INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Exploring the Use of Significant Words Language Modeling for Spoken Document Retrieval

Ying-Wen Chen[1], *Kuan-Yu Chen*[2], *Hsin-Min Wang*[2], *Berlin Chen*[1]

[1]National Taiwan Normal University, Taiwan
[2]Academia Sinica, Taiwan

{cliffchen, berlin}@ntnu.edu.tw, {kychen, whm}@iis.sinica.edu.tw

Abstract

Owing to the rapid global access to tremendous amounts of multimedia associated with speech information on the Internet, spoken document retrieval (SDR) has become an emerging application recently. Apart from much effort devoted to developing robust indexing and modeling techniques for spoken documents, a recent line of research targets at enriching and reformulating query representations in an attempt to enhance retrieval effectiveness. In practice, pseudo-relevance feedback is by far the most prevalent paradigm for query reformulation, which assumes that top-ranked feedback documents obtained from the initial round of retrieval are potentially relevant and can be exploited to reformulate the original query. Continuing this line of research, the paper presents a novel modeling framework, which aims at discovering significant words occurring in the feedback documents, to infer an enhanced query language model for SDR. Formally, the proposed framework targets at extracting the essential words representing a common notion of relevance (i.e., the significant words which occur in almost all of the feedback documents), so as to deduce a new query language model that captures these significant words and meanwhile modulates the influence of both highly frequent words and too specific words. Experiments conducted on a benchmark SDR task demonstrate the performance merits of our proposed framework.

Index Terms: Query Model, Significant Words, Pseudo Relevance Feedback

1. Introduction

Along with the rapid proliferation of multimedia associated with spoken content, spoken document retrieval (SDR) has become a pivotal application with cross-fertilization of ideas between the speech and natural language communities [1-4]. A great amount of research effort has been devoted to developing robust indexing techniques to extract probable spoken terms or phrases inherent in a spoken document that could match the query words or phrases literally [5-7]. In addition, several effective retrieval models originally put forward in a wide variety of information retrieval (IR) tasks, such as the vector space model (VSM) [8, 9], the Okapi BM25 model [10], the representation learning methods [11] and among others, have been applied with good success to SDR. Recently, an emerging stream of thought is to employ a statistical language model (LM) for IR and SDR, which has become an attractive choice due to its simplicity and clear probabilistic meaning, as well as state-of-the-art performance [12-14]. In practice, each text or spoken document is framed as a generative model composed of a mixture of multinomial (or *n*-gram) distributions for observing a query, while the query is regarded as observations, expressed by a sequence of words. Accordingly, documents can be ranked according to their likelihoods of generating the query, viz. the query-likelihood measure (QLM). Another popular formulation is the Kullback-Leibler divergence measure (KLM) [15], where both the query and the documents are modeled by a unigram language model, respectively. The relevance degree between a pair of query and document is recast as the divergence distance between the two respective unigram models. It is easy to show that KLM can reduce to QLM when the query language model is simply estimated based on the empirical query word distribution and the maximum likelihood (ML) estimator.

One critical issue in IR and SDR is that the input text or spoken query is usually too short to address the information need of a user. In order to mitigate the problem, pseudo-relevance feedback has become a promising strategy to enrich the query statistics so as to boost the retrieval performance [8, 16, 17]. Following the line of research, query reformation methods can be broadly grouped into two distinct classes. One is to leverage external resources, such as the Wikipedia or the WordNet, to expand and reorganize the original input query. The other is to reformulate the user query by referring to a small set of top-ranked feedback documents locally collected from an initial round of retrieval (i.e., the pseudo-relevance feedback process). Since the former requires more sophisticated natural language processing techniques such as semantic representation and understanding, much more effort has been put into launching query reformulation methods with automatic feedback documents [18]. The relevance model (RM) [16, 17] and the simple mixture model (SMM) [15, 17] are two well-practiced representatives.

In addition to the existing query reformulation methods, a novel framework of significant words language modeling is explored in this paper. The core idea, which is inspired from the Luhn's theory [19, 20], is that the top-ranked feedback documents is presumably composed of three components: the general background information, the specific information and the relevance information. Specifically, the proposed framework targets at extracting the *significant words* which occur in almost all of these feedback documents, meanwhile diminishing the influence of both generally common words and too specific words. By doing so, the inferred language model can convey only the relevance information and effectively supplement the original query. To recap, the main contribution of this paper is two-fold. On one hand, we explore to leverage a significant words language modeling framework to enhance the query language model involved in the LM-based SDR. On the other hand, the utilities of the proposed framework and several state-of-the-art methods are analyzed and compared thoroughly.

http://dx.doi.org/10.21437/Interspeech.2017-612

The remainder of this paper is structured as follows. We briefly review the mathematical formulations of the classic query language models in Section 2. In Section 3, we describe the proposed framework that seeks for an accurate query language model by excluding the generally common words and eliminating the words reoccurring concentratedly in only a few feedback documents. After that, the experimental settings and a series of retrieval experiments are presented in Sections 4, respectively. Finally, Section 5 concludes our presentation and discusses avenues for future work.

2. Classic Query Language Models

Due to the fact that a query usually consists of only a few words, the true query model $P(w|Q)$ of a query Q for predicting an arbitrary word w might not be accurately estimated by the simple ML estimator. With the alleviation of this deficiency as motivation, there are several studies devoted to achieve more accurate query modeling, saying that this can be approached with a pseudo-relevance feedback process. Such integration seems to hold promise for query reformulation. However, its success depends largely on the assumption that a set of top-ranked feedback documents, $\mathbf{D}_F = \{D_1, \cdots, D_{|\mathbf{D}_F|}\}$, obtained from an initial round of retrieval, are relevant and can be used to estimate a more accurate query language model. Representative methods developed so far include the relevance model and the simple mixture model, just to name a few.

2.1. Relevance Model (RM)

Under the notion of relevance model (RM) [16], 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 the query are samples drawn from the same relevance class R_Q. However, in reality, since there is no prior knowledge about R_Q, we may use the top-ranked feedback documents $\mathbf{D}_F$ to approximate R_Q. The corresponding relevance model, on the grounds of a multinomial view of R_Q, thus can be estimated using the following equation:

$$P_{\mathrm{RM}}(w|Q) = \frac{\sum_{D\in\mathbf{D}_F} P(D)P(w|D)\prod_{w'\in Q} P(w'|D)^{c(w',Q)}}{\sum_{D'\in\mathbf{D}_F} P(D')\prod_{w'\in Q} P(w'|D')^{c(w',Q)}}, \quad (1)$$

where $c(w', Q)$ is the occurrence count of a word w' in Q. The prior probability $P(D)$ of each document can be simply kept uniform, while the document models $P(w|D)$ are estimated on the basis of the occurrence counts of w in each document D with the ML estimator, respectively.

2.2. Simple Mixture Model (SMM)

Another perspective of estimating an enhanced query model with the feedback documents is the simple mixture model (SMM) [15], which assumes that words in $\mathbf{D}_F$ 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 general background language model $P_{\mathrm{BG}}(w)$. By doing so, the SMM model $P_{\mathrm{SMM}}(w|Q)$ can be estimated by maximizing the likelihood over all the feedback documents:

$$L = \prod_{D\in\mathbf{D}_F}\prod_{w\in V}[\alpha\cdot P_{\mathrm{SMM}}(w|Q) + (1-\alpha)\cdot P_{\mathrm{BG}}(w)]^{c(w,D)}, \quad (2)$$

where α is a pre-defined weighting parameter used to control the degree of reliance between $P_{\mathrm{SMM}}(w|Q)$ and $P_{\mathrm{BG}}(w)$. This estimation will enable more specific words (i.e., words in $\mathbf{D}_F$ that are not well-explained by the general background language model) to receive more probability mass, thereby leading to a more discriminative query model $P_{\mathrm{SMM}}(w|Q)$. Simply put, the SMM model is anticipated to extract useful word usage cues from $\mathbf{D}_F$, which are not only probably relevant to the query Q, but also external to those already well captured by the general background language model.

3. Significant Words Language Modeling

Inspired from the Luhn's theory [19], we explore a significant words language modeling (SWLM) framework to estimate an accurate query language model by parsimonizing the estimation toward not only the generally common words, but also the too specific words [20]. More formally, we seek for a SWLM model that is "significant" enough to distinguish the feedback documents from others by removing frequent words in the background, and meanwhile "general" enough to aggregate the shared characteristics of the feedback documents that underlie the notion of relevance by ruling out too specific words. To crystalize the idea, we assume words in the feedback documents are samples drawn from a three-component mixture model: the general background language model $P_{\mathrm{BG}}(w)$, the specific language model $P_{\mathrm{S}}(w)$ and the desired SWLM model $P_{\mathrm{SW}}(w)$. As such, the probability of a word occurring in a feedback document D can be defined by:

$$P(w|D) = \alpha\cdot P_{\mathrm{BG}}(w) + \beta\cdot P_{\mathrm{S}}(w) + (1-\alpha-\beta)\cdot P_{\mathrm{SW}}(w), \quad (3)$$

where α and β are empirical parameters used to modulate the contributions between $P_{\mathrm{BG}}(w)$, $P_{\mathrm{S}}(w)$ and $P_{\mathrm{SW}}(w)$.

In practice, the general background language model is employed to represent the frequent words in general, which can be estimated from a large collection of documents beforehand. Consequently, in order to infer an accurate query language model (i.e., SWLM $P_{\mathrm{SW}}(w)$), this study proposes three modeling mechanisms for estimating a reasonable and suitable specific language model $P_{\mathrm{S}}(w)$.

3.1. Inverse Document Frequency-based Method

Since the specific language model is used to indicate the words that reoccur concentratedly in only few feedback documents, a straightforward and intuitive mechanism is to leverage the inverse document frequency (IDF) calculated from the set of feedback documents [9]:

$$IDF(w) = \log\left(\frac{|\mathbf{D}_F|}{\varepsilon_{IDF}+|\{D\in\mathbf{D}_F: w\in D\}|}\right), \quad (4)$$

where $|\mathbf{D}_F|$ denotes the number of feedback documents, $|\{D\in\mathbf{D}_F: w\in D\}|$ denotes the number of documents that contain the word w, and ε_{IDF} is a constant used to avoid division-by-zero. By doing so, the higher the inverse document frequency of a word w, the more probability mass should be given to. The IDF statistics of words thereby can be obtained from the feedback documents and in turn be used to infer a specific language model $P_{\mathrm{S}}(w)$. To be concrete, the specific language model is computed by:

$$P_{\mathrm{S}}(w) = \frac{IDF(w)}{\sum_{w'\in V} IDF(w')}. \quad (5)$$

where V denotes the vocabulary. This method is hereafter denoted as the inverse document frequency-based (IDF-based) method.

Although the IDF-based method proposes a systematic and theoretical way to derive the specific language model, a weakness is that the method treats all of the feedback documents equally important. In order to remedy the possible shortcoming, a natural extension, named the weighted inverse document frequency-based (wIDF-based) method, is proposed.

More formally, we integrate the normalized similarity degree between a query and a document into the estimation of the inverse document frequency:

$$wIDF(w) = \log\left(\frac{\sum_{D\in\mathbf{D}_F} sim(Q,D)}{\varepsilon_{wIDF}+\sum_{D'\in\mathbf{D}_F \& w\in D'} sim(Q,D')}\right), \quad (6)$$

where $sim(Q,D)$ denotes the normalized similarity degree between Q and D, and ε_{wIDF} is a tunable constant used to avoid division-by-zero. After that, the specific language model can be obtained by normalizing the weighted inverse document frequency of all words (i.e., $wIDF(w)$). By doing so, the relevance degree of a feedback document can be taken into account, thereby contrasting the burstiness of words occurring in a small portion of less relevant feedback documents.

3.2. Inverse Entropy-based Method

Instead of the intuitive mechanisms proposed above (i.e., the IDF-based and wIDF-based methods), we manage to develop a more mathematical way to estimate the specific language model. Based on the original objective, i.e., a specific model is formulated to capture those words that reoccur concentratedly in only a few feedback documents, we first define a distribution over the top-ranked feedback documents for an arbitrary word w:

$$P(D|w) = \frac{P(D,w)}{\sum_{D'\in\mathbf{D}_F} P(D',w)} = \frac{P(w|D)P(D)}{\sum_{D'\in\mathbf{D}_F} P(w|D')P(D')}. \quad (7)$$

where $P(D)$ is the prior probability of document D; we leverage the normalized similarity degree between Q and D to approximate this measure here. Accordingly, an inverse entropy score can be calculated for a word w to quantify its occurrence behaviour in the top-ranked feedback documents:

$$IE(w) = \frac{1}{\varepsilon_{IE}-\sum_{D\in\mathbf{D}_F} P(D|w)\log P(D|w)}, \quad (8)$$

where ε_{IE} is a tunable constant used to avoid division-by-zero. Finally, the specific language model can be obtained by normalizing the inverse entropy score of an arbitrary word with respect to that of all distinct words in the vocabulary:

$$P_{\mathrm{S}}(w) = \frac{IE(w)}{\sum_{w'\in V} IE(w')}. \quad (9)$$

We term this mechanism as the inverse entropy-based (IE-based) method. Such a method capitalizes on the entropy statistics to indicate the specific words occurring in only few feedback documents and may not bear relevance to the query. Furthermore, the IE-based method also takes into account the document important degree naturally. It thus can emphasize the burstiness of words in relatively few feedback documents.

3.3. Mutual Exclusion-based Method

Inspired from the research on search result diversification in the context of information retrieval [20, 21], we device a novel mechanism to estimate the specific language model. We first define the specific words as those being supported by part of the feedback documents but not all. Thus, the specific language model can be derived by:

$$P_{\mathrm{S}}(w) \propto \sum_{D\in\mathbf{D}_F} P(w|D) \prod_{D'\in\mathbf{D}_F, D'\neq D} \big(1 - P(w|D')\big). \quad (10)$$

Obviously, the former term (i.e., $P(w|D)$) in (10) is used to characterize the occurrence degree of a word w with respect to a feedback document D of interest, while the latter (i.e., $\prod_{D'\in\mathbf{D}_F \& D'\neq D}\big(1 - P(w|D')\big)$) instead provides a measure to determine the non-occurrence degree of the word w in the other feedback documents. By doing so, a word that occurs only in few feedback documents tends to have a higher probability and is considered to be a specific word. Actually, this is similar to the notion of the "mutual exclusion" [23]; we thus refer to this mechanism as the mutual exclusion-based (ME-based) method henceforth.

3.4. The Significant Words Language Model

Based on the inferred specific language model and the assumption, i.e., words in a feedback document are treated as sample instances drawn from three models, the SWLM model $P_{\mathrm{SW}}(w)$ can be estimated by maximizing the likelihood of all the feedback documents:

$$L = \prod_{D\in\boldsymbol{D}_F} \prod_{w\in V} \left[\begin{array}{c} \alpha\cdot P_{\mathrm{BG}}(w) + \beta\cdot P_{\mathrm{S}}(w) \\ +(1-\alpha-\beta)\cdot P_{\mathrm{SW}}(w) \end{array} \right]^{c(w,D)}. \quad (11)$$

where α and β are interpolation factors. Here both the general background language model (estimated on a large collection) and the specific language model (estimated with one of the proposed methods elucidated above) are kept fixed, while the interpolation factors are empirically set. The significant words language model $P_{\mathrm{SW}}(w)$ therefore can be inferred with the Expectation-Maximization (EM) algorithm:

E-step:

$$\sigma = \frac{(1-\alpha-\beta)\cdot P_{\mathrm{SW}}(w)}{\alpha\cdot P_{\mathrm{BG}}(w)+\beta\cdot P_{\mathrm{S}}(w)+(1-\alpha-\beta)\cdot P_{\mathrm{SW}}(w)}, \quad (12)$$

M-step:

$$P_{\mathrm{SW}}(w) = \frac{\sum_{D\in\boldsymbol{D}_F} c(w,D)\sigma}{\sum_{w'\in V}\sum_{D'\in\boldsymbol{D}_F} c(w,D')\sigma}. \quad (13)$$

The notion of leveraging Luhn's theory for estimating an enhanced query language model has been applied with some success to an IR-related task recently [20]. However, as far as we are aware, the notion of SWLM and the development of the associated component models have never been thoroughly explored for SDR.

3.5. The Retrieval Model

In the retrieval phase, each query Q will have its own enhanced query language model (i.e., SWLM) based on one of the estimators for the specific language model. As such, the KLM method is subsequently employed to distinguish relevant documents from irrelevant ones.

4. Experiments

4.1. Experimental Setup

We employed the Topic Detection and Tracking collection (TDT-2) for our experiments [24]. The Mandarin news stories from Voice of America news broadcasts were taken as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. The average word error rate (WER) obtained for the spoken documents is about 35% [25]. The title of Chinese news stories from Xinhua News Agency were used as the test queries. The retrieval performance is evaluated with the commonly-used non-interpolated mean average precision (MAP) [26] metric.

4.2. Experimental Results

To begin with, we compare several well-practiced and/or state-of-the-art IR models for SDR, including the vector space-based methods and the language model-based methods. The results are summarized in Table 1. The best result within each column (corresponding to a specific evaluation condition) is type-set boldface. For the vector space-based methods, including the

Table 1. *Retrieval results (in MAP) achieved by various state-of-the-art baseline models*.

	TD	SD
VSM	0.339	0.275
DM	0.344	0.302
DBOW	**0.362**	**0.345**
KLM	0.368	0.317
LDA	0.401	0.341
RM	0.402	0.364
SMM	0.420	**0.395**
TRM	**0.442**	0.394

Table 2. *Retrieval results (in MAP) achieved by the proposed framework.*

	TD	SD
IDF	**0.478**	0.427
wIDF	0.473	**0.443**
IE	**0.478**	0.425
ME	0.368	0.359

vector space model (VSM) [9], the distributed memory model (DM) [7-11] and the distributed bag-of-words model (DBOW) [7-11], both query and documents are represented by vectors, while the relevance degree is computed by the cosine similarity measure. In contrast to the vector space-based methods, KLM is a language model-based retrieval system, the query and document language models of which are derived by the maximum likelihood estimator. LDA denotes latent Dirichlet allocation, in which each document language model is estimated by leveraging a probabilistic topic modeling paradigm [27]. In addition, two well-practiced query language models, namely the relevance model (RM) and the simple mixture model (SMM), are also compared here. Furthermore, the results of a recent extension of the RM model, i.e., the topic-based relevance model (TRM) [4], are listed for reference as well. Also noteworthy is that RM, SMM and TRM are employed to reformulate the original query language model by pairing with the pseudo-relevance feedback, while the document language model is derived by the maximum likelihood estimator as in the KLM system. Several observations can be drawn from Table 1. First, the language model-based methods in general outperform the vector space-based methods. Such results evidence that the language model-based methods indeed represent a school of efficient and effective mechanisms for SDR. Second, both the celebrated paragraph embedding methods (i.e., DM and DBOW) outperform VSM, and DBOW consistently outperforms DM by a large margin when being applied to either text documents (i.e., the TD case) or spoken ones (i.e., the SD case). Third, LDA outperforms KLM, while RM, SMM and TRM all outperform LDA in both cases. These results also indicate that deriving a more accurate query language model tends to be more effective than building enhanced document models. The reason might be that a document usually contains relatively sufficient word usage statistics to estimate a reliable language model, in contrast to a short query.

Next, we make a step forward to evaluate the proposed SWLM framework for SDR. The results are highlighted in Table 2. In order to compare the proposed framework with other models, the number of feedback documents is set to 15, unless stated otherwise. Inspection of the results reveals three noteworthy points. At first glance, the IDF-based, the wIDF-based and the IE-based methods have comparable performance

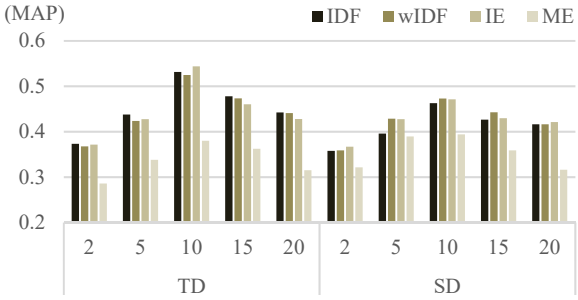

Figure 1: *Retrieval results (in MAP) for short queries with respect to the number of feedback documents.*

and they all outperform the ME-based method by a considerable margin in both the TD and SD cases. The reason might be that the criterion of the ME-based method is too rigorous to define the specific words and the derived specific language model thus can only capture a few specific words in the feedback documents. A detailed analysis of this phenomenon still awaits further investigation. Second, when compared with the state-of-the-art retrieval models (*cf.* Table 1), all of the proposed methods outperform both the vector space-based methods and the previous LM-based methods substantially in both the TD and SD cases, except that the ME-based method only outperforms the vector space-based methods. Third, Table 3 also signals that the wIDF-based method appears to be most robust against the recognition errors. To sum up, SWLM can offer effective query language models and further enhance the retrieval performance when pairing with the KLM measure.

In the last set of experiments, we look into the impact of the number of feedback documents on the various SWLM modeling mechanisms. As revealed by the results depicted in Figure 1, leveraging 10 feedback documents seems to benefit the performance for both the TD and SD cases the most. Nevertheless, the way to systemically determine the optimum number of feedback documents for each query reformulation method remains an open issue for further exploration. The results also notice that the naïve IDF-based method seems to be the best choice for the TD case, while the wIDF-based method is the most robust one to the recognition errors.

5. Conclusion and Future work

In this paper, we have presented a novel significant words language modeling framework that can be exploited to infer an enhanced query language model for a given query. All of the proposed variants have been thoroughly evaluated on a benchmark SDR corpus. Experimental results demonstrate the superiority of this framework in relation to other strong baselines compared in the paper, thereby indicating its good potential in query reformulation. For future work, we will explore the integration of extra cues, such as acoustic statistics and sub-word indexing strategies, into the proposed framework for the SDR task. We also plan to evaluate the framework on other large-scale IR corpora and NLP-related tasks like automatic summarization [28, 29].

6. 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 104-2911-I-003-301, MOST 104-2221-E-003-018-MY3 and MOST 105-2221-E-003-018-MY3.

7. References

[1] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 25(3), pp. 39-49, 2008.

[2] L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 22(5), pp. 42-60, 2005.

[3] C. L. Huang, B. Ma, H. Li, and C. H. Wu, "Speech indexing using semantic context inference," in *Proceedings of INTERSPEECH*, 2011.

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

[5] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Word co-occurrence matrix and context dependent class in LSA based language model for speech recognition", *International Journal of Computers*, pp. 85-95, 2009.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society of Information Science*, 41(6), pp. 391-407, 1990.

[7] K. Y. Chen, S. H. Liu, B. Chen, and H. M. Wang, "A locality-preserving essence vector modeling framework for spoken document retrieval," in *Proceedings of ICASSP*, 2017.

[8] C. D. Manning, P. Raghavan and H. Schtze, *Introduction to Information Retrieval*, New York: Cambridge University Press, 2008.

[9] G. Salton , A. Wong , and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, 18(11), pp. 613-620, Nov. 1975

[10] K. S. Jones, S. Walker, and S. E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments (parts 1 and 2)," *Information Processing and Management*, 36(6), pp. 779-840, 2000.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of ICML*, 2014.

[12] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of SIGIR*, 1998.

[13] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of CIKM*, 1999.

[14] W. B. Croft and J. Lafferty (eds.), "Language modeling for information retrieval," *Kluwer International Series on Information Retrieval*, Volume 13, Kluwer Academic Publishers, 2003.

[15] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of CIKM*, 2001.

[16] V. Lavrenko and W. Bruce Croft, "Relevance based language models," in *Proceedings of SIGIR*, 2001.

[17] K. Y. Chen, S. H. Liu, B. Chen, H. M. Wang, and H. H. Chen, "Exploring the use of unsupervised query modeling techniques for speech recognition and summarization," *Speech Communication*, Vol. 80, pp. 49-59, 2016.

[18] H. Zamani and W. B. Croft, "Embedding-based query language models," in *Proceedings of CIKM*, 2016.

[19] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development* 2.2, pp. 159-165, 1958.

[20] M. Dehghani, H. Azarbonyad, J. Kamps, D. Hiemstra, and M. Marx, "Luhn revisited: significant words language models," in *Proceedings of CIKM*, 2016.

[21] W. Zheng and H. Fang, "A comparative study of search result diversification methods," in *Proceedings of DDR*, 2011.

[22] R. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in *Proceedings of WWW*, 2010.

[23] W. C. Navidi, *Statistics for engineers and scientists*, McGraw-Hill (2 edition, 2007)

[24] LDC, "Project topic detection and tracking," *Linguistic Data Consortium*, 2000.

[25] H. Meng, S. Khudanpur, G. Levow, D. Oard, and H. M. Wang, "Mandarin–English information (MEI): investigating translingual speech retrieval," *Computer Speech and Language*, 18(2), pp. 163-179, 2004.

[26] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval: the concepts and technology behind search*, ACM Press, 2011.

[27] X. Wei and W. Bruce Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of SIGIR*, 2006.

[28] K. Y. Chen, S. H. Liu, B. Chen, H. M. Wang, E. E. Jan, W. L. Hsu, and H. H. Chen, "Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8), pp. 1322-1334, 2015.

[29] S. H. Liu, K. Y. Chen, B. Chen, H. M. Wang, H. C. Yen, and W. L. Hsu, "Combining relevance language modeling and clarity measure for extractive speech summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6), pp. 957-969, 2015.