FBK @ IWSLT 2013 - ASR tracks

D. Falavigna, R. Gretter, F. Brugnara, D. Giuliani, R. H. Serizel

HLT research unit, FBK, 38123 Povo (TN), Italy

(falavi,gretter,brugnara,giuliani,serizel)@fbk.eu

Abstract

This paper reports on the participation of FBK at the IWSLT2013 evaluation campaign on automatic speech recognition (ASR): precisely on both English and German ASR track. Only primary submissions have been sent for evaluation.

For English, the ASR system features acoustic models trained on a portion of the TED talk recordings that was automatically selected according to the fidelity of the provided transcriptions. Two decoding steps are performed interleaved by acoustic feature normalization and acoustic model adaptation. A final step combines the outputs obtained after having rescored the word graphs generated in the second decoding step with 4 different language models. The latter are trained on: out-of-domain text data, in-domain data and several sets of automatically selected data.

For German, acoustic models have been trained on automatically selected portions of a broadcast news corpus, called "Euronews". Differently from English, in this case only two decoding steps are carried out without making use of any rescoring procedure.

1. Introduction

The IWSLT 2013 Evaluation Campaign, similarly to the one carried out for IWSLT2012 [1], addresses the automatic transcription/translation of TED Talks [1]: a collection of public speeches on a variety of topics.

This year, for the transcription of English audio tracks we have focused on automatic selection and exploitation of training data, both audio and text.

We have trained acoustic models (AMs) on both in-domain audio data, extracted from videos downloaded from TED talk WEB site (i.e. http:// www.ted.com/talk/), and out-of-domain data including the broadcast news speech corpus "HUB4" provided by linguistic data consortium (LDC). Since audio recordings of TED talks have only associated "non-exact" transcriptions, a lightly supervised training approach [2] has been applied in order to select reliable data for AM training.

For language model (LM) training, out-of-domain data come from several sources and contain about 5 billions (5G) of words. In addition, a set of in-domain data, containing about 2.7 millions (2.7M) of words, has been provided by organizers. Then, similarly to what done in our ASR submission of last year [3], we have used the automatic transcription of each given English TED talk for automatically selecting from the out-of-domain text data a set of 100M words. From each text source a corresponding LM was trained and used for rescoring word graphs (WGs) generated in the second decoding step. In addition, an interpolated LM, resulting from the linear interpolation of all of the different LMs, has been used for rescoring. Our primary submission has been obtained after having combined, using the ROVER approach [4], all of the different rescored ASR hypotheses.

German AMs were trained using "Euronews" videos downloaded in the last few years from the portal http://de.euronews.com/. Since each video has associated a reference text, that doesn't not contain the exact transcription of the corresponding audio track, we have applied also in this case a lightly supervised approach [2] for AM training. Doing this, about 256 hours of audio data, including silences, were selected. Cut-off date for the latter data was March 2013.

German data for LM has been first normalized applying a procedure that split numbers and compound words automatically found inside training documents. One 4-gram LM has been trained on about 1.7G of words, coming from news, European Parliament, IWSLT13 training data. Cut-off of training data date was end of June 2012.

2. Automatic transcription systems

In this section we summarize the main features of the FBK primary systems used for transcribing TED talks delivered in English and German. This year, differently from previous evaluation campaigns, time boundaries of speech segments to be transcribed are not given. Hence, automatic speech segmentation has to be carried out.

2.1. Automatic speech segmentation

The input audio signal is first divided into segments by a start-end point detector module. The obtained segmentation is refined using an acoustic classifier, based on Gaussian mixture models (GMMs), which also performs classification of segments into several classes including non-speech classes [5]. Then, the obtained homogeneous non-overlapping speech segments are clustered by using a method based on the Bayesian information criterion. At the end of this process, each audio file to transcribe has assigned a set of temporal segments, each having associated a label that indicates the cluster to which it belongs (e.g. "female_1", "male_1", etc). The resulting segmentation and clustering is then exploited by the recognition system to perform cluster-wise feature normalization and acoustic models adaptation

[1] http://www.ted.com/talks

during two decoding passes described below.

3. English transcription system

3.1. Acoustic data selection

For AM training, HUB4 speech corpus was initially used. It contains around 164 hours of broadcast news speech with related word transcriptions, that include also "filler-words". These latter ones have been mapped into 6 different "spontaneous speech" models. After having trained triphone Hidden Markov Models (HMMs) on HUB4, domain specific acoustic data (i.e. a certain number of TED talks recordings) were exploited for lightly supervised training [2].

Recordings of TED talks released before the cut-off date, 31 December 2010, were downloaded with the corresponding subtitles which are content-only transcriptions of the speech. In content-only transcriptions anything irrelevant to the content is ignored, including most non-verbal sounds, false starts, repetitions, incomplete or revised sentences and superfluous speech by the speaker. A simple but robust procedure was implemented to select only audio data with an accurate transcription.

The collected data consisted in 820 talks, for a total duration of $\sim$216 hours, with $\sim$166 hours of actual speech. The provided subtitles are not a verbatim transcription of the speeches, hence the following procedure was applied to extract segments that can be deemed reliable. The approach is that of selecting only those portions in which the human transcription and an automatic transcription agree. To this end, a "background" 4-gram language model was first trained on all the talk transcriptions. Subsequently, a specific Language Model (LM) was built for each talk by adapting the language model to the human transcription of the talk. A preliminary automatic transcription was performed on the talks with the pre-trained HUB4 AM and the talk-specific LM (note that, in doing this, optional "spontaneous speech" models were allowed among words). The output of the system was aligned with the reference transcriptions, and the matching segments were selected, resulting in an overlap of $\sim$120 hours of actual speech out of the total of 166. By using these segments together with the segments labeled as silence, a TED-specific acoustic model was trained, as detailed in the following section. The label/select/train procedure was repeated two more times, resulting in a portion of selected actual speech that grew to $\sim$142 hours and then to $\sim$144 hours. Given the modest improvement in the third iteration, the procedure was not repeated further. In conclusion, the method made available 87% of the training speech, which was considered satisfactory.

In total, after automatic selection, we get around 307 hours ($\sim$164 hours from HUB4 plus $\sim$144 hours from TED recordings) of transcribed speech for training acoustic models.

3.2. AM training

Thirteen Mel-frequency cepstral coefficients, including the zero order coefficient, are computed every 10ms using a Hamming window of 20ms length. First, second and third order time derivatives are computed after segment-based cepstral mean subtraction to form 52-dimensional feature vectors. Acoustic features are normalized and HLDA-projected to obtain 39-dimensional feature vectors as described below.

AMs were trained exploiting a variant of the speaker adaptive training method based on Constrained Maximum Likelihood Linear Regression (CMLLR) [6]. In our training variant [7, 8, 9] there are two sets of AMs: the target models and the recognition models. The training procedure makes use of an affine transformation to normalize acoustic features on a cluster by cluster basis with respect to the target models. For each cluster of speech segments, an affine transformation is estimated through CMLLR [6] with the aim of minimizing the mismatch between the cluster data and the target models. Once estimated, the affine transformation is applied to cluster data in order to normalize acoustic features with respect to the target models. Recognition models are then trained on the normalized data. Leveraging on the possibility that the structure of the target and recognition models can be determined independently, a Gaussian Mixture Model (GMM) can be adopted as the target model for training AMs used in the first decoding pass [7]. This has the advantage that, at recognition time, word transcriptions of test utterances are not required for estimating feature transformations. Instead, target models for training recognition models used in a second or third decoding pass are usually triphones with a single Gaussian per state [8]. In all cases, the same target models are used for estimating cluster-specific transformations during training and recognition.

In the current version of the system, a projection of the acoustic feature space based on Heteroscedastic Linear Discriminant Analysis (HLDA) is embedded in the feature extraction process as follows. A GMM with 1024 Gaussian components is first trained on an extended acoustic feature set consisting of static acoustic features plus their first, second and third order time derivatives. For each cluster of speech segments, a CMLLR transformation is then estimated w.r.t. the GMM and applied to acoustic observations. After normalizing the training data, an HLDA transformation is estimated w.r.t. a set of state-tied, cross-word, gender-independent triphone HMMs with a single Gaussian per state, trained on the extended set of normalized features. The HLDA transformation is then applied to project the extended set of normalized features in a lower dimensional feature space, that is a 39-dimensional feature space. Recognition models used in both the first and second decoding pass are trained from scratch on normalized HLDA-projected features. HMMs for the first decoding pass are trained through a conventional maximum likelihood procedure. Recognition models used in the second decoding pass are speaker-adaptively trained, exploiting as target-models triphone HMMs with a single Gaussian density per state.

For each phone set and decoding pass, a set of state-tied, cross-word, gender-independent triphone HMMs were trained for recognition. Around 170,000 Gaussian densities, with diagonal covariance matrices, were allocated for each model set.

3.3. LM training

Text data used for training LMs are those released for the IWSLT2013-SLT Evaluation Campaign. Before training LMs, texts were cleaned, normalized (punctuation was removed, numbers and dates were expanded) and double lines were removed. Then, they have been grouped into the following three sets, on which a corresponding LM was trained:

- **giga5** GIGAWORD 5-th edition. Contains documents stemming from seven distinct international sources of English newswire. It is released from the Linguistic Data Consortium (see http://www.ldc.upenn.edu/). In total it contains about 4G words.
- **wmt13** Formed by documents in WMT12 news crawl, news commentary v7 and Europarl v7 (see IWSLT2013 official web site for some more details about these corpora). In total it contains about 1G words.
- **ted13** An in-domain set of texts extracted from TED talks transcriptions used for training. It contains about 2.7M words.

For each of the three sources listed above, we trained a 4-gram backoff LM using the modified shift beta smoothing method as supplied by the IRSTLM toolkit [10]. The three LMs CONTAIN, respectively, about:

- **giga5** 130M bigrams, 231M 3-grams, 422M 4-grams;
- **wmt13** 64M bigrams, 69M 3-grams, 92M 4-grams;
- **ted13** 687K bigrams, 223K 3-grams, 132K 4-grams.

Word pronunciations in the lexicon are based on a set of 45 phones. They were generated by merging different source lexica for American English (LIMSI '93, CMU dictionary, Pronlex). In addition, phonetic transcriptions for a number of missing words were generated by using the phonetic transcription module of the Festival speech synthesis system.

The **wmt13** LM was used to compile a static Finite State Network (FSN) which includes LM probabilities and lexicon for the first two decoding passes. The latter LM was pruned in order to obtain a network of manageable size, resulting in a recognition vocabulary of 200K words and into about: 42M bigrams, 34M 3-grams and 31M 4-grams.

As seen in section 1 the ASR hypotheses generated in the second decoding step were used to automatically select documents from all of the available out-of-domain data, i.e. **giga5** and **wmt13**. To do this we employed a similarity measure based on the well known TFxIDF (term frequencies x inverse document frequencies) [11] features. More specifically, we selected 100M of words for each given TED talk and trained a corresponding talk-dependent LM (in the following we will refer the latter with **aux100M**). Details of the automatic selection approach can be found in [12].

3.4. Word graphs rescoring

Word graphs are generated in the second decoding step. To do this, all of the word hypotheses that survive inside the trellis during the Viterbi beam search are saved in a word lattice containing the following information: initial word state in the trellis, final word state in the trellis, related time instants and word log-likelihood. From this data structure and given the LM used in the recognition steps, WGs are built with separate acoustic likelihood and LM probabilities associated to word transitions. To increase the recombination of paths inside the trellis and consequently the densities of the WGs, the so called word pair approximation [13] is applied. In this way the resulting graph error rate was estimated to be 6.0% on the development set used for IWSLT2011 evaluation campaign (i.e. 19 TED talks), about $\frac{1}{3}$ of the corresponding WER, that resulted to be 17.6%.

WGs are rescored using an interpolated LM that combine all of the four LMs described above, **giga5**, **wmt13**, **aux100M** and the in-domain LM **ted13**. To do this, the original LM probability on each arc of each WG is substituted with the linearly interpolated probability. Note that the development set used to train the interpolation weights is again the ASR output of the second decoding step and, therefore, talk specific interpolation weights are estimated. Note also that acoustic model probabilities associated to arcs of WGs remain unchanged.

In addition WGs were rescored using singularly each one of the above mentioned LMs, thus obtaining 5 different outputs for each automatically transcribed talk (including the ones obtained with the interpolated LM). These latter ASR output hypotheses have been combined, using ROVER, in order to produce the final submission. Note that the latter final ROVER combination makes use of word confidence measures.

4. German transcription system

German ASR makes only use of first and second decoding passes described for English ASR. For German we didn't perform any data selection, in order to build focused LMs, as well as any WG rescoring step.

4.1. AM training

German AMs were trained using Euronews videos downloaded in the last few years from the portal http://de.euronews.com/. Each video has associated a reference text, that could be just a summary, an accurate transcription of the news, or the transcription of a part of the news. We apply lightly supervised training, in a way similar to that described for English ASR, to select segments for training. Three iterations have been used before stopping the selection process, resulting into about 256 hours of training audio data, including silences.

4.2. Linguistic processing and LM training

In German, compound words are a significant percentage of the common lexicon, and should be taken into account to avoid unacceptable out-of-vocabulary (OOV) rate. We built an automatic system that, given a lexicon of German words ordered by frequency, decides which words have to be considered as compounds and propose a splitting.

We extracted form the lexicon a set of words that can

be considered "basewords". These latter words are shorter than a predefined threshold (e.g. 15 characters) and exhibit a frequency higher than another threshold (e.g. greater than 2).

Then we defined, by hand, a "falsebasewords" file which contain some acronyms (17 in the actual version, namely: der die das er es sc sch fts ic des wal sge ger cht ati rwe ler) than cannot be basewords but that were frequently observed. The defined acronyms are used to form wrong decompositions.

Finally, an algorithm was implemented that detects if a suspected compound word can be obtained by concatenating basewords. Among the possible decompositions, the one is chosen which minimizes a cost function favoring longer words. Some German compound rules were added to the algorithm, that basically allow the insertion of the suffixes "s","n","es","en". A sample of decompositions is given in Table 1.

compound word	decomposition
Krankenversicherung	kranken+ +Versicherung
Ministerpräsidenten	Minister+ +Präsidenten
Bundesgeschäftsführer	Bundes+ +Geschäfts+ +Führer
Sicherheitskonferenz	Sicherheits+ +Konferenz
Auseinandersetzungen	auseinander+ +Setzungen
Bundesverfassungsgericht	Bundes+ +Verfassungs+ +Gericht
Oberstaatsanwaltschaft	Oberstaatsanwaltschaft

Table 1: Example of compound word decomposition.

Finally, a method was implemented to join compound words after ASR.

A German 4-gram LM was trained after the split of numbers and compound words on a corpus, formed by crawled news and European Parliament transcriptions, containing about 1.6G of words. Cut-off date was end of June 2012. In-domain text data have also been used for LM adaptation.

5. Official results

Final results (%WER), after adjudication, of the English system for:
tst2011, primary 13.6%
tst2012, primary 16.2%
tst2013, primary 23.2%.

Final result, after adjudication, of the German system for:
tst2013, primary 37.5%.

6. Conclusions

We presented descriptions of our ASR systems used to submit runs to the IWSLT2013 Evaluation Campaign for both English and German audio track. Both systems were trained applying lightly supervised training to audio data that do not have associated "accurate" transcriptions.

English ASR system makes also use of a procedure that allows to rescore WGs with a combination of several LMs, some of them trained on sets of automatically selected data.

7. Acknowledgements

This work was partially supported by the European project EU-BRIDGE, under the contract FP7-287658.

8. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[2] L. Lamel, J. Gauvain, and G. Adda, "Investigating lightly supervised acoustic model training," in *Proc. of ICASSP*, vol. 1, Salt Lake City, USA, 2001, pp. 477–480.

[3] D. Falavigna, G. Gretter, F. Brugnara, and D. Giuliani, "Fbk @ iwslt 2012 - asr track," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[4] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, Santa Barbara, CA, 1997, pp. 347–352.

[5] M. Cettolo, "Segmentation, classification and clustering of an italian broadcast news corpus," in *Proc. of Content-Based Multimedia Inf. Access Conf. (RIAO)*, Paris, France, 2000, pp. 372–381.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[7] G. Stemmer, F. Brugnara, and D. Giuliani, "Using Simple Target Models for Adaptive Training," in *Proc. of ICASSP*, vol. 1, Philadelphia, PA, March 2005, pp. 997–1000.

[8] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization." *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.

[9] D. Giuliani and F. Brugnara, "Experiments on Cross-System Acoustic Model Adapatation," in *ASRU Workshop 2007*, Kyoto, Japan, Dec. 2007, pp. 117–122.

[10] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proc. of INTERSPEECH*, Brisbane, Australia, September 2008, pp. 1618–1621.

[11] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in *First International Conference on Machine Learning*, New Brunswick:NJ, USA, 2003.

[12] D. Falavigna and G. Gretter, "Focusing language models for automatic speech recognition," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[13] X. Aubert and H. Ney, "A word graph algorithm for large vocabulary continuous speech recognition," in *Proc. of ICSLP*, 1994, pp. 1355–1358.