CROSS DOMAIN AUTOMATIC TRANSCRIPTION ON THE TC-STAR EPPS CORPUS

Christian Gollan, Maximilian Bisani, Stephan Kanthak, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{gollan,bisani,kanthak,schluter,ney}@informatik.rwth-aachen.de

ABSTRACT

This paper describes the ongoing development of the British English European Parliament Plenary Session corpus. This corpus will be part of the speech-to-speech translation evaluation infrastructure of the European TC-STAR project.

Furthermore, we present first recognition results on the English speech recordings. The transcription system has been derived from an older speech recognition system built for the North-American broadcast news task. We report on the measures taken for rapid cross-domain porting and present encouraging results.

1. INTRODUCTION

A speech-to-speech translation (SST) system consits of three components: automatic speech recognition (ASR), machine translation and speech synthesis. The development of new approaches for SST demands corpora from a single domain to improve and evaluate all of these components in interaction.

The optimization of a statistical ASR system requires large task representive databases for acoustic and language model training. Today's ASR systems can be still improved by the use of 1,000 hours of speech data in the acoustic model training [1]. But the manual transcription of speech data is very time consuming and costly. Depending on the difficulty of the transcription task and on the level of consistency and correctness one wants to achieve, the effort ranges is between 10 and 100 man hours for one hour of recorded speech. However, it was reported that the impact of transcription errors on the performance of thereby trained ASR systems is not as bad as one might expect. Sundaram and Picone [2] report that 16% falsely labeled training data leads to an performance loss of 8.5% relative to the baseline on the Switchboard corpus. It can be assumed that the use of additional cheap but inaccurately transcribed speech data can redeem this performance loss.

This work was partially funded by the European Commission Union under the Human Language Technologies project TC-STAR (FP6-506738).

Unsupervised training relies on methods for the automatic transcription of speech data. Such methods are based on an initial ASR system which is then iteratively improved by its auto-generated transcripts. The transcription errors can be reduced by filtering the most likely error-prone data. Such filter methods can be based on confidence measures [3] or if available on aligned closed captions [1]. The initial automatic transcription system can be trained by a small manually transcribed subset of the audio data [3]. An attractive alternative is using an existing ASR system which was developed for a similar task. Due to mismatching conditions (domain, dialect, acoustic environments, transmission channel, etc.) the accuracy of such an initial automatic transcript may be poor, but some steps can be taken to lighten this problem. This paper describes such a cross-domain porting of an existing ASR system for initial automatic transcription.

2. TC-STAR

The TC-STAR project (Technology and Corpora for Speech to Speech Translation) is envisioned as a long term effort focused on advanced research in all core technologies for SST [4]. The project will target a selection of unconstrained conversational speech domains - i.e. broadcast news, political speeches, and discussion forums - and a few languages relevant for Europe's economy and society: European English, European Spanish and Chinese. The technical challenges and objectives of the project will focus on the development of new algorithms and methods, integrating relevant human knowledge which is available at translation time into a data-driven framework. Examples of such new approaches are the extension of statistical machine translation models to handle multiple sentence hypothesis produced by the speech recognizer, the integration of linguistic knowledge in the statistical approach of spoken language translation, the statistical modeling of pronunciation of unconstrained conversational speech in automatic speech recognition, and new acoustic and prosodic models for generating expressive speech in speech synthesis.

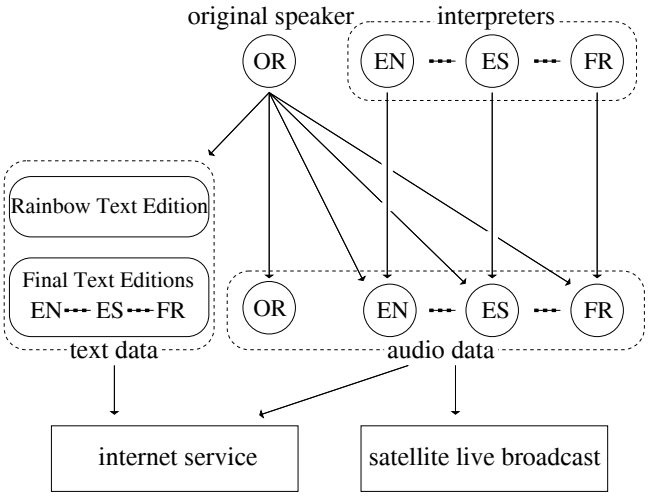

Fig. 1. Overview of the available EPPS resources

One of the project goals is the implementation of an evaluation infrastructure based on competitive evaluation, in order to achieve the desired breakthroughs in SST. The European Parliament Plenary Session (EPPS) is an attractive domain for the development of such a new evaluation infrastructure. The following section describes this domain and our ongoing data collection effort. At the end of the project, the resulting language resources will be made available to the public through ELDA/ELRA [5].

3. THE EPPS CORPUS

The *European Parliament* (EP) holds plenary sessions usually six days each month. The major part of the sessions take place in Strasbourg (France) while the residual sessions are held in Brussels (Belgium). Today the European Parliament consists of members from 25 countries, and 20 official languages are spoken. The sessions are chaired by the President of the European Parliament. Typically when the president hands over to a member of the parliament, the speaker's microphone is activated. Interjections from the Parliament are therefore softened in the recording. Simultaneous translations of the original speech are provided by interpreters in all official languages of the EU. Figure 1 gives an overview of the structure of existing EPPS data.

It is possible to categorize speakers in two ways: Firstly there are native speakers well as non-native speakers who have more or less pronounced accent. Secondly there are original speakers and interpreters. Although most of the speeches are planned, almost all speakers exhibit the usual effects known from spontaneous speech (hesitations, false starts, articulatory noises). The interpreters' speaking style is somewhat choppy: dense speech intervals ("bursts") alternate with pauses when the interpreter is listening to the original speech.

Rainbow Text Edition	**Verbatim Transcription**
It is for our Parliament, as we have already marked in a symbolic ceremony *outside*, a special and extraordinary moment. *In Dublin last Saturday,* Ireland's Nobel literature laureate Seamus Heaney *captured this special event with the words ...*	It is for our Parliament, as we have already marked in a symbolic ceremony *outdoor*, a special and extraordinary moment. *It was described in Dublin last Saturday captured in the words of* Ireland's Nobel literature laureate Seamus Heaney, *he talked about and I quote ...*

Fig. 2. Excerpt of a Rainbow Text Edition and the corresponding transcript out of the EPPS early test data

The European Union's TV news agency, *Europe by Satellite (EbS)* [6], provides Europe related information via internet and satellite. EbS broadcasts the EP Plenary Sessions live in the original language and the simultaneous translations via satellite on different audio channels: one channel for each official language of the EU and an extra channel for the original untranslated speeches. These channels are additionally available as 30 minute long internet streams for one week after the session. The audio transmissions are monaural. The internet audio streams have a sample rate of 16 kHz and are encoded with the RealAudio Sipro codec at a bit rate of 16 kbit/s. The satellite audio streams have a sample rate of 48 kHz and are encoded with the MPEG 1 layer II codec at a bit rate of 64 kbit/s.

In May 2004 we have started recording the EPPS broadcasts in five languages (English, Spanish, German, French, and Italian). These *EPPS recordings* from May to July were made from the EbS transmissions. While in May the recordings were made from internet audio streams, both sources (internet and satellite) were recorded in July. In this period we have recorded a total of 25 hours for each language. The internet streams from May and the satellite streams from July have been selected for transcription. We are currently in progress of manually segmentating and transcribing the English audio streams. As of this writing an early non-validated transcription of the EPPS recording of May the 3rd has been produced. This labeled subset has a duration of one hour and will be referred to as the *EPPS early test data* in the remainder of the paper.

The compilation of texts of the speeches given by members of the European Parliament in plenary sessions is known as the *Rainbow Text Edition* (RTE). Every speech in these reports appears in the language used by the speaker who is allowed to make corrections to the text afterwards.

The reports are published on the EUROPARL web site [7] on the day after the EPPS. The Final Text Edition (FTE) in all official languages of the EU is accessible about two months later. The web site also provides all previous reports since April 1996. We currently work with the available reports to build an English-Spanish parallel text corpus for the TC-STAR project. The RTE and FTE aim for high readability, and therefore do not provide a strict word-by-word transcript. Notable deviations from the original speech include removal of hesitations, false starts and word interruptions. Furthermore transposition, substitution, deletion and insertion of words can be observed in the reports. An example is given in Figure 2.

Table 1. Statistics of speech corpora

	Hub-4		EPPS
	training	test	early test
acoustic data	96.5h	2.9h	1.0h
silence portion	14%	12%	8%
# speakers	$\approx 3,157$	≈ 116	≈ 22
# utterances	26,136	728	442
# running words	1,053,050	32,834	8,782

4. SYSTEM DESCRIPTION

We have conducted initial recognition experiments with two objectives in mind: use of automatic transcriptions to assist human transcribers and evaluate the potential of applying unsupervised training methods.

The experiments to be reported in this paper were performed with our single-pass across-word, trigram recognizer. The recognition vocabulary comprises 65k words. More details about the system are given in [8]. The baseline system was setup for the the 1997 Hub-4 data from the DARPA benchmark evaluation. This corpus consists of transcribed American English broadcast news recordings. Table 1 gives an overview of the corpus statistics.

The acoustic vectors are computed by applying a linear discriminant analysis on several adjacent vectors consisting of 16 mel-frequency cepstral coefficients without derivatives. The gender dependent acoustic models consists of triphones which are represented by 6-state HMMs with skip, forward and loop transitions. Gaussian mixtures with a globally pooled diagonal covariance are used for modeling the HMM states which are tied using a decision tree. Silence is modeled using a single state HMM which is separated from the state of the other HMMs and not included in state tying. During training maximum approximation is applied.

Our baseline system achieves a word error rate (WER) of 19.3% on the Hub-4 1997 evaluation test corpus after NIST scoring. Best published results are around 14% WER, however these systems use multiple passes, speaker adaptation and 4-gram language models [9] [10].

5. LANGUAGE MODEL

The first step in improving performance on the EPPS data was building a new in-domain language model (LM). Therefore we used the data of the preprocessed EPPS reports from our English-Spanish EPPS parallel text corpus, which was further normalized: e.g. abbreviations and numbers were written out in full. The English text contains 30 million running words and does not contain any EPPS reports covering the time period of the recordings. (These are not available yet.) From this data a trigram LM was built using the SRI Language Modeling Toolkit [11] applying absolute discounting with interpolation (modified Kneser-Ney smoothing) [12].

6. VOCABULARY PORTING

Clearly the OOV rate of the EPPS data with the Hub-4 language model is rather high. To alleviate this problem we have added the most frequent 7,000 words from the EPPS data that were missing in the recognition vocabulary. To provide the phonetic transcriptions for the new words we have used the data-driven grapheme-to-phoneme conversion approach described in [13]: a grapheme-to-phoneme conversion model was trained on the existing Hub-4 lexicon, and then used to generate pronunciations for the newly added words. This approach is domain and language independent and requires no human expertise in phonetics or English pronunciation.

There is a remaining mismatch is in the pronunciation lexicon: The Hub-4 pronunciation lexicon and acoustic model are designed for North American English. The EPPS data however include native British Speakers as well as non-english speakers who approximate British or American English pronunciation at various levels of competence. Since acoustic models are strongly tied to the pronunciation dictionary they were trained with, it is not possible to simply change the latter to better match the observed pronunciations.

7. EXPERIMENTS

As a first test, we ran the Hub-4 ASR system as it was on the EPPS recordings. Considering the large mismatch in domain and speaking style, the recognition results were

surprisingly good. Table 2 gives an overview of the experimental results on the one hour long labeled EPPS early test data. The word error rate (WER) of the system was measured with respect to the recently made manual transcripts. The unmodified Hub-4 baseline system achieves a WER of 39.1% on this test data. As expected the perplexity of the out-domain Hub-4 LM on the EPPS transcripts is quite high.

Table 2. Experimental results with the Hub-4 system on the EPPS early test data

acoustic model	lexicon	language model	WER	perpl.	OOV
Hub-4	Hub-4	Hub-4	39.1%	207.1	1.7%
Hub-4	Hub-4	EPPS	34.4%	168.8	1.7%
Hub-4	EPPS	EPPS	33.9%	167.4	1.0%

To improve the performance we built a new LM from the English FTE documents restricting the vocabulary to that of the existing Hub-4 system. The perplexity of this LM is significantly lower than that of the Hub-4 LM, leading to a 12% relative reduction in WER.

In the next experiment we enlarged the Hub-4 lexicon with the most frequent 7,000 missing words from the English FTE documents vocabulary. This was done with grapheme-to-phoneme conversion. This enlarged vocabulary was used for building a new LM. The so modified Hub-4 system achieves 33.9% WER on the EPPS early test data.

8. CONCLUSION

We have reported on the EPPS corpus currently being developed in the TC-STAR project. One project goal in developing this corpus is the creation of an evaluation infrastructure for SST.

We generate manual transcripts of the acoustic data. To alleviate this time consuming task we consider the use of automatic transcription systems to aid human transcribers as well as the use of unsupervised training methods. This article describes the first steps in this direction. We presented the rapid and cheap cross-domain porting of an existing Hub-4 system, which has improved the recognition performance by 13% WER relative (from 39% to 34% WER).

9. REFERENCES

[1] L. Nguyen and B. Xiang, "Light Supervision in Acoustic Model Training," in *2005 IEEE International Conference on Speech, Acoustics, and Signal Processing*, March 2004, pp. 185–188.

[2] R. Sundaram and J. Picone, "Effects on Transcription Errors on Supervised Learning in Speech Recognition," in *2005 IEEE International Conference on Speech, Acoustics, and Signal Processing*, March 2004, pp. 169–172.

[3] F. Wessel and H. Ney, "Unsupervised Training for Broadcast News Speech Recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, December 2001.

[4] "TC-STAR: Technology and Corpora for Speech to Speech Translation," http://www.tc-star.org.

[5] Evaluations and Language resources Distribution Agency, "ELDA," http://www.elda.fr.

[6] European Union's TV news agency, "Europe by satellite," http://europa.eu.int/comm/ebs/.

[7] The Secretariat of the European Parliament, "EUROPARL: Plenary Session reports," http://www.europarl.eu.int/plenary/.

[8] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, and H. Ney, "Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2000, pp. 1671–1674.

[9] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and Michèle Jardino, "Recent advances in transcribing television and radio broadcasts," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 2, pp. 655–658.

[10] Long Nguyen, Spyros Matsoukas, Jason Davenport, Daben Liu, Jay Billa, Francis Kubala, and John Makhoul, "Further advances in transcription of broadcast news," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 2, pp. 667–670.

[11] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. Intl. Conf. Spoken Language Processing*, September 2002.

[12] Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359 – 394, Oct. 1999.

[13] Maximilian Bisani and Herman Ney, "Multigram-based grapheme-to-phoneme conversion for LVCSR," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, vol. 2, pp. 933 – 936.