Dynamic Language Modeling for Broadcast News

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

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

ABSTRACT

This paper describes some recent experiments on unsupervised language model adaptation for transcription of broadcast news data. In previous work, a framework for automatically selecting adaptation data using information retrieval techniques was proposed. This work extends the method and presents experimental results with unsupervised language model adaptation. Three primary aspects are considered: (1) the performance of 5 widely used LM adaptation methods using the same adaptation data is compared; (2) the influence of the temporal distance between the training and test data epoch on the adaptation efficiency is assessed; and (3) show-based language model adaptation is compared with story-based language model adaptation. Experiments have been carried out for broadcast news transcription in English and Mandarin Chinese. A relative word error rate reduction of 4.7% was obtained in English and a 5.6% relative character error rate reduction in Mandarin with story-based MDI adaptation.

1. INTRODUCTION

While n-gram models are successfully used in speech recognition, their performance is influenced by any mismatch between the training and test data [7]. The idea of language model (LM) adaptation is to use a small amount of domain specific data to adjust the LM to reduce the impact of linguistic differences between the training and testing data. Different schemes for LM adaptation have been proposed, such as the cache model based on the observation that a word which occurred in a recent text has a higher probability to be seen again [9]; the trigger model which uses a trigger word pair to get at semantic information [10]; and structured LMs [1].

Broadcast news (BN) transcription is a complicated task for both acoustic and language modeling. The linguistic attributes of BN data are complex, arising from the many different speaking styles, from spontaneous conversation to prepared speech (close in style to written texts). The content of BN data is open and any given BN show covers multiple topics.

As a consequence, it is difficult to predict the topics of a BN show without looking at the data itself. The only information that is available for the show are the hypotheses output from the speech recognizer. However, for any given broadcast, the number of words in the hypothesized transcript is quite small and contains recognition errors. Therefore the transcripts are not sufficient for use as an adaptive corpus. Information retrieval (IR) methods provide a means to address this problem. Instead of directly using the ASR hypotheses for LM adaptation, they can be used as queries to an IR system in order to select additional on-topic adaptation data from a large general corpus. This approach reduces the effect of transcription errors in the hypotheses and at the same time provides substantially more textual data for LM estimation.

In this paper, a series of experiments are presented exploring the general framework of unsupervised LM adaptation using IR methods [3]. The performances of a variety of popular techniques for LM adaptation using automatically selected adaptation data are compared. The investigated techniques are linear interpolation, maximum a posteriori (MAP) adaptation, mixture models, dynamic mixture models, and minimum discrimination information (MDI) adaptation. The effect of the temporal distance between the epoch of the adaptation corpus and of the epoch of the test data is also assessed. As mentioned above, a given BN show typically covers several stories, with each story being related to a different topic. To address the changing property of BN data, static and dynamic models for LM adaptation are investigated. In static modeling the LM is updated once for the whole show, which means that the LM must be simultaneously fit to multiple topics. Dynamic modeling updates the LM at each automatically detected story change, which entails estimating multiple story-based LMs for each BN show. Experiments are carried out for BN transcription in American English and Mandarin Chinese.

2. ADAPTATION DATA RETRIEVAL

The basic idea of the adaptation method is to use the hypotheses produced by the speech recognizer as query to extract adaptive data from a large general corpus. As described in [3], there are 3 steps in extracting an adaptation corpus:

1. Initial hypothesis segmentation: The recognition hypotheses almost always include texts covering multiple topics, which need to be segmented into individual stories, each associated with a single topic. The segment boundaries located by the audio partitioner [6] are used to initialize the process. 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: The topic information of a story can be represented by the co-occurrence of keywords. In this step, keywords that are most relevant to the specific story are selected. The relevance of each content word w_i in story s_j, is given by the following score:

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

where $\Pr(w_i, v)$ is the probability that w_i and v appear in same story and S_j is the set of content words in story s_j which form the trigger pair with w_i. All words having a relevance score higher than an empirically determined threshold are selected. We assume that most of the topic information has been captured by the selected keywords, and only these story relevant keywords are kept for the next step, while the story itself is discarded.

3. Retrieving relevant articles: This procedure is the inverse of keyword selection. The selected N content words for each story are used as query to retrieve relevant texts in a large general text corpus. The relevance between candidate adaptation texts and the query is calculated in a similar manner to that used in keyword selection:

$$\frac{K(q,s_j)^\gamma}{N_j}\sum_{i=1}^{N}\sum_{k=1}^{N_j} idf(w_k)\log\frac{\Pr(\mathit{keyword}_i, w_k)}{\Pr(\mathit{keyword}_i)\Pr(w_k)} \quad (2)$$

where $K(q, s_j)$ is the number of distinct keywords that occur both in the query and the candidate article s_j, and γ is a tuning parameter. $K(q, s_j)$ can be seen as a confidence score, the importance of each word being measured by its *idf* (inverse document frequency) value. All articles with a score exceeding an empirically determined threshold are selected to be part of the adaptation data for the story.

3. ADAPTIVE LANGUAGE MODELS

A number of popular approaches to adaptive language models are investigated using the automatically selected adaptation data. In this section, these methods are briefly reviewed.

1. Linear interpolation: Linear interpolation is a simple and general method. The adaptive corpus is used to train a LM, and the n-gram probabilities of this LM are linearly combined with those of the general LM. The interpolation weight is estimated using the Estimation-Maximization (EM) algorithm [8] to maximize the likelihood of the development data.

2. Maximum a posteriori based method: MAP [4] is another convenient method for language model adaptation. Given a set of adaptation data, a general language model is tuned to the special topic according to the maximum a posteriori criterion. The MAP based method is also an interpolation method. The linear interpolation interpolates the LM at the model level, whereas the MAP based method interpolates the LM at the frequency (word count) level.

3. Mixture models: Mixture modeling clusters the training corpus into subcorpora corresponding to different topics, and train topic dependent models on each of the topic subsets. The interpolation weights are estimated using the EM algorithm to maximize the likelihood of the adaptation data. The model components for the mixture models are trained in advance and the adaptive corpus is only used to modify the interpolation weights. In the experiments reported in this paper, the English and Mandarin mixture models contain 276 and 198 component models respectively.

4. Dynamic mixture models: Dynamic mixture models are an extension of mixture models. For mixture models, the different model components are independent of the current task and the adaptive information is only contained in the interpolation weights. In dynamic mixture models, the different model components are trained dynamically according to the topics determined from the recognition hypotheses.

Dynamic mixture model training consists of training the component models and the mixture weights. The adaptation data is split into two portions. Three hundred articles are reserved for training the mixture weights, and the remaining articles are used to train the component models. Since task-specific information is contained in both the mixture weights and the models, the dynamic mixture models are expected to be more accurate than mixture models.

5. Minimum discrimination information adaptation: MDI adaptation has been investigated in [5]. MDI adaptation can be expressed as follows: given a background model $P_b(h, w)$ and an adaptive corpus A, we aim to find a model $P(h, w)$ satisfying a set of linear constraints minimizing the Kullback-Leibler distance between $P(h, w)$ and $P_b(h, w)$. The MDI model can be trained by using the GIS (Generalized Iterative Scaling) algorithm. In this work, a simplified MDI algorithm was used in which only the unigram model is considered and only one iteration is performed.

4. EXPERIMENTAL RESULTS

The experiments reported here aim to improve the performance of the LIMSI baseline 10xRT broadcast news transcription systems for American English and Mandarin Chinese. The five IR-based LM adaptation methods have been applied to both languages, using the LIMSI English 1999 evaluation system described in [6] and the LIMSI Mandarin system described in [2] as baselines.

The LM training data for the English system contains three sources: over 340 M words of newspaper and newswire texts, 1.5 M words of accurate transcriptions of BN acoustic training data, and 200 M words of commercial transcripts of various BN shows from January 1992 to May 1998. The baseline 4-gram LM is obtained by interpolating component LMs trained on each of the three data sources. For Mandarin Chinese, the LM training data is comprised of about 460 k characters of detailed manual transcriptions of the acoustic training material and 186 M characters of texts from newspaper sources. The baseline 4-gram LM is obtained by interpolating a 4-gram trained on the newspaper text corpus with a trigram trained on the transcripts of the audio data.

The NIST BN 1999 test data was used to evaluate the different adapted LMs for the English system. This test set consists of 3 hours of audio data split in two subsets. The first set (bn99en_1) was taken from episodes broadcast in June 1998, and the second set (bn99en_2) was taken from a different set of shows broadcast in August/September of 1998. For Mandarin, the 1997 NIST Hub4 Mandarin evaluation data (h4ne97ma) containing 1h of speech was used for test purposes. All experimental results (in terms of perplexity and decoding error rates) are given for the individual test sets. The results of the baseline systems for both languages are shown in Table 1.

The 10x BN system [6] has 3 decoding passes: 1) initial hypotheses generation, 2) word graph generation, 3) final hypothesis generation. In the experiments of this paper, the hypotheses of the first decoding pass is used to generate a topic dependent LM, which is used in the second and third decoding passes.

	English WER			*Mandarin CER*
System	*bn99en_1*	*bn99en_2*	*avg.*	*h4ne97ma*
baseline 10x	18.3%	16.3%	17.1%	17.8%

Table 1: Results for the baseline 10x systems: Word error rate (WER) for English and character error rate (CER) for Mandarin.

Test	*base*	*MDI*	*MAP*	*Linear*	*Mixture*	*Dynamic*
Set	*line*			*interp.*	*models*	*mixture*
bn99en_1	280.9	260.2	262.0	250.7	249.7	238.1
bn99en_2	269.6	250.7	252.2	244.0	242.1	235.9
bn99en_1	18.3	18.1	18.3	18.2	18.3	17.9
bn99en_2	16.3	15.9	16.3	16.0	16.0	16.1
average	17.1	16.8	17.1	16.9	16.9	16.8

Table 2: Comparison of perplexity and word error rate (%) for the 5 different adaptation methods for the English BN system.

Test Set	*Base*	*MDI*	*MAP*	*Linear*	*Mixture*	*Dynamic*
h4ne97ma	*line*			*interp.*	*models*	*mixture*
perplexity	447.0	381.3	412.8	389.7	376.4	388.8
CER	17.8	17.2	17.7	17.4	17.4	17.4

Table 3: Comparison of perplexity and character error rate (%) for the 5 different adaptation methods for the Mandarin BN system on the NIST 1997 evaluation data (h4ne97ma).

4.1 Comparing adaptation methods

A set of experiments were carried out to compare the performance of the different adaptation methods. The training text corpora were automatically divided into different stories. The average length of a story in the English corpus is about 700 words and the average story in the Mandarin Chinese corpus has about 400 characters. Recognition experiments were carried out using the same adaptation corpus containing 3000 articles for the four of the five adaptation methods. Since 3000 articles were too many for training the mixture weights, for the mixture models, 300 articles were used as adaptation data. The top part of Table 2 gives the perplexities and the bottom part the word error rates for the English BN system on the NIST 1999 test data.

It can be seen that the MDI and dynamic mixture adaptation methods result in the best average performance, bringing a 0.3% absolute gain in WER. The MAP model is seen to give no gain over the baseline system. As discussed above, dynamic mixture models can be considered as a combination of linear interpolation and mixture models, incorporating the information contained in the adaptive data in both the mixture weights and the model components. This is why the dynamic mixture models are more accurate than the mixture models.

Table 3 gives the results in terms of perplexity and character error rate for the Mandarin BN system on the 1997 NIST evaluation data. The size of the adaptation corpus is the same as for the English system, i.e., 300 articles for the mixture model method and 3000 articles for the other 4 adaptation methods. The results are similar to those observed for the English BN system, that is MDI adaptation gives the largest improvement, and the smallest gain is with the MAP adaptation. However, for the Mandarin system, dynamic mixture models, linear interpolation and mixture models all yield the same result in terms of CER, even though the dynamic mixture model gives the lowest perplexity.

time period	*bn99en_1*	*bn99en_2*	*average*
may98 (all)	18.1	15.9	16.8
jan98	18.1	15.9	16.8
jul97	18.2	15.8	16.8
jan97	18.1	16.0	16.9
jan96	18.2	16.0	16.9
jul95	18.3	16.0	16.9
jan95	18.5	16.1	17.1

Table 4: English WER (%) with MDI adapted language models as a function of the temporal distance of the adaptation data and the test data.

time period	*h4ne97ma*
dec96 (all)	17.2
may96	17.6
jan96	17.6
jul95	17.6
jan95	17.3
jan94	17.2
jan93	17.4

Table 5: Mandarin CER (%) with MDI adapted language models as a function of the temporal distance of the adaptation data and the test data.

4.2 Impact of the data epoch

Broadcast news is time dependent data. In this subsection, the influence of the epoch of the adaptation data on the recognition performance is investigated. In contrast to the approach proposed in [11] which used external time dependent data sources for adaptation, in this work the training data is fixed (that is no additional data are available). Therefore in order to adapt the general LM temporally, the training text corpus is divided into different time periods, and the adaptation data selection procedure is carried out separately for each subset corresponding to the different periods. Since the training corpus is not evenly distributed over time, the amount of texts for some periods can be quite small. Therefore the text corpus was not divided into equal time periods, but rather into periods with roughly the same quantity of training texts. In these experiments, the epoch of the corpus is annotated as the date of the last articles in the corpus for this period. For example "jan98" indicates that the corpus contains data through January 1998.

Table 4 gives the results in terms of WER using MDI adapted LMs trained on the adaptation corpora for different time periods. The evaluation data is from June 1998, therefore, the distance between the adaptation data and the evaluation data varies from 5 months to 40 months. Although the average WER increases as the temporal distance between the adaptation data and evaluation data increases, the WER is seen to fluctuate over time (particularly for the bn99en_1 data). The results for the Mandarin system given in Table 5 are seen to be more irregular than for the English system. The irregularity shows that the old data contains useful information for adaptation and sometimes is even better than more recent data. It may be that the possible gain is limited since all of the available texts were used to train the original language model and no additional data were used. If additional adaptation data were available, there may be larger differences observed over time.

4.3 Story-based LM adaptation

In section 2, the idea of story-based language model adaptation was introduced. The initial hypotheses of the speech recognizer (the result of the first decoding pass) are automatically segmented into individual stories, where each story concerns a single topic. Then a topic-related corpus is extracted from the training data and used to build a story-specific adaptive LM. Given the topic-related corpus, two ways of building the adaptive LM are explored. In the first one, the story-specific adap-

LM	*bn99en_1*	*bn99en_2*	*average*
baseline model	18.3	16.3	17.1
show based MDI model	18.1	15.9	16.8
story based models	17.7	15.4	16.3

Table 6: WER (%) of the BN English system with story based LMs.

LM	*h4ne97ma*
baseline	17.8
show based MDI model	17.2
story based models	16.8

Table 7: CER (%) of the BN Mandarin system with story based LMs.

tation corpora for all stories in the BN show are combined together to estimate a single show-based LM. The second method builds separate adaptive LMs for each of the stories, and the speech segments are decoded using the corresponding story dependent LM.

The experiments reported in the previous subsections all made use of show-based adaptive language models, i.e., the entire show which contains stories on a variety of topics, were decoded using a common adaptive LM. Therefore, the adaptive LM includes the adaptation information for multiple topics. However, for any particular story, it is generally the case that only the information on a single topic is particularly useful for language modeling and the information from other topics may even have a negative effect.

The IR based adaptation corpus selection method is used to extract story specific adaptation data for each story. Then the MDI adaptation is carried out to train an LM for every story. Since the story segmentation method starts from the result of audio partitioner, the story boundaries are aligned with the audio speech segments. Therefore, it is straightforward to use a different story based LM to decode the speech segments corresponding to the story.

The results for English and Mandarin systems using story based LMs in the second and the third decoding passes are given in Tables 6 and 7 respectively. It can be seen that the story-based language models result in better performance than the baseline systems and the show-based adapted LMs. For the English system, the improvement with the best show-based LM (the MDI model and dynamic models) has an absolute gain of 0.3% whereas the story based LM gives a 0.8% absolute gain. For the Mandarin system, the gain of the best show based LM (the MDI model) is 0.6% absolute, while the story based LM bring a gain of 1.0%.

5. CONCLUSIONS

The work reported in this paper extends our previous work on information retrieval based unsupervised language model adaptation for a broadcast news transcription system, where IR techniques are used to select the adaptation data. A series of experiments have been carried out exploring how to use the selected adaptation data. Five adaptation methods were investigated: linear interpolation, MAP adaptation, mixture models, dynamic mixture models and MDI adaptation. All five methods have been tested for American English and Mandarin Chinese transcription. The experimental results show that MDI method yields the best performance for both languages. The dynamic mixture model gives the same performance as MDI adaptation for the English system, but was inferior to MDI models for the Mandarin data. The MAP-based method was the least successful for both languages.

The influence of the temporal period of the adaptation data on recognition performance was also assessed. While the recognition results were seen to degrade slightly as the time distance increased between the training and test epochs for English, the results on Mandarin were quite irregular. This may be in part due to the use of the same text corpus for training the general language model and to select the adaptation data.

Language model adaptation at the show level was compared with adaptation at the story level. The experiments show that story-based models trained on only topic-specific adaptation data outperform show-based models.

The IR based LM adaptation methods are seen to improve the performance on the BN transcription task. The results of our experiments suggest that the best way to build an adaptive LM within our adaptation framework is to estimate an adapted LM with the MDI method on a story basis. This approach led to a 0.8% absolute gain for the English 10x BN system and 1.0% absolute gain for Mandarin 10x BN system.

REFERENCES

[1] C. Chelba, F. Jelinek, "Structured Language Modeling," *Computer speech and language,* **14**(4):283-332, 2000.

[2] L. Chen, L. Lamel, G. Adda and J.L. Gauvain, "Broadcast News Transcription in Mandarin," *ICSLP'00*,**II**:1015-1018. 2000.

[3] L. Chen, L. Lamel, J.L. Gauvain and G. Adda, "Unsupervised Language Model Adaptation for Broadcast News." *ICASSP'03*, 2003.

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

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

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

[7] R. Iyer and M. Ostendorf, "Relevance Weighting for Combining Multi-Domain Data for N-Gram Language Modeling," *Computer Speech and Language*, **13**:267-282, 1999.

[8] R. Kneser and V. Steinbiss, "On the Dynamic Adaptation of stochastic Language Modeling," *ICASSP'93*, **2**:586-598.

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

[10] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive statistical Language Modeling," *Computer speech and language,* **10**:187-228, 1996.

[11] E. W. D. Whittaker, "Temporal Adaptation of Language Models", *ISCA Adaptation Methods for Speech Recognition Workshop,* 203-206, 2001.