by Satellite (EbS)*. The recordings include the original audio from the speakers as well as simultaneous translations into all official languages of the European Union [3]. For the second TC-STAR evaluation roughly 100h of transcribed debates for each, Spanish and English, were available as training material. For Spanish additional 40h of transcribed speech from the Spanish parliament, the CORTES data, was available.

Also available within TC-STAR is a corpus of parallel data consisting of the final text editions from the European Parliament available through the EuroParl website (http://www.europarl.europa.eu), processed and aligned by RWTH Aachen.

For the evaluation within TC-STAR the consortium provided among other resources a development set (dev2006) of 3 hours of speech for each language. No utterance level segmentation or speaker labels were given. The development sets, however, were divided into seven sessions for English and fourteen sessions for Spanish.

3. SYSTEMS USED

The experiments reported in this paper require the use of two speech recognition and two machine translation systems which we describe in this section.

3.1. ASR

The speech recognition systems used for our experiments were trained with the help of the Janus Recognition Toolkit (JRTk) which features the Ibis one-pass decoder [4]. As resources for training and testing we made use of the corpora provided within TC-STAR as described in 2.1 above. The English and Spanish evaluation systems consist of several left-right Hidden Markov Models (HMMs) without state skipping with three HMM states per phoneme. The training of the acoustic model involved applying an incremental splitting of Gaussians training followed by estimating one global semi-tied covariance matrix after LDA and several iterations of Viterbi training. In addition to that Constrained Feature Space Maximum Likelihood Linear Regression training (fMLLR)was applied to the models in the last stage of the systems. Training of the language models was performed with the help of the SRILM Toolkit [5].

3.1.1. English ASR

The English recognition system used in the experiments for this paper is the ISL system for the TC-STAR Spring 2005 evaluation [6]. For the purpose of cross-system adaptation we trained systems with models of different sizes, based on two different phoneme sets, and on two different kinds of front-ends [7]. One front-end is based on the traditionally used Mel-frequency scaled Cepstral Coefficients (MFCC), the other one on the Minimum Variance Distortion-less Response (MVDR)[8]. One phoneme set is based on the CMU dictionary, the other one on the Pronlex phoneme set.

For the language model we first trained separate 4-gram language models on the following corpora: the EPPS transcriptions, the EPPS final text editions, Hub4 Broadcast News data, and the English part of the UN Parallel Text Corpus v1.0. The resulting language models were then interpolated into one language model while tuning the interpolation weights on the 2005 EPPS development data. The resulting model yields a perplexity of 93 on the EPPS 2006 development set.

The system in the form used in this paper is made up of four stages. Each stage consists of two systems, one based on the MFCC front-end, the other on the MVDR. The output of both systems is combined via Confusion Network Combination (CNC)[9] to the final output of the stage. Stage 1 uses speaker independent acoustic models, all other stages use acoustic models that were unsupervised adapted on the output from the previous stage using Vocal Tract Length Normalization (VTLN), Maximum Likelihood Linear Regression (MLLR), and feature space constrained MLLR (fMLLR). The systems of stage 1, 2, and 4 are based on the Pronlex phoneme set, the systems in stage 3 on the CMU dictionary phoneme set. The frame shift of the acoustic front-ends is 10ms in the first stage and 8ms in all other stages. The system achieves an WER of 12.6% on the EPPS 2006 development set.

3.1.2. Spanish ASR

The Spanish ASR system consists of two stages with two systems per stage, one based on an MFCC front-end and one based on an MVDR front-end. The outputs of the two systems in a stage are again combined via CNC. The models of the second stage are adapted on the output from the first stage via VTLN, MLLR, and fMLLR. The frame shift in both front-ends for the first stage is 10 ms, and 8 ms for the second stage.

The dictionary and language model were created using the EPPS final text editions, the CORTES texts, and the EPPS + CORTES transcriptions.For each of the before mentioned corpora, a case sensitive 4-gram LM was computed and a final LM was created by interpolations of these. The interpolations weights were chosen to minimize the perplexity on the 2006 TC-STAR development set. The final 4-gram LM yielded a perplexity of 83 on the 2006 development set. The pronunciation dictionary has a size of 77.9K entries over a case sensitive vocabulary of 63.3K. The OOV rate is 0.67% on the development set. The word error rate after the second pass was 8.4% case-insensitive and 9.5% case-sensitive.

3.2. MT

The statistical machine translation system used for our experiments is based on phrase-to-phrase translations and was trained on the data mentioned in 2.1. Extraction of phrase translation candidate pairs is done by the PESA method, which views phrase alignment as a sentence splitting approach [10]. To allow for the use of phrases of arbitrary length, we do not build a static phrase table containing all possible phrase pairs up to a certain length, but extract phrase pairs from the bilingual corpus at decoding time [11].

The decoder used in the translation experiments is a beam search decoder which allows for restricted word reordering and uses the following models: 1. The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus. 2. A trigram language model, trained with the SRI language model toolkit [5] using modified Kneser-Ney smoothing. 3. A word reordering model, which assigns higher costs to longer distance reordering. 4. Simple word and phrase count models which compensate the tendency of the language model to prefer shorter translations, and favor longer phrases over shorter ones, potentially improving fluency. Each model score is multiplied by a scaling factor which can be modified to tune the overall system. More details can be found in [11].

The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations are used to generate a translation lattice. The second step is then a modified shortest path search through this lattice. Shortest, as we use the negative logarithms of the model probabilities. Modified, as we allow for word reordering. Decoding proceeds essentially along the source sentence. At each step, however, the next word or phrase to be translated may be selected from all words or phrases starting within a given look-ahead window from the current position [12].

4. INFLUENCE OF THE ASR ON THE TRANSLATION PERFORMANCE

When compared to translating written text, machine translation of spontaneous speech faces several new challenges. Other than in text automatic speech recognition usually does not deliver punctuation. The segments on which ASR often operates and passes on also do not necessarily correspond to sentences. Often ASR systems do not provide case-sensitive output. Moreover, spontaneous speech is filled with malformed utterances due to spontaneous speech effects, such as disfluencies, repetitions, grammatically incorrect and/or incomplete sentences etc. On top of that adds the problem of partially wrong transcriptions due to recognition errors. These errors then lead to subsequent errors in the translation.

In order to examine the influence of the ASR errors on the MT result we took several passes out of the Spanish and English recognition systems described before, that cover a reasonable range of WERs. The hypotheses with the different WERs were then automatically translated and the quality of the translation was measured

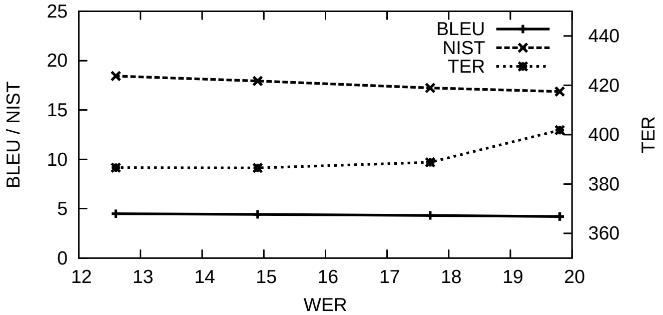

Fig. 1. WER vs. Translation Quality on the English 2006 Development Set.

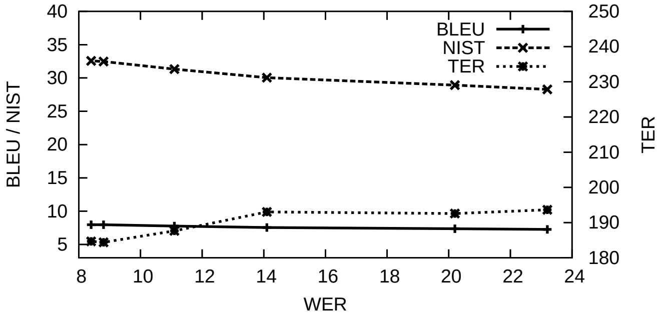

Fig. 2. WER vs. Translation Quality on the Spanish 2006 Development Set.

using the BLEU, NIST, and TER score. Figure 1 shows the correspondence between WER and translation score for the English system, Figure 2 for the Spanish system. Both figures display a roughly linear influence of the WER on the three translation scores.

5. SPEECH TRANSLATION ENHANCED ASR

In earlier work we experimented with ASR and MT in human mediated translation scenarios[13]. These scenarios are characterized by the presence of one or more human interpreters that translate the speech from a speaker into one or several other languages. In these scenarios it is often desirable to have a written transcript of the original speech from the speaker as well as the speech that is the result of the interpreter's translation, e.g. for archiving or publication purposes.

The sessions of the European parliament are an excellent example of such a scenario in which the speech of a speaker is simultaneously translated into 21 languages. Further, transcriptions of the original speech and its translation are being kept, published, and archived. Currently the transcriptions and translations are done by humans. Here automatic speech transcription systems can be a valuable tool.

The goal of *speech translation enhanced ASR (STE-ASR)* as introduced in [1] and [14] is to improve the speech recognition performance in one language, regardless of whether the speech comes from the original speaker or an interpreter, by making use of all available parallel speech and other information (e.g. in the form of documents) in all available languages. This is done by automatically translating the multilingual information into the language of the ASR system and then biasing the ASR system toward the gained knowledge. Figure 3 gives an overview of the setup for the case human speech in the foreign language is available.

In our previous work STE-ASR techniques were successfully applied to the bilingual Basic Travel Expression Corpus (BTEC) [15]. In these experiments we assumed that for every spoken source sentence, the respective target sentence audio data is available and fully aligned with the source sentence. Under this assumption it was possible to directly bias the source language ASR system for each sentence.

However, in more complex translation scenarios, such as it is the case for the European Parliament, this assumption no longer holds. Here, the source language speech and the speech in the various target languages is only loosely aligned and parallel multilingual information only occurs in a variable time frame at approximately the same time. In addition, the simultaneous translations from the human translators suffer from frequent synopses, omissions and self-corrections.

In this work we present the results of our first experiments in extending the STE-ASR approach to the EPPS task. These experiments partially factor out the above mentioned problems of real simultaneous translations by using parallel bilingual information that is aligned on a per session basis only.

5.1. STE-ASR Experiments

The 2006 Spanish and English ASR TC-STAR development sets only contain speech from politicians and no speech from the simultaneous translators. This means that both development sets do not contain any simultaneous translations of each other, as it would be necessary for directly applying speech translation enhanced ASR techniques.

For the evaluation of machine translation on the two development sets, human reference translations of the development sets into the opposite language were produced. We took these reference translations instead of simultaneous translations. In order to perform the STE experiments we automatically translated the English reference translations back to Spanish and the Spanish reference translations back to English using the MT systems described above.

As a baseline ASR systems we took the MFCC pass from the last stage of the English system and the last stage of the Spanish ASR system. In order to bias the systems toward the automatic translations, we interpolated the baseline 4-gram language models of the English and Spanish ASR system respectively with small 4-gram language models computed on these translations. Two sets of experiments for this method were performed.

First we estimated language models on the complete 2006 development set and then interpolated them with the language model of the recognition systems. The interpolation weight was calculated by minimizing the perplexity on the development set. For English the calculation of the interpolation weight on the development set was straight forward and resulted in an interpolation weight of 0.1 for the LM trained on the translation. The perplexity on the development decreased from 93.5 to 90.0, the WER dropped from 13.2% to 13.0%.

For Spanish, the automatic translations were provided in lowercase only. Since the Spanish ASR system utilizes a case sensitive language model we therefore had to deal with the question of how to chose the interpolation weight of the small, lowercase only translation LM. To address this question, we conducted two experiments. At first, we chose the interpolation weight as to achieve a minimal perplexity on the case sensitive development set. This resulted in an interpolation weight of 0.04 for the language model estimated on the translation, and the perplexity of the language model on the devel-

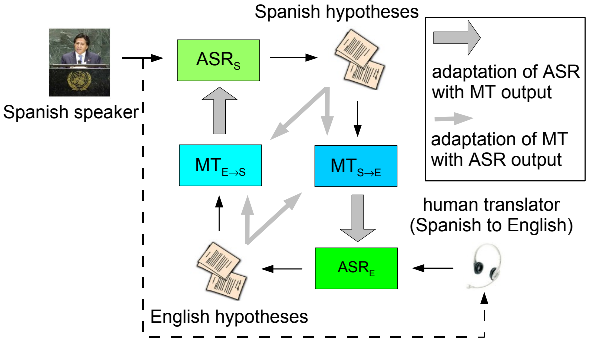
Spanish hypotheses
adaptation of ASR with MT output
adaptation of MT with ASR output
ASR$_S$
Spanish speaker
MT$_{E→S}$
MT$_{S→E}$
human translator (Spanish to English)
ASR$_E$
English hypotheses

Fig. 3. STE-Enhanced ASR

opment set decreased from 83.23 to 82.45. Secondly, we converted the baseline LM into a lowercase only LM and then chose the interpolation weight as to minimize the perplexity on the lowercase only development set. The resulting interpolation weight was 0.07 and the perplexity on the lowercase only development set decreased from 85.3 to 83.0. However, the final LM used for recognition was again an interpolation of the case sensitive baseline LM and the lowercase only translation LM weighted with 0.07. This final LM yielded a perplexity of 82.53 on the case sensitive development set. Although the decrease in perplexity was only minimal, we observed a reduction in WER of 0.2% absolute for both interpolated LMs, from 8.4% to 8.2%. However, the case sensitive WER of 9.5% could not be decreased.

In the second set of experiments, we now calculated separate LMs for the individual sessions in the development set. We did this by interpolating the original language model of the recognition systems with language models that were calculated on the translations of the respective sessions only. The interpolation weights were not calculated a new but rather taken from the first set of experiments and kept the same for a all session dependent language models.

For English the word error rate with the session dependent LMs dropped further, down to 12.8%, for Spanish down to 8.0%. However, again the case sensitive WER for Spanish could not be decreased. While with the baseline LM the average perplexity per session was 109.9 for English and 94.1 for Spanish, these drop to 105.0 for English and 89.5 for Spanish, when using the LM interpolated on the whole development set, and to 91.6 and 72.2 when using the session dependent LMs.

6. CONCLUSION

In this paper we have addressed the influence of the quality of the automatic speech recognition component on the machine translation quality in a speech-to-speech translation setup. Experiments on the EPPS task show an approximately linear influence of the WER onto three automatic translation error scores.

We further demonstrated, how speech translation enhanced automatic speech recognition techniques can be extended to improve the automatic transcription of European Parliamentary Plenary Speeches. In our current work we limited ourselves to the languages English and Spanish, thereby reducing the WER on English by 3.0% relative and 4.8% on Spanish. By incorporating additional languages into the setup and by refining the selection and alignment of the multilingual knowledge we anticipate even larger reductions in WER in our future work.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] M. Paulik, C. Fügen, S. Stüker, T. Schultz, T. Schaaf, and A. Waibel, “Document Driven Machine Translation Enhanced ASR,” in *Eurospeech*, Lisbon, Portugal, 2005.

[2] H. v. d. Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, “TC-STAR: New language resources for ASR and SLT purposes,” in *ICASSP*, Philadelphia, PA, USA, 2005.

[3] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, “Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus,” in *ICASSP*, Philadelphia, PA, USA, 2005.

[4] H. Soltau, F. Metze, C. Fügen, and A. Waibel, “A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment,” in *ASRU*, Madonna di Campiglio Trento, Italy, 2001.

[5] A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit,” in *ICSLP*, Denver, Colorado, USA, 2002.

[6] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y.-C. Tam, and M. Wölfel, “The ISL TC-STAR Spring 2006 ASR Evaluation Systems,” in *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[7] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, “Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End,” in *INTERSPEECH*, Pittsburgh, USA, 2006.

[8] M.C. Wölfel and J.W. McDonough, “Minimum Variance Distortionless Response Spectral Estimation, review and refinements,” *IEEE Signal Processing Magazine*, pp. 117–126, 2005.

[9] L. Mangu, E. Brill, and A. Stolcke, “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400.

[10] S. Vogel, “PESA: Phrase Pair Extraction as Sentence Splitting,” in *Machine Translation Summit*, 2005.

[11] M. Kolss, B. Zhao, S. Vogel, A. Venugopal, and Y. Zhang, “The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluations,” in *TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[12] S. Vogel, “SMT Decoder Dissected: Word Reordering,” in *Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003.

[13] M. Paulik, S. Stüker, and C. Fügen, “Speech Recognition in Human Mediated Translation Scenarios,” in *MELECON*, Malaga, Spain, 2006.

[14] M. Paulik, S. Stüker, C. Fügen, T. Schultz, Thomas Schaaf, and A. Waibel, “Speech Translation Enhanced Automatic Speech Recognition,” in *ASRU*, San Juan, Puerto Rico, 2005.

[15] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, “Creating Corpora for Speech-to-Speech Translation,” in *EUROSPEECH*, 2003.