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

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

STATISTICAL LANGUAGE MODEL ADAPTATION FOR MANDARIN BROADCAST NEWS TRANSCRIPTION

Berlin Chen, Wen-Hung Tsai, Jen-Wei Kuo

Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei
{berlin, louis, rogerkuo}@csie.ntnu.edu.tw

ABSTRACT

This paper investigates statistical language model adaptation for Mandarin broadcast news transcription. A topical mixture model was proposed to explore the long-span latent topical information for dynamic language model adaptation. The underlying characteristics and various kinds of model complexities were extensively investigated, while their performance was verified by comparison with the conventional MAP-based adaptation approaches, which are devoted to extracting the short-span *n*-gram information. The speech recognition experiments were conducted on the broadcast news collected in Taiwan. Very promising results in both perplexity and word error rate reductions were initially obtained.

1. INTRODUCTION

Statistical language modeling, which aims to capture the regularities in a natural language and quantify the acceptance of a given word sequence, has continuously been an important research issue in speech and language processing over the past two decades [1]-[2]. The *n*-gram modeling (especially the bigram and trigram modeling) approach, which determines the probability of a word given the previous *n*-1 word history, has been shown very powerful and is most prominently used in practice [3]-[5]. However, for complicated speech recognition tasks such as broadcast news transcription, it is still extremely difficult to build well-estimated language models, because the subject matters and lexical characteristics for the linguistic contents of news articles are very diverse and are often changing with time. Various attempts have been made to adapt the language model by taking advantage of either the contemporary or in-domain text articles [6]-[7]. Two of the most widespread approaches to language model adaptation are count merging and model interpolation, which can be respectively viewed as a maximum *a posteriori* (MAP) language model adaptation with a different parameterization of the prior distribution and can be easily integrated into the *n*-gram modeling framework to capture the local regularities of word usage in the new task domain [8]. In contrast, the latent semantic analysis (LSI) approach originally formulated for relevance measures in a vast variety of IR tasks also has been proposed to explore the latent topical factors for language model adaptation [9]-[10]. LSI transforms the high-dimensional vector representations of a word and a document (or a search history) into a lower dimensional space (the so-called latent semantic space). The relevance measure can be estimated in the reduced space and then be transformed into an approximate probability measure. However, LSI is mainly based on linear algebra operations, and thus is much more deterministic and lacks for a solid statistical foundation for automatic model refinement or optimization. With these observations in mind, in this paper, a topical mixture model (TMM) previously proposed for information retrieval [11] is investigated to dynamically explore the long-span latent topical information for language model adaptation. The TMM model was first trained beforehand on a set of contemporary or in-domain text articles and then can be gradually optimized when being applied to speech recognition. Structures similar to the presented approach have also been investigated recently [12]-[15]. The main differences between the presented approach and the previous ones are that we explicitly interpret the document (or the search history) as a mixture model used to predict the newly occurring word, and both the perplexity and word error rate experiments are simultaneously investigated with very good potential indicated. Besides, various kinds of model complexities are extensively tested and their performance is compared with the conventional MAP-based adaptation approaches, which are devoted to extracting the short-span *n*-gram information. The speech recognition experiments were carried out on the broadcast news collected in Taiwan.

2. THE NTNU BROADCAST NEWS SYSTEM

The major constituent parts of the broadcast news system developed at National Taiwan Normal University (NTNU) as well as the speech and language data used in this paper will be described in this section [7].

2.1. Front-End Processing

The front-end processing is conducted with the data-driven LDA-based (Linear Discriminant Analysis) feature extraction approach. The states of each HMM were taken as the unit for class assignment. The outputs of 18 filter banks are chosen as the basic vectors. The basic vectors from every nine successive frames were spliced together to form the time-frequency supervectors for the construction of the LDA transformation matrix, which was then used to project the supervectors to a lower feature space. The dimension of the resultant vectors was set to 39. Utterance-based Cepstral mean subtraction and variance normalization were further applied.

2.2. Speech Corpus and Acoustic Training

The speech data set consists of about 112 hours of FM radio broadcast news, which were collected from several radio stations located at Taipei in the 1998-2002 period [16]. All the speech materials were manually segmented into separate stories, and each of them is a news abstract pronounced by one anchor speaker. Some stories contain background noise and music. Only 7.7 hours of speech data is equipped with

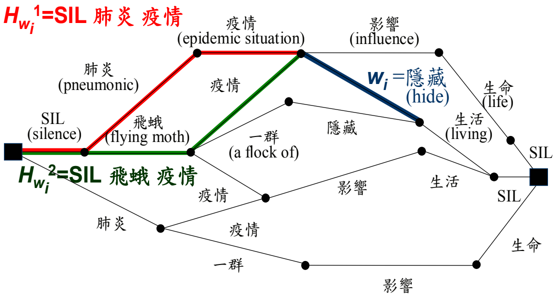


Figure 1: An illustration of the word graph, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis.

corresponding orthographic transcripts, in which about 4.0 hours of data collected during 1998 to 1999 is used to bootstrap the acoustic training and the other 3.7 hours of data (506 stories) collected in September 2002 is for testing. The rest 104.3 hours of untranscribed speech data (about 18,000 stories) is reserved for unsupervised acoustic training [7]. The acoustic models chosen for speech recognition are 112 right-context-dependent INITIAL's and 38 context-independent FINAL's, specially considering the phonetic structure of Mandarin syllables [16]. Each INITIAL is represented by an HMM with 3 states while each FINAL with 4 states. Gender-independent models were used.

2.3. Lexicon and *N*-gram Language Modeling

The recognition lexicon initially consists of 67K words. A set of about 5K compound words was automatically derived using the forward and backward bigram statistics and was then added to the lexicon to form a new lexicon of 72K words [7]. The background language models used in this paper consist of trigram and bigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2000 and 2001 (the Chinese Gigaword Corpus released by LDC). On the other hand, another text corpus of about 39,000 news articles (20 million Chinese characters) collected from CNA during August to October 2002 is taken as the contemporary corpus for language model adaptation. The *n*-gram language models were trained with Katz backoff smoothing using the SRI Language Modeling Toolkit (SRILM) [17].

2.4. Speech Recognition

The speech recognizer was implemented with a left-to-right frame-synchronous tree search as well as a lexical prefix tree organization of the lexicon [18]. At each speech frame, the so-called word-conditioned method grouped the path hypotheses that shared the same history of predecessor words to the same copies of the lexical tree, and expanded and recombined them according to the tree structure until reaching a possible word ending. At word boundaries, the path hypotheses among the tree copies that had the equivalent search history were recombined, and were then propagated into the existing tree copies or used to start up new ones in case that they did not exist yet. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their unigram language model look-ahead and syllable-level acoustic look-ahead scores [7], was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores higher than a predefined threshold, their associated decoding information, such as the start and end speech frames, the identities of current and predecessor words, and the acoustic score, will be kept in order to build a word graph for further language model rescoring. Once the word graph had been built, as illustrated in Figure 1, the Viterbi beam search with a more sophisticated language model was conducted on it to generate the most likely word sequence. In the baseline system, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word graph rescoring procedure.

3. TOPICAL MIXTURE MODEL (TMM)

We first explain the structural characteristics of the topical mixture model (TMM) previously investigated in information retrieval and then show how it can be applied to speech recognition.

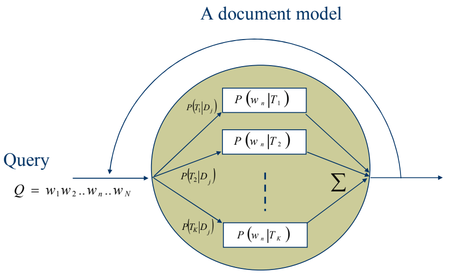
In information retrieval, the relevance measure between a query Q and a document D_j can be expressed as $P(D_j|Q)$; i.e., the probability that the document D_j is relevant given that the query Q was posed. Based on Bayes' theorem and some assumptions, this measure can be approximated by $P(Q|D_j)$, which stands for the probability of the query Q being posed, under the hypothesis that document D_j is relevant [11]. The query Q is treated as a sequence of input observations (terms or words), $Q = w_1 w_2 .. w_n .. w_N$, where the query terms are assumed to be conditionally independent given the document D_j. Therefore, the relevance measure $P(Q|D_j)$ can be decomposed as a product of the probabilities of the query terms generated by the document:

$$P(Q|D_j) = \prod_{n=1}^{N} P(w_n|D_j). \tag{1}$$

Each individual document D_j can be interpreted as a mixture model as shown in Figure 2, which is just a special case of HMM. In the model, a set of K latent topical distributions characterized by unigram language modeling are used to predict the query terms, and each of the latent topics is associated with a document-specific weight. That is, each document can belong to many topics. The relevance measure therefore can be further expressed as:

$$P(Q|D_j) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(w_n|T_k) P(T_k|D_j), \tag{2}$$

where $P(w_n|T_k)$ denotes the probability of the query term w_n occurring in a specific latent topic T_k, and $P(T_k|D_j)$ is the posterior probability (or weight) of topic T_k conditioned on the document D_j, with the constraint $\sum_{k=1}^{K} P(T_k|D_j) = 1$ imposed. More precisely, the topical unigram distributions, e.g. $P(w_n|T_i)$, are tied among the entire document collection, while each document D_j has its own probability distribution over the latent topics, e.g. $P(T_k|D_j)$. The key idea we wish to illustrate here is that the relevance measure of a query term w_n and a document D_j is not computed directly based on the



$$P(Q|D_j) = \prod_{n=1}^{N} P(w_n|D_j)$$
$$= \prod_{n=1}^{N} \sum_{k=1}^{K} P(w_n|T_k) P(T_k|D_j)$$

Figure 2: The TMM model for a specific document D_j .

frequency of w_n occurring in D_j , but instead based on the frequency of w_n in the latent topic T_k as well as the likelihood that D_j generates the respective topic T_k, which in fact exhibits some sort of concept matching. During training, the K-means algorithm [19] is first used to partition the entire document collection into K topical classes. Hence, the initial topical unigram distribution for a cluster topic can be estimated according to the underlying statistical characteristics of the documents being assigned to it, and the probabilities for each document generating the topics are measured according to its proximity to the centroid of each respective cluster as well. Then, given a training set of query exemplars with the corresponding query-document relevance information, each document mixture model can be optimized in a supervised manner by the expectation-maximization (EM) algorithm [20].

While the TMM retrieval model is applied to language model adaptation, a set of contemporary (or in-domain) articles are first collected and used to train their corresponding mixture models. However, because there is no any query exemplar provided for the document model to be trained, we simply use each individual document in the collection as a query to train its own mixture model in an unsupervised manner. In speech recognition, for each newly occurring word w_i (as $w_i =$ 隱藏 shown in Figure 1), its corresponding search history H_{w_i} (as $H^1_{w_i}$ or $H^2_{w_i}$ shown in Figure 1), can be treated as a document. The corresponding document TMM model can be optimized using the EM algorithm, throughout the whole search process. In this work, we keep the topic factors $P(w_i|T_k)$ unchanged, but let the search history's probability distribution over the latent topics, $P(T_k|H_{w_i})$, be gradually updated as path extension is performed during the search process. Once the TMM model for a search history is estimated, it can thus be used to predict the occurrence probability of the newly occurring word w_i (acting here as a single query word):

$$P_{TMM}(w_i|H_{w_i}) = \sum_{k=1}^{K} P(w_i|T_k) P(T_k|H_{w_i}). \tag{3}$$

Such a kind of language model probability to some extent dynamically captures the underlying global topical information of the path hypothesis and can be further combined with the background n-gram (e.g. trigram) language probability, which provides the general constraint information of lexical regularities, to form an adapted language model for guiding the search process:

$$\widetilde{P}_{Adapt-1}(w_i|w_{i-2}w_{i-1}) = \lambda \cdot P_{TMM}(w_i|H_{w_i}) + (1-\lambda) \cdot P_{Back.}(w_i|w_{i-2}w_{i-1}), \tag{4}$$

where $P_{TMM}(w_i|H_{w_i})$ and $P_{back}(w_i|w_{i-2}w_{i-1})$ are, respectively, the TMM probability and background trigram probability, and λ is a tunable weighting parameter.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

As mentioned earlier in Section 2.3, a text corpus (about 39,000 news articles) collected from CNA during August to October 2002 is taken as the contemporary data, which is postulated to be temporally consistent with the broadcast news speech to be tested and therefore can be used to explore the latent topical and local contextual information which might be helpful for speech recognition. For the TMM approach, its training and special utilization for language model adaptation have been described earlier in Section 3. While for the two conventional MAP-based adaptation approaches to be used here for comparison [8], i.e., count merging and model interpolation, the adaptation formulae (e.g. for trigram modeling) can be respectively written as:

$$\widetilde{P}_{Adapt-2}(w_i|w_{i-2}w_{i-1}) = \frac{\alpha \cdot C_{d,Cont}(w_{i-2}w_{i-1}w_i) + \beta \cdot C_{d,Back}(w_{i-2}w_{i-1}w_i)}{\alpha \cdot C_{Cont}(w_{i-2}w_{i-1}) + \beta \cdot C_{Back}(w_{i-2}w_{i-1})}, \tag{5}$$

and

$$\widetilde{P}_{Adapt-3}(w_i|w_{i-2}w_{i-1}) = \gamma \cdot P_{Cont}(w_i|w_{i-2}w_{i-1}) + (1-\gamma) \cdot P_{Back}(w_i|w_{i-2}w_{i-1}). \tag{6}$$

For the count merging formula in Equation (5), $C_{d,Cont}(w_{i-2}w_{i-1}w_i)$ and $C_{d,Back}(w_{i-2}w_{i-1}w_i)$ are respectively the discounted trigram counts [3] accumulated from the contemporary and background text corpora, $c_{Cont}(w_{i-2}w_{i-1})$ and $c_{Back}(w_{i-2}w_{i-1})$ are respectively the bigram counts accumulated from the contemporary and background text corpora as well, and α and β are tunable weighting parameters; while for the model interpolation formula in Equation (6), $P_{Cont}(w_i|w_{i-2}w_{i-1})$ and $P_{Back}(w_i|w_{i-2}w_{i-1})$ are the trigram probabilities respectively estimated from the contemporary and background text corpora, and γ is a tunable weighting parameter. Detailed derivation of Equations (5) and (6) can be found in [8].

In this paper, the language model adaptation experiments were performed in the word graph rescoring procedure, as described in Section 2.4. A set of 506 broadcast news stories collected in September 2002 is used for testing. For each broadcast news story to be processed, its associated word graph was built beforehand by the tree search and using the background bigram language model.

4.2. Experiments on TMM-based Adaptation Approach

All experimental results are reported in Table 1, in which the baseline result (in Row 2) was obtained by performing word graph rescoring with the background trigram language model alone. A character error rate (CER) of 15.22% and a perplexity (PP) of 752.49 were initially obtained. We first evaluate the performance of the TMM adapted language model shown in Equation (4) by varying the model complexities (the number of latent topics is ranged from 16 to 256). The weighting parameter λ in Equation (4) was initially set to 0.1 in this research. The results by using the TMM adapted language models are respectively shown in Rows 3 to 7. As can be seen from the table, both the CER and PP are steadily reduced as the topic mixture number increases. A best

	CER (%)	PP
Baseline	15.22	752.49
+ TMM(16 Topics)	14.83 (2.56%)	576.95 (23.33%)
+ TMM(32 Topics)	14.73 (3.22%)	555.93 (26.12%)
+ TMM(64 Topics)	14.58 (4.20%)	529.49 (29.63%)
+ TMM(128 Topics)	14.53 (4.53%)	492.40 (34.56%)
+ TMM(256 Topics)	14.47 (4.93%)	457.74 (39.17%)
+ Count Merging	13.70 (9.99%)	458.79 (39.03%)
+ Model Interpolation	13.74 (9.72%)	430.59 (42.78%)
+ TMM(256 Topics) + Count Merging	13.27 (12.81%)	306.75 (59.24%)
+ TMM(256 Topics) + Model Interpolation	13.37 (12.16%)	311.28 (58.63%)

Table 1: The language model adaptation results expressed in terms of character error rates (CERs) and perplexities (PPs).

result of CER of 14.47% (4.93% relative reduction) and PP of 457.74 (39.17% relative reduction) is obtained when the TMM topic number is set to 256. Though the performance seems not to be saturated yet, these results clearly demonstrate the effectiveness of the TMM-based approach for dynamic language model adaptation.

4.3. Comparisons with MAP-based Adaptation Approaches

Then, the experimental results as the count merging and model interpolation adaptation approaches are respectively applied are listed in Rows 9 and 10 for comparison. The parameter settings for them are: $\alpha=1$ and $\beta=3$ in Equation (5) and $\gamma=0.5$ in Equation (6), respectively. As can be seen, these two approaches are quite comparable to each other. Count merging is slightly better than model interpolation in CER reduction, while model interpolation is slightly better in PP reduction. If we further compare the TMM-based approach with the MAP-based ones, it can be found that the TMM approach is competitive with the MAP approaches in PP reduction, but only reaches half of the CER reduction as that provided by the MAP approaches, which also implies that the local word regularity (or contextual) information obtained from the contemporary corpus is still vital for speech recognition and should be taken into account when performing language model adaptation.

4.4. Fusion of Topical and Contextual Information

Thus, we combine the TMM-based approach with the MAP-based approaches in an attempt to explore both the subject domain and lexical regularity information embedded in the contemporary corpus for language model adaptation. The results are shown in the last two rows of Table 1. As can be observed, the fusion of these two kinds of information sources does provide additional gains. For example, the combination of the TMM-based approach with the count merging approach achieves the best reduction of 12.81% in CER and 59.24% in PP.

5. CONCLUSIONS

In this paper we presented a topical mixture model for dynamic language model adaptation. The underlying characteristics and different kinds of model complexities were extensively investigated and tested. We compared it with two conventional MAP-based approaches. The fusion of global topical and local contextual information has been investigated as well. Very promising results in both perplexity and word error rate reductions were initially obtained. More in-deep investigation and analysis of the TMM-based approach as well as comparison to other approaches are currently undertaken.

6. REFERENCES

[1] R. Rosenfeld, “Two Decades of Statistical Language Modeling: Where Do We Go from Here,” *Proc. IEEE*, 88 (8), 2000.

[2] J. R. Bellegarda, “Statistical Language Model Adaptation: Review and Perspectives,” *Speech Communication 42*, 2004.

[3] S. F. Chen, J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” *Computer Speech and Language 13*, 1999.

[4] W. B. Croft (editor), J. Lafferty (editor), *Language Modeling for Information Retrieval*, Kluwer-Academic Publishers, 2003.

[5] P. Beyerlein et al., “Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH approach,” *Speech Communication 37*, 2002.

[6] M. Federico, N. Bertoldi, “Broadcast News LM adaptation Using Cotemporary Texts,” in *Proc. Eurospeech 2001*.

[7] B. Chen et al., “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription, ” in *Proc. ICASSP 2004*.

[8] M. Bacchiani, B. Roark, “Unsupervised Language Model Adaptation” in *Proc. ICASSP 2003*.

[9] W. Chou (editor),. B.H. Juang (editor), *Pattern Recognition in Speech and Language Processing*, Chapter 9, CRC Press, 2003.

[10] W. Kim, S. Khudanpur, “Cross-Lingual Latent Semantic Analysis for Language Modeling,” in *Proc. ICASSP 2004*.

[11] B. Chen et al., “Statistical Chinese Spoken Document Retrieval Using Latent Topical Information,” in *Proc. ICSLP 2004*.

[12] D. Gildea, T. Hoffmann, “Topic-based Language Models Using EM,” in *Proc. Eurospeech 1999*.

[13] T. Hoffmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” *Machine Learning 42*, 2001.

[14] S. Wang et al., “Semantic N-gram Language Modeling with the Latent Maximum Entropy Principle,” in *Proc. ICASSP 2003*.

[15] D. Mrva and P. C. Woodland, “A PLSA-based Language Model for Conversational Telephone Speech,” in *Proc. ICSLP2004*.

[16] B. Chen et al., “Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese,” *IEEE Trans. on Speech and Audio Processing 10(5)*, 2002.

[17] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[18] X. L. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language* 16, 2002.

[19] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley Sons, 1973.

[20] A. P. Dempster et al., “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. Royal Statistical Society B*, Vol. 39, 1977.