10.21437/Eurospeech.2001-86

USING INFORMATION RETRIEVAL METHODS FOR LANGUAGE MODEL ADAPTATION [*]

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

In this paper we report experiments on language model adaptation using information retrieval methods, drawing upon recent developments in information extraction and topic tracking. One of the problems is extracting reliable topic information with high confidence from the audio signal in the presence of recognition errors. The work in the information retrieval domain on information extraction and topic tracking suggested a new way to solve this problem. In this work, we make use of information retrieval methods to extract topic information in the word recognizer hypotheses, which are then used to automatically select adaptation data from a very 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% with this adaptation method.

1. INTRODUCTION

Language model adaptation recognized to be an important research domain in speech recognition. However, despite numerous efforts to improve upon the commonly used n-gram language models, for complex tasks such as broadcast news transcription, language model (LM) adaptation has been only moderately successful. This can be attributed at least partially to the wide domain of broadcast news data which reduces the efficiency of adaptation techniques. Previous work on language model adaptation has focused mainly on the algorithms used for adaptation. The difficulty of getting reliable topic information from complicated texts in the presence of recognition errors is one of the reasons that language model adaptation is challenging.

In this paper, we present a method to address the problem of language model adaptation in broadcast news (BN) transcription which is based on automatically detecting topics in the audio data. Topic information is important for language modeling, and various attempts have been made to use topic information in the speech recognizer's models. Kuhn and De Mori[1] developed a cache model based on the observation that word a which occurred in a recent text has a higher probability to be seen again. Niesler and Woodland [2] simulated the word co-occurrence in texts with trigger-target model. Iyer and Ostendorf [3] combined training data from different domains using a similarity weight to improve the performance of an in-domain model.

Of the proposed methods, mixture based language model [4] is one of most widely used. The approach is to interpolate all of the topic-dependent language models using a set of the 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 [5]. Given an adaptation corpus and a general model, we can obtain a domain dependent model according to maximum a posteriori criterion. (The MAP based model is similar to Iyer's work.) In this paper, both of these models are explored.

Although traditional language model adaptation methods have been shown to improve the word error rate and/or perplexity for specific tasks, such approaches have been less successful on the broadcast news transcription task. In previous work [6] it was reported that even though an adapted language model decreased the perplexity, it did not improve the performance of the BN transcription system. There are at least three reasons that can account for this observation. Firstly, because of the wide variety of test data, a given audio segment is almost always related to more than one topic; secondly, the content of test data is open, that is new stories appear without advance warning, which makes it impossible to obtain adaptation data in advance; and thirdly, there can be large differences between the training and test data.

We propose a new method using information retrieval (IR) technology to solve the linguistic problem in adaptive language modeling for BN transcription. IR methods provide powerful tools to extract different types of information from text data [7]. Many problems brought from complicated language environment of broadcast news can be addressed using IR methods. For example, the algorithms developed for topic detection and tracking [8] can be used to deal with multi-topic data. IR methods have been used in text corpus clustering for large general corpora, topic information extraction from multi-topic, open-content texts and adaptation corpus selection. We use these methods to extract topic information from the initial recognizer hypotheses when an adaptation corpus is unavailable.

The remainder of this paper is organized as follows. In the next section the two approaches used for language model adaption are described briefly. Section 3 describes the method for topic selection and topic dependent language model training, and Section 4 reports on the extraction of adaptation data, followed by the experimental results and conclusions.

2. ADAPTIVE LANGUAGE MODEL

Two approaches to adaptive language models have been investigated, one is based on mixture models, the other is based on a maximum a posteriori adaptation.

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

The mixture based models cluster the training corpus into subcorpora corresponding to different topics, and train topic dependent models on each of the topic 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.

MAP 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 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 adaption data.

3. TEXT CLUSTERING

In order to build the mixture based model, the first problem is to cluster the training texts. Previously reported work on text clustering usually made use of a hill-climbing algorithm or k-means algorithm to maximum the likelihood of training data [6, 9]. In this work, we want to find a clustering method which groups the most topic specific information. We look at text clustering as a process of extracting articles on a specific topic from the large general corpus. Our clustering algorithm is based on a list of topic dependent keywords, where a keyword is a word that is representative of a particular topic. A keyword usually occurs frequently in articles related to the specific topic, and seldomly occurs in others articles. In an information retrieval system, the *idf*tf* is a commonly used measure to evaluate how much topic information a word has in a given document. 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. 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, about 3000 keywords were selected using this method.

All the keywords are clustered into different topics, using a hierarchical clustering method 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 process 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, we can select the articles related to the topic according to these keywords. Although we apply information retrieval techniques, there are some important differences for this work. In an 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. Small amounts of out-of-topic articles are not expected to 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 enough. Considering these reasons, 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 this method, the general corpus was clustered into 198 topics and the respective topic dependent language models were trained.

4. ADAPTATION DATA SELECTION

An adaptation corpus serves as the basis for LM adaptation both for the mixture based model and the MAP based model. However for many tasks, it is not possible to obtain the adaptation data in advance. Also, when the content of speech data can relate to multiple topics, it is more difficult to get the adaptation data. In the case of BN transcription, the only text from which the topic information can be derived is the speech recognition hypotheses. The recognition hypotheses are quite small 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 degraded recognition performance.

In order to solve this problem, a method similar to corpus clustering has been used. Instead of using the initial hypotheses directly, we use them as a query as in an information retrieval system to select similar texts from 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 has 3 steps:

1. Initial hypothesis segmentation: The recognition hypotheses almost always 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 [10], 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 regroup 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 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 topic information are selected. Unlike the keyword selection in text clustering which selects keywords from the whole general corpus, here we select the keywords that are most relevant to the specific story. 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 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 topic of story 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 selected N 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 we use a more accurate criterion to reduce the false alarm probability as much as possible. The on-topic articles are selected according to the following score:

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

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. Likewise language model training texts are selected for each story.

The selected articles are used as an adaptation data 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 [11]. The acoustic and language model training data used to develop the system were distributed by the LDC. 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). 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[10] has two main phases: audio partitioning and speaker independent continuous speech recognition. The continuous speech recognizer makes use of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. 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 4-gram static language model originally in the final decoding step 3.

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 with the topic adaptive models are compared with this reference language model. The mixture models are obtained by clustering all of the texts into 198 clusters, and training trigram models for each cluster. The perplexity is used as a first comparison of the adaptive models. While the perplexity is not equivalent to 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 selected automatically, some off-topic articles will inevitably be introduced into the data set, so it is important to get an idea of the relationship between the size of 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 models. A general 4-gram was interpolated with 198 topic dependent language models and a trigram model trained on the transcriptions of the acoustic training data. It can be seen that even a small corpus pf adaptation data can decrease the perplexity of test data. As the size of the adaptation corpus increases from 100 to 600 articles, the perplexity of test data is seen to flucuate. We think the reason is that as the size of the corpus increases, there are more errors (off-topic articles) introduced by information retreival process which influences the model's performance.

Amount of adaptation data	*Perplexity*
0 articles (0 words)	406
100 articles (103k words)	382
200 articles (203k words)	381
300 articles (289k words)	376
400 articles (377k words)	383
500 articles (459k words)	379
600 articles (541k words)	380

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

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.

Table 2 gives the perplexity of test data as a function of the amount of data available for MAP adaptation. Compared with Table 1, we find that unlike the mixture models, as the amount of adaptation data increases, the perplexity of the MAP based adaptive model decreases continuously. 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 has been assessed for a selected set of adapted LMs. The performance is reported in Table 3 in terms of the character error rate (CER) for the following 6 configurations:

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

2- Mixture models+hypotheses: the reference LM interpolated with 198 topic dependent LMs, where the interpolation weights were trained directly with the recognition hypotheses from the first decoding pass.

3- Mixture models+reference transcript: the reference LM interpolated with 198 topic dependent LMs, but the interpolation weights were trained with the manual reference transcriptions of the test data (remove recognition errors).

4- Mixture models+selected corpus: the reference LM interpolated with 198 topic dependent LMs, where the interpolation weights were trained using the automatically selected adaptation data (300 articles).

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

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

language model	*perplexity*	*CER*
reference BN model	447	18.5%
mixture models+hypotheses	388	18.6%
mixture models+reference transcript	369	18.0%
mixture models+selected corpus (300)	376	18.0%
MAP+selected corpus(30000)	376	18.0%
mixture models+MAP+selected corpus	375	17.7%

Table 3: CER results on the 1997 NIST Hub4 Mandarin eval set.

It can be seen in Table 3 that using the automatic transcripts to directly train the interpolation weights degraded the system performance despite a substantial decrease in perplexity. The degradation can be attributed to the 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 provide a better result than the reference model, reducing the CER by about 2.7% relative. This performance is the same as with supervised adaptation. Although the adaptive mixture models and MAP model are trained data selected from the same general text corpus, combining the two methods results in the largest relative error reduction of 4.3%.

6. CONCLUSION

In this paper we have proposed a method for building topic dependent language models using information retrieval methods. This method can be used when the content of the speech is related to multiple topics and when adaptation data is not available in advance. IR methods are powerful for extracting useful information from texts and for tracking topics in large corpora. We use these methods to extract on-topic articles from a large general text corpus, which are in turn used to train topic specific language models. These methods are used with the initial recognition hypotheses after automatic segmentation into (approximate) single topic stories based on an automatic audio partition to automatically select an the corpus used for adaptation.

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 have large effect on the mixture models, 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.

Both the mixture based and MAP based models reduced the recognition relative character error rate by 2.7%. Although both of these methods make use of the same source for the adaptation data, by combining them better performance was obtained than with either method alone. The combine 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. Kuhn, R. De Mori, "A Cache-Based Natural Language Model for Speech Reproduction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(6):570-583, 1990.

[2] T.R. Niesler, P.C. Woodland, "Modeling Word-Pair Relations in a Category-Based Language Model," *ICASSP-97*, 795-798.

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

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

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

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

[7] 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.

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

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

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

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