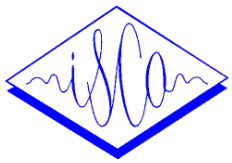
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

LANGUAGE MODEL ADAPTATION FOR BROADCAST NEWS TRANSCRIPTION *

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

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

ABSTRACT

This paper reports on language model adaptation for the broadcast news transcription task. Language model adaptation for this task is challenging in that the subject of any particular show or portion thereof is unknown in advance and is often related to more than one topic. One of the problems in language model adaptation is the extraction of reliable topic information from the audio signal, particularly in the presence of recognition errors. In this work, we draw upon techniques used in information retrieval to extract topic information from the word recognizer hypotheses, which are then used to automatically select adaptation data from a large general text corpus. Two adaptive language models, a mixture-based model and a MAP-based model, have been investigated using the adaptation data. Experiments carried out with the LIMSI Mandarin broadcast news transcription system gives a relative character error rate reduction of 4.3% by combining both adaptation methods.

1. INTRODUCTION

Adaptation techniques have been used widely in speech recognition. Some methods for acoustic model adaptation, such as MLLR and MAP, have been quite successfully used for a variety of tasks to account for mismatches in training and testing conditions. Unfortunately, reported attempts at language model adaptation have been less successful [1]. On the broadcast news transcription task, some previous work has reported that language model adaptation actually decreased performance [2] There are at least three reasons that make language model adaptation difficult for the broadcast news transcription task. Firstly, because of the wide variety of broadcast news data, a given show is almost always related to more than one topic; secondly, the content of such news data is open, that is new stories appear without advance warning, which makes it impossible to get adaptation data in advance; thirdly, there can be large differences between training and test data (for example, the majority of the training text materials come from newspaper sources which can differ significantly from transcripts of spoken language).

This paper reports recent work on language model adaptation for the broadcast news transcription task, applying some methods widely used in the information retrieval (IR) area. Since the content of news broadcasts can vary widely and there is no readily available set of adaptation data, it is hard to directly use traditional adaptation methods. We found that using the hypothesis of an initial decoding pass as adaptation data for modifying the mixture weights did not improve the speech recognition performance. We attribute this to the small amount of data in each hypothesis and to the presence of transcription errors. However, the recognizer hypothesis is the only text available for topic determination. We therefore decided to adopt the same type of approach commonly used in information retrieval, that is to use the hypothesis to select a subset of texts from the training corpus which can then serve for language model adaptation. This approach reduces the effect of transcription errors in the hypothesis and at the same time provides substantially more textual data for LM estimation.

Various attempts have been made to use topic information in the speech recognizer's models. Of all these methods, the mixture-based language model [3] is one of most widely used. The approach is to interpolate all of the topic-dependent language models using a set of mixing factors. By tuning the mixing factors, the interpolated language model can be adapted to new domains. Another effective and easy to use method is based on MAP (*maximum a posteriori*) adaptation [4]. Given a general model and a corpus of adaptation data, a domain-dependent model can be obtained according to the maximum a posteriori criterion.

The remainder of this paper is organized as follows. The next section provides a description of the two approaches explored for language model adaption. Section 3 describes the method for topic selection and topic-dependent language model training. Section 4 reports on how the adaptation data is selected, and is followed by the experimental results section and some conclusions.

2. ADAPTIVE LANGUAGE MODEL

Two approaches for estimating adaptive language models have been investigated. The first method is based on mixture models, the second is based on maximum a posteriori adaptation.

In the mixture-based model approach, the training corpus is clustered into subcorpora corresponding to different topics, and topic-dependent models are trained on each of the topic-specific subsets. For a trigram, mixture based models can be expressed as:

$$\Pr(w_i|w_{i-2},w_{i-1}) = \sum_{k=0}^{N} \lambda_k * \Pr(w_i|w_{i-2},w_{i-1},M_k)$$

where N is number of the topic models and model M_0 is a general model trained on the entire corpus. The interpolation weights are trained using the EM algorithm so as to maximize the likelihood of the adaptation corpus.

*This work was partially financed by the European Commission under the IST-1999 Human Language Technologies project Coretex.

MAP is another efficient method for language model adaptation. Given a set of adaptation data, a general language model is tuned to the specific topic according to a maximum a posteriori criterion. The MAP based model can be expressed as:

$$\Pr(w_i|w_{i-1}, w_{i-2}) = \frac{C_I(w_{i-2}, w_{i-1}, w_i) + \varepsilon * C_A(w_{i-2}, w_{i-1}, w_i)}{C_I(w_{i-2}, w_{i-1}) + \varepsilon * C_A(w_{i-2}, w_{i-1})}$$

where ε is the weight of the contribution of the adaptation data.

3. CLUSTERING THE TEXTS

In order to build the mixture based model, the first problem is solve is clustering the training texts. Most of the previously reported work on text clustering make use of either a hill-climbing algorithm or a k-means algorithm to maximum the likelihood of the training data [2, 7]. In this work, we aim to find a clustering method which groups the most topic-specific information. The text clustering is seen as a process of extracting articles on a specific topic from a large general text corpus. Our clustering algorithm is based on a list of topic-dependent keywords, where a keyword is a word judged to be representative of a particular topic. A keyword usually occurs frequently in articles related to the specific topic, and seldomly occurs in others articles. In information retrieval systems, the commonly used measure to evaluate how much topic information a word has in a given document is the *idf*∗*tf*. *idf* is the inverse document frequency defined as:

$$idf(w) = log_{10}(\frac{N}{df(w)})$$

where N is the total number of documents and $df(w)$ is the number of documents containing the word w. *tf* is the term frequency in the article. In this work, we are not interested in the words in any particular article, but want to select all the keywords from the whole training corpus. The following function is used:

$$idf(w) * \frac{1}{N_w} \sum_{k=1}^{N_w} df_k(w)$$

where 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 selected as keywords. In our experiments, this method was used to select a set of about 3000 keywords.

All the keywords are clustered into different topics. Two methods for keyword clustering have been explored. In the first method, 8 topics were manually selected. These correspond to very broad categories: international politics, national politics, economics, sports, legal issues, history, arts & leisure, and science & technology. The second method is an automatic one which uses a hierarchical clustering with no a priori selection of the total number of topics. The keywords are represented as a feature vector in an n-dimensional space, where n is the number of unique keywords. The coefficients of the vectors are the frequencies of co-occurrence of the given keyword and other keywords in an article. A bottom-up, tree-based hierarchical clustering procedure was used to cluster keywords into different topics. At first, every keyword belongs to a different cluster. The similarity between different clusters is calculated according to cosine coefficient:

$$sim(a, b) = \frac{\sum_{i=1}^{n} C_a(w_i) * C_b(w_i)}{\sqrt{\sum_{i=1}^{n} C_a^2(w_i)} * \sqrt{\sum_{i=1}^{n} C_b^2(w_i)}}$$

If cluster a is the nearest neighbor of cluster b and at the same time, cluster b is also the nearest neighbor of cluster a, then clusters a and b are merged to a new cluster and their associated feature vectors are merged. After all possible merges are made, the result is the new cluster set. This process is iteratively carried out until the desired keyword clusters are obtained. Each keyword cluster represents a topic which will be used to build the topic-dependent language models.

Since every keyword cluster contains the most representative keywords for a given topic, articles related to the partiular topic can be selected by using these keywords. Although information retrieval techniques are applied, there are some important differences for this work. In an typical information retrieval task, there is a high demand for both a low miss probability and a low false alarm probability. In our case, we favor a lower miss probability in the hope of getting as many topic-specific articles as possible. This is because including a small number of out-of-topic articles is expected to not be too damaging for language modeling. Another factor is that we need to deal with a very large corpus containing about 500,000 articles, so the process must be fast. Given these considerations, the following simple method is used to collect topic-specific articles: if an article contains a number of topic-specific keywords exceeding an empirically determined threshold, it is included in the topic dependent corpus. Using the two methods described above, the general corpus was clustered into 8 topics with manual selection and 198 topics by automatic clustering and the corresponding topic-dependent language models were trained.

4. SELECTING THE ADAPTATION DATA

An adaptation corpus serves as the basis for LM adaptation both for the mixture-based model and the MAP-based model. When the content of speech data can relate to multiple topics as is the case for broadcast news, it is difficult to obtain appropriate adaptation data. In addition, in the case of BN transcription, the only text from which the topic information can be derived are the speech recognition hypotheses. These recognition hypotheses are typically quite short and are subject to recognition errors, so it is not advisable to use these directly as adaptation data. In our experiments, using the initial hypotheses directly for LM adaptation slightly degraded the recognition performance.

In order to address the problem of obtaining appropriate adaptation data, a method similar to that used to cluster the text has been used. Instead of using the initial recognizer hypotheses directly, they are used as a query to information retrieval system to select similar texts from a general corpus. This approach reduces the effect of transcription errors and at the same time provides substantially more textual data for LM estimation. The method includes 3 steps:

1. Initial hypothesis segmentation: The recognition hypotheses often include texts on multiple topics, which should be segmented into individual stories each associated with a single topic. As a first approximation to paragraphs we make use of the segment boundaries located by the audio partitioner [8], which are the points of speaker change found by a maximum likelihood segmentation/clustering iterative procedure based on Gaussian mixture models. Since these paragraphs are usually shorter than real stories, it is necessary to combine successive paragraphs. The cosine coefficient is used to measure the similarity of two paragraphs, and neighboring paragraphs with a high similarity are merged. This process is iterated until no more merges are possible. The result of this step is the 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 topic information are selected. Unlike the keyword selection for text clustering where keywords are selected from the whole general corpus, here only the keywords that are most relevant to the specific story are chosen. The relevance of each content word w_i in story s_j, is given by the following score:

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

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

3. Retrieving relevant articles: The N selected content words for each story are used to retrieve relevant texts in the training corpus. This process is also similar to the text clustering described above, but here a more accurate criterion is used so as to reduce the false alarm probability as much as possible. The on-topic articles are selected according to the following score:

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

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. In this manner, language model adaptation texts are selected for each story.

The selected articles are used as adaptation data for both methods, that is they are used to train the interpolation weights of mixture-based models and to adapt a general model to specific topics according to the MAP criterion.

5. EXPERIMENTAL RESULTS

Experimental results are reported for the Mandarin Chinese broadcast news transcription system recently developed at LIMSI [9]. The acoustic training material consists of about 24 hours of manually transcribed broadcasts from two radio sources VOA and KAZN-AM (Los Angeles), and one TV source CCTV (Beijing). The language models were trained on text corpora containing about 186M characters, and the transcriptions of the acoustic training data (460k characters). The texts come from three sources: China Radio International (1994-1996, 86.7M characters); People's Daily (1991-1996, 89.2M characters); Xinhua News Agency (1994-1996, 9.9M characters). Both the acoustic and language model training data used to develop the system were distributed by the LDC. The transcription system was evaluated on the 1997 NIST Hub4 Mandarin test data, containing 1h of speech from the same sources as the acoustic training data.

The LIMSI transcription system[8] has two main phases: audio partitioning and speaker-independent continuous speech recognition. The speech recognizer makes use of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling and n-gram language models. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. Unsupervised acoustic model adaptation is carried out between decoding passes. Our framework of language model adaptation uses the initial hypotheses generated in step 1 to select the adaptation data to train an adaptive language model which replaces the static 4-gram language model used in the final decoding (step 3).

Amount of adaptation data	*198 models*	*8 models*
0 articles (0 words)	406	431
100 articles (103k words)	382	417
200 articles (203k words)	381	397
300 articles (289k words)	376	396
400 articles (377k words)	383	396
500 articles (459k words)	379	396
600 articles (541k words)	380	396

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

The reference word based 4-gram LM was trained on all the text data and interpolated with a trigram model trained on the transcriptions of the acoustic training data. Results obtained with the topic adaptive models are compared to those obtained with this reference language model. The two mixture-based models are obtained by clustering all of the texts into the 8 clusters (manual) and 198 clusters (automatic), respectively, training trigram models for each cluster. The perplexity is used as a first measure for comparing the different adaptive models. While the perplexity is evidently not the same as measuring LM performance in speech recognition, it still gives some information about the accuracy of the language model. In particular, when the computation time is long for recognition, the perplexity can provide a preliminary idea of the expected model performance. Since in this work the adaptation data is automatically selected, some off-topic articles will inevitably be introduced into the data set, thus it is important to get an idea of the relationship between the size of the adaptation corpus and the model accuracy.

Table 1 shows the perplexity of the test data as a function of the amount of adaptation data for the mixture model. The first column gives the results of a general 4-gram interpolated with a trigram model trained on the transcriptions of the acoustic training data (the reference model) and the 198 topic-dependent language models. The second column shows the results when reference model is interpolated with the 8 topic-dependent models. It can be seen that even a small corpus (100 articles) of adaptation data can decrease the perplexity of the test data. As the size of the adaptation corpus increases from 100 to 600 articles, the perplexity of the test data is seen to flucuate with the 198 topic models. We think the this may be due to the introduction of more off-topic articles as the size of the corpus increases, which affects the model's performance. The results with the 8 topic models are not as good as with the 198 models, but they are more stable – the perplexity of the test data does not flucuate as the amount of adaptation data increases.

Table 2 gives the perplexity of test data as a function of the amount of data available for MAP adaptation. The perplexity of the MAP-based adaptive model decreases continuously, as the amount of adaptation data increases. This implies that large corpora are well-suited to the MAP-based model, and that the MAP-based model is robust to errors introduced by the IR system.

The recognition performance was assessed for a selected set of adapted LMs as reported in Table 3 in terms of the character error rate (CER) for the following 10 configurations:

1- Reference BN Mandarin language model (text 4-gram and acoustic transcripts 3-gram)

2,3- Mixture models+hypotheses: the reference LM interpolated

Amount of adaptation data	*Perplexity*
0 articles (0 words)	447
300 articles (289k words)	424
3000 articles (2698k words)	413
6000 articles (4250k words)	406
9000 articles (5878k words)	401
12000 articles (8059k word	398
15000 articles (9950k words)	396
20000 articles (12854k words)	393
30000 articles (18338k words)	391

Table 2: Perplexity of MAP models vs. adaptation corpus size.

Language Model	*PPX*	*CER (%)*
reference BN model	447	18.5
198 mixture models+hypotheses	388	18.6
8 mixture models+hypotheses	399	18.6
198 mixture models+reference transcript	369	18.0
8 mixture models+reference transcript	388	18.1
198 mixture models+selected corpus (300)	376	18.0
8 mixture models+selected corpus (300)	396	18.1
MAP+selected corpus(30000)	376	18.0
198 mixture models+MAP+selected corpus	375	17.7
8 mixture models+MAP+selected corpus	370	18.1

Table 3: Perplexity (ppx) and character recognition error rate (CER) results on the 1997 NIST Hub4 Mandarin evaluation set.

with 198 and 8 topic-dependent LMs respectively. The interpolation weights were trained directly with the recognition hypotheses from the first decoding pass.

4,5- Mixture models+reference transcript: the reference LM interpolated with 198 and 8 topic-dependent LMs respectively. The interpolation weights were trained with the manual reference transcriptions of the test data (no recognition errors).

6,7- Mixture models+selected corpus: the reference LM interpolated with 198 and 8 topic-dependent LMs respectively. The interpolation weights were trained using the automatically selected adaptation data (300 articles).

8- MAP+selected corpus: the reference LM tuned to the topic using the automatically selected adaptation data (30000 articles) and the MAP criterion.

9,10- Mixture models+MAP+selected corpus: the reference LM tuned to the topic model by MAP, then interpolated with 198 and 8 topic-dependent LMs respectively. The interpolation weights were trained using the automatically selected adaptation data.

It can be seen in Table 3 that using the automatic transcripts to directly train the interpolation weights degraded (slightly) the system performance despite a substantial decrease in perplexity. This degradation can be attributed to the presence of recognition errors since adaptation using the reference transcripts improves performance. Using an automatically selected corpus of adaptation data, both the mixture model and MAP adaptation approaches are seen to outperform the reference model. This performance is the same as with supervised adaptation via the reference transcripts. Although the adaptive mixture models and MAP model are trained on data selected from the same general text corpus, combining the 198 mixture models with the MAP model yields the largest relative error reduction of 4.3%. Interpolating the 8 mixture models with MAP-tuned models does not improve performance, even though this LM configuration has the lowest perplexity.

6. CONCLUSIONS

In this paper we have proposed using information retrieval methods to select adaptation data for language model adaptation in broadcast news transcription. IR methods are powerful for extracting useful information from texts and for tracking topics in large corpora. These methods have been used to retrieve on-topic articles from a large general text corpus, which are in turn used to train topic-specific language models. The initial recognition hypotheses are automatically segmented into (approximate) single topic stories based on an automatic audio partition, and then used to automatically select the adaptation data.

Two approaches to building topic-dependent language models have been investigated: mixture-based models and MAP-based models. Experiments based on perplexity showed that the off-topic data affect the stability of the mixture models (with 198 topics), while a large set of adaptation data is well-suited to the MAP-based model, even though the large corpus contains some off-topic articles.

When the mixture-based and MAP-based models are combined, differences in performance are observed between 8 and 198 mixtures. Combining the 8 mixture models with MAP-based model results in the lowest test data perplexity, but does not improve the CER compared to interpolating the reference LM with the 8 mixtures. For 198 mixture models, combining the mixture model with the MAP-tuned models reduces the character error rate by an additional 2.7% relative. Although both of these methods make use of the same text source for the adaptation data, by combining them better performance was obtained than with either method alone. The combined relative reduction in CER is 4.3% which represents a significant gain for this kind of system with n-gram language models trained on large amounts of data.

REFERENCES

[1] R. Rosenfeld, "Two decades of Statistical Language Modeling: Where Do We Go From Here?" *Proc. of the IEEE,* Aug 2000.

[2] P. Clarkson, T. Robinson, "The Applicability of Adaptive Language Modeling for the Broadcast News Task," *ICSLP-98*, **1**, 233-236.

[3] R. Kneser, V. Steinbiss. "On the Dynamic Adaptation of Stochastic Language Modeling," *ICASSP-93*, **2**:586-589.

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

[5] D.E. Appelt, D. Martin, "Named Entity Extraction From Speech: Approach and Results Using the Textpro System," *Proc. DARPA Broadcast News Workshop*, 51-54, 1999.

[6] J.M. Schultz, M. Liberman, "Topic Detection and Tracking using idf-Weighted Cosine Cefficient," *Proc. DARPA Broadcast News Workshop*, 189-192, 1999.

[7] R. Kneser, J. Peters, "Semantic Clustering for Adaptive Language Modeling, *ICASSP-97*, 779-782.

[8] J.L. Gauvain L. Lamel G.Adda, M. Jardino, "The LIMSI 1998 Hub-4E Transcription System," *Proc. DARPA Broadcast News Workshop*, 99-104, 1999.

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