Leveraging Effective Query Modeling Techniques for Speech Recognition and Summarization

Kuan-Yu Chen[*†], Shih-Hung Liu[*], Berlin Chen[#], Ea-Ee Jan[+], Hsin-Min Wang[*], Wen-Lian Hsu[*], and Hsin-Hsi Chen[†]

[*]Institute of Information Science, Academia Sinica, Taiwan
[†]National Taiwan University, Taiwan
[#]National Taiwan Normal University, Taiwan
[+]IBM Thomas J. Watson Research Center, USA

{kychen, journey, whm, hsu}@iis.sinica.edu.tw,
berlin@ntnu.edu.tw, hhchen@csie.ntu.edu.tw, ejan@us.ibm.com

Abstract

Statistical language modeling (LM) that purports to quantify the acceptability of a given piece of text has long been an interesting yet challenging research area. In particular, language modeling for information retrieval (IR) has enjoyed remarkable empirical success; one emerging stream of the LM approach for IR is to employ the pseudo-relevance feedback process to enhance the representation of an input query so as to improve retrieval effectiveness. This paper presents a continuation of such a general line of research and the main contribution is threefold. First, we propose a principled framework which can unify the relationships among several widely-used query modeling formulations. Second, on top of the successfully developed framework, we propose an extended query modeling formulation by incorporating critical query-specific information cues to guide the model estimation. Third, we further adopt and formalize such a framework to the speech recognition and summarization tasks. A series of empirical experiments reveal the feasibility of such an LM framework and the performance merits of the deduced models on these two tasks.

1 Introduction

Along with the rapidly growing popularity of the Internet and the ubiquity of social web communications, tremendous volumes of multimedia contents, such as broadcast radio and television programs, digital libraries and so on, are made available to the public. Research on multimedia content understanding and organization has witnessed a booming interest over the past decade. By virtue of the developed techniques, a variety of functionalities were created to help distill important content from multimedia collections, or provide locations of important speech segments in a video accompanied with their corresponding transcripts, for users to listen to or to digest. Statistical language modeling (LM) (Jelinek, 1999; Jurafsky and Martin, 2008; Zhai, 2008), which manages to quantify the acceptability of a given word sequence in a natural language or capture the statistical characteristics of a given piece of text, has been proved to offer both efficient and effective modeling abilities in many practical applications of natural language processing and speech recognition (Ponte and Croft, 1998; Jelinek, 1999; Huang, *et al.*, 2001; Zhai and Lafferty, 2001[a]; Jurafsky and Martin, 2008; Furui *et al.*, 2012; Liu and Hakkani-Tur, 2011).

The LM approach was first introduced for the information retrieval (IR) problems in the late 1990s, indicating very good potential, and was subsequently extended in a wide array of follow-up studies. One typical realization of the LM approach for IR is to access the degree of relevance between a query and a document by computing the likelihood of the query generated by the document (usually referred to as the query-likelihood approach) (Zhai, 2008; Baeza-Yates and Ribeiro-Neto, 2011). A document is deemed to be relevant to a given query if the corresponding document model is more likely to generate the query. On the other hand, the Kullback-Leibler divergence measure (denoted by KLM for short hereafter), which quantifies the degree of relevance between a document and a query from a more rigorous information-theoretic perspective, has been proposed (Lafferty and Zhai, 2001; Zhai and Lafferty, 2001[b]; Baeza-Yates and Ribeiro-Neto, 2011). KLM not only can be thought as a natural generalization of the query-likelihood approach, but also has the additional merit of being able to accommodate extra information cues to improve the performance of document ranking. For example, a main challenge facing such a measure is that since a given query usually consists of few words, the true information need is hard to be inferred from the surface statistics of a query. As such, one emerging stream of thought for KLM is to employ the

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1474–1480,
October 25-29, 2014, Doha, Qatar.

pseudo-relevance feedback process to construct an enhanced query model (or representation) so as to achieve better retrieval effectiveness (Hiemstra *et al.*, 2004; Lv and Zhai, 2009; Carpineto and Romano, 2012; Lee and Croft, 2013).

Following this line of research, the major contribution of this paper is three-fold: 1) we analyze several widely-used query models and then propose a principled framework to unify the relationships among them; 2) on top of the successfully developed query models, we propose an extended modeling formulation by incorporating additional query-specific information cues to guide the model estimation; 3) we explore a novel use of these query models by adapting them to the speech recognition and summarization tasks. As we will see, a series of experiments indeed demonstrate the effectiveness of the proposed models on these two tasks.

2 Language Modeling Framework

2.1 Kullback-Leibler Divergence Measure

A promising realization of the LM approach to IR is the Kullback-Leibler divergence measure (KLM), which determines the degree of relevance between a document and a query from a rigorous information-theoretic perspective. Two different language models are involved in KLM: one for the document and the other for the query. The divergence of the document model with respect to the query model is defined by

$$\mathrm{KL}(Q \| D) = \Sigma_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|D)}. \quad (1)$$

KLM not only can be thought as a natural generalization of the traditional query-likelihood approach (Yi and Allan, 2009; Baeza-Yates and Ribeiro-Neto, 2011), but also has the additional merit of being able to accommodate extra information cues to improve the estimation of its component models in a systematic way for better document ranking (Zhai, 2008).

Due to that a query usually consists of only a few words, the true query model $P(w|Q)$ might not be accurately estimated by the simple ML estimator (Jelinek, 1991). There are several studies devoted to estimating a more accurate query modeling, saying that it can be approached with the pseudo-relevance feedback process (Lavrenko and Croft, 2001; Zhai and Lafferty, 2001[b]). However, the success depends largely on the assumption that the set of top-ranked documents, $\mathbf{D}_{Top}=\{D_1,D_2,...,D_r,...\}$, obtained from an initial round of retrieval, are relevant and can be used to estimate a more accurate query language model.

2.2 Relevance Modeling

Under the notion of relevance modeling (RM, often referred to as RM-1), each query Q is assumed to be associated with an unknown relevance class R_Q, and documents that are relevant to the semantic content expressed in query are samples drawn from the relevance class R_Q. Since there is no prior knowledge about R_Q, we may use the top-ranked documents $\mathbf{D}_{Top}$ to approximate the relevance class R_Q. The corresponding relevance model can be estimated using the following equation (Lavrenko and Croft, 2001; Lavrenko, 2004):

$$P_{\mathrm{RM}}(w|Q) = \frac{\Sigma_{D_r \in \mathbf{D}_{Top}} P(D_r) P(w|D_r) \Pi_{w' \in Q} P(w'|D_r)}{\Sigma_{D_r'' \in \mathbf{D}_{Top}} P(D_r'') \Pi_{w' \in Q} P(w'|D_r'')}. \quad (2)$$

2.3 Simple Mixture Model

Another perspective of estimating an accurate query model with the top-ranked documents is the simple mixture model (SMM), which assumes that words in $\mathbf{D}_{Top}$ are drawn from a two-component mixture model: 1) One component is the query-specific topic model $P_{\mathrm{SMM}}(w|Q)$**,** and 2) the other is a generic background model $P(w|BG)$. By doing so, the SMM model $P_{\mathrm{SMM}}(w|Q)$ can be estimated by maximizing the likelihood over all the top-ranked documents (Zhai and Lafferty, 2001[b]; Tao and Zhai, 2006):

$$L = \prod_{D_r \in \mathbf{D}_{Top}} \prod_{w \in V} \left(\alpha \cdot P_{\mathrm{SMM}}(w|Q) + (1-\alpha) \cdot P(w|BG)\right)^{c(w,D_r)}, \quad (3)$$

where α is a pre-defined weighting parameter used to control the degree of reliance between $P_{\mathrm{SMM}}(w|Q)$ and $P(w|BG)$. This estimation will enable more specific words to receive more probability mass, thereby leading to a more discriminative query model $P_{\mathrm{SMM}}(w|Q)$.

Although the SMM modeling aims to extract extra word usage cues for enhanced query modeling, it may confront two intrinsic problems. One is the extraction of word usage cues from $\mathbf{D}_{Top}$ is not guided by the original query. The other is that the mixing coefficient α is fixed across all top-ranked documents albeit that different documents would potentially contribute different amounts of word usage cues to the enhanced query model. To mitigate these two problems, the regularized simple mixture model has been proposed and can be estimated by maximizing the likelihood function (Tao and Zhai, 2006; Dillon and Collins-Thompson, 2010)

$$L = \prod_{w \in V} P_{\mathrm{RSMM}}(w|Q)^{\mu \cdot P(w|Q)} \times \prod_{D_r \in \mathbf{D}_{Top}} \prod_{w \in V} \left(\alpha_{D_r} \cdot P_{\mathrm{RSMM}}(w|Q) + (1-\alpha_{D_r}) \cdot P(w|BG)\right)^{c(w,D_r)}, \quad (4)$$

where μ is a weighting factor indicating the confidence on the prior information.

3 The Proposed Modeling Framework

3.1 Fundamentals

It is obvious that the major difference among the

representative query models mentioned above is how to capitalize on the set of top-ranked documents and the original query. Several subtle relationships can be deduced through the following in-depth analysis. First, a direct inspiration of the LM-based query reformulation framework can be drawn from the celebrated Rocchio's formulation, while the former can be viewed as a probabilistic counterpart of the latter (Robertson, 1990; Ponte and Croft, 1998; Baeza-Yates and Ribeiro-Neto, 2011). Second, after some mathematical manipulation, the formulation of the RM model (*c.f.* Eq. (2)) can be rewritten as

$$P_{\mathrm{RM}}(w|Q) = \sum_{D_r \in \mathbf{D}_{Top}} P(w|D_r) \frac{P(Q|D_r)P(D_r)}{\sum_{D_r'' \in \mathbf{D}_{Top}} P(Q|D_r'')P(D_r'')}. \quad (5)$$

It becomes evident that the RM model is composed by mixing a set of document models $P(w|D_r)$. As such, the RM model bears a close resemblance to the Rocchio's formulation. Furthermore, based on Eq. (5), we can recast the estimation of the RM model as an optimization problem, and the likelihood (or objective) function is formulated as

$$L = \prod_{w \in V} \left(\sum_{D_r \in \mathbf{D}_{Top}} P(w|D_r)P(D_r|Q) \right)^{c(w,Q)}, \quad (6)$$
$$s.t. \sum_{D_r \in \mathbf{D}_{Top}} P(D_r|Q) = 1$$

where the document models $P(w|D_r)$ are known in advance; the conditional probability $P(D_r|Q)$ of each document D_r is unknown and leave to be estimated. Finally, a principled framework can be obtained to unify all of these query models, including RM (*c.f.* Eq. (6)), SMM (c.f. Eq. (3)) and RSMM (*c.f.* Eq. (4))), by using a generalized objective likelihood function:

$$L = \prod_{w \in V} \prod_{E_i \in \mathbf{E}} \left(\sum_{M_r \in \mathbf{M}} P(w|M_r)P(M_r) \right)^{c(w,E_i)}, \quad (7)$$
$$s.t. \sum_{M_r \in \mathbf{M}} P(M_r) = 1$$

where **E** represents a set of observations which we want to maximize their likelihood, and **M** denotes a set of mixture components.

3.2 Query-specific Mixture Modeling

The SMM model and the RSMM model are intended to extract useful word usage cues from $\mathbf{D}_{Top}$, which are not only relevant to the original query Q but also external to those already captured by the generic background model. However, we argue in this paper that the "generic information" should be carefully crafted for each query due mainly to the fact that users' information needs may be very diverse from one another. To crystallize the idea, a query-specific background model $P_Q(w|\mathrm{BG})$ for each query Q can be derived from $\mathbf{D}_{Top}$ directly. Another consideration is that since the original query model $P(w|Q)$ cannot be accurately estimated, it thus may not necessarily be the best choice for use in defining a conjugate Dirichlet prior for the enhanced query model to be estimated. We propose to use the RM model as a prior to guide the estimation of the enhanced query model. The enhanced query model is termed query-specific mixture model (QMM), and its corresponding training objective function can be expressed as

$$L = \prod_{w \in V} P_{\mathrm{QMM}}(w|Q)^{\mu \cdot P_{\mathrm{RM}}(w|Q)} \times \prod_{D_r \in \mathbf{D}_{Top}} \prod_{w \in V} \left(\alpha_{D_r} \cdot P_{\mathrm{QMM}}(w|Q) + (1-\alpha_{D_r}) \cdot P_Q(w|BG) \right)^{c(w,D_r)}. \quad (8)$$

4 Applications

4.1 Speech Recognition

Language modeling is a critical and integral component in any large vocabulary continuous speech recognition (LVCSR) system (Huang *et al.*, 2001; Jurafsky and Martin, 2008; Furui *et al.*, 2012). More concretely, the role of language modeling in LVCSR 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. Once the various aforementioned query modeling methods are applied to speech recognition, for a search history H, we can conceptually regard it as a query and each of its immediately succeeding words w as a (single-word) document. Then, we may leverage an IR procedure that takes H as a query and poses it to a retrieval system to obtain a set of top-ranked documents from a contemporaneous (or in-domain) corpus. Finally, the enhanced query model (that is $P(w|H)$ in speech recognition) can be estimated by RM, SMM, RSMM or QMM, and further combined with the background n-gram (e.g., trigram) language model to form an adaptive language model to guide the speech recognition process.

4.2 Speech Summarization

On the other hand, extractive speech summarization aims at producing a concise summary by selecting salient sentences or paragraphs from the original spoken document according to a predefined target summarization ratio (Carbonell and Goldstein, 1998; Mani and Maybury, 1999; Nenkova and McKeown, 2011; Liu and Hakkani-Tur, 2011). Intuitively, this task could be framed as an ad-hoc IR problem, where the spoken document is treated as an information need and each sentence of the document is regarded as a candidate information unit to be retrieved according to its relevance to the information need. Therefore, KLM can be used to quantify how close the document D and one of its sentences S are: the closer the sentence model $P(w|S)$ to the document model $P(w|D)$, the more

likely the sentence would be part of the summary. Due to that each sentence S of a spoken document D to be summarized usually consists of only a few words, the corresponding sentence model $P(w|S)$ might not be appropriately estimated by the ML estimation. To alleviate the deficiency, we can leverage the merit of the above query modeling techniques to estimate an accurate sentence model for each sentence to enhance the summarization performance.

5 Experimental Setup

The speech corpus consists of about 196 hours of Mandarin broadcast news collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 (Wang *et al.*, 2005), which is publicly available and has been segmented into separate stories and transcribed manually. Each story contains the speech of one studio anchor, as well as several field reporters and interviewees. A subset of 25-hour speech data compiled during November 2001 to December 2002 was used to bootstrap the acoustic model training. The vocabulary size is about 72 thousand words. The background language model was estimated from a background text corpus consisting of 170 million Chinese characters collected from the Chinese Gigaword Corpus released by LDC.

The dataset for use in the speech recognition experiments is compiled by a subset of 3-hour speech data from the corpus within 2003 (1.5 hours for development and 1.5 hours for test). The contemporaneous (in-domain) text corpus used for training the various LM adaptation methods was collected between 2001 and 2003 from the corpus (excluding the test set), which consists of one million Chinese characters of the orthographic broadcast news transcripts. In this paper, all the LM adaptation experiments were performed in word graph rescoring. The associated word graphs of the speech data were built beforehand with a typical LVCSR system (Ortmanns *et al.*, 1997; Young *et al.*, 2006).

In addition, the summarization task also employs the same broadcast news corpus as well. A subset of 205 broadcast news documents compiled between November 2001 and August 2002 was reserved for the summarization experiments (185 for development and 20 for test). A subset of about 100,000 text news documents, compiled during the same period as the documents to be summarized, was employed to estimate the related summarization models compared in this paper. We adopted three variants of the widely-used ROUGE metric (i.e., ROUGE-1, ROGUE-2 and ROUGE-L) for the assessment of summarization performance (Lin, 2003). The summarization ratio, defined as the ratio of the number of words in the automatic (or manual) summary to that in the reference transcript of a spoken document, was set to 10% in this research.

6 Experimental Results

In the first part of experiments, we evaluate the effectiveness of the various query models applied to the speech recognition task. The corresponding results with respect to different numbers of top-ranked documents being used for estimating their component models are shown in Table 1. Also worth mentioning is that the baseline system with the background trigram language model, which was trained with the SRILM toolkit (Stolcke, 2005) and Good-Turing smoothing (Jelinek, 1999), results in a Chinese character error rate (CER) of 20.08% on the test set. Consulting Table 1 we notice two particularities. One is that there is more fluctuation in the CER results of SMM than in those of RM. The reason might be that, for SMM, the extraction of relevance information from the top-ranked documents is conducted with no involvement of the test utterance (i.e., the query; or its corresponding search histories), as elaborated earlier in Section 2. When too many feedback documents are being used, there would be a concern for SMM to be distracted from being able to appropriate model the test utterance, which is probably caused by some dominant distracting (or irrelevant) feedback documents. The other interesting observation is that RSMM only achieves a comparable (even worse) result when compared to SMM. A possible reason is that the prior constraint of the RSMM may contain too much noisy information so as to bias the model estimation. Furthermore, it is evident that the proposed QMM is the best-performing method among all the query models compared in the paper. Although the improvements made by QMM are not as pronounced as expected, we believe that QMM has demonstrated its potential to be applied to other related applications. On the other hand, we compare the various query models with two well-practiced language models, namely the cache model (Cache) (Kuhn and Mori, 1990; Jelinek *et al.*, 1991) and the latent Dirichlet allocation (LDA) (Liu and Liu, 2007; Tam and Schultz, 2005). The CER results of these two models are also shown in Table 1, respectively. For the cache model, bigram cache was used since it can yield better results than the unigram and trigram cache models in our experiments. It is worthy to notice that the LDA model was trained with the entire set of contemporaneous text document collection (*c.f.* Section 4), while all of the query models explored in the paper were estimated based on a subset of the corpus selected by an initial round of retrieval. The results reveal that most of these query models can achieve superior performance over the two conventional language models.

In the second part of experiments, we evaluate the utilities of the various query models as applied to the speech summarization task. At the outset, we assess the performance level of the baseline KLM method by comparison with two well-practiced unsupervised methods, viz. the vector space model (VSM) (Gong and Liu, 2001), and its extension, maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998). The corresponding results are shown in Table 2 and can be aligned with several related literature reviews. By looking at the results, we find that KLM outperforms VSM by a large margin, confirming the applicability of the language modeling framework for speech summarization. Furthermore, MMR that presents an extension of VSM performs on par with KLM for the text summarization task (TD) and exhibits superior performance over KLM for the speech summarization task (SD). We now turn to evaluate the effectiveness of the various query models (viz. RM, SMM, RSMM and QMM) in conjunction with the pseudo-relevance feedback process for enhancing the sentence model involved in the KLM method. The corresponding results are also shown in Table 2. Two noteworthy observations can be drawn from Table 2. One is that all these query models can considerably improve the summarization performance of the KLM method, which corroborates the advantage of using them for enhanced sentence representations. The other is that QMM is the best-performing one among all the formulations studied in this paper for both the TD and SD cases.

Going one step further, we explore to use extra prosodic features that are deemed complementary to the LM cue provided by QMM for speech summarization. To this end, a support vector machine (SVM) based summarization model is trained to integrate a set of 28 commonly-used prosodic features (Liu and Hakkani-Tur, 2011) for representing each spoken sentence, since SVM is arguably one of the state-of-the-art supervised methods that can make use of a diversity of indicative features for text or speech summarization (Xie and Liu, 2010; Chen *et al.*, 2013). The sentence ranking scores derived by QMM and SVM are in turn integrated through a simple log-linear combination. The corresponding results are shown in Table 2, demonstrating consistent improvements with respect to all the three variants of the ROUGE metric as compared to that using either QMM or SVM in isolation. We also investigate using SVM to additionally integrate a richer set of lexical and relevance features to complement QMM and further enhance the summarization effectiveness. However, due to space limitation, we omit the details here. As a side note, there is a sizable gap between the TD and SD cases, indicating room for further improvements. We may seek remedies, such as robust indexing schemes, to compensate for imperfect speech recognition.

Table 1. The speech recognition results (in CER (%)) achieved by various language models along with different numbers of latent topics/pseudo-relevance feedback documents.

	16	32	64	128
Baseline	20.08			
Cache	19.86			
LDA	19.29	19.30	19.28	19.15
RM	19.26	19.26	19.26	19.26
SMM	19.19	19.00	19.14	19.10
RSMM	19.18	19.14	19.15	19.19
QMM	**19.05**	**18.97**	**19.00**	**18.99**

Table 2. The summarization results (in F-scores) achieved by various language models along with text and spoken documents.

	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
VSM	0.347	0.228	0.290	0.342	0.189	0.287
MMR	0.407	0.294	0.358	0.381	0.226	0.331
KLM	0.411	0.298	0.361	0.364	0.210	0.307
RM	0.453	0.335	0.403	0.382	0.239	0.331
SMM	0.439	0.320	0.388	0.383	0.229	0.327
RSMM	0.472	0.365	0.423	0.381	0.235	0.329
QMM	**0.486**	**0.382**	**0.435**	**0.395**	**0.256**	**0.349**
SVM	0.441	0.334	0.396	0.370	0.222	0.326
QMM+ SVM	**0.492**	**0.395**	**0.448**	**0.398**	**0.261**	**0.358**

7 Conclusion and Outlook

In this paper, we have presented a systematic and thorough analysis of a few well-practiced query models for IR and extended their novel applicability to speech recognition and summarization in a principled way. Furthermore, we have proposed an extension of this research line by introducing query-specific mixture modeling; the utilities of the deduced model have been extensively compared with several existing query models. 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 so as to improve the empirical effectiveness of such query modeling.

Acknowledgements

This research is supported in part by the "Aim for the Top University Project" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, and by the Ministry of Science and Technology, Taiwan, under Grants MOST 103-2221-E-003-016-MY2, 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 103-2911-I-003-301.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. Modern information retrieval: the concepts and technology behind search, ACM Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, pp.993–1022.

David M. Blei and John Lafferty. 2009. Topic models. In A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversitybased reranking for reordering documents and producing summaries. In *Proc. SIGIR*, pp. 335–336.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, vol. 44, pp.1–56.

Stephane Clinchant and Eric Gaussier. 2013. A theoretical analysis of pseudo-relevance feedback models. In *Proc. ICTIR.*

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. SIGIR*, pp. 243–250.

Berlin Chen, Shih-Hsiang Lin, Yu-Mei Chang, and Jia-Wen Liu. 2013. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing & Managemen*t, 49(1), pp. 1cess

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39(1), pp. 1–38.

Joshua V. Dillon and Kevyn Collins-Thompson. 2010. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *Proc. CIKM*, pp. 1069–1078.

Sadaoki Furui, Li Deng, Mark Gales, Hermann Ney, and Keiichi Tokuda. 2012. Fundamental technologies in modern speech recognition. *IEEE Signal Processing Magazine*, 29(6), pp. 16–17.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. SIGIR*, pp. 19–25.

Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. 2004. Parsimonious language models for information retrieval. In *Proc. SIGIR*, pp. 178–185.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. SIGIR*, pp. 50–57.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, pp. 177–196.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. 2001. Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Frederick Jelinek, Bernard Merialdo, Salim Roukos, and M. Strauss. 1991. A dynamic language model for speech recognition. In *Proc. the DARPA workshop on speech and natural language*, pp. 293–295.

Frederick Jelinek. 1999. Statistical methods for speech recognition. MIT Press.

Daniel Jurafsky and James H. Martin. 2008. Speech and language processing. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Roland Kuhn and Renato D. Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6), pp. 570–583.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), pp. 79–86.

Chin-Yew Lin. 2003. ROUGE: Recall-oriented Understudy for Gisting Evaluation. Available: http://haydn.isi.edu/ROUGE/.

Feifan Liu and Yang Liu. 2007. Unsupervised language model adaptation incorporating named entity information. In *Proc. ACL*, pp. 672–769.

Yang Liu and Dilek Hakkani-Tur. 2011. Speech summarization. *Chapter 13* in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, G. Tur and R. D. Mori (Eds), New York: Wiley.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proc. SIGIR*, pp. 111–119.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proc. SIGIR*, pp. 120–127.

Victor Lavrenko. 2004. A Generative Theory of Relevance. PhD thesis, University of Massachusetts, Amherst.

Shasha Xie and Yang Liu. 2010. Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech & Language*, 24(3), pp. 495–514.

Yuanhua Lv and Chengxiang Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. In *Proc. CIKM*, pp. 1895–1898.

Yuanhua Lv and Chengxiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proc. SIGIR*, pp. 579–586.

Kyung Soon Lee, W. Bruce Croft, and James Allan. 2008. A cluster-based resampling method for pseudo-relevance feedback. In *Proc. SIGIR*, pp. 235–242.

Kyung Soon Lee and W. Bruce Croft. 2013. A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback. *Inf. Process. Manage. 49(4)*, pp. 792–806.

Inderjeet Mani and Mark T. Maybury (Eds.). 1999. Advances in automatic text summarization. Cambridge, MA: MIT Press.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3), pp. 103–233.

Stefan Ortmanns, Hermann Ney, and Xavier Aubert. 1997. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language*, pp. 43–72.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. SIGIR*, pp. 275–281.

Stephen E. Robertson. 1990. On term selection for query expansion. *Journal of Documentation*, 46(4), pp. 359–364.

Andreas Stolcke. 2005. SRILM - An extensible language modeling toolkit. In *Proc. INTERSPEECH*, pp.901–904.

Tao Tao and Chengxiang Zhai. 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proc. SIGIR*, pp. 162–169.

Yik-Cheung Tam and Tanja Schultz. 2005. Dynamic language model adaptation using variational Bayes inference. In *Proc. INTERSPEECH*, pp. 5–8.

Xuanhui Wang, Hui Fang, and Chengxiang Zhai. 2008. A study of methods for negative relevance feedback. In *Proc. SIGIR*, pp. 219–226.

Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng. 2005. MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, 10(2), pp. 219–236.

Xing Yi and James Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *Proc. ECIR*, pp. 29–41.

Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. 2006. The HTK book version 3.4. Cambridge University Press.

Chengxiang Zhai and John Lafferty. 2001[a]. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pp. 334–342.

Chengxiang Zhai and John Lafferty. 2001[b]. Model-based feedback in the language modeling approach to information retrieval. In *Proc. CIKM*, pp. 403–410.

Chengxiang Zhai. 2008. Statistical language models for information retrieval: a critical review. Foundations and Trends in Information Retrieval, 2 (3), pp. 137–213.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proc. SIGIR*, pp. 81–88.