Incorporating Proximity Information for Relevance Language Modeling in Speech Recognition

Yi-Wen Chen[*], *Bo-Han Hao*[*], *Kuan-Yu Chen*[†] *and Berlin Chen*[*]

[*]National Taiwan Normal University, Taipei, Taiwan
[†]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[*]{699470462, 60047082s, berlin}@ntnu.edu.tw, [†]kychen@iis.sinica.edu.tw

Abstract

Language modeling (LM), aiming to provide a statistical mechanism to associate quantitative scores to sequences of words, has long been an interesting yet challenging problem in the field of speech and language processing. Although the *n*-gram model remains the predominant one, a number of disparate LM methods have been developed to complement the *n*-gram model. Among them, relevance modeling (RM) that explores the relevance information inherent between the search history and an upcoming word has shown preliminary promise for dynamic language model adaptation. This paper continues this general line of research in two significant aspects. First, the so-called "*bag-of-word*s" assumption of RM is relaxed by incorporating word proximity evidence into the RM formulation. Second, latent topic information is additionally explored in the hope to further enhance the proximity-based RM framework. A series of experiments conducted on a large vocabulary continuous speech recognition (LVCSR) task seem to demonstrate that the various language models deduced from our framework are very comparable to existing language models.

Index Terms: speech recognition, language model, relevance model, proximity evidence, latent topic information

1. Introduction

Language modeling is a critical and integral component in any large vocabulary continuous speech recognition (LVCSR) system [1] 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 the speech recognizer. The *n*-gram model [2, 3] that follows a statistical modeling paradigm is the most prominently used in speech recognition because of the inherent simplicity and predictive power. Nevertheless, the *n*-gram language model, aiming at capturing the local contextual information or lexical regularity of a language, has frequently been criticized on two fronts. First, it is brittle across domains, since the performance is sensitive to changes in the genre or topic of the text on which it is trained. Second, it fails to capture the information (either semantic or syntactic information) conveyed in the search history beyond the immediately preceding *n*-1 words of a newly decoded word.

In view of these problems, the latent topic modeling approaches, which were originally formulated in information retrieval (IR) [4], have been introduced to dynamic language model adaptation and investigated to complement the *n*-gram models with varying degrees of success [5, 6, 7]. To exemplify, latent Dirichlet allocation (LDA) [8] and its precursor, probabilistic latent semantic analysis (PLSA) [9], are two good instantiations. A commonality among these methods is that they introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics. The dependence between an upcoming 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. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes the model parameters are fixed and unknown, whereas LDA places additional a priori constraints on the model parameters, i.e., thinking of them as random variables that follow some Dirichlet distributions. In addition, there are other approaches developed to complement the *n*-gram models, such as the trigger-based language model (TBLM) [10], for which word trigger pairs are automatically generated to capture the co-occurrence information among words. By using TBLM, the associations between the words in the search history and an upcoming word can be modeled. On the other hand, the recurrent neural network language model (RNNLM) [11] and the discriminative language model (DLM) [12] have attracted considerable attention of researchers and practitioners over the years. The former tries to estimate the probability of an upcoming word given its corresponding search history through mapping both of them into a continuous space in a recursive fashion, while the latter can utilize a rich set of lexical and/or syntactic features and a wide variety of training algorithms in an attempt to correctly discriminate the recognition hypotheses for obtaining better recognition results rather than just fitting the distribution of training data.

Apart from the above efforts, a more recent school of thought is to build language models for speech recognition on top of the notion of relevance modeling (RM) [13]. Towards this end, the RM approach explores relevance cues in order to render the co-occurrence relationship between words in the search history and an upcoming word. The key idea of RM is that each search history is assumed to be associated with an unknown relevance class, and each possible upcoming word can be regarded as a sample drawn from the relevance class [4, 14]. This paper continues this general line of research in two significant aspects. First, the so-called "*bag-of-word*s" assumption of RM is relaxed by further incorporating word proximity evidence into the RM formulation. Second, latent topic information is additionally explored to further enhance the proximity-based RM language model.

The rest of this paper is organized as follows. In Section 2, we shed light on the basic foundation of the RM modeling framework that can leverage lexical co-occurrence in a systematic way for language modeling in speech recognition, followed by an elaboration of how to incorporate word

proximity evidence into the formulation of RM in Section 3. Then, the experimental settings and a series of speech recognition experiments are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper and suggests avenues for future work.

2. Relevance Modeling for Speech Recognition

2.1. Fundamentals

In the relevance modeling (RM) approach to IR [4, 7, 14], each query is assumed to be associated with an unknown relevance class R, and the query and documents that are relevant to the information need expressed in the query are samples drawn from R. The document ranking problem then can be reduced to the problem of finding a mechanism to determine the relevance model or, more specifically, the probability $P_{\mathrm{RM}}(w)$ of observing words w in the query (and in the relevant documents) with respect to a particular information need. The relevance model $P_{\mathrm{RM}}(w)$, as a multinomial view of R, can be defined as the probability distribution which gives the probability that we would observe a word if we were to randomly select a document from the relevant class and select the word from that document. But in reality, since there is no prior knowledge about the subset of relevant documents in the collection for each query, a local feedback-like procedure [4] is performed to take advantage of the top-ranked documents obtained from an initial round of retrieval to approximate R. As such, the relevance model can be estimated in an unsupervised manner, and then works in tandem with various IR measures to distinguish relevant documents from irrelevant ones.

The notion of relevance modeling has recently attracted much attention and been applied with success to many IR-related tasks. However, to our best knowledge, there still is a dearth of work investigating the effectiveness of relevance modeling for language modeling in speech recognition [13].

2.2. RM for Speech Recognition

The task of language modeling in speech recognition can be interpreted as calculating the conditional probability $P(w|H)$, in which H is a search history, usually expressed as a sequence of words $H=h_1,h_2,\ldots,h_L$, and w is one of its possible immediately succeeding words (i.e., an upcoming word) [1, 2, 3]. When RM is applied to language modeling in speech recognition, we hypothesize that each search history H has a relevance class R_H associated with it, which can serve as a basis for predicting its immediately succeeding words w (the more relevant w to H the more likely that w is drawn alongside H from the relevance class R_H of H). The joint probability of H and w being generated by R_H, i.e., $P_{\mathrm{RM}}(H,w)$, thus can be used to derive the conditional probability $P(w|H)$ for speech recognition [13].

However, because the relevance class R_H of each search history H is not known in advance, we may leverage a local feedback-like procedure that takes H as a query and poses it to an IR system to obtain a top-ranked list of M relevant documents from the contemporaneous (or in-domain) corpus to approximate R_H, denoted by $\mathbf{D}_H=\{D_1,D_2,\ldots,D_M\}$. Then, the joint probability of observing H together with w is given by

$$P_{\mathrm{RM}}(H,w)=\sum_{m=1}^{M}P(D_m)P(h_1,h_2,\ldots h_L,w|D_m), \quad (1)$$

where $P(D_m)$ is the probability that we would randomly select D_m and $P(h_1,h_2,\ldots h_L,w|D_m)$ is the joint probability of simultaneously observing H and w in D_m. If we further assume that words are conditionally independent given D_m and their order is of no importance (i.e., the so-called "*bag-of-words*" assumption), then the joint probability can be decomposed as a product of unigram probabilities of words generated by D_m:

$$P_{\mathrm{RM}}(H,w)=\sum_{m=1}^{M}P(D_m)P(w|D_m)\prod_{l=1}^{L}P(h_l|D_m). \quad (2)$$

The probability $P(D_m)$ can be simply kept uniform or determined in accordance with the relevance of D_m to H, while $P(w|D_m)$ and $P(h_l|D_m)$ are estimated based on the word occurrence frequencies in a document and refined with the Bayesian or Jelinek-Mercer smoothing method [7]. Accordingly, the conditional probability $P(w|H)$ is expressed by

$$\begin{aligned} P_{\mathrm{RM}}(w|H)&=\frac{P_{\mathrm{RM}}(H,w)}{P_{\mathrm{RM}}(H)}\\ &=\frac{\sum_{m=1}^{M}P(D_m)P(w|D_m)\prod_{l=1}^{L}P(h_l|D_m)}{\sum_{m=1}^{M}P(D_m)\prod_{l=1}^{L}P(h_l|D_m)}. \end{aligned} \quad (3)$$

The implementation of (3) can be quite efficient, if the composition of language model probabilities is realized in the logarithmic domain [13]. On the other hand, since the baseline *n*-gram language model trained on a large general corpus can provide the generic constraint information of lexical regularities, there is a good reason to combine RM with the background *n*-gram (e.g., trigram) language model to form an adaptive language model for guiding the speech recognition process:

$$\widetilde{P}(w|H)=\lambda\cdot P_{\mathrm{RM}}(w|H)+(1-\lambda)\cdot P_{n\text{-gram}}(w|H), \quad (4)$$

where the interpolation parameter λ controls the degree of reliance on the RM model rather than on the baseline *n*-gram language model.

2.3. Implementation

Since the search histories typically are not known in advance and their number could be enormous and varying during speech recognition, we may further assume that all search histories share the same relevance class (i.e., define the relevance class at the utterance level). In order to construct an RM model for representing the relevance class of the test utterance, the top-one word sequence hypothesis, output by the baseline speech recognition with the background *n*-gram language model, is taken as a query and posed to the IR system to obtain a set of M relevant documents from the

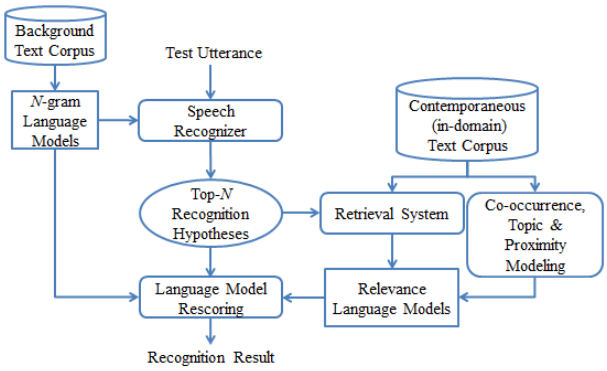

Figure 1: *A schematic illustration of the RM language modeling framework for speech recognition.*

contemporaneous (or in-domain) text collection. Empirical observations made on the development set revealed that this simplification can greatly reduce the language model lookup time and make almost negligible effects on the final performance of language model adaptation [13]. We, therefore, adopt such simplification for the following evaluations of the proposed various RM methods. Note also that we implement the IR system with the LM retrieval approach [7], where each document is respectively formulated as a unigram language model that can offer a probability distribution for generating words in the query. Such a unigram probability distribution is estimated based on the word occurrence frequencies in a document and further combined with a background unigram language model using the Jelinek-Mercer smoothing method to model the general properties of the language as well as to avoid the problem of zero probability. The documents (or document clusters) having higher probabilities of generating the query are deemed to be more relevant to the test utterance. A schematic illustration of the RM language modeling framework for speech recognition is depicted in Figure 1.

3. RM with Proximity Information

While the "*bag-of-words*" assumption can facilitate the derivation and estimation of the RM model, it seems to be an oversimplification for the problem of language modeling in speech recognition. To mitigate such deficiency, one possible remedy is to account for the constraints of word order and adjacency relationships among history words and the upcoming word in the formulation of the RM model (*cf.* Section 2). For this idea to work, the joint probability of simultaneously observing $H = h_1, h_2, \ldots, h_L$ and w in a pseudo-relevant document D_m can be alternatively decomposed as follows, by considering the pairwise word order and (immediate or intermediate) adjacency relationships:

$$\begin{aligned}\widetilde{P}_{\mathrm{RM}}(H, w \mid D_m) \\ = P(h_1 \mid D_m)\left[\prod_{l=2}^{L} P(h_l \mid h_{l-1}, D_m)\right] P(w \mid h_L, D_m),\end{aligned} \tag{5}$$

where $P(h_l \mid h_{l-1}, D_m)$ and $P(w \mid h_L, D_m)$ have similarity functionality like the conventional bigram language model to render the pairwise proximity (more specifically, word order and adjacency) relationships among history words and the upcoming word inside a pseudo-relevant document. As an instantiation, the conditional probability $P(w \mid h_L, D_m)$ is estimated through the following formulation:

$$P(w \mid h_L, D_m) = \frac{C_\tau(h_L, w, D_m)}{\sum_{w'} C_\tau(h_L, w', D_m)}. \tag{6}$$

In (6), $C_\tau(h_L, w, D_m)$ denotes the number of times h_L and w co-occur within a fixed-length sliding window being applied to a pseudo-relevant document D_m, where the sliding window starts at each occurrence of h_L and has a span of τ words. It is worth pointing out that when the value of τ is set to two, $P(w \mid h_L, D_m)$ shown in (6) is actually equivalent (degenerated) to a conventional bigram language model that is simply estimated based on D_m. As such, we can to some extent modulate the impact of the closeness of word proximity on relevance modeling. The resulting language model is referred to hereafter as the proximity-based RM model (denoted by P-RM).

4. Experimental Setup

The speech corpus consists of about 196 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News) [15]. A subset of 25-hour speech data 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. Another subset of 3-hour speech data collected within 2003 is reserved for the development set (1.5 hours) and the test set (1.5 hours).

The vocabulary size is about 72 thousand words. The trigram language model used in this paper was 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) [16]. The adaptation (contemporaneous) text corpus used for training the proposed various RM models and the other adaptation methods was collected from MATBN 2001, 2002 and 2003 (excluding the test set), which consists of one million Chinese characters (3,643 documents) of the orthographic broadcast news transcripts.

In this paper, all the language model adaptation experiments were performed in word graph rescoring. The associated word graphs of the speech data were built beforehand with a typical large vocabulary continuous recognition (LVCSR) system [17, 18]. The baseline rescoring procedure with the background trigram language model results in a character error rate (CER) of 20.22% on the test set. Notice that the constants or weighting (interpolation) coefficients of all the language models compared in this paper were all tuned at optimum values. Albeit that, it is generally agreed upon that the way to systemically determine the values of the constants or weighting (interpolation) coefficients that the various language models incorporate is still an open issue and needs further investigation and proper experimentation.

Table 1. *The speech recognition results (in CER (%)) of various language models compared in this paper.*

RM	PLSA	LDA	TBLM	RNNLM	DLM (MERT)	DLM (GCLM)	DLM (WGCLM)
19.21	19.28	19.22	20.09	19.10	19.74	19.89	19.62

5. Experimental Results

In this section, we begin by comparing the performance of RM with several well-practiced language models, including PLSA, LDA, TBLM, RNNLM and DLM; their corresponding CER results are shown in Table 1. It should be noted that in this paper, RNNLM was implemented with the toolkit released by [19]. Furthermore, DLM utilizing features composed of word unigram, bigram and trigram counts was trained with different algorithms, including minimum error rate training (denoted by "MERT" for short), global conditional log-linear model (denoted by "GCLM" for short) and weighted global conditional log-linear model (denoted by "WGCLM" for short). Interested readers may refer to [12, 20] for a thorough and updated introduction to various training algorithms for DLM. Inspection of Table 1 reveals several noteworthy points. First, TBLM provides an almost negligible improvement as compared to the baseline trigram model. Second, RNNLM achieves the best performance, which leads to a relative CER improvement of 5.5% over the baseline trigram model. Third, RM, PLSA and LDA seem to perform slightly worse than RNNLM but apparently better than the three variants of DLM. Fourth, although RM is not the best performing one among these language models, the relevance information it tries to explore is orthogonal (complementary) to those discovered by the other language models.

In the second set of experiments, we evaluate the utility of additionally incorporating the proximity information (i.e., the word order and adjacency cues) to better describe the word-word co-occurrence relationships in a pseudo-relevant document for relevance modeling; the resulting model is designated as P-RM (*cf.* Section 3). Table 2 shows the corresponding CER results as a function of different window lengths being used to capture the word order and adjacency cues. Consulting Table 2 we notice two particularities. One is that the performance of P-RM is improved when the length of the window becomes larger; the improvements, however, seem to soon reach a plateau when the length of the sliding window is set to 3 words. The other is that P-RM performs better than RNNLM and RM, and yields a relative CER improvement of 6.6% over the baseline trigram model (when τ is set to 3 for P-RM). Significance tests based on the standard NIST MAPSSWE [21] indicate the statistical significance of such a CER reduction (note here that the statistical significance was determined at the 95% confidence level).

In the third set of experiments, on the basis of a similar motivation to PLSA and LDA, we explore the use of latent topic information for the RM modeling [13]. For this idea to work, pseudo-relevant documents of a search history $H = h_1, h_2, \ldots, h_L$ are assumed to share a same set of latent topic variables $\{T_1, T_2, \ldots, T_K\}$ describing the "*word-document*" co-occurrence characteristics. Therefore, the probability that the search history H and an upcoming word w are jointly sampled from a pseudo-relevant document D_m (*cf.* (3)) is not estimated directly based on the frequency of H and w co-occurring in the document, but rather based on the frequency of H and w co-occurring in the latent topics as well as the likelihood that D_m generates the respective topics:

$$\hat{P}_{\mathrm{RM}}(H, w \mid D_m) = \sum_{k=1}^{K} \left[\prod_{l=1}^{L} P(h_l \mid T_k) \right] P(w \mid T_k) P(T_k \mid D_m). \quad (7)$$

The component probabilities involved in (7) can be estimated using the expectation-maximization (EM) inference algorithm [22]. Substituting (7) into (3) in a sense provides a mechanism to render the proximity of the search history H and the upcoming word w in the latent topic space pertaining to a pseudo-relevant document. The resulting relevance model is referred to as the topic-based relevance model (T-RM). We can also pair P-RM with T-RM (through a simple linear interpolation) so as to capitalize on two different sources of proximity information simultaneously for RM modeling. As can be seen from Table 3, the improvement brought by T-RM is less pronounced as compared to that of P-RM, which to some extent confirms the intuition that proper modeling of word order and adjacency information is critical to the success of speech recognition. The combination of P-RM and T-RM, however, offers a moderate improvement over P-RM in isolation. As a final point, if we further linearly combine P-RM with PLSA, the CER can be ultimately reduced to 18.71%.

Table 2. *The speech recognition results (in CER (%)) of RM with proximity information.*

P-RM (τ =2)	P-RM (τ =3)	P-RM (τ =4)	P-RM (τ =5)	P-RM (τ =6)
18.91	18.89	18.97	18.98	19.07

Table 3. *The speech recognition results (in CER (%)) of T-RM and its combination with P-RM.*

T-RM	P-RM (τ =3) + T-RM
19.18	18.84

6. Conclusions

In this paper, we have presented a novel extension of the RM framework for language modeling in speech recognition. Our contribution is two-fold. First, the so-called "*bag-of-words*" assumption of RM is relaxed by incorporating word proximity evidence into the RM formulation. Second, latent topic information is additionally explored in an effort to further enhance the proximity-based RM framework. A series of experiments conducted on a large vocabulary continuous speech recognition (LVCSR) task seem to reveal that the various language models deduced from our framework are very comparable to existing language models. As to future work, we would like to adopt this LM framework for speech retrieval and summarization applications [23, 24].

7. References

[1] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, 1999.

[2] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE*, 88(8), pp. 1270 – 1278, 2000.

[3] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, 42(1), pp. 93–108, 2004.

[4] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, ACM Press, 2011.

[5] D.Gildea and T. Hofmann, "Topic-based language models using EM," in *Proceedings of the European Conference on, Speech Communication and Technology*, pp. 2167–2170, 1999.

[6] Y. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 5–8, 2005.

[7] C. X. Zhai, "Statistical language models for information retrieval: A critical review," *Foundations and Trends in Information Retrieval*, 2(3), pp. 137–213, 2008.

[8] T. Hoffmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, 42, pp. 177–196, 2001.

[9] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, pp. 993–1022, 2003.

[10] R. Lau, R. Rosenfeld and S. Roukos, "Trigger-based language models: a maximum entropy approach," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing,* pp. 45–48, 1993.

[11] T. Mikolov, M. Karafi´at, L. Burget, J. Cernock´y and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1045–1048, 2010.

[12] B. Roark, M. Saraclar and M. Collins, "Discriminative *n*-gram language modeling," *Computer Speech and Language*, 21(2), pp. 373–392, 2007.

[13] B. Chen and K.-Y. Chen, "Leveraging relevance cues for language modeling in speech recognition," *Information Processing & Management*, 49(4), pp. 807–816, 2013.

[14] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127, 2001.

[15] H.-M. Wang, B. Chen, J.-W. Kuo and S.-S. Cheng, "MATBN: a Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, 10(1), pp. 219–235, 2005.

[16] A. Stolcke, *SRI Language Modeling Toolkit* (http://www.speech.sri.com/projects/srilm/), 2000.

[17] B. Chen, J.-W. Kuo and W.-H. Tsai, "Lightly supervised and data-driven approaches to Mandarin broadcast news transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 777–780, 2004.

[18] H.-S. Lee, B. Chen, "Generalized likelihood ratio discriminant analysis," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, pp. 158–163, 2009.

[19] T. Mikolov, S. Kombrink, A. Deoras, L. Burget and J. Černocký, "RNNLM – Recurrent neural network language modeling toolkit," in *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, 2011.

[20] T. Oba, T. Hori and A. Nakamura, "A comparative study on methods of weighted language model training for reranking LVCSR *N*-best hypotheses," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 5126–5129, 2010.

[21] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 5126–5129, 1989.

[22] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, 39(1), pp. 1-38, 1977.

[23] Y.-W. Chen, K.-Y. Chen, H.-M. Wang and B. Chen, "Effective pseudo-relevance feedback for spoken document retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[24] B. Chen, H.-C. Chang and K.-Y. Chen, "Sentence modeling for extractive speech summarization," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, 2013.