COMPARISON OF A BIGRAM PLSA AND A NOVEL CONTEXT-BASED PLSA LANGUAGE MODEL FOR SPEECH RECOGNITION

Md. Akmal Haidar and Douglas O'Shaughnessy

INRS-EMT, 6900-800 de la Gauchetiere Ouest, Montreal (Quebec), H5A 1K6, Canada

haidar@emt.inrs.ca, dougo@emt.inrs.ca

ABSTRACT

We propose a novel context-based probabilistic latent semantic analysis (PLSA) language model for speech recognition. In this model, the topic is conditioned on the immediate history context and the document in the original PLSA model. This allows computing all the possible bigram probabilities of the seen history context using the model. It properly computes the topic probability of an unseen document for each history context present in the document. We compare our approach with a recently proposed unsmoothed bigram PLSA model where only the seen bigram probabilities are calculated, which causes computing the incorrect topic probability for the present history context of the unseen document. The proposed model requires a significantly less amount of computation time and memory space requirements than the unsmoothed bigram PLSA model. We carried out experiments on a continuous speech recognition (CSR) task using the Wall Street Journal (WSJ) corpus. The proposed approach shows significant reduction in both perplexity and word error rate (WER) measurements over the other approach.

Index Terms— Topic models, bigram PLSA models, speech recognition, word co-occurrence, statistical language model

1. INTRODUCTION

Statistical n-gram LMs suffer from shortages of long-range information, which limit performance. To capture the long-range information, one of the earliest attempts was a cache-based LM that took advantage that a word observed earlier in a document could occur again. This helps to increase the probability of the seen words when predicting the next word [1]. A similar idea was used in trigger-based LM adaptation, which uses a maximum entropy approach [2] to raise the probability of unseen but topically related words. In addition recently, latent topic analysis has been used broadly to compensate for the weaknesses of n-gram models. Several techniques such as Latent Semantic Analysis (LSA) [3], PLSA [4], and Latent Dirichlet Allocation (LDA) [5] have been studied to extract the latent semantic information from a training corpus. These methods have been used successfully for speech recognition [3, 4, 6, 7, 8, 9, 10, 11]. A bigram LDA topic model, where the word probabilities are conditioned on their preceding context and the topic probabilities are conditioned on the documents, has been recently investigated [12]. A similar model but in the PLSA framework called bigram PLSA model was introduced recently [13]. An updated bigram PLSA model (UBPLSA) was proposed in [14] where the topic is further conditioned on the bigram history context.

In the UBPLSA model [14], the bigram probabilities for each topic are modeled and the topic is conditioned on the bigram history and the document. For each topic, it requires M distributions, where M is the size of vocabulary. So, it needs high computation time and huge memory space. However, this approach is not practical as it assigns zero probability to the unseen bigrams. Furthermore, in testing, the model computes the topic probabilities for the bigram histories that are present in the test document. However, it cannot compute the topic probabilities for some bigram history contexts that are present in both the training and test set as the bigram probabilities for the corresponding bigram histories are zero because the model assigns zero probability to the unseen bigrams. Therefore, the model cannot compute some bigram probabilities of the test document that should be computed by the training model. However, those bigram probabilities of the test document are computed later by the smoothing process.

In this paper, we propose a context based PLSA (CPLSA) model where the topic is further conditioned on the history context in the original PLSA model. It allows computing all the possible bigram probabilities for the seen history context in the training set. Therefore, the topic probabilities for the history contexts of the test document can be computed properly. We have seen that the proposed approach gives significantly better results over the UBPLSA model [14]. In addition, it reduces the complexity and memory requirements as it uses unigram probabilities for topics.

The rest of this paper is organized as follows. Section 2 is used for reviewing the PLSA and the UBPLSA models. The proposed CPLSA model is described in section 3. The UBPLSA and the proposed CPLSA models are compared in section 4. In section 5, the time complexity and memory requirements of the UBPLSA and CPLSA models are analysed.

The experimental details are explained in section 6. Finally the conclusions and future work are described in section 7.

2. REVIEW OF PLSA AND UBPLSA MODELS

2.1. PLSA MODEL

The PLSA model [4] can be described in the following procedure. First a document d_j $(j = 1, 2, \ldots, N)$ is selected with probability $p(d_j)$. A topic z_k $(k = 1, 2, \ldots, K)$ is then chosen with probability $p(z_k|d_j)$, and finally a word w_i $(i = 1, 2, \ldots, M)$ is generated with probability $p(w_i|z_k)$. The probability of word w_i given a document d_j can be estimated as:

$$p(w_i|d_j) = \sum_{k=1}^{K} p(w_i|z_k)p(z_k|d_j). \quad (1)$$

The model parameters $p(w_i|z_k)$ and $p(z_k|d_j)$ are computed by using the expectation maximization (EM) algorithm [4].

2.2. UBPLSA MODEL

Before describing the UBPLSA model, the previous bigram PLSA model is briefly explained. A combination of PLSA and bigram models was introduced in [13]. Instead of $P(w_i|z_k)$ in Equation 1, the bigram PLSA model uses $p(w_j|w_i, z_k)$ in computing the probability of word w_j given the bigram history w_i and the document d_l:

$$p(w_j|w_i, d_l) = \sum_{k=1}^{K} p(w_j|w_i, z_k)p(z_k|d_l). \quad (2)$$

The model parameters were computed using the EM procedure [13].

The UBPLSA model was recently proposed in [14] which outperform the previous bigram PLSA model [13]. Here, the topic probability is further conditioned on the bigram history. It can model the topic probability for the document given a context, using the word co-occurrences in the document. In this model, the probability of the word w_j given the document d_l and the word history w_i is computed as:

$$p(w_j|w_i, d_l) = \sum_{k=1}^{K} p(w_j|w_i, z_k)p(z_k|w_i, d_l). \quad (3)$$

The EM procedure for training the model takes the following two steps: E-step:

$$p(z_k|w_i, w_j, d_l) = \frac{p(w_j|w_i, z_k)p(z_k|w_i, d_l)}{\sum_{k'} p(w_j|w_i, z_{k'})p(z_{k'}|w_i, d_l)}, \quad (4)$$

M-step:

$$p(w_j|w_i, z_k) = \frac{\sum_l n(w_i, w_j, d_l)p(z_k|w_i, w_j, d_l)}{\sum_{j'} \sum_l n(w_i, w_{j'}, d_l)p(z_k|w_i, w_{j'}, d_l)}, \quad (5)$$

$$p(z_k|w_i, d_l) = \frac{\sum_{j'} n(w_i, w_{j'}, d_l)p(z_k|w_i, w_{j'}, d_l)}{\sum_{k'} \sum_{j'} n(w_i, w_{j'}, d_l)p(z_{k'}|w_i, w_{j'}, d_l)}. \quad (6)$$

where $n(w_i, w_j, d_l)$ is the number of times the word pair $w_i w_j$ occurs in the training document d_l.

3. PROPOSED CPLSA MODEL

The CPLSA model is similar to the original PLSA model except the topic is further conditioned on the history context as like the UBPLSA model. To better understand the model, the matrix decomposition of the CPLSA model is described in Figure 1. Using this model, we can compute the bigram

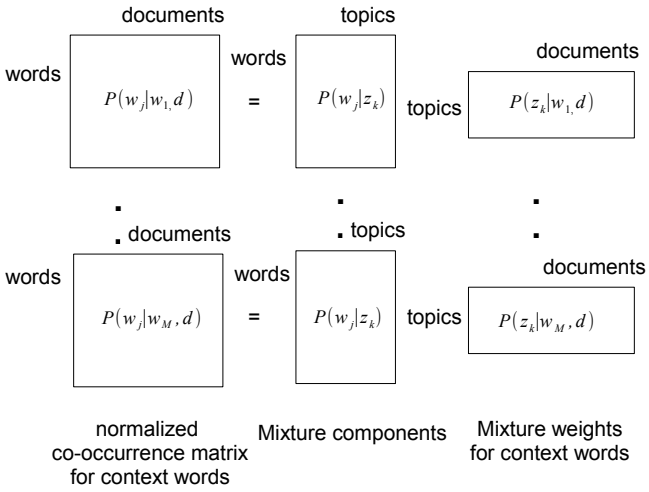

Fig. 1. Matrix decomposition of the CPLSA model

probability using the unigram probabilities of topics as:

$$p(w_j|w_i, d_l) = \sum_{k=1}^{K} p(w_j|z_k)p(z_k|w_i, d_l). \quad (7)$$

The parameters of the model are computed as: E-step:

$$p(z_k|w_i, w_j, d_l) = \frac{p(w_j|z_k)p(z_k|w_i, d_l)}{\sum_{k'} p(w_j|z_{k'})p(z_{k'}|w_i, d_l)}, \quad (8)$$

M-step:

$$p(w_j|z_k) = \frac{\sum_i \sum_l n(w_i, w_j, d_l)p(z_k|w_i, w_j, d_l)}{\sum_{j'} \sum_{i'} \sum_l n(w_{i'}, w_{j'}, d_l)p(z_k|w_{i'}, w_{j'}, d_l)}, \quad (9)$$

$$p(z_k|w_i, d_l) = \frac{\sum_{j'} n(w_i, w_{j'}, d_l)p(z_k|w_i, w_{j'}, d_l)}{\sum_{k'} \sum_{j'} n(w_i, w_{j'}, d_l)p(z_{k'}|w_i, w_{j'}, d_l)}. \quad (10)$$

From Equations 8 and 10, we can see that the model can compute all the possible bigram probabilities of the seen history context in the training set. Therefore, the model can overcome the problem of computing topic probability of the test

document using the UBPLSA model, which causes the problem in the computation of the bigram probabilities of the test document.

4. COMPARISON OF UBPLSA & CPLSA MODELS

In our proposed CPLSA model, the topic is conditioned to the bigram history context and the document as like the UBPLSA model [14]. The UBPLSA model upgrades the previous bigram PLSA model [13] by conditioning the history context to the topic probability in addition to the document. In the UBPLSA model, the bigram probabilities of the topics are unsmoothed in the training procedure. So, using the UBPLSA model, the topic weights of the unseen test document cannot be computed properly as for some history contexts; the topic probabilities are assigned zeros as the bigram probabilities in the training model are not smoothed. Therefore, some of the bigram probabilities of the test document cannot be computed by using the training model. However, they are later smoothed in the test phase. That approach is not practical as for the corresponding history context some other bigrams may be present in the training set. Our proposed CPLSA model can solve the problem of finding the topic probabilities for the test set as the model can assign probabilities to all possible bigrams of the seen history context in the training set. Therefore, the possible bigram probabilities of the test set can be computed by using the CPLSA model. Moreover, the model needs the unigram probabilities for topics that can reduce a vast amount of memory requirements and the complexity over the UBPLSA model. As the UBPLSA model, the proposed CPLSA model can also be extended to the n-gram case with increasing complexity and memory space requirements.

5. COMPLEXITY ANALYSIS OF THE CPLSA AND UBPLSA MODELS

The number of free parameters in our proposed CPLSA model are $(M-1)K+(K-1)MN$, where M, K, and N are the number of words, the number of topics and the number of documents, respectively. In contrast, the number of free parameters for the UBPLSA model are $M(M-1)K+(K-1)MN$, which is greater than our proposed CPLSA model. So, the proposed CPLSA model requires smaller memory space than the UBPLSA model [14].

For the E-step of the EM algorithm, the time complexity of the proposed CPLSA model and the UBPLSA model [14] are $O(MNK)$ and $O(M^2NK)$ respectively. The time complexity for the M-step are $O(MNK)$ and $O(KNB)$ for the proposed CPLSA and the UBPLSA models respectively. Here, B is the average number of word pairs in the training documents [14]. The size of B is obviously greater than the size of M. Therefore, our proposed CPLSA model also needs less training time than the UBPLSA model [14].

6. EXPERIMENTS

6.1. Data and experimental setup

We randomly selected 500 documents from the '87-89 WSJ corpus [15] for training the UBPLSA and the CPLSA models. The total number of words in the documents is 224,995. We used the 5K non-verbalized punctuation closed vocabulary from which we removed the MIT stop word list [16] and the infrequent words that occur only once in the training documents. After these removals, the total number of vocabulary is 2628. We could not consider more training documents due to higher computational cost and huge memory requirements for the UBPLSA model [14]. For valid comparison, we used the same number of documents for the PLSA and CPLSA models. To capture the lexical regularity, the topic models are interpolated with a back-off trigram background model. The trigram background model is trained on the '87-89 WSJ corpus using the back-off version of the Witten-Bell smoothing; 5K non-verbalized punctuation closed vocabulary and the cutoffs 1 and 3 on the bi-gram and tri-gram counts respectively are incorporated. The interpolation weights are computed by optimizing on the held-out data. We used the acoustic model from [17] in our experiments. The acoustic model is trained by using all WSJ and TIMIT [18] training data, the 40 phones set of the CMU dictionary [19], approximately 10000 tied-states, 32 gaussians per state and 64 gaussians per silence state. The acoustic waveforms are parameterized into a 39-dimensional feature vector consisting of 12 cepstral coefficients plus the 0^{th} cepstral, delta and delta delta coefficients, normalized using cepstral mean subtraction ($MFCC_{0-D-A-Z}$). We evaluated the results on the evaluation test, which is the Nov'93 Hub 2 5K test data from the ARPA November 1993 WSJ evaluation (215 sentences from 10 speakers) [15, 20].

6.2. Experimental Results

We used the folding-in procedure [4] to compute $p(w_j|d)$ (Equation 1) and $p(w_j|w_i,d)$ (Equations 3 and 7) for the test document d. We keep the unigram (Equations 1 and 7) and bigram (Equation 3) probabilities for topics unchanged, and used them to compute the test document's topic probabilities $p(z_k|d)$ for PLSA and $p(z_k|w_i,d)$ for the UBPLSA and CPLSA models for those w_i present in d. In the UBPLSA model [14], the topic probability $p(z_k|w_i,d)$ for some histories w_i are assigned zeros (Equations 4 and 6), as the training model gives zero probabilities to the unseen bigrams in the training model. Therefore, some bigrams of the test document with history context w_i are assigned zero probabilities. The problem is solved by the proposed CPLSA model, which is able to assign probabilities to all the bigrams of the seen history context in the training set. The remaining zero probabilities of the obtained matrix $p(w_j|w_i,d)$ are computed by using the Witten-Bell back-off smoothing. The model is then inter-

polated with a back-off trigram background model to capture the short-range information.

We tested the proposed approach for various sizes of topics. We performed the experiments five times and the results are averaged. The perplexity results are described in Table 1. From Table 1, we can see that the perplexities are decreased

Table 1. Perplexity results of the topic models

Language Model	20 Topics	40 Topics
Background	83.39	83.39
PLSA	613.64	605.73
UBPLSA	275.42	265.82
CPLSA	258.63	196.90
Background+PLSA	71.09	71.05
Background+UBPLSA	68.56	68.29
Background+CPLSA	66.03	63.79

with increasing topics. The proposed CPLSA model outperforms both the PLSA and the UBPLSA models. Both the UBPLSA and CPLSA models outperform the PLSA model significantly.

We performed the paired t-test on the perplexity results of the UBPLSA and CPLSA models and their interpolated form with the significance level of 0.01. The p-values for different topic sizes are described in Table 2. From Table 2,

Table 2. p-values obtained from the paired t test on the perplexity results

Language Model	20 Topics	40 Topics
UBPLSA and CPLSA	$2.35E$-08	$5.21E$-14
Background+UBPLSA and Background+CPLSA	$1.39E$-12	$1.99E$-13

we can note that all p-values are less than the significance level 0.01. Therefore, the perplexity improvements of CPLSA model over UBPLSA model are statistically significant.

We evaluated the WER experiments using lattice rescoring. In the first pass, we used the back-off trigram background language model for lattice generation. In the second pass, we applied the interpolated model of the LM adaptation approaches for lattice rescoring. The experimental results are explained in Figure 2. From Figure 2, we can note that the proposed CPLSA model yields significant WER reductions of about 8.63% (8.11% to 7.41%), 5.48% (7.84% to 7.41%), and 2.75% (7.62% to 7.41%) for 20 topics and about 9.49% (8.11% to 7.34%), 6.37% (7.84% to 7.34%), and 4.17% (7.66% to 7.34%) for 40 topics, over the background model, PLSA model [4], and the UBPLSA [14] approaches respectively.

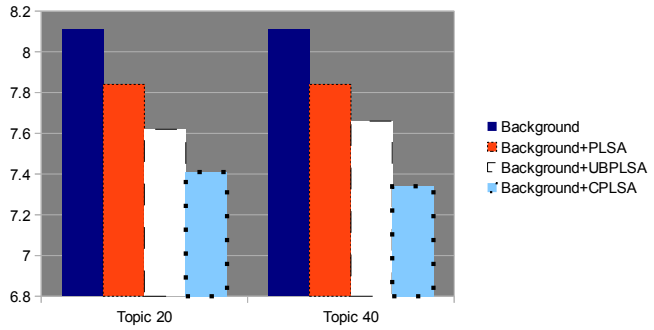

Fig. 2. WER results for different topic sizes

We also performed a paired t test on the WER results for the interpolated form of the UBPLSA and CPLSA models with a significance level 0.01. The p-values of the test are explained in Table 3. From Table 3, we can see that the p-

Table 3. p-values obtained from the paired t test on the WER results

Language Model	20 Topics	40 Topics
Background+UBPLSA and Background+CPLSA	$3.6E$-04	$1.41E$-05

values are smaller than the significance level 0.01. Therefore, the WER improvements are statistically significant.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new context-based PLSA model where the topic is further conditioned on the history context to the original PLSA model. Since the recently proposed updated bigram PLSA model assigns probabilities to the seen bigrams only, the model yields zero topic probabilities to the history context of the test document that are seen in the training set. This causes that some of the bigram probabilities of the test document cannot be computed using the training model, which is not practical. Our proposed model gives a way to find all the possible bigram probabilities of the seen history context in the training set, which helps to find the topic weights of the unseen test documents correctly and thereby gives the correct bigram probabilities to the test document. Moreover, the proposed approach saves complexity and memory space requirements over the other approach as the proposed approach uses unigram probabilities instead of bigram probabilities for topics.

For future work, we will apply the proposed approach in the LDA framework.

8. REFERENCES

[1] R. Kuhn and R. D. Mori, "A Cache-Based Natural Language Model for Speech Recognition", IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 12(6), pp. 570-583, 1990.

[2] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling", Computer, Speech and Language, vol. 10(3), pp. 187-228, 1996.

[3] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling", IEEE Trans. on Speech and Audio Proc, vol. 88, No. 8, pp. 1279-1296, 2000.

[4] D. Gildea and T. Hofmann, "Topic-Based Language Models Using EM", Proc. of EUROSPEECH, pp. 2167-2170, 1999.

[5] D. M. Blei, A. Y.Ng, and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[6] D. Mrva and P. C. Woodland, "A PLSA-based Language Model for conversational telephone speech", in Proc. of ICSLP, pp. 2257-2260, 2004.

[7] Y.-C. Tam and T. Schultz, "Dynamic Language Model Adaptation Using Variational Bayes Inference", Proc. of INTERSPEECH, pp. 5-8, 2005.

[8] Y.-C. Tam and T. Schultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals", Proc. of INTERSPEECH, pp. 2206-2209, 2006.

[9] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using N-gram weighting", in Proc. of CCECE, pp. 857-860, 2011.

[10] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation and Dynamic Marginals", in Proc. of EUSIPCO, pp. 1480-1484, 2011.

[11] M. A. Haidar and D. O'Shaughnessy, "Topic N-gram Count Language Model for Speech Recognition", Proc. of IEEE SLT workshop, pp. 165-169, 2012.

[12] H. M. Wallach, "Topic Modeling: Beyond bag-of-words", Proc. of the 23rd International Conference of Machine Learning (ICML'06), pp. 977-984, 2006.

[13] J. Nie, R. Li, D. Luo, and X. Wu, "Refine bigram PLSA model by assigning latent topics unevenly", Proc. of the IEEE workshop on ASRU, pp. 141-146, 2007.

[14] M. Bahrani and H. Sameti, "A New Bigram PLSA Language Model for Speech Recognition", Research Article, Euraship Journal on Signal Processing, pp. 1-8, 2010.

[15] "CSR-II (WSJ1) Complete", Linguistic Data Consortium, Philadelphia, 1994.

[16] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[17] K. Vertanen, "HTK Wall Street Journal Training Recipe",http://www.inference.phy.cam.ac.uk/kv227/htk/

[18] John S. Garofolo, et al,"TIMIT Acoustic-Phonetic Continuous Speech Corpus" Linguistic Data Consortium, Philadelphia, 1993.

[19] "The Carnegie Mellon University (CMU) Pronounciation Dictionary",http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[20] P.C. Woodland, J.J. Odell, V. Valtchev and S.J. Young, "Large Vocabulary Continuous Speech Recognition Using HTK", Proc. of ICASSP, pp. II:125-128, 1994.