WORD TOPICAL MIXTURE MODELS FOR DYNAMIC LANGUAGE MODEL ADAPTATION

Hsuan-Sheng Chiu and Berlin Chen

Department of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
{g93470240, berlin}@csie.ntnu.edu.tw

ABSTRACT

This paper considers dynamic language model adaptation for Mandarin broadcast news recognition. A word topical mixture model (TMM) is proposed to explore the co-occurrence relationship between words, as well as the long-span latent topical information, for language model adaptation. The search history is modeled as a composite word TMM model for predicting the decoded word. The underlying characteristics and different kinds of model structures were extensively investigated, while the performance of word TMM was analyzed and verified by comparison with the conventional probabilistic latent semantic analysis-based language model (PLSALM) and trigger-based language model (TBLM) adaptation approaches. The large vocabulary continuous speech recognition (LVCSR) experiments were conducted on the Mandarin broadcast news collected in Taiwan. Very promising results in perplexity as well as character error rate reductions were initially obtained.

***Index Terms*—**word topical mixture model, probabilistic latent semantic analysis, trigger-based language model, language model adaptation, speech recognition

1. INTRODUCTION

As we know, for complicated speech recognition tasks, such as broadcast news transcription, it is difficult to build well-estimated language models, because the subject matters and lexical characteristics for the linguistic contents of broadcast news speech to be transcribed are very diverse and are often changing with time. Hence, there is always good reason to dynamically adapt the language models for better speech recognition results. On the other hand, the conventional *n*-gram modeling approach is inadequate. The *n*-gram modeling approach can only capture the local contextual information or the word regularity, and it will be a problem for *n*-gram modeling when there is mismatch of word usage between training and test conditions.

In the recent past, the latent topic modeling approaches, which were originally formulated in information retrieval (IR) [1], have been introduced to dynamic language model adaptation and investigated to complement the *n*-gram models as well. Among them, the latent semantic analysis (LSA) [2] and the probabilistic latent semantic analysis (PLSA) [3] have been widely studied. LSA has been demonstrated effective in a few of speech recognition tasks; however, its derivation is based on linear algebra operations. Though PLSA-based language model (PLSALM) has a probabilistic framework for model optimization, it merely targets on maximizing the collection likelihood but not directly on its language model prediction capability, and it also suffers from the problem that part of its model parameters have to be dynamically estimated on the fly during the speech recognition process, which would be time-consuming and makes it impractical for real-world speech recognition applications. Moreover, there are also other approaches developed to complement the *n*-gram models, such as the trigger-based language model (TBLM) [4, 5], for which word trigger pairs are automatically generated to capture the co-occurrence information among words. By using TBLM, the long-distance relationship between the words in the search history and the currently predicted word can be captured to some extent.

Based on these observations, in this paper a word topical mixture model (TMM), using words as the modeling units, is proposed to explore the co-occurrence relationship between words, as well as the latent topical information inherent in the search histories, for language model adaptation. The underlying characteristics and different kinds of model structures were extensively investigated, while the performance of word TMM was analyzed and verified by comparison with the conventional PLSALM and TBLM models. The large vocabulary continuous speech recognition (LVCSR) experiments were carried out on the Mandarin broadcast news transcription task [6].

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work with the PLSALM and TBLM models. In Section 3, we present our proposed word TMM and elucidate its difference with the other models. Then, the experimental settings and a series of speech recognition experiments conducted are presented in Sections 4 and 5, respectively. Finally, conclusions and future work are given in Section 6.

2. RELATED WORK

2.1. PLSA-based Language Model (PLSALM)

PLSA is a general machine learning technique for modeling the co-occurrences of words and documents, and it evaluates the relevance between them in a low-dimensional factor space [3]. When PLSA is applied to language model adaptation in speech recognition, for a decoded word w_i, we can interpret each of its corresponding search histories H_{w_i} as a history (or document) model M_{Hw_i} used for predicting the occurrence probability of w_i:

$$P_{PLSA}\left(w_i \middle| M_{H_{w_i}}\right) = \sum_{k=1}^{K} P\left(w_i \middle| T_k\right) P\left(T_k \middle| M_{Hw_i}\right), \tag{1}$$

where T_k is one of the latent topics and $P(w_i \mid T_k)$ is the probability of the word w_i occurring in T_k. The latent topic distributions $P(w_i \mid T_k)$ can be estimated beforehand by maximizing the total log-likelihood of the training (or adaptation) text document collection. However, the search histories are not known in advance and their number could be

1-4244-0728-1/07/$20.00 ©2007 IEEE

enormous and varying during speech recognition. Thus, the corresponding PLSA model of a search history has to be estimated on the fly. For example, during the speech recognition process, we can keep the topic factors $P(w_i \mid T_k)$ unchanged, but let the search history's probability distribution over the latent topics $P(T_k \mid M_{Hw_i})$ be gradually updated as path extension is performed, by using the expectation-maximization (EM) updating formulae [7]. Then, the probabilities of the PLSALM and background *n*-gram (e.g., trigram) language models can be combined through a simple linear interpolation:

$$\begin{aligned} P_{Adapt}(w_i|w_{i-2}w_{i-1}) \\ = \lambda_1 \cdot P_{PLSA}(w_i|H_{w_i}) + (1-\lambda_1) \cdot P_{n-gram}(w_i|w_{i-2}w_{i-1}), \end{aligned} \quad (2)$$

where λ_1 is a tunable interpolation weight.

2.2. Trigger-based Language Model (TBLM)

To capture long-distance information, we also can use trigger pairs. A trigger pair $(A \rightarrow B)$ denotes that unit A is significantly and semantically related to unit B within a given context window. The complexity can be reduced by using words as the modeling units. Instead of using the average mutual information (MI) for the selection of trigger pairs [4], the TF/IDF measure which captures both local and global information respectively from the document and the collection can be used for this purpose [5]. Then, the trigger pairs whose constituent words have the average MI scores or the TF/IDF scores higher than a predefined threshold can be selected for language modeling, and the associated conditional probability of the selected trigger pair (w_j, w_i) can be estimated using the following equation:

$$P_{Trig}(w_i \mid w_j) = \frac{n(w_j, w_i)}{\sum_{w_l} n(w_j, w_l)}, \quad (3)$$

where $n(w_j, w_i)$ is the count of co-occurrences of word w_j and w_i within a given context window in the training (or adaptation) text collection, and w_l is an arbitrary word that co-occurs with w_j within the context window. Therefore, the search history H_{w_i} for a decoded word w_i can be viewed as a series of words w_j and the probability of the search history H_{w_i} predicting word w_i can be expressed by linearly combining the conditional probabilities of the trigger pairs (w_j, w_i) as follows:

$$P_{Trig}(w_i \mid H_{w_i}) = \frac{1}{|H_{w_i}|} \sum_{w_j \in H_{w_i}} P_{Trig}(w_i|w_j), \quad (4)$$

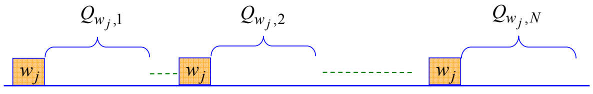
where $|H_{w_i}|$ is the length of the search history H_{w_i} ; w_j is the word in H_{w_i} . The TBLM probability $P_{Trig}(w_i \mid w_j)$ can also be combined with the background *n*-gram model through a simple linear interpolation similar to Eq.(2).

3. WORD TOPICAL MIXTURE MODEL

In this paper, we present an alternative probabilistic latent topic approach by treating each word w_j of the language as a topical mixture model (TMM) M_{w_j} for predicting the occurrences of the other word w_i :

$$P_{TMM}(w_i \mid M_{w_j}) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid M_{w_j}) \quad (5)$$

3.1. Training of Word TMMs



$$Q_{w_j} = Q_{w_j,1}, Q_{w_j,2}, \cdots, Q_{w_j,N}$$

Figure 1: A schematic depiction of the occurrences of a word w_j and its corresponding training observation sequence Q_{w_j} .

Each word TMM M_{w_j} can be trained by concatenating those words occurring within a context window of size N (for simplicity, N is set to 3 in this study) around each occurrence of w_j , which are postulated to be relevant to w_j , in the training (or adaptation) text document collection to form the observation Q_{w_j} for training M_{w_j} . Figure 1 is a schematic depiction of the occurrences of a word w_j and its corresponding training observation sequence Q_{w_j} . The words in Q_{w_j} are assumed to be conditionally independent given M_{w_j} . Therefore, the TMM models of the words of the language can be estimated by maximizing the total log-likelihood of their corresponding training observations respectively generated by themselves:

$$\begin{aligned} \log L_{\mathbf{Q}_{TrainSet}} &= \sum_{Q_{w_j}} \log P\left(Q_{w_j} \middle| M_{w_j}\right) \\ &= \sum_{Q_{w_j}} \sum_{w_n \in Q_{w_j}} n(w_n, Q_{w_j}) \log P_{TMM}\left(w_n \middle| M_{w_j}\right), \end{aligned} \quad (6)$$

where $n(w_n, Q_{w_j})$ is the number of times a word w_n occurring in Q_{w_j} . The word TMM parameters are estimated using the following three EM updating formulae [7]:

$$\hat{P}\left(T_k \middle| M_{w_j}\right) = \frac{\sum_{w_s \in Q} n(w_s, Q_{w_j}) P\left(T_k \middle| w_s, M_{w_j}\right)}{\sum_{w_l \in Q_{w_j}} n(w_l, Q_{w_j})}, \quad (7)$$

$$\hat{P}(w_n | T_k) = \frac{\sum_{w_j} n(w_n, Q_{w_j}) P\left(T_k \middle| w_n, M_{w_j}\right)}{\sum_{w_l} \sum_{w_{n'} \in Q_{w_l}} n(w_{n'}, Q_{w_l}) P(T_k | w_{n'}, M_{w_l})}, \quad (8)$$

$$P\left(T_k \middle| w_n, M_{w_j}\right) = \frac{P\left(T_k \middle| M_{w_j}\right) P(w_n | T_k)}{\sum_{l=1}^{K} P\left(T_l \middle| M_{w_j}\right) P(w_n | T_l)}. \quad (9)$$

3.2. Speech Recognition using Word TMMs

During the speech recognition process, for a decoded word w_i , we can again interpret it as a (single-word) observation. While for each of its search histories $H_{w_i} = w_1, w_2, \ldots, w_{i-1}$, we can linearly combine the associated TMM models of the words involved in H_{w_i} to form a composite word TMM model for predicting w_i :

$$\begin{aligned} P_{TMM}\left(w_i \middle| M_{H_{w_i}}\right) &= \sum_{j=1}^{i-1} \alpha_j P_{TMM}\left(w_i \middle| M_{w_j}\right) \\ &= \sum_{j=1}^{i-1} \alpha_j \sum_{k=1}^{K} P(w_i | T_k) P\left(T_k \middle| M_{w_j}\right) \\ &= \sum_{k=1}^{K} P(w_i | T_k) P'\left(T_k \middle| M_{H_{w_i}}\right), \end{aligned} \quad (10)$$

where the values of the nonnegative weighting coefficients α_j are empirically set to be exponentially decayed as the word w_j is being apart from w_i and summed to 1 ($\sum_{j=1}^{i-1} \alpha_j = 1$); and the search history's probability distribution over the latent topics $P'(T_k | M_{H_{w_i}})$ is thus represented as:

$$P'\left(T_k \middle| M_{H_{w_i}}\right) = \sum_{j=1}^{i-1} \alpha_j P\left(T_k \middle| M_{w_j}\right). \quad (11)$$

Figure 2 is a schematic representation of speech recognition using a composite word TMM model.

It is noteworthy that unlike PLSALM where the topic mixture weights trained with the training (or adaptation) collection are entirely discarded during the speech recognition process, the topic mixture weights of word TMM are instead retained and exploited. A nice feature of word TMM is that given a training set of speech utterances equipped with corresponding correct and recognized transcripts, we can further explore the use of discriminative training algorithms, such as the minimum word error (MWE) training [8] and so forth, to train the mixture weights of the word TMM model to correctly discriminate the recognition hypotheses for the best recognition results rather than just to fit the model distributions. We also can combine the word TMM and the background *n*-gram probabilities through a simple linear interpolation similar to Eq.(2).

3.3. Issue on the Amount of Training Observations

As mentioned previously, during the training of a word TMM M_{w_j}, those words w_i that occur within a context window of size N around each occurrence of w_j in the training (or adaptation) text document collection will be included in the training observation Q_{w_j} for training M_{w_j}. However, the training observation Q_{w_j} will become considerably larger as the context window increases. Therefore, in this paper, we investigated the use of two statistical measures to remove words w_i that are not quite related or relevant to word w_j from the training observation Q_{w_j} of the word TMM M_{w_j} for the purpose of speeding up the training process. The first statistical measure we used in the paper is mutual information (MI). The MI score of a word pair (w_j, w_i) co-occurring within the context window can be computed as follows:

$$Score_{MI}(w_j, w_i) = \log \frac{P(w_j, w_i)}{P(w_j)P(w_i)}. \tag{12}$$

The second statistical measure is the geometrical average of the forward-backward (FB) bigrams [9] of a word pair (w_j, w_i) co-occurring within the context window. For example, the FB score of a word pair w_j and w_i can be computed as follows:

$$Score_{FB}(w_j, w_i) = \sqrt{P_f(w_i \mid w_j) P_b(w_j \mid w_i)}. \tag{13}$$

Based on these two scores, respectively, we can rank the total word pairs of the training collection and then select different portions of the word pairs with higher statistical scores for the training of the word TMM models.

3.4. Comparison of Word TMM, PLSALM and TBLM Models

Word TMM, PLSALM and TBLM can be compared from four perspectives. First, PLSALM models the co-occurrence relationship between words and documents (or search histories), while word TMM and TBLM directly model the co-occurrence relationship between words. Second, PLSALM needs to update the distributions of the search histories over the latent topics on the fly; however, word TMM and TBLM can be trained in an offline manner. Third, PLSALM and word TMM model the topics of sentences or words with explicit distributions, while TBLM models topics implicitly. Finally, word TMM has $V \times K \times 2$ parameters, PLSALM has $V \times K + K \times D$ parameters and TBLM at most has $V \times V$ parameters; where V is the size of the vocabulary, K is the number of the latent topics and D is the number of the documents used for training (or adaptation).

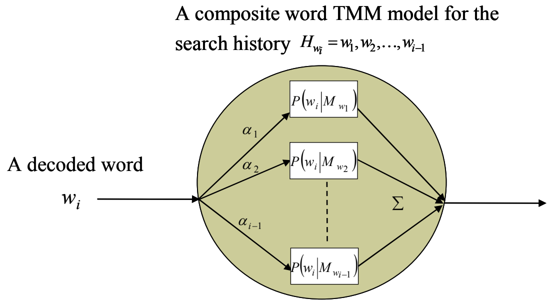


Figure 2: Speech recognition using a composite word TMM model.

4. EXPERIMENTAL SETUP

The speech corpus consists of about 200 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News), which were collected by Acdemia Sinica and Public Television Service Foundation of Taiwan [10]. All the speech materials were manually segmented into separate stories. About 25 hours of gender-balanced speech data of the field reporters collected during November 2001 to December 2002 were used to bootstrap the acoustic training. Another set of 1.5-hour speech data of the field reporters collected within 2003 were reserved for testing. On the other hand, the acoustic models chosen here for speech recognition are 112 right-context-dependent INITIALs and 38 context-independent FINALs. The acoustic models were first trained with the maximum likelihood (ML) criterion and then by the minimum phone error rate (MPE) criterion [8].

The *n*-gram language models used in this paper consist of trigram and bigram models, which were estimated using a background text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The vocabulary size is about 72 thousand words. The *n*-gram language models were trained with the Katz backoff smoothing using the SRI Language Modeling Toolkit (SRILM) [11]. The adaptation text corpus used for training word TMM, PLSALM and TBLM is collected from MATBN 2001 and 2002, which consists of one million Chinese characters of the orthographic transcripts.

In this paper, the language model adaptation experiments were performed in the word graph rescoring procedure. The associated word graphs of the 1.5-hour speech test data were built beforehand by a tree search procedure [6] and using the background bigram language model.

5. EXPERIMENTAL RESULTS

The baseline trigram system results in a character error rate (CER) of 20.79% and a perplexity (PP) of 667.23. We first compare the performance of word TMM with those of PLSALM and TBLM. The weights respectively for the interpolation of word TMM, PLSALM and TBLM with the background trigram were all tuned at optimum values. As can be seen from Table 1, the performance of both word TMM and PLSALM tends to be better as the topic mixture number

	CER (%)	PP
Baseline (Trigram)	20.79	667.23
Word TMM	CER (%)	PP
16 topics	19.80	520.31
32 topics	19.76	510.26
64 topics	19.69	507.13
128 topics	19.55	499.30
PLSALM	CER (%)	PP
16 topics	20.13	540.52
32 topics	20.06	533.07
64 topics	19.99	527.82
128 topics	19.95	519.71
TBLM	CER (%)	PP
MI (51,594)	20.11	507.54
MI (465,722)	19.99	467.62
TF/IDF (88,310)	20.63	614.70
TF/IDF (914,159)	20.25	501.01

Table 1: The results of word TMM, PLSALM and TBLM.

increases, and word TMM slightly outperforms PLSALM in all cases. The best PP and CER results for word TMM are 19.55% (5.96% relative reduction) and 499.30 (25.17% relative reduction), respectively, while for PLSALM are 19.95% (4.04% relative CER reduction) and 519.71 (22.11% relative PP reduction), respectively. Our postulation for the superiority of word TMM over PLSALM is that, for word TMM, both the latent topic distributions and the distributions of words over the latent topics, estimated from the adaptation corpus, were fully exploited during the speech recognition process, while for PLSALM, only the latent topic distributions were exploited. On the other hand, the results of TBLM achieved by using either MI or TF/IDF for the selection of word trigger pairs are also listed in the lower part of Table 1, where the numbers in the parentheses represent the actual size of TBLM when different selection thresholds were applied. The CER performance of word TMM is always better than that of TBLM with different selection criteria and thresholds; however, the best PP result of word TMM is worse than the best result (467.62) of TBLM with the MI criterion.

We also try to investigate the influence of the amount of training observations on the performance of word TMM. The results are depicted in Figure 3. As can be seen, by appropriately using the FB scores for training observation selection, we can obtain almost the best CER results even though the word TMM models of different complexities (16 or 128 mixtures) were trained with reduced observations. Moreover, when the ratio of training observations exploited during training is in the interval of 0.4 to 0.8, the results achieved by using the FB scores consistently outperform that done by using the MI scores in most cases. Therefore, it is concluded here that with the aid of the FB scores for training data selection, the word TMM models trained with reduced observations still can retain the same performance level as the word TMM trained with the whole training observations.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a word topical mixture model (TMM) for dynamic language model adaptation. The underlying characteristics and different kinds of model structures were extensively investigated and tested. We compared it with the PLSA- and TBLM-based approaches. Very promising results in both perplexity and character error rate reductions were initially obtained. More in-deep investigation and analysis of the word TMM-based approaches, such as the discriminative training of the word TMM models, and their possible applications to spoken document summarization and organization, are currently undertaken [12].

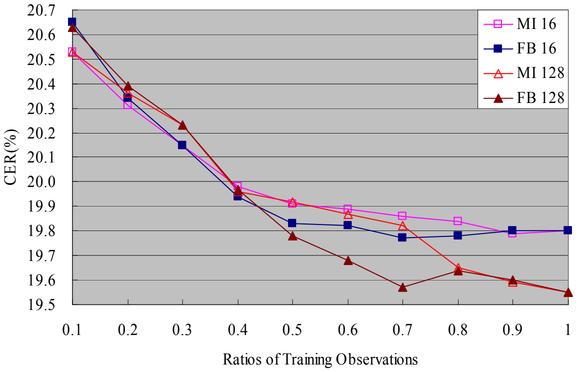


Figure 3: The relationship between the performance, ratios of training observations and selection strategies for word TMM.

7. REFERENCES

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

[2] J. R. Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine* 22(5), 2005.

[3] D. Gildea, T. Hofmann, "Topic-based language models using EM," in Proc. *Eurospeech 1999*.

[4] R. Lau et al., "Trigger-based language models: A maximum entropy approach," in Proc. *ICASSP 1993*.

[5] C. Troncoso, T. Kawahara, "Trigger-Based Language Model Adaptation for Automatic Meeting Transcription," in Proc. *Eurospeech 2005*.

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

[7] A. P. Dempster et al., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B* 39(1), 1977.

[8] J. W. Kuo, B. Chen, "Minimum Word Error Based Discriminative Training of Language Models," in Proc. *Eurpspeech 2005*.

[9] Saon, G. and M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing* 9(4), 2001.

[10] H. M. Wang et al., "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing* 10(1), 2005.

[11] A. Stolcke, "SRI language Modeling Toolkit," version 1.5, http://www.speech.sri.com/projects/srilm/.

[12] B. Chen et al., "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," in Proc. *ICASSP 2006*.