LATENT TOPIC MODELING OF WORD VICINITY INFORMATION FOR SPEECH RECOGNITION

Kuan-Yu Chen¹, Hsuan-Sheng Chiu¹,², Berlin Chen¹

¹National Taiwan Normal University, Taipei, Taiwan
²Delta Electronics, Inc., Taipei, Taiwan
{g96470203, g93470240, berlin}@csie.ntnu.edu.tw

ABSTRACT

Topic language models, mostly revolving around the discovery of “word-document” co-occurrence dependence, have attracted significant attention and shown good performance in a wide variety of speech recognition tasks over the years. In this paper, a new topic language model, named word vicinity model (WVM), is proposed to explore the co-occurrence relationship between words, as well as the long-span latent topical information for language model adaptation. A search history is modeled as a composite WVM model for predicting a decoded word. The underlying characteristics and different kinds of model structures are extensively investigated, while the performance of WVM is thoroughly analyzed and verified by comparison with a few existing topic language models. Moreover, we also present a new modeling approach to our recently proposed word topic model (WTM), and design an efficient way to simultaneously extract “word-document” and “word-word” co-occurrence characteristics through the sharing of the same set of latent topics. Experiments on broadcast news transcription seem to demonstrate the utility of the presented models.

Index Terms—speech recognition, topic language model, word vicinity model, broadcast news transcription

1. INTRODUCTION

Language modeling is an indispensable ingredient in any speech recognizer since it can be used to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output from a speech recognizer. The n-gram language model [1, 2, 3] that follows a statistical modeling paradigm is the most prominently-used in speech recognition because of its simplicity and predictive power. Nevertheless, the n-gram language model, aiming at capturing only the local contextual information or the lexical regularity of a language, is inevitably faced with the problem of missing the information (either semantic or syntactic information) conveyed in the history before the immediately preceding n-1 words of a newly decoded word.

In the past decade, the latent topic modeling approaches, originally formulated in information retrieval (IR) [4, 5], have been introduced to speech recognition and investigated to complement the n-gram language model as well. Among them, the probabilistic latent semantic analysis (PLSA) [4, 6] and the latent Dirichlet allocation (LDA) [5, 7] are often considered two basic representatives of this category. They both introduce a set of latent topic variables to describe the “word-document” co-occurrence characteristics [8]. The dependence between a decoded word and its search history (regarded as a document) is based on the frequency of the word in the latent topics as well as the likelihood that the search history generates the respective topics, which in fact exhibits some sort of concept matching. On the other hand, instead of treating each search history as a whole as a document topic model (DTM), such as PLSA and LDA, we have recently proposed a word topic model (WTM) [9, 10] that attempts to discover the long-span “word-word” co-occurrence dependence through a set of latent topics; a given search history consequently can be represented as a composite WTM model for predicting an observed word.

This paper serves two main purposes. First, a new topic language model, named word vicinity model (WVM), is proposed to explore the co-occurrence relationship between words, as well as the long-span latent topical information for language model adaptation. WVM is close in spirit to WTM, but has a more concise parameterization leading to more reliable model estimation, particularly when available training data is of limited size. Second, we present a new modeling approach to WTM, and also design an efficient way to simultaneously extract “word-document” and “word-word” co-occurrence dependence inherent in the adaptation corpus through the sharing of a common set of latent topics.

The rest of this paper is organized as follows. In Section 2, we briefly review the existing topic models. Section 3 introduces WVM and elucidates its difference with the other models, while an approach to simultaneously capturing “word-document” and “word-word” natural relationships is presented in Section 4. Then, the experimental settings and a series of speech recognition experiments are presented in Sections 5 and 6, respectively. Finally, conclusions and future work are given in Section 7.

2. RELATED WORK

2.1. Document Topic Model (DTM)

DTM introduces a set of latent topic variables to describe the “word-document” co-occurrence characteristics. The dependence between a decoded word and one of its search histories (regarded as a document) is not computed directly based on the frequency of the word occurring in the history, but instead based on the frequency of these words in the latent topics as well as the likelihood that the history generates the respective topics, which in fact exhibits some sort of concept matching. PLSA and LDA are often considered to be two basic representatives of this category.
To take PLSA \([4, 6]\) for example, 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 \(H_j\) as a document (or history) topic model used for predicting the occurrence probability of \(w_i\):

\[
P_{\text{PLSA}}(w_i \mid H_j) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid H_j)
\]

where \(T_k\) is one of the latent topics; \(P(w_i \mid T_k)\) and \(P(T_k \mid H_j)\) are respectively the probability of the word \(w_i\) occurring in \(T_k\) and the probability of \(T_k\) conditioned on the history \(H_j\). The latent topic distribution \(P(w_i \mid T_k)\) can be estimated beforehand by maximizing the total log-likelihood of the training (or adaptation) text document collection \([4]\). However, the search histories are not known in advance and their number could be enormous and varying during speech recognition. Thus, the corresponding topic mixture weight \(P(T_k \mid H_j)\) of a search history has to be estimated on the fly, using inference algorithms like expectation-maximization (EM). Then, the probabilities of PLSA and the background \(n\)-gram (e.g., trigram) language model can be combined through a simple linear interpolation:

\[
P_{\text{Adapt}}(w_i \mid w_{i-2} \ldots w_{i-1}) = \lambda \cdot P_{\text{PLSA}}(w_i \mid H_j) + (1 - \lambda) \cdot P_{\text{n-gram}}(w_i \mid w_{i-2} \ldots w_{i-1})
\]

On the other hand, LDA \([5, 7]\), having a formula analogous to PLSA for language modeling in speech recognition, is regarded as an extension to PLSA and has enjoyed much success for various speech recognition tasks. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes the model parameters are fixed and unknown; while LDA places additional a priori constraints on the model parameters, i.e., thinking of them as random variables that follow some Dirichlet distributions. Since LDA has a more complex form for model optimization, which is hardly to be solved by exact inference, several approximate inference algorithms, such as the variational approximation algorithm \([5]\), the expectation propagation method and the Gibbs sampling algorithm \([11]\), hence have been proposed for estimating the parameters of LDA. In this paper, the parameters of LDA are estimated by the variational approximation algorithm.

### 2.2. Word Topic Model (WTM)

Instead of treating each search history as a document topic model, we can regard each word \(w_i\) of the language as a word topic model (WTM). To get to this point, all words are assumed to share the same set of latent topic distributions but have different weights over these topics. The WTM model of each word \(w_i\) for predicting the occurrence of a particular word \(w_j\) can be expressed by \([9, 10]\)

\[
P_{\text{WTM}}(w_j \mid M_{w_i}) = \sum_{k=1}^{K} P(w_j \mid T_k) P(T_k \mid M_{w_i})
\]

Each WTM model \(M_{w_j}\) can be trained in a data-driven manner by concatenating those words occurring within the vicinity of \(w_j\) in a training or adaptation corpus, which are postulated to be relevant to \(w_j\). To this end, a sliding window with a size of \(S\) words is placed on each occurrence of \(w_j\), and a pseudo-document associated with such vicinity information of \(w_j\) is aggregated consequently. The WTM model of each word can be estimated by maximizing the total log-likelihood of words occurring in their associated “vicinity documents,” using the EM algorithm. Notice that words in such a document are assumed to be independent of each other (the so-called “bag-of-words” assumption).

During speech recognition, for a search history \(H_s = w_i, w_{i+1}, \ldots, w_{j-1}\) of a decoded word \(w_j\), we can linearly combine the associated WTM models of the words occurring in \(H_s\) to form a composite WTM model for predicting \(w_j\):

\[
P_{\text{WTM}}(w_j \mid H_s) = \sum_{k=1}^{K} \alpha_k \cdot P_{\text{WTM}}(w_j \mid M_{w_k})
\]

where the values of the nonnegative weighting coefficients \(\alpha_k\) are empirically set to be exponentially decayed as the word \(w_j\) is being apart from \(w_i\) and summed to one \([9]\). We can combine the search history’s composite WTM model and the background \(n\)-gram language model through a simple linear interpolation similar to Eq.\((2)\).

### 3. WORD VICINITY MODEL (WVM)

#### 3.1. Principle

WVM bears a certain similarity to WTM in its motivation of modeling “word-word” co-occurrences, but has a more concise parameterization. WVM explores the word vicinity information by directly modeling the joint probability of any word pair in the language, rather than modeling the conditional probability of one word given the other word as done by WTM. In this regard, the joint probability of any word pair that describes the associated word vicinity information can be expressed by the following equation, using a set of latent topics:

\[
P_{\text{WVM}}(w_i, w_j) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid T_j) P(w_j \mid T_k)
\]

where \(P(T_k)\) is the prior probability of a given topic \(T_k\). Notice that the relationship between words, originally expressed in high-dimensional probability space, are now projected into a low-dimensional probability space characterized by the shared set of topic distributions.

Along a similar vein, WVM is trained by maximizing the probabilities of all word pairs, respectively, co-occurring within a sliding window of \(S\) words in the training corpus, using the EM algorithm. During speech recognition, we can convert Eq. \((5)\), on the fly, into a conditional probability form, representing one historical word \(w_i\) predicting another decoded word \(w_j\), through a simple mathematical manipulation:

\[
P_{\text{WVM}}(w_i \mid w_j) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid T_j) P(w_j \mid T_k)
\]

A composite WVM model \(P_{\text{WVM}}(w_i \mid H_s)\) of a search history for predicting \(w_i\) can be thus obtained by linearly combining \(P_{\text{WVM}}(w_i \mid w_j)\) of all words \(w_j\) in the history, analogously to WTM (cf. Eq.\((4)\)). The WVM model of the search history also can be further combined with the background \(n\)-gram language model using an equation like Eq.\((2)\).
Figure 1: A schematic illustration for the hybrid of PLSA and WTM.

3.2. Theoretical Comparison with Other Topic Models

DTM (PLSA or LDA), WTM and WVM can be analyzed from several perspectives. First, DTM models the co-occurrence relationship between words and documents, while WTM and WVM directly model the co-occurrence relationship between words in the collection.

Second, for DTM, the topic mixture weights of a new search history have to be estimated online using EM or other more sophisticated algorithms, which would be time-consuming; on the contrary, for WTM and WVM, the topic mixture weights of a new search history can be simply obtained on the basis of the topic mixture weights of all words involved in the search history without using any complex inference procedure.

Third, PLSA has $V \times K + M \times K$ parameters to be estimated, WTM has $2 \times V \times K$ parameters, and both LDA and WVM have the smallest parameter size of $K + V \times K$; $V$ denotes the size of the vocabulary set, while $M$ denotes the number of the documents and $K$ is the number of the latent topics. To take a step forward, the number of the parameters of PLSA to be estimated grows linearly with the number of documents used for training, whereas that of LDA, WTM and WVM instead remains the same regardless of the number of training documents, as the speech recognition system adopts a closed set of vocabulary. Since WVM has a more concise parameterization than WTM, it is expected to obtain a more reliable model estimation given a limited amount of training data.

Finally, for WTM and WVM, we can further assume their model parameters are governed by some Dirichlet distributions, in analogy with that of LDA (see Section 6).

4. HYBRID OF DTM AND WTM

In this paper, we also present a novel approach to simultaneously capturing “word-document” and “word-word” co-occurrence relationship inherent in the corpus through using a common set of latent topics. Here we investigate the hybrid of PLSA and WTM, as graphically illustrated by probabilistic matrix decomposition in Figure 1. On the left-hand side, each column of the matrix can denote either a probability vector of a document in the collection which offers a probability for every word occurring in the document (PLSA), or a probability vector of a word’s “vicinity document” which offers a probability for every other word occurring in its vicinity (WTM). This matrix can be further decomposed into two matrices on the right-hand side of Figure 1, respectively representing the topic mixture components $p(w|\hat{D})$ for both PLSA and WTM, and the topic mixture weights $p(\hat{D}|M_w)$ for PLSA and $p(\hat{D}|M_w)$ for WTM. These model parameters can be estimated by jointly maximizing the total log-likelihood of words occurring in the document collection and the total log-likelihood of words occurring in their associated “vicinity documents.”

During speech recognition, we can linearly combine PLSA and WTM to form a hybrid topic language model:

$$P_{\text{hybrid}}(w|H_w) = \gamma \cdot \hat{P}_{\text{PLSA}}(w|H_w) + (1 - \gamma) \cdot \hat{P}_{\text{WTM}}(w|H_w)$$

where $\hat{P}_{\text{PLSA}}(w|H_w)$ and $\hat{P}_{\text{WTM}}(w|H_w)$ share the same set of latent topic distributions $p(w|\hat{D})$.

5. EXPERIMENTAL SETUP

The speech corpus consists of about 200 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News) [12]. A subset of 25-hour speech compiled during November 2001 to December 2002 was used to bootstrap the acoustic training with the minimum phone error rate (MPE) criterion and the training data selection scheme [12]. Another subset of 3-hour speech data collected within 2003 is reserved for development (1.5 hours) and testing (1.5 hours).

The vocabulary size is about 72 thousand words. The $n$-gram language models used in this paper consist of trigram and bigram models, which were estimated from 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) using the SRI Language Modeling Toolkit (SRILM) [13]. The adaptation text corpus used for training PLSA, LDA, WTM and WVM was collected from MATBN 2001, 2002, and 2003, which consists of one million Chinese characters of the orthographic broadcast news transcripts.

The speech recognition system [14] with the background trigram language model results in a character error rate (CER) of 20.08% and a perplexity (PPL) of 682.10 on the test set.

6. EXPERIMENTAL RESULTS

In the first set of experiments, we compare the performance of WVM with those of PLSA, LDA and WTM on the test set, as a function of different topic numbers. Notice that the constants or weighting (interpolation) coefficients used for language modeling were all tuned at optimum values by the held-out (development) set. As can be seen from Table 1, the performance of all topic language models is apt to become better as the topic number increases. However, such a tendency is more obvious for PPL than for CER. This is mainly because that the language model probability is not the only dominant factor affecting CER, in contrast to PPL, on the test set. To recap, PLSA, LDA, WTM, and WVM have, respectively, 23.1%, 21.8%, 26.1% and 24.2% PPL reductions, and have, respectively, 4.1%, 5.5%, 5.5% and 4.6% CER reductions over the baseline system when using their best model settings. Significance tests based on the standard NIST MAPSSWE [15] indicate the statistical significance of such reductions over the baseline system; namely, it reveals the effectiveness of the latent topic approaches (including WVM) for dynamic language model adaptation. On the other hand, although WVM does not offer additional performance gains as compared to the other topic language models, its inherent simplicity and neat

5396
formulation (cf. Section 3), as well as its efficiency in implementation, still make WVM attractive for speech recognition. It was experimentally shown that for comparable recognition performance, the language model access effort required by WVM or WTM was more than 30 times less than that of PLSA or LDA for language model adaptation.

In the next set of experiments, we first investigate the hybrid of PLSA and WTM in both model training and testing (cf. Section 4), and the corresponding results are shown in the upper left part of Table 2. As is evident, the joint exploration of “word-document” and “word-word” latent topic information is beneficial to CER or PPL reduction for all model complexities, which leads to the best CER of 18.81% and perplexity of 494.91 when the topic number is set to 8. Then, we study the hybrid of LDA and WDTM (Word Dirichlet Topic Model), which is a nature extension of WTM by further assuming that the model parameters of WTM are governed by some Dirichlet distributions. The corresponding results are shown in the lower part of Table 2, which reveal that the hybrid of LDA and WDTM can achieve better results than the hybrid of PLSA and WTM in PPL reduction, but it is not always the case in CER reduction. As we further look into the results achieved by using WDTM alone, WDTM achieves a CER of 19.23%, 19.08% and 18.87%, as well as a PPL of 539.73, 523.11, and 498.44, when the topic number is set to 8, 32 and 128, respectively, which in fact is the best among all topic language models (cf. Table 1).

<table>
<thead>
<tr>
<th></th>
<th>CER (%)</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Trigram)</td>
<td>20.08</td>
<td>682.10</td>
</tr>
<tr>
<td>PLSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.26</td>
<td>553.92</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.26</td>
<td>535.93</td>
</tr>
<tr>
<td>128 topics</td>
<td>19.34</td>
<td>524.60</td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.29</td>
<td>555.34</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.15</td>
<td>536.94</td>
</tr>
<tr>
<td>128 topics</td>
<td>18.97</td>
<td>533.18</td>
</tr>
<tr>
<td>WTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.17</td>
<td>542.65</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.14</td>
<td>521.27</td>
</tr>
<tr>
<td>128 topics</td>
<td>19.13</td>
<td>503.74</td>
</tr>
<tr>
<td>WVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.31</td>
<td>531.63</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.15</td>
<td>523.11</td>
</tr>
<tr>
<td>128 topics</td>
<td>19.23</td>
<td>519.33</td>
</tr>
</tbody>
</table>

Table 2: Comparisons among different topic models.

<table>
<thead>
<tr>
<th></th>
<th>CER (%)</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA+WDTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.32</td>
<td>534.76</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.08</td>
<td>503.60</td>
</tr>
<tr>
<td>128 topics</td>
<td>18.97</td>
<td>478.61</td>
</tr>
<tr>
<td>PLSA+WDTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 topics</td>
<td>19.12</td>
<td>538.18</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.11</td>
<td>513.13</td>
</tr>
<tr>
<td>128 topics</td>
<td>18.81</td>
<td>494.91</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS

In this paper, we have presented a new topic language model, named word vicinity model (WVM), which not only owns a concise parameterization but also has shown effective in the broadcast news transcription task. Moreover, we have explored a new modeling approach to WTM, and have proposed an efficient way to combine different kinds of topic models. Our future research directions include: 1) investigating more elaborate window functions for WVM and WTM, 2) seeking the use of discriminative training algorithms for training WVM and WTM [16], and 3) applying WVM to speech retrieval and summarization tasks.

8. ACKNOWLEDGEMENT

This work was supported in part by the National Science Council, Taiwan, under Grants: 98-2221-E-003-011-MY3, 96-2628-E-003-015-MY3, and 98-2631-S-003-002.

9. REFERENCES