UNSUPERVISED LANGUAGE MODEL ADAPTATION FOR BROADCAST NEWS

Langzhou Chen, Jean-Luc Gauvain, Lori Lamel, and Gilles Adda

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{clz,gauvain,lamel,gadda}@limsi.fr

ABSTRACT

Unsupervised language model adaptation for speech recognition is challenging, particularly for complicated tasks such the transcription of broadcast news (BN) data. This paper presents an unsupervised adaptation method for language modeling based on information retrieval techniques. The method is designed for the broadcast news transcription task where the topics of the audio data cannot be predicted in advance. Experiments are carried out using the LIMSI American English BN transcription system and the NIST 1999 BN evaluation sets. The unsupervised adaptation method reduces the perplexity by 7% relative to the baseline LM and yields a 2% relative improvement for a 10xRT system.

1. INTRODUCTION

Unsupervised language model (LM) adaptation is an outstanding challenge for speech recognition, especially for complex tasks such as broadcast news (BN) transcription where the content of any given show is open and is related to multiple topics. Due to this and the dynamic nature of the task, it is not possible to select adaptation data in advance. Therefore selecting the adaptive data is one of the main problems in unsupervised LM adaptation. The data available are the speech recognition hypotheses and the large general text corpus which usually has been used to estimate the recognizer's LMs. Unsupervised LM adaptation can be seen as tuning a general LM to some special topics without domain specific training data. A straightforward approach for unsupervised adaptation is to use the speech recognizer hypothesis directly as adaptation data. This approach is not very successful since the hypothesis is quite small and may contain recognition errors. Information retrieval techniques have been proposed to address this problem [1]. In this case, the speech recognition hypothesis is used as a query to extract articles or text segments with related topics. If a large text corpus is available, the selected texts can provide sufficient adaptation data. At the same time, this adaptive data is not subject to recognition errors. We previously presented a method for unsupervised LM adaptation based on IR methods which was successfully applied to the transcription of Mandarin BN data [1]. Unfortunately, this approach was not successful for the transcription of American English BN data. We attribute this to the fact that the American English BN data come from a larger number of sources than the Mandarin data (the data we used come from only three sources, two radio and one television). It is therefore more difficult to get adaptive data from the text corpus.

This work builds upon the approach developed for Mandarin BN transcription. The IR procedure has been modified to extract a more accurate adaptive corpus. A series of experiments were performed to compare different adaptation methods, which resulted in choosing a minimum discrimination information (MDI) adaptation as our prefered method. We also carried out experiments to find a proper way to integrate the unsupervised LM adaptation into the decoding procedure.

The remainder of this paper is organized as follows. The next section provides a description of our unsupervised language model adaptation method. Section 3 describes the MDI LM adaptation method and Section 4 describes the modifications to the data selection method. These are followed by some experimental results and conclusions.

2. ADAPTATION METHOD OVERVIEW

The basic idea of our adaptation method is to use the hypothesis of the speech recognizer as query to extract adaptive data from a large general corpus. The adaptive algorithm can be divided into two parts: extraction of the adaptation corpus and LM adaptation. The adaptive corpus extraction method has 3 steps:

1. Initial hypothesis segmentation: The recognition hypotheses almost always include texts on multiple topics, which need to be segmented into individual stories, each associated with a single topic. As a first approximation to stories, we used the segment boundaries located by the audio partitioner [5]. Since these segments are usually shorter than true stories, neighboring paragraphs which have a similar content are iteratively regrouped until no more merges are possible. The result of this procedure is a hypothesized transcription with hypothesized story boundaries, where each story ideally concerns a single topic.

2. Keyword selection: For each story, the content words with the most relevant topic information are selected. The relevance of word w_i and story s_j, is given by the following score:

$$R(w_i, s_j) = \sum_{v \in k_j} \log \frac{\Pr(w_i, v)}{\Pr(w_i)\Pr(v)} \qquad (1)$$

where $p(w_i, v)$ is the probability that w_i and v appear in the same story and k_j is the set of all words (excluding function words) in story s_j. All words having a relevance score higher than an empirically determined threshold are selected. In practice the selected words represent the topic of a story quite well, since many of the recognition errors are unrelated to the story, and any words co-occurring with other highly relevant words in the story also help provide reliable information.

3. Retrieving relevant articles: The selected N content words for each story are used to retrieve relevant texts in the training corpus

(200M words of BN data from January 1992 to May 1998). The on-topic articles are selected by computing the following score:

$$\frac{1}{N_j}\sum_{i=1}^{N}\sum_{k=1}^{N_j}\log \Pr(\textit{keyword}_i, w_k)\Pr(\textit{keyword}_i)\Pr(w_k) \qquad (2)$$

where N_j is the number of content words in article A_j. All articles with a score exceeding an empirically determined threshold are extracted. Training texts are selected in this way for each story. The selected articles are used as adaptation data to train the adaptive LM.

In our original method, two kinds of adaptive language models were trained, a mixture model and a MAP (maximum a posteriori) based LM. In this work a series experiments were carried out to select the proper adaptation method.

3. MDI ADAPTATION

We have done a series experiments to select a proper LM adaptation method. We considered using a mixture model, a MAP adaptive model, a minimum discrimination information (MDI) model and a dynamic mixture model.

We first investigated using a MDI [3] adaptation method. MDI adaptation can be expressed as follows. Given a background model $P_b(h, w)$ and a adaptive corpus *A*, we have to find a model $P(h, w)$ satisfying a set of linear constraints for which the Kullback-Leibler distance between $P(h, w)$ and $P_b(h, w)$ is minimized. MDI model can be trained by using the GIS (Generalized Iterative Scaling) algorithm. For the special case where we consider only the unigram model, and perform only one iteration, the GIS algorithm can be simplified as :

$$P(h, w) = P_b(h, w)\frac{P_a(w)}{P_b(w)} \qquad (3)$$

where $P_a(w)$ is a unigram estimated on the adaptive data and $P_b(w)$ is the unigram from the background model. The conditional probability can be expressed as:

$$P(w|h) = \frac{P_b(w|h)\alpha(w)}{Z(h)} \qquad (4)$$

where $Z(h)$ is a normalization factor,

$$Z(h) = \sum_{v} P_b(v|h)\alpha(v) \qquad (5)$$

and

$$\alpha(w) = \left(\frac{P_a(w)}{P_b(w)}\right)^{\gamma} \qquad (6)$$

where γ is a parameter ranging from 0 to 1 used to adjust the weight of the adaptation data.

Comparing the MDI-based model and MAP-based model, we found that MDI based model to be a little better than MAP model (about 0.1% absolute in WER). Therefore, we decided to use MDI based LM adaptation instead of MAP based adaptation.

Another experiment explored dynamic modeling, in which dynamic weighting of the mixture model replaces the original fixed modeling. In the original mixture model, the different model components are trained in advance. The new adaptive corpus is only used to tune the interpolation weights. The different model components are fixed and all of the topic-specific information is contained in the mixture weights. In this work, the different model components are trained dynamically according to the topics from the recognition hypotheses. The hypotheses are also used to train the mixture weights via the EM algorithm. Using this method, both the mixture weights and the models themselves include topic-specific information. It should therefore be more accurate and more robust to the recognition errors. Using the new mixture models without MDI yields an absolute WER reduction of about 0.2%. When the mixture model is combined with the MDI model there is no additional gain, the WER is equal to that obtained using the MDI model alone. So we decided to adopt the MDI adaptation method in out system.

4. DATA SELECTION

Although the previously reported adaptive corpus selection method worked well for the Mandarin language 2, it was less successful for American English. Here we present some modifications to improve this approach. As describe in section 2, the original data selection method mentioned has 3 steps: initial hypothesis segmentation, keyword selection and retrieving relevant articles. Our modifications concern the steps 2 and 3, i.e. keyword selection and article retrieving.

As described above, our procedure of selecting the adaptive data is to use keywords extracted from the recognition hypotheses as a query to extract adaptive data from a large general corpus. The keywords are words judged to be representative of a particular topic. Usually such keywords occur frequently in articles related to the specific topic, and seldomly occur in others articles. In order to filter out words without topic information, a keyword set was defined and only the words included in the keyword set can be extracted.

In our previous work [1], the keyword set was determined by its *idf*∗*tf* value using the following function:

$$idf(w)\frac{1}{N_w}\sum_{k=1}^{N_w} tf_k(w) \qquad (7)$$

where $idf(w)$ is the inverse document frequency of word w, N_w is the number of articles which contain the word w, and $tf_k(w)$ is the term frequency of word w in article k. All words with a score exceeding an empirically determined threshold are included in the keyword set.

In order to get an accurate keyword set, we used a new way of selecting keyword directly from different articles. Firstly, the BN corpus was segmented into different stories, and for each story a keyword that represents the topic of the story was selecting according to Eqn. 1. The keywords from all the stories comprise the keyword set. This method ensures coverage of the topics of all stories occurring in the large general corpus, and is more accurate and more complete than the previous method.

The second modification concerns the procedure of extracting the adaptive corpus based on the keywords. The previous method was based on Eqn. 2. However, in our experiments we found that some keywords such as *Clinton*, occur frequently and are often related to many topics. Given only the keyword *Clinton*, it is not possible to accurately determine the topic of the story. If one article contains 5 occurrences of *Clinton*, while another article contains only one occurrence of *Clinton* and one occurrence of *Lewinsky*, the topic of the second article is clearer than the first one. Using our original scoring method the first article will have higher score, because *Clinton* appears more frequently. In order to solve this problem, the following new score has been used:

$$\frac{K(q,s_j)^{\gamma}}{N_j}\sum_{i=1}^{N}\sum_{k=1}^{N_j} idf(w_k)\log\frac{\Pr(\textit{keyword}_i, w_k)}{\Pr(\textit{keyword}_i)\Pr(w_k)} \qquad (8)$$

where $K(q, s_j)$ is number of distinct keywords that occur both in the query and the candidate article s_j , and γ is a parameter used for tuning. This factor can be seen as a confidence score, the importance of each word being measured by its *idf* value.

5. EXPERIMENTAL RESULTS

Experimental results are reported for the LIMSI American English broadcast news transcription system [4, 5]. We used the NIST BN 1999 test data to evaluate the LMs. This test set is consist of 3 hours of speech divided into two data sets. The first set (bn99en_1) was taken from episodes broadcast in June 98, and the second set (bn99en_2) was taken rom a different variety of shows broadcast in August/September of 1998. Different LMs are built for each test subset.

The baseline 4-gram LM model is trained on three sources of data [5]:

1. NEWS: Over 340M words news text from various sources, distributed by the LDC. A 4-gram LM is trained for each of the 3 years covered by the corpus (95-97).

2. BNA: 1.5M words of accurate BN acoustic training data transcriptions with some non lexical items such as breath noise. A 4-gram LM is trained on this data.

3. BNC: 200M words of commercial transcripts of various BN shows from January 1992 to May 1998. A 4-gram LM is trained on this data.

Our baseline 4-gram LM is obtained by interpolating these 5 LMs (BNC, BNA, NEW95, NEWS96, NEWS97). The interpolation weights are 0.47, 0.15, 0.14, 0.11, 0.13 respectively. The perplexity of this baseline LM is 214.6 for bn99en_1 and 207.3 for bn99en_2.

From the weights of the different components of the baseline model, it can be seen that the BNC model is the most important component. The BNC corpus is also sufficiently large to use for LM adaptation. We use the recognizer hypotheses as query to extract adaptive data from the BNC corpus, and use this adaptive corpus to carry out MDI adaptation of the original BNC 4-gram model. All the other components in the baseline model and the interpolation weights remain unchanged.

Experiments with MDI

Our first experiment investigates MDI adaptation. Given the imperfect recognizer output and imperfect information retrieval methods, false alarms are inevitable, i.e. some off-topic stories will be selected as part of the adaptation data. In order to get an idea about the tradeoff between the size of adaptive data set and its accuracy, MDI adaptation was performed for the BNC model using different size of adaptation data automatically extracted from the BNC corpus. The adapted LMs (only using the BNC corpus) are compared in terms of perplexity.

Table 1 shows the perplexity for the second show (bn99en_2) as a function of the amount of adaptation data for the MDI based models. The parameter γ in Eqn. 6 is set to 1. It can be seen in this table, that as the size of the adaptive corpus increased, the perplexity of the test data decreases continuously even for a very large adaptive corpus (more than 6000 articles). This is similar to the results with MAP adaptation reported in [1]. When the adaptive corpus is small (fewer than 300 articles) the perplexity of the adaptive model is higher than the topic independent model.

Amount of adaptation data	*perplexity*
0 articles	269.6
120 articles	279.3
600 articles	254.3
1200 articles	246.4
3600 articles	241.5
4800 articles	239.9
6600 articles	239.2

Table 1: Perplexity of MDI models vs. adaptation corpus size.

	adaptive corpus size		
γ	*120 articles*	*600 articles*	*6600 articles*
0	269.6	269.6	269.6
0.1	256.9	256.9	257.0
0.3	249.3	246.2	246.3
0.5	253.4	243.2	241.1
0.7	266.3	244.9	238.4
0.9	291.4	252.0	238.3
1.0	309.8	254.3	239.2

Table 2: Perplexity of MDI models vs. γ and adaptation corpus size.

The second experiment assesses the impact of γ which givesus a way to reduce the weight of the adaptation distribution. Previous work [3, 6] proposed to set this parameter to 0.5, but our experiments suggest a different setting. Table 2 shows the perplexity for bn99en_2 as a function of γ based on different sized adaptive corpora. It can be seen that when the adaptive corpus is small, the value of γ corresponding to the lowest perplexity is also small, and as the size of adaptive corpus increases the best value of γ also increases. When the adaptive corpus is very large (over 6000 articles), the optimal value of γ approaches 1. As a matter of fact when the adaptive corpus is small, the estimation of the adaptation distribution $P_a(w)$ is not very reliable, and therefore the weight should be considerably reduced. When the adaptive corpus is large, the estimation of the adaptation distribution is more reliable, and the best value of γ is high. In our experiments, the adaptive corpus is always large. However, since the adaptation is unsupervised, it is difficult to tune the γ value using test data, so the value of γ has been fixed to 1.

Experiments with data selection

We also compared the performance of the adaptive model in function of the adaptive corpus extracted by different IR methods. The results are shown in Table 3. In this table, the results labelled "old method" were obtained with our original IR method corresponding to Eqn. 2, and the "new method" represents the improved method corresponding to Eqn. 8. The results are based on a 10xRT system, where the adapted LM is used in a second 10xRT decoding (resulting in a 20xRT system). The new method is seen to improve the performance of adaptive LM.

Experiments with 10x and 50x systems

We also investigate the effect of the adapted language models at two decoding speeds: for a 10xRT system and for a 50xRT system [4, 5].

LIMSI 10x BN system has 3 decoding passes. The first pass uses

	WER		*perplexity*	
IR method	*bn99en_1*	*bn99en_2*	*bn99en_1*	*bn99en_2*
baseline 10x system	18.3%	16.3%	214.6	207.3
old method	18.2%	16.1%	202.1	196.5
new method	17.9%	15.8%	198.8	192.7

Table 3: Word error rates and test perplexities for the two data subsets for the baseline 10x system and with the adaptive LM using the two data selection procedures.

a trigram LM and small HMM (5.6k phone contexts, 16 Gaussians per state, 92k Gaussians total) to generate initial hypothesis. These hypotheses are used to carry out MLLR acoustic model adaptation which are use to generate word graphs with a trigram LM and a larger HMM (28k phone contexts, 16 Gaussians per state, 180k Gaussians total). The third pass carries out MLLR acoustic model adaptation and the word graph decoding with a 4-gram LM and a large HMM (28k phone contexts, 32 Gaussians per state, 350k Gaussians total).

Several ways of integrating the adapted LMs are investigated: the first one uses the hypotheses of the first decoding pass to generate a topic dependent LM, which is used in the second and third decoding passes. This still results in a 10xRT system. Another possibility is to use the second pass hypotheses to generate the topic dependent LM, which is only used in the third decoding pass. Because the word graph generation is an important step, it was found to be more effective to use the adapted LM in the second pass for word graph generation. The experimental results are summarized in Table 4. For the 10xRT system, LM unsupervised adaptation brings an absolute improvement of 0.3%. If the final hypotheses of the original 10xRT system are used to extract the adaptive corpus, and a 10xRT decoding is then carried out with the adapted LM (resulting in a 20xRT system), a 0.45% absolute improvement is obtained.

		baseline LM			*adaptive LM*		
decoding	*xRT*	*set 1*	*set 2*	*AVG*	*set 1*	*set 2*	*AVG*
pass#1	0.9	30.3	28.6	29.3	–	–	–
pass#2	6.5	19.8	17.6	18.5	19.8	17.4	18.4
pass#3	1.5	18.3	16.3	17.1	18.1	15.9	16.8

Table 4: WER of 10xRT system.

A final experiment was carried out using the adapted LMs in a slower, 50xRT system with 5 decoding passes. The first 3 passes are the same as the 10xRT system. The third pass hypotheses are used to create the topic dependent LM. The 4th pass consists of MLLR adaptation and word graph generation using the adapted 4-gram LM and the largest acoustic model set. The 5th pass carries out another MLLR acoustic model adaptation, and decoding is done using the word graph. The recognition results are given in Table 5. Unfortunately, the adaptive LM gives only 0.1% improvement over the baseline 50xRT system.

		baseline LM			*adaptive LM*		
decoding	*xRT*	*set 1*	*set 2*	*AVG*	*set 1*	*set 2*	*AVG*
pass#1-3	9.3	18.3	16.3	17.1	–	–	–
pass#4	38.1	17.2	15.1	16.0	17.0	14.9	15.8
pass#5	5.7	16.9	14.5	15.5	16.6	14.5	15.4

Table 5: WER of 50x system.

6. CONCLUSIONS

In this paper we have proposed a method using information retrieval techniques for the unsupervised language model adaptation of broadcast news transcription system. This work is an extension of our previous work on the transcription of Mandarin broadcast news data.

LM unsupervised adaptation is a difficult topic, especially for the BN transcription task, each show containing speech on multiple topics and the topic information beeing not known in advance. In this work we have used IR methods to divide the recognizer hypotheses into different topics and to extract adaptive data on these topics from a large general corpus of broadcast news.

We have improved our method used to extract topic relevant articles, and we have carried out a series of experiments to compare different adaptation methods within the LIMSI American English BN transcription system. The results on the NIST BN 1999 evaluation sets shows that the proposed method reduces the test perplexity by 7% and reduces by 2% relative the word eror rate for a 10xRT system. The adaptive language model was found to be less efficient for a very slow system as there is evidently less room for improvement.

REFERENCES

[1] L. Chen J.L. Gauvain L. Lamel G. Adda M. Adda, "Using Information Retrieval Methods for Language Model Adaptation" *Proc. EUROSPEECH'01*, pp. 255-258, 2001

[2] M. Federico, "Bayesian Estimation Methods for N-Gram Language Model Adaptation," *ICSLP'96*, pp. 240-243, 1996.

[3] M. Federico, "Efficient Language Model Adaptation through MDI Estimation." *EUROSPEECH'99*, pp. 1583-1586, 1999.

[4] J.L. Gauvain L. Lamel, "Fast Decoding for Indexation of Broadcast Data" *Proc. ICSLP'00*, pp. 794-797, 2000.

[5] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.

[6] R. Kneser, J. Peters D. Klakow, "Language Model Adaptation Using Dynamic Marginals" *Proc. EUROSPEECH'97*, pp. 1971-1974, 1997.

[7] R. Kuhn, R. de Mori, "A Cache-Based Natural Language Model for Speech Reproduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 12(6) pp. 570-583, 1990.