EFFECTIVE PSEUDO-RELEVANCE FEEDBACK FOR LANGUAGE MODELING IN SPEECH RECOGNITION

Berlin Chen[†] *, Yi-Wen Chen*[†] *, Kuan-Yu Chen*[#], *Ea-Ee Jan*[*]

[†] National Taiwan Normal University, Taiwan
[#]Institute of Information Science, Academia Sinica, Taiwan
[*]IBM Thomas J. Watson Research Center, USA
E-mail: [†] {berlin, 699470462}@ntnu.edu.tw, [#]kychen@iis.sinica.edu.tw, [*]ejan@us.ibm.com

ABSTRACT

A part and parcel of any automatic speech recognition (ASR) system is language modeling (LM), which helps to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output hypothesis given an input utterance. Despite the fact that the *n*-gram model remains the predominant one, a number of novel and ingenious LM methods have been developed to complement or be used in place of the *n*-gram model. A more recent line of research is to leverage information cues gleaned from pseudo-relevance feedback (PRF) to derive an utterance-regularized language model for complementing the *n*-gram model. This paper presents a continuation of this general line of research and its main contribution is two-fold. First, we explore an alternative and more efficient formulation to construct such an utterance-regularized language model for ASR. Second, the utilities of various utterance-regularized language models are analyzed and compared extensively. Empirical experiments on a large vocabulary continuous speech recognition (LVCSR) task demonstrate that our proposed language models can offer substantial improvements over the baseline *n*-gram system, and achieve performance competitive to, or better than, some state-of-the-art language models.

Index Terms **—** Speech recognition, language modeling, pseudo-relevance feedback, information retrieval, relevance

1. INTRODUCTION

Among many language modeling (LM) methods developed so far, the *n*-gram model is arguably the most prominently used one and has shown its practical effectiveness on many automatic speech recognition (ASR) tasks [1, 2, 3]. Nevertheless, the *n*-gram model with the goal of capturing the local contextual information or lexical regularity of a language has frequently been criticized on two fronts. On one hand, it is fragile across domains, since the performance is susceptible to changes in the genre or topic of the text on which it is trained. On the other hand, it fails to capture the information (either lexical or semantic information) conveyed in the search history beyond the immediately preceding *n*-1 words of a newly decoded word.

With the alleviation of the aforementioned two deficiencies as motivation, the speech recognition community has recently witnessed a flurry of research activity aimed at the development of novel and ingenious LM methods to be used complementarily to (or as a replacement for) the *n*-gram model with varying degrees of success [4]. The topic modeling approach [5, 6], which was originally formulated in information retrieval (IR) [7], constitutes one most prominent and successful school of thought for dynamic language model adaptation in ASR [8, 9]. To exemplify, latent Dirichlet allocation (LDA) [10] and its precursor, probabilistic latent semantic analysis (PLSA) [11], 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 entire search history (virtually 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, viz. thinking of them as random variables that follow some Dirichlet distributions [5].

In addition, there are other methods developed to complement the *n*-gram models, such as the trigger-based language model (TBLM) [12, 13], 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. Alternatively, the recurrent neural network language model (RNNLM) [14] and the discriminative language model (DLM) [15, 16] has also garnered 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 stream of research, orthogonal to the foregoing LM methods, is to capitalize on information cues gathered from pseudo-relevance feedback (PRF) [7, 17] for dynamic language model adaptation in ASR. PRF is arguably the most effective and commonly-used paradigm for improving query modeling in the IR community, which assumes that a small number of top-ranked documents obtained from an initial round of retrieval is relevant to the query and can be utilized for query reformulation. Subsequently, the system performs a second round of retrieval with the enhanced query representation to search for more relevant documents. As the notion of PRF is adopted and formalized for dynamic language model adaptation in ASR, for each test utterance, search histories (or, for example, the initial top-one output hypothesis) generated by the baseline recognition system, is taken as a query and posed to an IR system to obtain a set of top-ranked relevant documents from the contemporaneous (or in-domain) text collection. These documents are in turn used to estimate an utterance-regularized language model for complementing the *n*-gram model. However, dynamic language model adaptation on top of the PRF process may confront two intrinsic challenges. One is how to purify the top-ranked feedback documents obtained from the retrieval so as to remove redundant and non-relevant information [18]. The other is how to effectively utilize the selected set of representative feedback documents for estimating a more accurate utterance-regularized language model. For the latter, we have recently explored the use of the relevance model (RM) to render the lexical co-occurrence relationships between a search history and the word to be predicted in the feedback documents [19]. Our work in this paper continues this general line of research and the main contribution includes two aspects. First, we explore an alternative and more efficient formulation to construct such an utterance-regularized language model for ASR. Second, the utilities of various utterance-regularized language models are analyzed and compared extensively.

The rest of this paper is organized as follows. In Section 2, we start by illustrating the intuition underlying the mathematical formulation of our previous proposed relevance model (RM), and then shed light on the principal idea of leveraging a simple mixture model (SMM) in concert with PRF for dynamic language model adaptation in ASR. After that, the experimental settings and a series of LVCSR experiments are presented in Sections 3 and 4, respectively. Finally, Section 5 concludes the paper alongside further avenues of future research.

2. UTTERANCE-REGULARIZED LANGUAGE MODELING

2.1. Relevance Model (RM)

The task of language modeling in speech recognition can be framed 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 the relevance model (RM) is applied to language modeling in speech recognition, we hypothesize that each search history H has a relevance class R_H [21, 22] 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, viz. $P_{\mathrm{RM}}(H,w)$, accordingly can be used to derive the conditional probability $P(w|H)$ for speech recognition [19].

Nevertheless, because the relevance class R_H of each search history H is not known in advance, we may exploit a PRF process 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

$$P_{\mathrm{RM}}(H,w)=\sum_{m=1}^{M}P(D_m)P(H,w|D_m), \tag{1}$$

where $P(D_m)$ is the probability that we would randomly select D_m and $P(H,w|D_m)$ (or $P(h_1,h_2,\ldots h_L,w|D_m)$) 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:

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

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 [6, 19]. As such, the conditional probability $P(w|H)$ formulated with the RM model is expressed by

$$P_{\mathrm{RM}}(w \mid H) = \frac{P_{\mathrm{RM}}(H,w)}{P_{\mathrm{RM}}(H)} = \frac{\sum_{m=1}^{M} P(D_m) P(w \mid D_m) \prod_{l=1}^{L} P(h_l \mid D_m)}{\sum_{m'=1}^{M} P(D_{m'}) \prod_{l=1}^{L} P(h_l \mid D_{m'})}. \quad (3)$$

In addition, 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 the RM model with the baseline *n*-gram (e.g., trigram) language model to form an utterance-regularized (adaptive) language model for guiding the speech recognition process:

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

where the interpolation parameter λ encodes the trade-off between the RM model and the baseline *n*-gram language model.

2.2. Simple Mixture Model (SMM)

In this paper, we explore an alternative and more efficient formulation to extract relevance information from PRF for language model adaptation in ASR, which is referred to hereafter as the simple mixture model (SMM). The basic idea of SMM is to assume that the set of feedback documents $\mathbf{D}_H = \{D_1, D_2, \ldots, D_M\}$ are relevant and a feedback model $P(w \mid FB)$ estimated from these documents can potentially benefit ASR. Specifically, SMM assumes that words in $\mathbf{D}_H$ are drawn from a two-component mixture model [23, 24]: 1) one component is the feedback model $P(w \mid FB)$ **,** and 2) the other is a background model $P(w \mid BG)$, which is set to be the baseline unigram language model in this study. The feedback model $P(w \mid FB)$ is estimated by maximizing the log-likelihood of the set of feedback documents $\mathbf{D}_H$ expressed as follows, using the expectation-maximization (EM) algorithm [25]:

$$LL_{\mathbf{D}_H} = \sum_{D_m \in \mathbf{D}_H} \sum_{w \in V} c(w, D_m) \cdot \log[\alpha \cdot P(w \mid FB) + (1-\alpha) \cdot P(w \mid BG)], \quad (5)$$

where V denotes the set of all the words in the vocabulary, $c(w, D_m)$ is the occurrence count of w in D_m, and α is the interpolation parameter used to control the degree of reliance on $P(w \mid FB)$ rather than on $P(w \mid BG)$. The maximization of (5) can be conducted iteratively via the following two EM update equations:

$$P^{(l)}(FB \mid w) = \frac{\alpha \cdot P^{(l)}(w \mid FB)}{\alpha \cdot P^{(l)}(w \mid FB) + (1-\alpha) \cdot P(w \mid BG)} \quad (6)$$

and

$$P^{(l+1)}(w \mid FB) = \frac{\sum_{D_m \in \mathbf{D}_H} c(w, D_m) \cdot P^{(l)}(FB \mid w)}{\sum_{w' \in V} \sum_{D_j \in \mathbf{D}_H} c(w', D_j) \cdot P^{(l)}(FB \mid w')}, \quad (7)$$

where l denotes the l-th iteration of the EM algorithm. This estimation will enable more specific words (i.e., words in $\mathbf{D}_H$ that are not well-explained by the background model) to receive more probability mass, thereby leading to a more discriminative feedback model $P(w \mid FB)$. Simply put, the feedback model $P(w \mid FB)$ is anticipated to extract useful word usage cues from $\mathbf{D}_H$, which are not only relevant to the search history H, but also external to those already captured by the background model. Accordingly, the feedback model $P(w \mid FB)$ of SMM can be combined with the baseline *n*-gram language model through a simple linear interpolation similar to (4).

The notion of jointly leveraging PRF and SMM has recently attracted much attention and been applied with success to many IR tasks [7, 23, 24]. However, this notion has never been extensively explored for language modeling in speech recognition, as far as we are aware.

2.3. Model 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 would share the same relevance information. In order to construct the component models of RM and SM respectively for representing a given test utterance, the top-one word sequence hypothesis, output by the baseline ASR system with the background *n*-gram language model, is taken as a query and posed to an IR system to obtain a set of M relevant documents from the 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 [19]. We, therefore, adopt such simplification for the following evaluations of the RM and SMM methods. One thing to note is that since RM has to dynamically compose the conditional probability $P_{\mathrm{RM}}(w \mid H)$ (*cf.* (3)) for each new search history H, which inevitably involves a denominator term accumulated from all history words, it would be less efficient than SMM. It was experimentally shown that SMM delivered a 20-fold speed increase in language model access compared to RM for dynamic language model adaptation. In addition, the access time of SMM is at least two orders of magnitude faster than that of PLSA and LDA; the latter two models have to estimate their component probability distributions on-the-fly for a new search history using EM or other more sophisticated algorithms, which is indeed more time-consuming.

TABLE I
The speech recognition results (in CER (%)) of various language models compared in this paper.

SMM	RM	PLSA	LDA	TBLM	RNNLM	DLM (MERT)	DLM (GCLM)	DLM (WGCLM)	Baseline Trigram
18.94	19.21	19.28	19.22	20.09	19.10	19.74	19.89	19.62	20.22

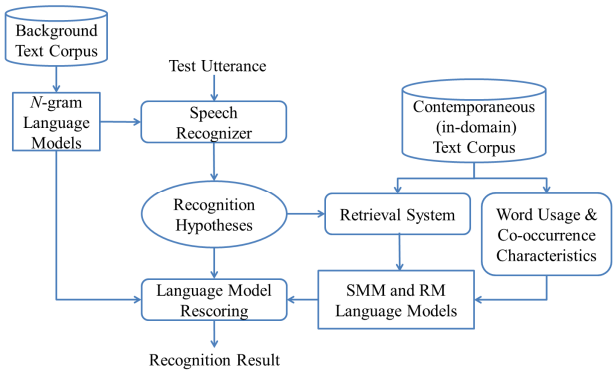

Fig. 1. A schematic illustration of the exploration of pseudo-relevance feedback in concert with SMM and RM for ASR.

Also worth mentioning is that we implement the IR system with the LM retrieval approach [6, 20], 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 having higher probabilities of generating the query are deemed to be more relevant to the test utterance. Furthermore, such an IR system can be very efficiently implemented with inverted files [6, 7]. A schematic illustration of how to explore pseudo-relevance feedback in concert with SMM and RM for ASR is depicted in Figure 1.

3. EXPERIMENTAL SETUP

The speech corpus consists of about 196 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News) [26]. 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 [27]. Another subset of 1.5-hour speech data collected within 2003 is reserved as the test set.

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) [28]. The adaptation (contemporaneous) text corpus used for training the RM and SMM models and the other adaptation methods compared in this paper (*cf.* Section 4) 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 (viz. language model rescoring). The associated word graphs of the speech data were built beforehand with a typical large vocabulary continuous recognition (LVCSR) system [29, 30]. 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.

4. EXPERIMENTAL RESULTS

At the outset, we assess the utility of SMM for ASR, by comparing it with RM and several well-practiced, state-of-the-art language models, including PLSA, LDA, TBLM, RNNLM and DLM. The corresponding CER results are shown in Table I, where the result of the baseline trigram model is also listed for comparison. It should be noted that in this paper, RNNLM was implemented with the toolkit released by [31]. Furthermore, DLM, utilizing features such as acoustic model probability, baseline trigram probability, and unigram and bigram counts, was trained with different algorithms, including minimum error rate training (referred to as "MERT" for short), global conditional log-linear model (referred to as "GCLM" for short) and weighted global conditional log-linear model (referred to as "WGCLM" for short). Interested readers may refer to [15, 32] for a thorough and updated introduction to various training algorithms designed and developed for DLM. Several noteworthy observations can be drawn from Table I. First, SMM achieves the best performance, which leads to a relative CER improvement of about 6% over the baseline trigram model. Next, the second best performance is

TABLE II
The perplexity results (in CER (%)) of some language models compared in this paper.

SMM	RM	PLSA	LDA	Baseline Trigram
347.47	524.10	519.50	521.12	667.23

obtained with RNN, while RM, PLSA and LDA are marginally inferior compared to SMM and RNN, but are apparently better than the three variants of DLM. Finally, TBLM provides an almost negligible improvement as compared to the baseline trigram model.

In the next set of experiments, we compare among SMM, RM and two representative topic models (viz. PLSA and LDA) in terms of perplexity reduction. This is because that these four LM methods bear close resemblance to each other in the sense that they dynamically perform utterance-regularized language modeling during word graph rescoring, and thereby are the best potentials that would offer substantial perplexity reductions over the baseline trigram model. The corresponding results of these four LM methods, alongside that of the baseline trigram model, are shown in Table II. As can be seen, all these four LM methods indeed perform quite well in perplexity reduction, which confirms our postulation. In particular, SMM is again the best performing one, yielding a significant perplexity reduction of 48% over the baseline trigram model, while RM, PLSA and LDA tend to be on par with each other.

In the third set of experiments, we evaluate the CER performance levels of SMM and RM with respect to different numbers of top-ranked feedback documents being used for estimating their component models, as shown in Figure 2. Consulting Figure 2 we notice two particularities. One is that there is more fluctuation in the CER curve of SMM than in that of RM. The CER performance of using SMM is steadily improved when the number of top-ranked documents being used becomes larger; the improvement, however, seems to reach a peak elevation when the number is set to 64, and then to degrade when the number is set to 128. For SMM, the extraction of relevance information from feedback documents is not guided by a test utterance (or its corresponding search histories), as elaborated earlier in Section 2.2. When too many feedback documents are being used, there would be a concern for SMM to be distracted from being able to appropriately model the test utterance, which is probably caused by some dominant distracting (or irrelevant) feedback documents. The other is that the CER performance of RM is in contrast less sensitive to the number of feedback documents being used. We might attribute this phenomenon to the ability of RM to explicitly and dynamically render the co-occurrence relationship between an entire search history and a word being predicted in each feedback document (*cf.* (3)). However, this is achieved at the price of longer LM access time. As a final

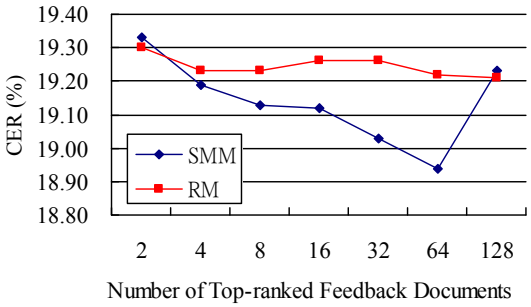

Fig. 2. The CER (%) results of SMM and RM with respect to different numbers of top-ranked documents being used for estimating their component models.

point, we have also investigated to integrate SMM and RM though a simple linear combination to complement the baseline trigram model, for which only moderate improvements with respect to CER and perplexity reductions have been evidenced. In the meantime, we are extensively experimenting on more principled ways to integrate SMM and RM, as well as their combinations with other LM methods like RNNLM, PLSA and LDA, to name a few.

5. CONCLUSIONS

In this paper, we have presented a more efficient and effective approach to leveraging pseudo-relevance feedback for language modeling in speech recognition. Our contribution is two-fold. First, we have explored various formulations to derive utterance-regularized language models. Second, the utilities of our proposed language models have been extensively analyzed and compared with several existing language models. Experimental evidence supports that the language models deduced from our modeling framework are very comparable to existing ones for LVCSR. As to future work, we would like to investigate jointly integrating proximity and other different kinds of relevance and lexical/semantic information cues into the process of feedback document selection [18, 33] so as to improve the empirical effectiveness of such utterance-regularized language modeling in ASR. In addition, we intend to further adopt and formalize the proposed LM methods for speech summarization and retrieval.

6. ACKNOWLEDGEMENTS

This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 101-2221-E-003-024-MY3, NSC 102-2221-E-003-014-, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 and NSC 99-2221-E-003-017-MY3.

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] X. Huang, A. Acero and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.

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

[5] D. Blei and J. Lafferty, "Topic models," in A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] C. Troncoso and T. Kawahara, "Trigger-based language model adaptation for automatic meeting transcription," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 1297–1300, 2005.

[14] 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.

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

[16] J.-W. Kuo and B. Chen, "Minimum word error based discriminative training of language models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1277–1280, 2005.

[17] J. Rocchio, "Relevance feedback in information retrieval," in G. Salton (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–23, Prentice Hall, 1971.

[18] 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*, pp. 8535–8539, 2013.

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

[20] B. Chen, K.-Y. Chen, P.-N. Chen, Yi-W. Chen, "Spoken document retrieval with unsupervised query modeling techniques," *IEEE Transactions on Audio, Speech and Language Processing*, 20(9), pp. 2602-2612, November 2012.

[21] 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.

[22] V. Lavrenko, *A Generative Theory of Relevance*, Springer, 2009.

[23] C. X. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of ACM SIGIR Conference on Information and knowledge management*, pp. 403–410, 2001.

[24] T. Tao and C. X. Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *Proceedings of ACM SIGIR Conference on Information and knowledge management*, pp. 162–169, 2006.

[25] 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.

[26] 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.

[27] S.-H. Liu, F.-H. Chu, S.-H. Lin, H.-S. Lee and B Chen, "Training data selection for improving discriminative training of acoustic models," in *Proceedings of* I*EEE workshop on Automatic Speech Recognition and Understanding*, pp. 284–289, 2007.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] Y.-W. Chen, B.-H. Hao, K.-Y. Chen and B. Chen, "Incorporating proximity information for relevance language modeling in speech recognition," to appear in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2013.