Word Relevance Modeling for Speech Recognition

Kuan-Yu Chen[1], *Hao-Chin Chang*[2], *Berlin Chen*[2], *Hsin-Min Wang*[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]National Taiwan Normal University, Taipei, Taiwan
[1]{kychen, whm}@iis.sinica.edu.tw, [2]{699470503, berlin}@ntnu.edu.tw

Abstract

Language models for speech recognition tend to be brittle across domains, since their performance is vulnerable to changes in the genre or topic of the text on which they are trained. A number of adaptation methods, discovering either lexical co-occurrence or topic cues, have been developed to mitigate this problem with varying degrees of success. Among them, a more recent thread of work is the relevance modeling approach, which has shown promise to capture the lexical co-occurrence relationship between the entire search history and an upcoming word. However, a potential downside to such an approach is the need of resorting to a retrieval procedure to obtain relevance information; this is usually complex and time-consuming for practical applications. In this paper, we propose a word relevance modeling framework, which introduces a novel use of relevance information for dynamic language model adaptation in speech recognition. It not only inherits the merits of several existing techniques but also provides a flexible yet systematic way to render the lexical, topical, and proximity relationships between the search history and the upcoming word. Experiments on large vocabulary continuous speech recognition demonstrate the performance merits of the methods instantiated from this framework when compared to several existing methods.

Index Terms: language model, relevance, lexical co-occurrence, topic cues, adaptation

1. Introduction

Language modeling (LM) is indispensable for most automatic speech recognition (ASR) systems. It can be used to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the recognition output. The *n*-gram language model [1, 2, 3] that follows a statistical modeling paradigm is most prominently used in ASR because of the inherent simplicity and predictive power. Nevertheless, the model, aiming at capturing the local contextual information or the lexical regularity of a language, is inevitably faced with two fundamental problems. 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) conveyed in the search history beyond the immediately preceding *n*-1 words when predicting a word.

In view of the problems, several latent topic modeling approaches, which were originally formulated in information retrieval (IR) [4, 5, 6], have been introduced to complement the *n*-gram models through dynamic language model adaptation. The latent Dirichlet allocation (LDA) method [6, 7] and its precursor, the probabilistic latent semantic analysis (PLSA) method [5, 8], are two well-known instances. A commonality among them is that they both introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics [4]. 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. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes that the model parameters are fixed and unknown; while LDA places additional a priori constraints on the model parameters, i.e., viewing them as random variables that follow some Dirichlet distributions. Therefore, LDA possesses fully generative semantics and could overcome the over-fitting problem to some extent.

Apart from topic models, the notion of relevance modeling (RM), stemming from the information retrieval (IR) community [12, 13, 14], have recently been introduced to complement the *n*-gram models for ASR [11]. The RM approach tries to explore relevance cues so as to induce the co-occurrence relationship between the words in the search history and the upcoming word. The major spirit of RM is that each search history is assumed to be topically or semantically associated with an unknown relevance class, and each possible upcoming word can be regarded as a sample drawn from the relevance class. Thus, the probability of the co-occurrence (in a multinomial point of view) can be estimated from the relevance class, which is orthogonal to those that build on the lexical or topical cues merely inferred from the search history.

Our work in this paper can be viewed as a novel extension of the RM approach to ASR. To counteract the shortcoming of the RM approach, viz. the need of resorting to a time-consuming retrieval procedure for relevance modeling, we propose a word relevance modeling (WRM) framework. This framework not only inherits the merits of several existing techniques but also provides a more general mechanism to render the lexical and/or topical relationships between the search history and the word to be predicted. The utility of the proposed modeling framework is verified by both analytical and empirical comparisons with several widely used LM methods.

The remainder of this paper is organized as follows. In Section 2, we briefly review the conventional RM approach to ASR, and then describe the fundamentals of the word relevance language modeling framework and its extensions for language model adaptation. We compare our framework with other existing models in Section 3. Then, the experimental settings and the ASR results are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper and suggests avenues for future work

2. Word relevance language modeling

2.1. RM for language model adaptation

In the RM approaches to IR [12, 13, 14, 15], each query is assumed to be associated with an unknown relevance class R, and the documents that are relevant to the information need expressed in the query are samples drawn from R. The relevance model $P_{\mathrm{RM}}(w)$, from 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 the document. However, 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 [12, 13, 15, 16] is performed to approximate R with the top-ranked documents obtained from an initial round of retrieval.

The task of language modeling in ASR can be interpreted as calculating the conditional probability $P(w|H)$, where 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., a newly decoded word). When RM is applied to language modeling in ASR [11], the search history H is conceptually regarded as a query, while w is regarded as a (single-word) document. Therefore, the probability of H and w being jointly generated by the relevance class R_H of H, $P_{\mathrm{RM}}(H,w)$, can serve as the basis for deriving the conditional probability $P(w|H)$.

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

$$\begin{aligned} P_{\mathrm{RM}}(H,w) &= \Sigma_{m=1}^{M} P(D_m) P(h_1, h_2, \ldots h_L, w | D_m) \\ &= \Sigma_{m=1}^{M} P(D_m) P(w|D_m) \Pi_{l=1}^{L} P(h_l | D_m), \end{aligned} \tag{1}$$

where $P(D_m)$ is the probability that we would select D_m from $\mathbf{D}_H$ and $P(h_1, h_2, \ldots h_L, w | D_m)$ is the probability of simultaneously observing H and w in D_m. We further assume that w and the words in H are conditionally independent given D_m and their order is no importance (i.e., the so-called "*bag-of-words*" assumption). Therefore, the joint probability can be decomposed as a product of unigram probabilities of words generated by D_m. The probability $P(D_m)$ can be simply assumed 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 frequency in D_m and refined with smoothing techniques.

2.2. Word relevance modeling

The most challenging aspect facing RM is how to efficiently infer the relevance class so as to model the co-occurrence relationship between words in a search history and any upcoming word. As discussed in Section 2.1, the relevance class of a search history H is commonly approximated by the top-ranked documents returned by an IR system in response to H (taken as a query) during the speech recognition process. Although RM represents a promising alternative to other existing language model adaptation methods by exploring the relevance cues, the perennial need of resorting to a time-consuming IR procedure would obscure its feasibility for ASR. In view of this, in this paper, we propose a novel word relevance modeling (WRM) framework. WRM tries to explore "*word-word*" relevance cues prior to the speech recognition process, and such relevance cures are distinct from those "*word-document*" relevance cues inferred by RM.

The important notion of WRM is that the words in a search history H determine the semantic meaning of H, and the distributional information of other words occurring immediately nearby each history word in the training corpus, to some extent exhibiting a kind of "*word-word*" relevance information, collectively can be used to approximate the relevance class R_H of H. To this end, in the training phase (before speech recognition), the WRM model $P(w'|M_w)$ of each word w in the language can be trained by concatenating those words occurring within a vicinity of (or a word context window of size S around) each occurrence of w (which are postulated to be relevant to w) to form a relevant observation sequence for estimating $P(w'|M_w)$. As such, during speech recognition, the composite WRM model of a search history can be efficiently derived, without the perennial need of resorting to an external IR procedure. Consequently, the joint probability of H and w can thus be alternatively computed by:

$$P_{\mathrm{WRM}}(H,w) = \Sigma_{l'=1}^{L} P(M_{h_{l'}}) P(w|M_{h_{l'}}) \Pi_{l=1}^{L} P(h_l | M_{h_{l'}}), \tag{2}$$

where $P(w|M_{h_{l'}})$ and $P(h_l|M_{h_{l'}})$ are estimated on top of the word occurrence frequencies and refined with the Bayesian or Jelinek-Mercer smoothing method in the training phase. Finally, the conditional probability $P_{\mathrm{WRM}}(w|H)$ can be expressed by:

$$\begin{aligned} P_{\mathrm{WRM}}(w|H) &= \frac{P_{\mathrm{WRM}}(H,w)}{P_{\mathrm{WRM}}(H)} \\ &= \frac{\Sigma_{l'=1}^{L} P(M_{h_{l'}}) P(w|M_{h_{l'}}) \Pi_{l=1}^{L} P(h_l|M_{h_{l'}})}{\Sigma_{l'=1}^{L} P(M_{h_{l'}}) \Pi_{l=1}^{L} P(h_l|M_{h_{l'}})}. \end{aligned} \tag{3}$$

2.3. Incorporating latent topic information into WRM

To take a step forward, we investigate to incorporate latent topic information into the WRM framework. We assume that the WRM models share a common set of latent topic variables $\{T_1, T_2, \ldots, T_K\}$. Then, the probability that a word w is sampled from a WRM model M_{h_l} is no longer estimated based on the frequencies of the word occurring within a vicinity of word h_l, but rather based on the frequencies of the word in the latent topics as well as the likelihoods that the WRM model generates the respective topics:

$$P(w|M_{h_l}) = \Sigma_{k=1}^{K} P(w|T_k) P(T_k|M_{h_l}). \tag{4}$$

As with PLSA and LDA, the probabilities $P(w|T_k)$ and $P(T_k|M_{h_l})$ can be estimated by using inference algorithms like the expectation-maximization (EM) algorithm or variational approximation on the complete set of WRM models. The probability of H and w being simultaneously observed in the relevance class R_H of H is thus decomposed as:

$$\begin{aligned} &P_{\mathrm{TWRM}}(H,w) \\ &= \Sigma_{l'=1}^{L} \Sigma_{k=1}^{K} P(M_{h_{l'}}) P(T_k|M_{h_{l'}}) P(w|T_k) \Pi_{l=1}^{L} P(h_l|T_k). \end{aligned} \tag{5}$$

We term (5) the topic-based word relevance model (TWRM) hereafter. By the same token, the conventional RM model (*cf.* Section 2.1) can also incorporate a set of latent topic variables to describe the co-occurrence relationship between the search history and the upcoming word. The resulting model is designated as the topic-based relevance model (TRM) [11, 12].

2.4. Incorporating proximity information into WRM

As discussed in Section 2.1, RM usually assumes that the document prior $P(D_m)$ follows a uniform distribution [11, 12, 13] because there is no prior knowledge to determine the probability. However, for WRM and TWRM, we may leverage the prior probability to capture the proximity information between each history word (regarded as a WRM model) and the newly decoded word. For this idea to go, an exponential decay function can be utilized to govern the prior probability [9, 10]:

$$P(M_{h_l}) = \phi_l \Pi_{j=l+1}^{L}(1-\phi_j), \quad (6)$$

where ϕ_1 is set to 1 and ϕ_l, $l = 2,\ldots,L$, is set to a fixed value between 0.5 and 1; and $P(M_{h_L})$ will be equal to ϕ_L. The probability exponentially decays with the distance that h_l is apart from w and sums to 1.

2.5. Language model adaptation

Since the background *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 various WRM models with the background *n*-gram (e.g., trigram) language model to form an adaptive language model to guide the speech recognition process. For example, the general trigram language model can be adapted by the WRM model as follows [2, 3, 15]:

$$P_{\text{Adapt}}(w \mid H) = \lambda \cdot P_{\text{WRM}}(w \mid H) + (1-\lambda) \cdot P_{\text{BG}}(w \mid h_{L-1}, h_L), \quad (7)$$

where λ is a tunable nonnegative weighting parameter.

3. Comparison with other models

Our proposed methods, namely WRM and TWRM, can be analyzed from several perspectives. First, like RM and TRM [11, 12], WRM and TWRM also consider the co-occurrence relationship between the entire search history and the newly decoded word. But it should be mentioned that our proposed framework in this paper explores different relevance cues, viz. the word-word co-occurrence relationships, from those word-document co-occurrence relationships derived by RM (TRM). The major advantage of WRM and TWRM is that they do not need to intensively perform a local feedback-like procedure to approximate the relevance class for the search histories generated during speech recognition. As such, WRM and TWRM are deemed more efficient (and thus feasible) than RM and TRM for ASR. Further, it is generally expected that WRM and TWRM could additionally take the distance (or proximity) between the history words and the decoded word into account through the proper use of weights assigned by an exponential decay function.

Second, analogous to topic models such as PLSA and LDA, TWRM makes use of a common set of latent topic variables to describe the co-occurrence relationships between a word and its pseudo-document (viz. those words occurring immediately nearby the word in the training corpus). These latent topics somehow address the important notions of synonymy and polysemy. However, PLSA and LDA have to estimate their component probability distributions on-the-fly for a new search history using the expectation-maximization or other more sophisticated algorithms, which would be time-consuming. In contrast, for WRM and TWRM, the language model probability can be easily composed from the component probability distributions that have been trained beforehand, without recourse to any complex inference procedure during the recognition (or rescoring) process.

Third, it should be mentioned that the size of the word context window plays an important role in our proposed WRM framework. A large window size can capture more document-level like information for a WRM or TWRM model, while a small window size can better model the local co-occurrence relationships between words in the search history and a newly decoded word. The selection of the window size will be discussed later (*cf.* Section 5).

4. Experimental setup

Our ASR experiments were conducted on the MATBN (Mandarin Across Taiwan Broadcast News) corpus [10, 11, 21]. 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 [18]. Another subset of 3-hour speech data collected in 2003 was used as the development set (1.5 hours) and the test set (1.5 hours).

The vocabulary size is about 72 thousand words. The baseline trigram language model was estimated from a background text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 (extracted from the Chinese Gigaword Corpus released by LDC) using the SRI Language Modeling Toolkit (SRILM) [17]. Another text corpus consisting of the orthographic transcripts of the MATBN corpus (excluding the test set) was used for training the proposed models and those of other adaptation methods. There are about one million Chinese characters from 3,643 news stories.

All the experiments were performed in a word graph rescoring way. The word graphs of the speech data were built beforehand with a typical large vocabulary continuous speech recognition (LVCSR) system [18]. The baseline rescoring procedure with the background trigram language model results in a character error rate (CER) of 20.08% on the test set. Note that the parameter ϕ_l of the exponential decay function was set to 0.6 directly and all the other constants or weighting (interpolation) coefficients used for language modeling were tuned by using the development set.

5. Experimental results

In the first set of experiments, we evaluated the utility of RM (*cf.* Section 2.1) and WRM (*cf.* Section 2.2). The CER results of RM and WRM with respect to the number of documents being retrieved to approximate the relevance class of the RM model, or the size of the window used to obtain the WRM model for a word, are shown in Table 1. Several observations can be made from the results. First, both RM and WRM bring remarkable improvements over the baseline trigram model in all cases. However, WRM is more attractive than RM since it does not need to perform a time-consuming additional run of retrieval to approximate the relevance class for each search history during the recognition (or rescoring) process. Second, RM achieves its best performance when the number of retrieved documents is set to 16. The result reveals that a small subset of relevant documents retrieved from the contemporaneous corpus is sufficient for dynamically constructing the RM model. Third, WRM achieves its best performance when the window size is set to 16 (with a uniform prior) and 32 (with a proximity-based prior (*cf.* Section 2.4)). To our surprise, WRM with a uniform prior seem to perform better than WRM with a proximity-based prior;

Table 1: *ASR results of RM and WRM (in CER (%))*.

Document No. / Window Size	RM	WRM (Uniform)	WRM (Proximity)
8	**19.40**	19.59	19.60
16	**19.40**	**19.39**	**19.54**
32	19.42	19.42	**19.54**

Table 2: *ASR results of PLSA, LDA, TRM, and TWRM (in CER (%))*.

Topic No.	PLSA	LDA	TRM	TWRM
16	19.21	19.29	**19.25**	19.36
32	19.22	19.30	19.27	19.26
64	19.17	19.28	19.31	**19.24**
128	**19.15**	**19.15**	19.30	19.45

the reason behind it is worthy of further investigation. Both WRM and RM yield a relative CER reduction of about 3.4% over the baseline system when using their best settings. The CER reduction is significant according to the standard NIST MAPSSWE test [19]. The experimental results validate the utility of WRM for dynamic language model adaptation.

In the next set of experiments, we evaluated TWRM (*cf.* Section 2.3), which is a natural extension of WRM by additionally introducing a set of latent topics to describe the "*word-word*" co-occurrence relationship. The window size was set to 16, and the uniform prior was adopted. As shown in Table 2, TWRM, which combines both relevance modeling and topic modeling, demonstrates consistent performance gains over WRM, which considers relevance modeling only. TWRM can further reduce the CER to 19.24%, which was 19.39% for WRM under the same setting. By combining these two extra information cues jointly, TWRM yields a relative CER reduction of 4.2% over the baseline system.

Furthermore, we compared TWRM with several well-practiced language model adaptation methods, including PLSA, LDA, and TRM. As shown in Table 2, PLSA, LDA, and TRM achieve the best CER of 19.15%, 19.15%, and 19.25%, respectively. Although TWRM preforms slightly worse than PLSA and LDA, the performance is in fact at the same level. An important advantage of TWRM over PLSA, LDA, and TRM is that TWRM is much more efficient than PLSA, LDA, and TRM. Therefore, TWRM is more suitable than others in dynamic language model adaptation.

Finally, we investigated the combination of WRM (or TWRM) and LDA. Unlike LDA, which considers the "*word-document*" co-occurrence relationships by using a set of latent topics, WRM and TWRM explore the "*word-word*" relevance cues. Therefore, it is expected that they can conspire to further boost the speech recognition performance. Our experimental results show that the combination of WRM and LDA and that of TWRM and LDA achieve the CER of 19.09% and 19.06%, respectively. These results indeed exhibit good complementarity between WRM (or TWRM) and LDA.

6. Conclusions and future work

In this paper, we have proposed a word relevance modeling framework for dynamic language model adaptation in speech recognition. The new framework suggests a promising avenue for the integration of relevance, topic and proximity information. However, the proximity information does not seem to show a clear gain, and further investigation is needed. Next, the utility of the methods instantiated from this framework, viz. WRM and TWRM, have been validated by extensive comparisons with several widely used language model adaptation methods. As to future work, we envisage several directions, including discriminative training [20] and the exploration of different granularities of semantic context for relevance modeling. Additionally, we will explore leveraging WRM and TWRM for speech retrieval and summarization.

7. Acknowledgment

This work was supported in part by the National Science Council, Taiwan, under Grants NSC 99-2221-E-001-009-MY3, NSC 99-2221-E-003-017-MY3, NSC 98-2221-E-003-011-MY3, NSC 100-2515-S-003-003 and NSC 99-2631-S-003-002, and "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan.

8. References

[1] F. Jelinek, "Up from trigrams! – the struggle for improved language models," in *Proc. Eurospeech 1991*.

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

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

[4] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of Latent Semantic Analysis*, 2007.

[5] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, 42(1/2):177-196, 2001.

[6] D.M. Blei et al., "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3:993-1022, 2003.

[7] Y. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. Interspeech 2005*.

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

[9] B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," in *Proc. ICASSP 2009*.

[10] K.Y. Chen et al., "Latent topic modeling of word vicinity information for speech recognition," in *Proc. ICASSP 2010*.

[11] K.Y. Chen and B. Chen, "Relevance language modeling for speech recognition," in *Proc. ICASSP 2011*.

[12] P.N. Chen et al., "Leveraging relevance cues for improved spoken document retrieval," in *Proc. Interspeech 2011*.

[13] V. Lavrenko and B. Croft, "Relevance-based language models," in *Proc. SIGIR 2001*.

[14] B. Croft and J. Lafferty (eds.), *Language Models for Information Retrieval*, Kluwer International Series on Information Retrieval, Volume 13, Kluwer Academic Publishers, 2002.

[15] C.X. Zhai, Statistical Language Models for Information Retrieval, Morgan & Claypool Publishers, 2008.

[16] J. Xu and B. Croft, "Query expansion using local and global document analysis," in *Proc. SIGIR 1996*.

[17] A. Stolcke, SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/.

[18] B. Chen, et al., "Lightly supervised and data-driven approaches to Mandarin broadcast news transcription," in *Proc. ICASSP 2004*.

[19] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP 1989*.

[20] B. Chen and J.W. Liu, "Discriminative language modeling for speech recognition with relevance information," in *Proc. ICME 2011*.

[21] H.M. Wang, et al., "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), pp. 219-236, June 2005.