OPEN DOMAIN SPEECH RECOGNITION & TRANSLATION: LECTURES AND SPEECHES

C. Fügen[1], *M. Kolss*[1], *D. Bernreuther*[1], *M. Paulik*[1], *S. Stüker*[1], *S. Vogel*[2], *A. Waibel*[1,2]

[1]{fuegen,kolss,bernreuther,paulik,stueker,waibel}@ira.uka.de
interACT, Universität Karlsruhe (TH), Karlsruhe, Germany

[2]{stephan.vogel,waibel}@cs.cmu.edu
interACT, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

For years speech translation has focused on the recognition and translation of discourses in limited domains, such as hotel reservations or scheduling tasks. Only recently research projects have been started to tackle the problem of open domain speech recognition and translation of complex tasks such as lectures and speeches. In this paper we present the on-going work at our laboratory in open domain speech translation of lectures and parliamentary speeches. Starting from a translation system for European parliamentary plenary sessions and a lecture speech recognition system we show how both components perform in unison on speech translation of lectures.

1. INTRODUCTION

Estimates for the number of existing languages today range from 4000 to 6000. At the same time the phenomenon of globalisation requires an active flow of information among people speaking a wide variety of languages. Lectures are an effective way of performing this dissemination. Personal talks are preferable to written publications because they allow the speaker to tailor his presentation to the needs of a specific audience, and in return allow the listeners to get access to the information relevant to them through interaction with the speaker. Currently, many lectures just cannot happen because human translators would be too expensive. However, the use of modern machine translation techniques can potentially provide affordable translation services to a wide audience, making it possible to overcome the language barrier for almost everyone.

So far speech translation research has focused only on limited domains such as the scheduling of meetings, basic tourist expressions or the pre-arrival reservation of hotel rooms. The next step we have started to take now is the development of open domain speech translation systems for spontaneous speech, that do not have to be individually tailored to every single domain one wants to cover. Therefore, we have combined a speech translation system and a speech recognition system that we developed under the European Commission integrated projects CHIL and TC-STAR. We evaluated them in a speech translation scenario of lectures, translating English lectures to Spanish and German, given special attention to the behavior of the system when used on a new, previously unseen domain.

2. SPEECH RECOGNITION

For speech recognition we used our own single-pass decoder *Ibis* [1]. The acoustic models were trained with the help of the Janus Recognition Toolkit, the language models with SRILM [2].

2.1. Test Data

Since we focus on open domain speech recognition of lectures, the most suitable development data we have is the CHIL lecture part of the NIST RT-05S development set (RT-05Sdev), which consists of approx. 130 minutes of speech. In order to have a comparison to current state-of-the-art systems, we also evaluated on the NIST RT-05S lecture evaluation set (RT-05Seval).

The data presents significant challenges to both models used in ASR, the language and acoustic model. With respect to the former, the data primarily concentrates on technical topics with a focus on speech research, and contains spontaneous and disfluent speech, due to the interactive nature of seminars and the varying degree of the speakers' familiarity with their topics. On the acoustic modeling side, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech.

2.2. Impact of different Acoustic Model Training Data

As relatively little training data for lecture recognition was available we selected different data sources, which were available in our lab for acoustic model training:

Meeting: 96h of transcribed speech, coming from different locations [3] together with an additional amount of 30h of not publicly available meetings collected at Carnegie Mellon.

BN: 180h broadcast news data (Hub-4) [3].

TED: 10h of transcribed speech of the Translanguage English Database (TED) [3] together with an additional 3h from that corpus transcribed at Carnegie Mellon.

EPPS: 80h of transcribed speech from European Parliamentary Plenary Sessions (EPPS) of native and non-native English speech, which was collected within TC-Star [4].

In order to gain better understanding of which acoustic training data are most suitable for the lecture scenario, we evaluated the impact of the different data sources. Therefore, starting from our close talking system for the NIST's RT-04S meeting evaluation [5], we trained new acoustic models on different training data combinations and evaluated them on RT-05Sdev.

Data Sources	# Gaussians	WER
Meeting	237k	36.5%
Meeting + BN	350k	35.9%
Meeting + TED	251k	34.9%
Meeting + EPPS	285k	35.0%
Meeting + TED + BN	352k	35.3%
Meeting + TED + EPPS	292k	34.3%

Table 1. Impact of different data sources on RT-05Sdev.

RT-05Sdev	WER	RTF
first pass	35.2%	5.6
adaptation		1.4
second pass 10ms	28.9%	3.0
second pass 8ms	28.0%	5.0
CNC	26.8%	0.1
CNC result on RT-05Seval	26.8%	

Table 2. ASR results RT-05Sdev and RT-05Seval.

For each data combination, we trained unadapted gender-independent acoustic models consisting of 24k distributions over 6k codebooks starting with an "incremental growing of Gaussians" training procedure limited to a maximum number of 64 Gaussians per codebook, followed by 4 iterations of STC training. After that, 2 iterations of viterbi training were performed in order to compensate the negative influence of the fixed state alignments used for the previous training passes. The decision tree, used from our RT-04S meeting evaluation system [5], was kept fixed through all experiments.

For decoding, we used a 25k vocabulary with an OOV-rate of 0.18% on RT-05Sdev. The LM was computed as an interpolation of separate 3 and 4-gram LMs build on the following corpora: transcribed TED talks (93k words), conference proceedings from ICSLP, Eurospeech, ICASSP or ASRU (49M words), broadcast news text material (140M words), transcriptions of the meeting corpus (1M words) and CHIL and TC-Star text documents (192k words). The interpolation weights were tuned on a development set. The resulting 4-gram LM was pruned by removing n-grams with low probabilities resulting in 3.3M bi-, 2.7M tri- and 1.5M four-grams and a perplexity of 122 on RT-05Sdev. The pre-processing was kept unchanged from the RT-04S meeting evaluation system [5].

As can be seen from Table 1 adding TED and EPPS to the meeting data gives the largest relative improvement, because both the TED and the EPPS data cover non-native English speech. The gain from using BN data disappears, when combining them with Meeting and TED.

2.3. Lecture Recognition

Like other state-of-the-art lecture recognition systems, we used a multi pass decoding strategy with different acoustic models. The first pass uses a speaker-independent, unadapted model, the second a speaker adaptive model using VTLN and constrained MLLR. Both models consist of 24k distributions over 6k codebooks resulting in nearly 350k Gaussians. After decoding with the first model an interleaved estimation of the VTLN and MLLR parameters for the second model is performed. The lattices produced by decoding with the adapted second model with two different frame shifts of 10ms and 8ms are combined by confusion networks combination, in order to deliver the final result. The language model and 25k vocabulary described above were used.

As can be seen in Table 2 the adaptation is quite effective. One explanation for this could be, that in a lecture scenario only one speaker is speaking most of the time, and therefore sufficient adaptation material is available. The described system distinguishes itself from the January 2005 CHIL evaluation system [6] by using more acoustic training material and only 3 instead of 4 decoding passes together with tighter beams. In addition to that, we used the new 25k vocabulary and LM. Despite the lower computational cost, the final WER on the RT05Sdev task could be reduced by 8.1% absolutely and performs now as well as those presented by other sites in e.g. the NIST RT-05S evaluation [7, 8], except that manual segmentation was used.

3. MACHINE TRANSLATION

The statistical machine translation (SMT) component in our lecture translator is a phrase-based beam search decoder. In contrast to many other SMT systems a different phrase alignment is used. Typically, phrase pairs are read off the Viterbi word alignment. In our approach we view phrase alignment as a sentence splitting approach.

3.1. Phrase Alignment

To find a translation for a source phrase $\tilde{f} = f_1 ... f_l$ we restrict the general word alignment: Words inside the source phrase align to words inside the target phrase, and words outside the source phrase align outside the target phrase. We calculate this constrained alignment probability by using the well-known IBM1 word alignment model, but restrict the summation of the target words to the appropriate regions in the target sentence. Also, the position alignment probabilities are adjusted accordingly [9]. Optimization is over the target side boundaries i_1 and i_2.

$$p_{i_1,i_2}(f|e) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{1}{k} p(f_j|e_i) \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \quad (1)$$

Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words. To find the optimal target phrase we interpolate the log probabilities and take the pair (i_1, i_2) that gives the highest probability. The interpolation factor c can be estimated on a development test set.

The scores calculated in the phrase alignment are alignment scores for the entire sentence. As phrase translation probabilities we use the second term in Eqn. 1.

3.2. Decoder

The decoder used in the translation experiments is a beam search decoder, which allows for restricted word reordering. The different models used in the decoder are: 1. The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus according to the new alignment method described in this paper. 2. A trigram language model. The SRI language model toolkit was used to train the models [2]. Modified Kneser-Ney smoothing was used throughout. 3. A word reordering model, which assigns higher costs to longer distance reordering. We use the jump probabilities $p(j|j')$ of the HMM word alignment model

[10] where j is the current position in the source sentence and j' is the previous position. 4. A very simple sentence length model, which gives a constant bonus for each word generated. This is essentially used to compensate for the tendency of the language model to prefer shorter translations. Each model score is multiplied by a scaling factor, which can be modified to tune the overall system.

The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations and, if available, other specific information like named entity translation tables 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, i.e. costs. 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 laying or phrases starting within a given look-ahead window from the current position [11].

3.3. Training Data

One of the core tasks of TC-STAR is the recognition and translation of European Parliamentary Plenary Session (EPPS). For the purpose of speech recognition and translation research, an English-Spanish EPPS corpus [4] was created by RWTH Aachen within TC-STAR. We used the above mentioned English-Spanish corpus to train the English-to-Spanish translation system. The English-to-German models were trained on the EPPS data as provided by Philipp Koehn [12]. The corpus statistics of the preprocessed EPPS training corpora are shown in Table 3.

	English-Spanish		English-German	
	EN	ES	EN	DE
Sentences	1,162,176		966,526	
Words	27,7m	28,9m	23,1m	21,4m
Vocabulary	93,157	130,473	81,902	247,337

Table 3. Training and data statistics for EPPS.

3.4. Results

Table 4 shows the translation results on the unseen EPPS test set with similar characteristics as the training data. The test set consists of 1093 sentences with 26,826 running words. Of the 3,781 different word types 146 where not seen in the English-Spanish training data, and 178 where not seen in the somewhat smaller English-German training corpus. Results are reported using the well known Bleu and NIST mteval scores. The lower scores for the English to German direction as compared with the English to Spanish direction mainly reflect that translating into German is more difficult, due to language characteristics such as compounds, morphology, and word order.

	Bleu	NIST
English-Spanish, text input	31.0	7.44
English-German, text input	18.5	5.91

Table 4. SMT results on EPPS with one reference translation.

4. SPEECH TRANSLATION OF LECTURES

In this section we focus on speech translation of lectures. For this we combined the ASR system for lectures with the SMT system for EPPS described above. While the ASR system is already trained on lectures, but in different domains, the speeches used for training the SMT system distinguish themselves from the lectures by a cleaner speaking style. Therefore, it is necessary to adapt both, the ASR and SMT system, to this new situation.

	WER	RTF	PPL	OOV
dev	14.2%	2.5	155	0.69%
t035	11.2%	2.1	140	0.36%
t036+	10.5%	2.1	165	0.46%

Table 5. ASR results on the three lectures.

4.1. Data

As development and evaluation data we selected three different lectures, which were held in non-native English by the same speaker on different topics and were recorded with close talking microphones.

dev: This 24min talk was held to give a broad overview of current research projects in our lab and therefore ideal as development set.

t035: A 35min talk held as a conference key-note, but only partly covered by the dev talk.

t036+: A 31min talk on the same topic as t035, but held in a different environment and situation.

4.2. Speech Recognition

Lectures are an ideal scenario for doing supervised adaptation, because a lecture is being held by only one speaker. Therefore, we estimated the VTLN and MLLR parameters on a small amount of data for that speaker. In order to compensate the difference in channels between the training and testing conditions, we use incremental constrained MLLR adaptation. As we focus on open domain speech recognition we adapted only the LM, by tuning the interpolation weights on the dev talk and kept the vocabulary fixed. Thereby the perplexity could be reduced from 185 to 155.

As can be seen in Table 5 we got a WER of 14.2% on the dev talk. It seems, that this talk is a little bit more difficult than the other two, first due to a more worse recoding quality, which we found out later and also maybe due to the broader domain of that talk, which also explains the higher OOV rate. Nevertheless, the OOV rates are still pretty low. Even though t036+ has a higher perplexity and OOV rate, the WER is the best among all three talks. The explanation therefore is again the higher background noise in t035. t035 was recorded in lecture hall, whereas t036+ in a smaller (quieter) room in our lab.

4.3. SMT

We used the translation systems trained on the EPPS data to translate the three lectures from English to Spanish and German. The results for German are not available for all talks, because the reference translations were not finished yet. Table 6 show the translation results on manual transcripts and ASR hypotheses. As expected, translating ASR hypotheses reduces translation scores and the performance degradation seems to be roughly linear to the rise in word error rate.

Overall, translation performance for these lectures is lower than for the EPPS test set reported in section 3.4. This appears to stem not so much from the slightly higher out-of-vocabulary rate,

talk	direction	input type	Bleu	NIST
dev	English-Spanish	text input	15.5	4.72
dev	English-Spanish	ASR input	11.9	4.22
dev	English-German	text input	14.7	5.26
dev	English-German	ASR input	11.7	4.61
t035	English-Spanish	text input	12.1	4.93
t035	English-Spanish	ASR input	10.2	4.19
t035	English-German	text input	13.8	4.87
t035	English-German	ASR input	11.0	4.06
t036+	English-Spanish	text input	15.5	4.83
t036+	English-Spanish	ASR input	11.2	4.14

Table 6. Lecture translation results.

which was about 5-6% for the lectures vs. 4-5% for the EPPS test set, but rather from the much more conversational style of these lectures. Disfluencies and spontaneous speech effects present an additional challenge versus the cleaned speech of EPPS. It should also be noted that the EPPS test set was manually segmented into sentences before being passed to the translation system. In contrast, for the lecture data automatic segmentation based on speaker pauses was used.

To compensate the more conversational style we adapted our LMs to the new situation, by collecting web data. Therefore we used the tools provided by the University of Washington and created search queries out of the dev talk by using the top 850 3-grams. The resulting data containing 175M words were further filtered by a tfidf and perplexity based selection method. Herewith, we first select an amount of x closest matching sentences to the dev talk with the help of tfidf. This set is then divided into two sub-sets, by incrementally adding n sentences to the first set, if they decrease the perplexity of that set compared to the dev talk or to the second set, if not. In our case this led to sets of 2.4M and 16.4M words. The final LM is a 4-fold interpolation of separate LMs build over the two selected sets, the complete web data, and EPPS tuned on the dev talk. Thereby the perplexity of the dev talk could be reduced from 578 with the original EPPS LM to 209 by only interpolating with the web data and further to 134 with the tfidf & PPL based selection.

We evaluated the adapted LM only on the Spanish part of t036+ to see, if even the best translation results could be further improved. As can been seen from Table 7 the Bleu score could be improved by approx. 15% relative and the NIST score by 6%, which is a result of the perplexity reduction from 510 to 366.

talk	direction	input type	Bleu	NIST
t036+	English-Spanish	text input	18.2	5.11
t036+	English-Spanish	ASR input	12.9	4.44

Table 7. Lecture translation results after LM adaptation.

5. CONCLUSION

In this paper we have presented our work in taking first steps towards building open domain speech translation systems. We have successfully combined the lecture recognition system from the CHIL project with the translation system used in TC-Star to translate lectures on a new domain from English to Spanish and German.

The performance on the recognition side outperforms our expectations showing the feasibility of designing open domain recognition systems. We achieved WERs in the range of 14% to 10% depending on the quality and domain of the talk.

For translation, lectures still pose a significant challenge. The results for English-to-Spanish lectures were worse than the results for parliamentary speeches, which could be an effect of the more spontaneous speaking style, but also the significant mismatch between training and testing data. Nevertheless, the results could be significantly improved, by doing tfidf & PPL based language model adaptation.

6. ACKNOWLEDGMENTS

We would like to thank Susanne Burger, Maria Kernecker, and Tim Notari for transcribing and translating the lecture evaluation data. This work was partly funded by the *European Union* (EU) under the integrated projects CHIL (Grant number IST-506909) and TC-Star (Grant number IST-506738).

7. REFERENCES

[1] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *ASRU*, Trento, Italy, 2001.

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

[3] "Linguistic data consortium," http://www.ldc.upenn.edu.

[4] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus," *ICASSP*, 2005.

[5] F. Metze, Q. Jin, C. Fügen, Y. Pan, and T. Schultz, "Issues in Meeting Transcription – The Meeting Transcription System," in *ICSLP*, Jeju Island, Korea, 2004, ISCA.

[6] M. Wölfel and J. McDonough, "Combining Multi-Source Far Distance Speech Recognition Strategies: Beamforming, Blind Channel and Confusion Network Combination," in *INTERSPEECH*, Lissabon, Portugal, 2005, ISCA.

[7] A. Stolcke, X. Anguera, K. Boakye, Ö. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C.Wooters, and J. Zheng, "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System," in *NIST RT-05 Meeting Recognition Workshop*, United Kingdom.

[8] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI System for the Transcription of Speech in Meetings," in *NIST RT-05 Meeting Recognition Workshop*, Edinburgh, United Kingdom, 2005.

[9] S. Vogel, "PESA: Phrase Pair Extraction as Sentence Splitting," in *Machine Translation Summit 2005*, Thailand.

[10] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Translation," in *COLING 96*, Copenhagen, 1996, pp. 836–841.

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

[12] P. Koehn, "Europarl: A Multilingual Corpus for Evaluation of Machine Translation," http://people.csail.mit.edu/koehn/publications/europarl.