POSITION INFORMATION FOR LANGUAGE MODELING IN SPEECH RECOGNITION

Hsuan-Sheng Chiu¹, Guan-Yu Chen¹, Chun-Jen Lee², Berlin Chen¹

¹ National Taiwan Normal University, Taipei

² Telecommunication Labs., Chunghwa Telecom Co., Ltd., Taoyuan

{g93470240, g96470203, berlin}@csie.ntnu.edu.tw, cjlee@cht.com.tw

ABSTRACT

This paper considers word position information for language modeling. For organized documents, such as technical papers or news reports, the composition and the word usage of articles of the same style are usually similar. Therefore, the documents can be separated into partitions consisting of identical rhetoric or topic styles by the literary structures, e.g., introductory remarks, related studies or events, elucidations of methodology or affairs, conclusions of the articles, and references, or footnotes of reporters. In this paper, we explore word position information and then propose two position-dependent language models for speech recognition. The structures and characteristics of these position-dependent language models were extensively investigated, while its performance was analyzed and verified by comparing it with the existing n -gram, mixture- and topic-based language models. The large vocabulary continuous speech recognition (LVCSR) experiments were conducted on the broadcast news transcription task. The preliminary results seem to indicate that the proposed position-dependent models are comparable to the mixture- and topic-based models.

Index Terms— Speech recognition, language model, position information, topic information, language model adaptation

1. INTRODUCTION

Language model (LM) plays a decisive role in many research fields of natural language processing, such as machine translation, information retrieval, speech recognition, etc. The n -gram language model that follows a statistical modeling paradigm is the most prominently used language model in speech recognition because of its simplicity and predictive power. Nevertheless, the n -gram model, which aims 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 the newly decoded word.

In the recent past, various language modeling approaches have been extensively investigated to extract information among the decoded word and its history to complement the conventional n -gram model. According to different levels of linguistic information being utilized, language models can be roughly classified into the following several categories:

1. Word-based language models: The n -gram model is usually the basic model and many other models of this type are designed to overcome the major drawback of n -gram models, i.e., to capture long-distance word dependence information without increasing the model complexity rapidly. For example, the mixed-order Markov model [1] and trigger-based language model [2] belong to this category.

2. Word class- or topic-based language models: These models are analogous to the n -gram model, but the co-occurrence relationship among words is constructed via (latent) word classes or topics instead. When the relationship is established, the probability of a decoded word given the history words can be efficiently computed. For example, the class-based n -gram model [3], aggregated Markov model [1] and our previously proposed word topical mixture model [4] are of this category.

3. Sentence structure-based language models: Because the constraints of grammars, rules for a sentence may be derived and represented by a parse tree. Then, we can select among candidate words on the basis of the sentence patterns or head words of the history. For example, the structured language model [5] falls into this category.

4. Document topic-based language models: Words are aggregated in a document to represent some topics (or concepts). During speech recognition, the history is considered as an incomplete document and the associated latent topic distributions can be discovered on the fly. The decoded words related to most of the topics that the history probably belongs to can be therefore selected. The mixture-based language model [6], latent semantic analysis (LSA) [7], probabilistic latent semantic analysis (PLSA) [8] and latent Dirichlet allocation (LDA) [9] are good examples of this category.

As it is known, there still exists much more information beyond n -grams that can be used for language modeling in speech recognition. Latent topic information has been widely explored by the document topic-based language models in recent years. Despite that, word position information was seldom mentioned before. We hence propose to explore word position information for speech recognition, and two kinds of position-dependent language models are designed to incorporate position and/or topic information. The structures and characteristics of the position-dependent language models were extensively investigated, while its performance was analyzed and verified by comparing it with the existing n -gram, mixture- and topic-based language models on the broadcast news transcription task [10].

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work with the mixture-based language model and the PLSA model, which are close in spirit to the models proposed in this paper. Section 3 presents the proposed position-dependent language models, and also tries to exemplify the position information inherent in the documents. Then, the experiment settings and a series of speech recognition results are presented in Sections 4 and 5, respectively. Finally, conclusions are drawn in Section 6.

Table 1. *Style words selected on four partitions of the broadcast news corpus.*

P1	P2	P3	P4
1 繼續 Continue	4 醫師 Doctor	7 學生 Student	10 公視 TV station name
2 現場 Locale	5 網路 Internet	8 老師 Teacher	11 綜合報導 Roundup
3 歡迎 Welcome	6 珊瑚 Coral	9 米酒 Rice wine	12 編譯 Edit and translate

2. RELATED WORK

2.1. Mixture-based language model

To capture the local word regularity and the global document style simultaneously, we can use the mixture-based language model [6]. For this model, the training corpus was separated into several groups according to the topics of documents. Moreover, a simple k -means clustering technique can be used to cluster the documents into different groups (topics) if the categorical labels of the training set are not clearly indicated. An n -gram language model is trained for each group and then a mixture-based language model is constructed with these n -gram language models through a simple linear combination, i.e.

$$P_{MIX}(w_i | w_{i-2}, w_{i-1}) = \sum_{j=1}^K \lambda_j P(w_i | w_{i-2}, w_{i-1}, M_j) \quad (1)$$

where K is the number of groups (topics), λ_j is the weight for each model M_j . During speech recognition, the set of weights $\lambda_j, j=1,2,\dots,K$ can be optimized using the history word sequences or the top scoring one, as well as the EM (Expectation-Maximization) algorithm. To avoid the data sparsity problem, the n -gram language model trained with a non-partitioned corpus can also be included as an additional mixture. Moreover, smoothing techniques like Good-Turing smoothing can be performed for each language model.

2.2. Probabilistic latent semantic analysis

PLSA is a general machine learning technique for modeling the co-occurrences of words and documents, and it evaluates the relevance between them through a low-dimensional factor space [8]. 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 $M_{H_{w_i}}$ used for predicting the occurrence probability of w_i :

$$P_{PLSA}(w_i | M_{H_{w_i}}) = \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{H_{w_i}}) \quad (2)$$

where T_k is one of the latent topics and $P(w_i | T_k)$ is the probability of the word w_i occurring in T_k . The latent topic distributions $P(w_i | T_k)$ 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 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 $P(w_i | T_k)$ unchanged, but let the search history's probability distribution over the latent topics $P(T_k | M_{H_{w_i}})$ be gradually updated as path extension is performed, by using the EM update formulas. Then, the probabilities of the PLSA and background n -gram (e.g., trigram) language models can be combined using a simple linear interpolation:

Figure 1. *The distributions of selected style words on four partitions of the broadcast news corpus.*

$$P_{Adapt}(w_i | w_{i-2} w_{i-1}) = \alpha \cdot P_{PLSA}(w_i | H_{w_i}) + (1 - \alpha) \cdot P_{n-gram}(w_i | w_{i-2} w_{i-1}) \quad (3)$$

where α is a tunable interpolation weight.

3. POSITION INFORMATION FOR LANGUAGE MODELING

3.1. Position information among documents

In order to verify our belief of the usefulness of word position information, we try to analyze the word usage of a broadcast news corpus partitioned by the structure of documents. A left-to-right HMM segmenter was trained and applied to separate every broadcast news document of the corpus into partitions. By using the HMM segmenter, each partition within a document is expected to be semantically cohesive. Table 1 and Figure 1 show the style words with higher rank of multiplication of term frequency (TF) and inverse document frequency (IDF) scores and their probabilities on four partitions of the corpus, respectively. The detailed description of the corpus will be given in Section 4. According to Table 1 and Figure 1, we can observe that the word usage with respect to different partitions (or positions) is apparently quite different. For the first partition (P1), most of the words are about introductory remarks, e.g., greetings, notices or conjunctives. For the second and third partitions (P2 and P3), most of the words are content words because the themes and events of documents are mostly described in the middle parts of a news report. For the final partition (P4), most of the words are actions, reporter names or TV station names. They signal an ending of the news report. Reporters usually make a short conclusion and a footnote about where the event took place. Moreover, the word probabilities are more specifically distributed for the final partition (P4), and some high-frequency words at this partition will have zero probabilities at the other partitions. Consequently, we can conclude that words in the marginal positions of documents are more specific while words in the middle positions are more comprehensive for the broadcast news documents.

3.2. Positional n -gram model

In this paper, we propose a positional n -gram model in attempt to explore the positional information inherent in the broadcast news documents. For this model, the training corpus might be partitioned according to the structure of documents, such as the introductory remarks, related studies or events, elucidations of methodology or affairs, conclusions of articles, and references or footnotes of reporters. In this preliminary study, we split each document of the corpus into

Figure 2. A schematic comparison of topic and position-based model.

partitions of an equal size, or into partitions of different sizes by using the HMM segmenter. The n -gram language model is then trained respectively for each partition, and finally a positional n -gram model is constructed as a composite n -gram language model:

$$P_{POS}(w_i | w_{i-2}, w_{i-1}) = \sum_{s=1}^S \beta_s P(w_i | w_{i-2}, w_{i-1}, L_s) \quad (4)$$

where S is the number of partitions, β_s is the weight for a specific position L_s . By comparison between the complete document length and the history length, we can easily use the word position information without any obstacle. Accordingly, the positional n -gram model can be reduced to an n -gram model of a certain position (or partition), denoted as deterministic model in Section 5. Whatever, a mixture-based model may be also used to alleviate the data sparsity problem, or each n -gram model can be smoothed with Good-Turing smoothing. The weights β_s are optimized for the search hypotheses during recognition as well.

3.3. Positional PLSA

We also try to integrate word position information into the PLSA model as a complement of the topic (or concept) information that has already been modeled by PLSA, and the resulting model is referred to as the positional PLSA model. The positional PLSA model can be expressed in a similar way as the original PLSA model:

$$P_{PosPLSA} \left(w_i | M_{Hw_i} \right) = \sum_{s=1}^S P(L_s | M_{Hw_i}) \sum_{k=1}^K P(T_k | M_{Hw_i}) P(w_i | T_k, L_s) \quad (5)$$

where T_k is a specific topic, L_s is a specific position, $P(w_i | T_k, L_s)$ is the probability of the word w_i occurring in T_k and L_s simultaneously, and $P(T_k | M_{Hw_i})$ and $P(L_s | M_{Hw_i})$ are the probability distributions of the latent topic T_k and position L_s given the search history Hw_i . If the position of a decoded word is observable, positional PLSA can be easily reduced to original PLSA with respect to a certain position, which means $P(L_s | M_{Hw_i})$ will be 1 for a certain position L_s . For instance, the ratio of history length over length of the top scoring word sequence obtained from the word lattice can be used to determine the current position of a decoded word in the lattice. Alternatively, $P(L_s | M_{Hw_i})$ can be directly estimated during speech recognition (like the procedure for estimating $P(T_k | M_{Hw_i})$ of the PLSA model), when the search history evolves over time.

3.4. Comparisons with other models

We compare the positional n -gram model with the mixture-based language models. For a given set of training documents, each n -gram language model of the mixture-based language model is trained with a topical subset of clustered documents, while each n -gram model of the positional n -gram model is trained with the document fragments collected from the same location (or position). That is, the mixture-based language model requires additional clustering being performed, while the positional n -gram model assumes that the documents in

the collection share the similar structure. The model complexity of both models are equal to $V^n \times U$, where V is vocabulary size, n denotes the length of the window of words considered by the n -gram model, and U is either the topic number or the position number. Figure 2 shows the major difference between topic- and position-based models, while they are conceptually orthogonal.

Then, we compare the positional PLSA model with the PLSA model. The model complexities for positional PLSA and PLSA are $V \times T \times S + (T + S) \times H$ and $V \times T + T \times H$, respectively, where V is the size of the vocabulary, T is the number of the topics, S is the number of partitions, and H is the number of various search histories. Generally, positional PLSA would be more complicated than PLSA.

4. EXPERIMENT SETUP

The speech corpus consists of about 200 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News) [4]. A subset of 25-hour speech collected during November 2001 to December 2002 was used to bootstrap the acoustic training. Another subset of 3-hour speech data collected within 2003 was reserved for development (1.5 hours) and evaluation (1.5 hours). The acoustic models were trained with the minimum phone error rate (MPE) criterion.

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 adaptation text corpus used for training the position-dependent language models, mixture-based language model and PLSA model was collected from MATBN 2001 and 2002, which consists of one million Chinese characters of the orthographic broadcast news transcripts.

The language model adaptation experiments were performed in the word lattice rescoring procedure, where the associated word lattices of 1.5-hour test speech data were built beforehand by a tree search procedure and using the background bigram language model [4]. The speech recognition system with background trigram language model resulted in a character error rate (CER) of 20.32% and a perplexity (PP) of 682.10.

5. EXPERIMENT RESULTS

In the first set of experiments, we evaluate the performance of our proposed positional n -gram model, for which the value of n is set to three. The corresponding CER and PP results are shown in the first and middle parts of Table 2. The application of the positional n -gram model for language model adaptation can be discussed from three aspects: (1) whether the language model training corpus is segmented uniformly or segmented by the HMM segmenter, (2) whether the word position is determined deterministically or non-deterministically, and (3) the number of partitions used. The word position information for test data could be known beforehand by segmenting top-1 hypothesis of speech utterances using trained HMM segmenter.

First, as we look into the results of both the deterministic and nondeterministic positional n -gram models, it can be found that for both models, using the HMM-segmenter for language model training can yield better results than using

Table 2. The CER and PP results of positional n -gram models and mixture-based models.

	CER(%)	PP
Background Trigram	20.32	682.10
+ Deterministic Positional n -gram	CER(%) Uniform/HMM Segmentation	PP Uniform/HMM Segmentation
1 partition	19.23	434.46
2 partitions	19.09/19.44	402.31/387.08
4 partitions	19.29/19.54	408.02/382.78
8 partitions	19.62/19.37	416.41/378.20
16 partitions	19.85/19.36	453.59/387.48
+ Nondeterministic Positional n -gram	CER(%) Uniform/HMM Segmentation	PP Uniform/HMM Segmentation
2 partitions	19.08/19.12	392.48/389.33
4 partitions	19.08/18.94	399.67/392.93
8 partitions	19.19/19.05	408.54/401.47
16 partitions	19.35/18.97	423.13/405.99
+Mixture-Based LM	CER(%)	PP
2 topics	19.12	388.00
4 topics	19.17	384.26
8 topics	18.95	377.64
16 topics	18.80	372.26

Table 3. The CER and PP results of PLSA models.

PLSA	CER(%)	PP
8 topics	19.76	563.70
16 topics	19.77	554.07
32 topics	19.60	545.14
64 topics	19.71	539.61
128 topics	19.55	533.29

Table 4. The CER and PP results of positional PLSA models (with uniform segmentation).

CER(%)	Topics		
	8	16	32
2 partitions	19.76	19.57	19.63
3 partitions	19.73	19.68	19.68
4 partitions	19.69	19.75	19.68
PP	Topics		
	8	16	32
2 partitions	555.97	546.27	538.73
3 partitions	547.90	544.28	537.77
4 partitions	552.22	554.66	557.70

uniform segmentation at most cases. One possible reason might be that the broadcast news documents intrinsically do not have paragraphs with equal length, so uniform segmentation fails to match the corpus characteristics well. Second, the CER performance of the nondeterministic positional n -gram model outperforms that of the deterministic one. Third, increasing the number of partitions will not always lead to the CER improvement. One possible explanation is that the document structure of the evaluation set is not complicated enough to be split into such many partitions. More precisely, the language models of all partitions behave quite the same, except that of the last partition. Although the use of a specific language model for the last partition might provide an additional benefit, its short duration (compared to the other positions) will make its contribution to the overall CER improvement insignificant.

We have experimentally observed that, when documents were being segmented into four partitions by the HMM segmenter, the resulting ratios for the durations of the four partitions (P1 to P4) in a document were 31%, 35%, 28% and 6% on average. On the other hand, as can be seen, the positional n -gram model performs comparably to the mixture-based language model (in the bottom part of Table 2), when the number of partitions is kept small.

In the next set of experiments, we compare the original PLSA language model with the positional PLSA language model (with uniform segmentation) under different numbers of topics and partitions. The results obtained from PLSA and positional PLSA are shown in Tables 3 and 4, respectively. The performance of CER and PP for positional PLSA tends to be better as the topic mixture number increases, while contrary results are obtained as the partition number increases (e.g., when the partition number is set equal to 4). The performance of positional PLSA seems not to be significantly different from that of PLSA.

Finally, although the performance gains are not very significant for our proposed positional n -gram model and positional PLSA model, we believe that the use of position information still has its potential, according to the observations exemplified in Section 3.1. In the meantime, we are also being devoted to the investigation of better ways to model the document styles and structures more precisely for speech recognition. Furthermore, position information might be taken as an additional feature in the n -best (or word-lattice) rescoring task that uses discriminatively trained language models.

6. CONCLUSIONS

In this paper, we have incorporated the position information into language modeling. We proposed two position-dependent language models, i.e., the positional n -gram model and the positional PLSA model. The corresponding structures and characteristics were extensively investigated, while the performance was also analyzed. The preliminary results seem to demonstrate that the proposed position-dependent models are comparable to the mixture- and topic-based models.

7. REFERENCES

- [1] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," *EMNLP 1997*.
- [2] P. Brown et al., "Class-based n -gram models of natural language," *Computational Linguistics*, 18(4), 1999.
- [3] C. Troncoso et al. "Trigger-based language model construction by combining different corpora," *IEICE Technical Report*, SP2004-100, 2004.
- [4] H. S. Chiu and B. Chen, "Word topical mixture models for dynamic language model adaptation," *ICASSP 2007*.
- [5] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, 14(4), 2000.
- [6] P. R. Clarkson and A. J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," *ICASSP 1997*.
- [7] J.R. Bellegarda, "A multispan language modeling framework for large vocabulary speech recognition," *IEEE Trans. on Speech and Audio Processing*, 6(5), 1997.
- [8] D. Gildea and T. Hofmann, "Topic-based language models using EM," *Eurospeech 1999*.
- [9] Y. C. Tam and T. Schultz, "Dynamic language model adaptation using variational bayes inference," *Interspeech 2005*.
- [10] B. Chen et al., "Lightly supervised and data-driven approaches to mandarin broadcast news transcription," *ICASSP 2004*.