FBK @ IWSLT 2012 - ASR track

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

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

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

Abstract

This paper reports on the participation of FBK at the IWSLT2012 evaluation campaign on automatic speech recognition: namely in the English ASR track. Both primary and contrastive submissions have been sent for evaluation.

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. Three decoding steps are performed interleaved by acoustic feature normalization and acoustic model adaptation.

A final rescoring step, based on the usage of an interpolated language model, is applied to word graphs generated in the third decoding step. For the primary submission, language models entering the interpolation are trained on both out-of-domain and in-domain text data, instead the contrastive submission uses both "general purpose" and auxiliary language models trained only on out-of-domain text data. Despite this fact, similar performance are obtained with the two submissions.

1. Introduction

The IWSLT 2012 Evaluation Campaign [1], similarly to the one carried out for IWSLT2011, focused on the automatic transcription/translation of TED Talks [1]: a collection of public speeches on a variety of topics.

This year, for automatic speech recognition (ASR) we mostly focused our work on language modeling, while acoustic models remained unchanged with respect to those used in the evaluation campaign of IWSLT2011. In particular, we propose a method for focusing the language models (LMs) used during a final linguistic rescoring of the word graphs produced by our ASR system, towards the ASR output of previous decoding stages, obtaining significant reduction in word error rate (WER) without the usage of in-domain text data. Although approaches similar to the one used for producing our contrastive submissions are also reported in the literature (see [2] and [3]), there are some substantial differences that make the method reported in this paper quite novel.

More specifically, we propose to apply an automatic selection procedure to the same texts employed to train the "general purpose" LMs used in the various decoding steps of the ASR system. Then, we use the set of selected documents to train auxiliary LMs which are linearly interpolated, on a talk specific basis, with the general ones in order to provide LM probabilities to a final decoding pass based on word-graphs rescoring. In this way, we are able to train LMs focused on the ASR output. We prefer to use the term "LM focusing", instead of LM adaptation, to underline the fact that we are not using new data to train auxiliary LMs but, on the contrary, a subset of existing text data is enhanced in order to better match the linguistic content of the audio to transcribe.

Since we want, in principle, to "frequently" focus LMs using the ASR output corresponding to a given (or automatically detected) segmentation of the audio stream to transcribe, we developed a technique that allows to efficiently select a subset of documents from the large set of available documents. This latter technique is based on a vector space model: each document is represented with a vector of coefficients, while a metric is defined that allows to estimate the distance between two vectors or, equivalently, the similarity between two documents. The "auxiliary" documents are hence obtained as the ones that are most similar to a given query document (i.e. to the ASR output of a piece of speech to transcribe).

The definition of the features and of the metrics have been inspired from TFxIDF (Term Frequency x Inverse Document Frequency) vector space model [4], however the employed features, the way adopted for storing them and the similarity metrics used, has allowed to improve both computation and memory efficiency with respect to TFxIDF approach.

2. Automatic transcription system

In this section we summarize the main features of the FBK primary system used in the IWSLT2012 Evaluation Campaign for transcribing TED talks delivered in English. For each talk, in addition to the audio file, time boundaries of speech segments to be transcribed are given. The word transcription of a talk is generated in three decoding passes. All the decoding passes make use of a 4-gram language model and are interleaved by acoustic feature normalization and Acoustic Model (AM) adaptation.

2.1. Acoustic data selection for training

For AM training, domain specific acoustic data were exploited. Recordings of TED talks released before the cut-off date, 31 December 2010, were downloaded with the corre-

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

sponding 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 a pre-trained general AM for English and the talk-specific LM. 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.

2.2. Acoustic model

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) [5]. In our training variant [6, 7, 8] there are two sets of AMs: the target models and the recognition models. For each cluster of speech segments, an affine transformation is estimated through CMLLR [5] 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 [6]. 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 [7]. 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 Hidden Markov Models (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 the first and subsequent decoding passes 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 or third decoding pass are speaker-adaptively trained, exploiting as target-models triphone HMMs with a single Gaussian density per state.

2.3. Lexica

Two different lexica were used to provide phonetic transcriptions of words:

- *USLex*: Pronunciations in the lexicon are based on a set of 45 phones. The lexicon was 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.

- *BEEPLex*: This lexicon was generated by exploiting the British English Example Pronunciations (BEEP) lexicon. Pronunciation models in this lexicon are based on a set of 44 phones. Transcription for a number of missing words were obtained by exploiting the pronunciation models in the *USLex* lexicon and mapping phonetic symbols into the BEEP phone set.

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.

2.4. Language models

Text data used for training the LMs are those released for the IWSLT2012-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.
- **wmt12** Formed by documents in WMT12 news crawl, news commentary v7 and Europarl v7 (see IWSLT2012 official web site for some more details about these corpora). In total it contains about 830M words.
- **ted12** An in-domain set of texts extracted from TED talks transcriptions used for training. It contains about 2.4M 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 [9]. The three LMs resulted, respectively, into about:

- **giga5** 128M bigrams, 231M 3-grams, 422M 4-grams;
- **wmt12** 44M bigrams, 50M 3-grams, 68M 4-grams;
- **ted12** 599K bigrams, 199K 3-grams, 125K 4-grams.

The **wmt12** LM is 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.

The non-pruned LMs, **giga5** and **wmt12**, are instead linearly interpolated (as explained below) in order to provide LM probabilities for expanding word graphs to be used in the third decoding step.

2.5. Word graphs generation

Word graphs (WGs) 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 [10] is applied. In this way the resulting graph error rate was estimated to be 8.8% on the development set used for IWSLT2012 evaluation campaign, less than $\frac{1}{2}$ of the corresponding WER (which resulted to be 18.9%, as reported in section 4).

2.6. Transcription process

In the IWSLT2012 ASR evaluation, time boundaries of speech segments to be transcribed are given for each audio file. These non-overlapping speech segments are clustered by using a method based on the Bayesian information criterion [11]. The resulting clustering is exploited by the transcription system to perform cluster-based acoustic feature normalization and AM adaptation.

The first decoding pass is carried out with acoustic models based on *BEEPlex*, while the second and third decoding passes make use of acoustic models based on *USLex*. This configuration was chosen based on preliminary experiments on development data. In addition, as previously seen, the **wmt12** LM has been used in both first and second decoding pass.

Cluster-based, text-independent acoustic feature normalization is first performed before HLDA projection. The output of the first decoding pass on these acoustic features is used as a supervision for conducting cluster-based CMLLR acoustic feature normalization and MLLR-based acoustic model adaptation [12] before the second decoding pass, where both the first-best output and word graphs are generated.

The search space employed in the third decoding pass is obtained after expansion of WGs produced in the second decoding pass. The LMs used for WG expansion is a combination of non pruned **giga5** and **wmt12** LMs.

The simplest way for combining LMs trained on different sources is to compute the probability of a word w, given its past history h, as:

$$P[w \mid h] = \sum_{j=1}^{j=J} \lambda_j P_j[w \mid h] \quad (1)$$

where $P_j[w \mid h]$ are LM probabilities trained on the j^{th} source, λ_j are weights estimated with the aim of minimizing the overall perplexity on a development set and J is the total number of LMs to combine. In this case, the development set on which weights λ_j are trained is the one given by the (second pass) ASR output of each TED talk. Note that, in this way, we estimate interpolation weights that depend on each given talk.

The expanded WGs are compiled into corresponding decoding networks using the *USLex* lexicon. Also in this case, the best recognition hypothesis generated in the second decoding pass is exploited for conducting cluster-based CMLLR acoustic feature normalization and MLLR-based acoustic model adaptation. Finally, WGs are again generated in the third decoding pass and successively rescored for providing both primary and contrastive submissions, as will be explained below.

2.7. Primary submission

WGs generated in the third decoding step are rescored using an interpolated LM that combine all of the three LMs described above, **giga5**, **wmt12** and the in-domain LM **ted12**. To do this, the original LM probability on each arc of each WG is substituted with the linearly interpolated probability given by equation 1. The development set used to train the interpolation weights is the ASR output of the third decoding step and, therefore, also in this case talk specific interpolation weights are estimated.

Note that in the latter WG based rescoring phase acoustic model probabilities associated to arcs of word graphs remain unchanged, i.e. a pure linguistic rescoring is implemented.

2.8. Contrastive submission

As mentioned in the introduction our contrastive submission involves the usage of focused LMs. Figure 1 shows a block diagram of the ASR system employing these LMs, emphasizing both the procedure for selecting auxiliary documents for LM training and the WG based rescoring pass.

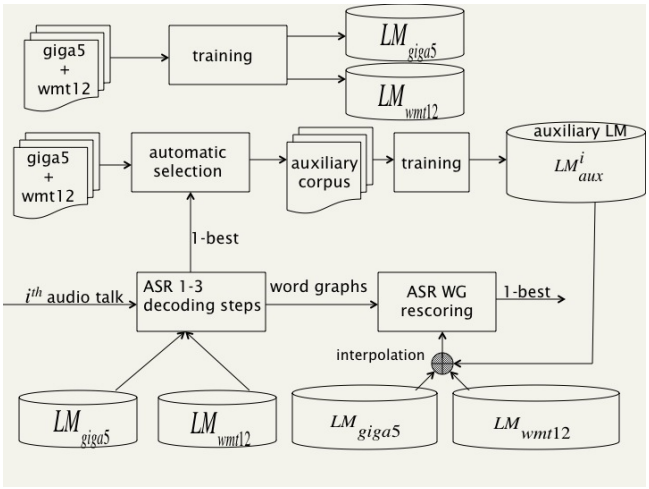

Figure 1: *Block diagram of the ASR system using focused LMs.*

The best word sequences generated in the third decoding pass are used to evaluate the baseline performance, as well as for selecting auxiliary documents. For each given i^{th} talk an auxiliary LM (LM^i_{aux}) is trained on data automatically selected from the out-of-domain text resources **giga5** and **wmt12**, with the selection method described below. The i^{th} query document used to score the out-of-domain text corpora consists of the 1-best output produced in the third ASR decoding step. Then, similarly to primary submission, the original LM probability on each arc of each WG is substituted with the probability given by the interpolation, using equation 1, of the three LMs: **giga5**, **wmt12** and LM^i_{aux}. Also in this case interpolation weights, $\lambda^i_j, 1 \leq j \leq 3$, associated to the three LMs are estimated so as to minimize the overall LM perplexity on the 1-best output (the same used to build the i^{th} query document), of the third ASR decoding step. For clarity reasons this latter procedure is not explicitly shown in Figure 1. The resulting WGs are rescored using the new interpolated LM probabilities.

Note that for this submission no LM trained on in-domain data is used in the last WG rescoring pass, actually the difference between contrastive and primary submission only relies on entering LM^i_{aux} instead of **ted12** in the LM probability interpolation.

3. Auxiliary data selection

In this section we describe the processes for selecting documents (rows in the corpus formed by **giga5** plus **wmt12** text resources) which are semantically similar to a given automatically transcribed document. In the following, N is the number of total rows in the corpus (about 42M for this work) and D is the total number of unique words in the corpus.

The result of this process is to obtain a sorted version of the whole corpus according to similarity scores. The most similar documents will be used to build talk-dependent auxiliary LMs.

3.1. Preprocessing stage

First, we build a table containing all the different words found in the corpus to select, each one with an associated counter of the related number of occurrences.

Then, a dictionary $\mathcal{V}$ is built containing the words that, according to inverse order of occurences, have an index $D^{''} \leq i \leq D^{'}$, where $D^{''} = 100$ and $D^{'} = 200,000$.

Then, every word in the corpus is replaced with its corresponding index in $\mathcal{V}$. Words outside $\mathcal{V}$ are discarded.

Indices of each row are then sorted to allow quick comparison (this point will be discussed later). The rationale behind this approach is the following:

- very common words only carry syntactic information, therefore they are useless if the purpose is to find semantically similar sentences;
- very uncommon words will be used rarely so they will just slow down the search process.

The choice for the reported values of D' and D'' has been done on the basis of preliminary experiments carried out on a development data set (see section 4) and did not result to be critical. With the chosen values about half of the words of the corpus were discarded: i.e. about 2.6M millions of indices survived. We keep alignment between the original corpus and its indexed version.

3.1.1. Searching stage

From the sequence of automatically recognized words $W^i = w^i_1, \ldots, w^i_{\texttt{len}(W^i)}$ of the given i^{th} query document (i.e. the i^{th} automatically transcribed talk) we derive a corresponding sequence of numerically sorted indices. Hence, both the i^{th} talk and the n^{th} document in the corpus are represented by two vectors (containing integer indices): $\mathbf{C}'^i$ and $\mathbf{R}'^n$, respectively. The similarity score is:

$$s'(\mathbf{C}'^i, \mathbf{R}'^n) = \frac{e(\mathbf{C}'^i, \mathbf{R}'^n)}{\mathtt{dim}(\mathbf{C}'^i) + \mathtt{dim}(\mathbf{R}'^n)} \quad (2)$$

where $e(\mathbf{C}'^i, \mathbf{R}'^n)$ is the number of common indices between the two vectors $\mathbf{C}'^i$ and $\mathbf{R}'^n$. Note that the two vectors $\mathbf{C}'^i$ and $\mathbf{R}'^n$ have dimensions exactly equal to the number of the corresponding indexed words survived after pruning of dictionary, as explained above.

The proposed approach is similar to the well known method based on TFxIDF [4]. However, while the latter allows to compare two documents by weighting same words both with their frequencies and with their relevance in the documents to select, the proposed approach is essentially a method to count the number of same words in the documents (word counters are not used in the similarity metric). However, since components of index vectors are numerically ordered, the computation of the similarity score $s'(\mathbf{C}'^i, \mathbf{R}'^n)$ results very efficient. This is essential given the large number of documents in the corpus to score.

In addition, differently from TFxIDF, the proposed approach doesn't require to load into memory of the computer any parameter related to the whole dictionary, instead only the sequence of indices (i.e. one sequence of integer values for each row in the corpus to select) entering equation 2 is needed. In our implementation the latter indices are conveniently stored and read from a file. Therefore, the memory requirements of the proposed approach are negligible. Furthermore, since the resulting document scores are not normalized, the estimate of the threshold to be used for selecting the subset of the documents to sort from the whole corpus is based on a preliminary computation of a histogram of scores.

Finally, in order to measure the complexities of proposed method and TFxIDF based one, we led three different selection runs using ASR output of a predefined TED talk. For processing the whole **giga5** + **wmt12** corpus the proposed method took on average about $16min$, with a memory occupation of about 10MB, while the TFxIDF based method took on average about $114min$, with a memory occupation of about 650MB. These runs were carried out on the same Intel/Xeon E5420 machine, free from other computation loads.

A more detailed comparison among: the proposed selection approach, the TFxIDF based one and another one based on perplexity minimization is reported in a companion paper.

4. System run

In order to tune some parameters of our automatic transcription system we carried out some preliminary experiments on the development set of IWSLT2012 evaluation campaign. The latter is made by 19 TED talks derived from the union of the IWSLT 2010 development and evaluation sets. In particular, we need to choose, for the contrastive submission, an optimal number of words on which to train auxiliary LMs as explained in section 3. To do this we evaluated, on the above mentioned dev set, both perplexity (PP) and WER as functions of the latter number of words. Results are given in Figures 2 and 3.

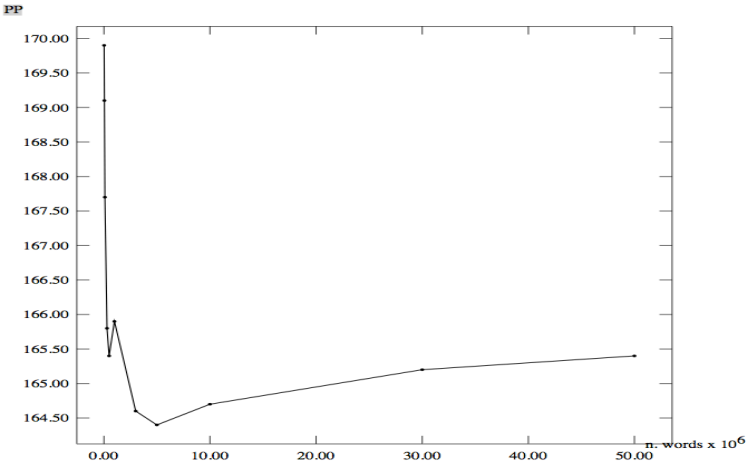

Figure 2: *Perplexity on dev set of focused LMs, as a function of the number of words used to train auxiliary LMs (the point corresponding to 0 words on the abscissa refers to the usage of the baseline LM).*

In the figures the point corresponding to 0 words on the abscissa indicates performance obtained with the baseline LM, i.e through the interpolation of **giga5** and **wmt12** without including auxiliary LMs.

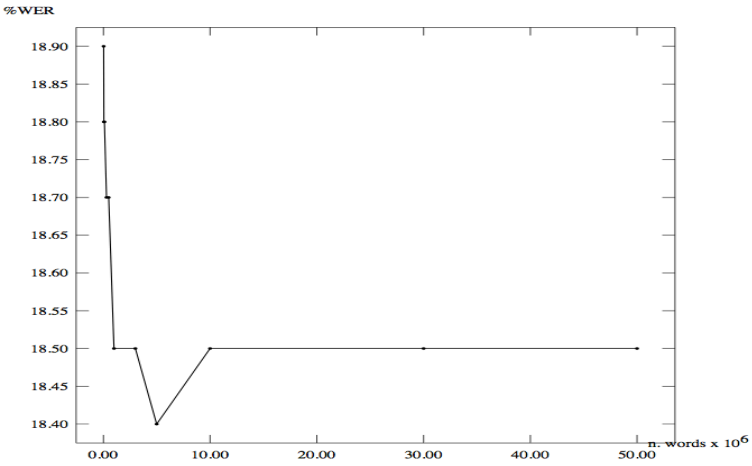

Figure 3: *%WER on dev set, using focused LMs in the final WG based rescoring step, as a function of the number of words used to train auxiliary LMs (the point corresponding to 0 words on the abscissa refers to the usage of the baseline LM).*

Note that the overall perplexity on the dev set PP_{dev} is computed summing the LM log-probabilities of each reference talk and dividing by the total number of words, according to the following equation:

$$PP_{dev} = 10^{\frac{\sum_{i=1}^{i=19} -\log_{10}(P^i_{LM}[W_i])}{NW}} \quad (3)$$

where $P^i_{LM}[W_i]$ is the probability of the reference word sequence in the i^{th} talk, computed using the i^{th} talk-

dependent interpolated LM, and NW is the total number of words in the dev set.

Performance, both in terms of PP and WER, obtained on test set 2011 are reported in Table 1. According to experiments led on dev set, the number of words used to train auxiliary LMs was chosen to be equal to 5M. In Table 1 performance are given for ASR decoding passes two and three and for the final WG based rescoring step. The latter, as explained in section 2.6, has been executed twice: once for producing the primary submission and once for generating the contrastive one. Primary submission is obtained through rescoring of WGs with interpolated LM $\mathbf{wmt12} \oplus \mathbf{giga5} \oplus \mathbf{ted12}$, where $\oplus$ denotes linear interpolation according to equation 1. Contrastive submission is obtained substituting auxiliary LMs LM^i_{aux}, as depicted in figure 1, to **ted12** in the interpolation.

Table 1: *Results obtained on test set 2011 in the various decoding steps, and on test set2012 for both primary and contrastive submissions.*

	test2011		**test2012**
	PP	%WER	%WER
step 2	160	17.1	
step 3	159	16.7	
WG rescoring (primary)	126	15.4	16.8
WG rescoring (contrastive)	146	15.7	17.3

In Table 1 the WERs obtained on test set 2012 are also given for both primary and contrastive submissions. Note that on both test sets the usage of focused LMs (contrastive submissions) allows to achieve performance comparable with that of primary submissions, but without using in-domain data for LM training.

5. Conclusions

We presented our submission runs to the IWSLT2012 Evaluation Campaign for the ASR English track. Our ASR system was trained on a significant portion of TED talk recordings, by exploiting an automatic data selection method evaluating the fidelity of the provided transcripts.

We have described a method for focusing LMs towards the output of the ASR system. The approach is based on the useful and efficient selection, according to a novel similarity score, of documents belonging to large sets of text corpora on which the general purpose LM, used along the various ASR decoding steps, was trained. Significant improvement on WER has been reached without making use of in-domain text data.

Future work will address domains different from TED, the usage of larger sets of text corpora and more efficient selection methods.

6. Acknowledgements

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

7. 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] S. Maskey and A. Sethy, "Resampling Auxiliary Data for Language Model Adaptation in Machine Translation for Speech," in *Proc. of ICASSP*, Taipei,Taiwan, April 2009, pp. 4817–4820.

[3] G. Lecorve, J. Dines, T. Hain, and P. Motlicek, "Supervised and unsupervised Web-based language model domain adaptation," in *Proc. of INTERSPEECH*, Portland, USA, September 2012.

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

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

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.