Incorporating Proximity Information in Relevance Language Modeling for Extractive Speech Summarization

Shih-Hung Liu[*¥], Hung-Shih Lee[*], Hsiao-Tsung Hung[†], Kuan-Yu Chen[*],
Berlin Chen[†], Hsin-Min Wang[*], Hsu-Chun Yen[¥], Wen-Lian Hsu[*]
[*]Academia Sinica, Taiwan
E-mail: {journey, hslee, kychen, whm, hsu}@iis.sinica.edu.tw
[¥]National Taiwan University, Taiwan
E-mail: hcyen@ntu.edu.tw
[†]National Taiwan Normal University, Taiwan
E-mail: {alexhung, berlin}@csie.ntnu.edu.tw

***Abstract*—Extractive speech summarization refers to automatic selection of an indicative set of sentences from a spoken document so as to offer a concise digest covering the most salient aspects of the original document. The language modeling (LM) framework alongside the pseudo-relevance feedback (PRF) technique has emerged as a promising line of research for conducting extractive speech summarization in an unsupervised manner, showing some preliminary success. This paper extends such a general line of research and its main contributions are two-fold. First, we explore several effective formulations of proximity-based cues for use in the sentence modeling process involved in the LM-based summarization framework. Second, the utilities of the methods instantiated from the LM-based summarization framework and several well-practiced state-of-the-art methods are analyzed and compared extensively. The empirical results suggest the effectiveness of our methods.**

I. INTRODUCTION

Research on speech summarization has witnessed a booming interest in the speech processing community over the past decade [1]-[4]. This is largely attributed to the advances in automatic speech recognition (ASR) as well as the popularity and ubiquity of multimedia associated with spoken documents [5], [6]. As one predominant branch of this research area, extractive speech summarization manages to select indicative sentences from an original spoken document according to a predefined summarization ratio and concatenate them together to form a compact summary that can represent the major theme of the original document. Consequently, it is capable of providing all locations of important speech segments along with their corresponding transcripts for users to access and assimilate by listening or by reading, so as to save most of their time.

The methodology of extractive speech summarization may be coarsely clustered in two groups: ASR-based and non-ASR-based approaches [19]. The former approach conducts summarization with the automatic (imperfect) transcripts generated by an ASR system, and thereby can harness and extend the power of a broad range of methods well-practiced in text summarization to the context of speech summarization. On the contrary, the latter approach endeavors to estimate the importance of a spoken sentence and/or its relevance to the spoken document to be summarized directly based on the acoustic or prosodic features derived from the raw speech signal, without recourse to any ASR system for generating the corresponding automatic transcript. In general, the empirical performance of the latter is usually inferior to that of the former; nevertheless, how to systematically and effectively combine the strengths of both approaches for better performance in speech summarization awaits further study.

More recently, for the ASR-based approach, an emerging stream of research is to capitalize on the language modeling (LM) framework along with the pseudo-relevance feedback technique in an unsupervised manner [9]-[12], which has demonstrated as a promising avenue to extractive speech summarization. Our work in this paper presents a continuation of this general line of research. The main contributions are at least two-fold. On one hand, we explore to leverage pseudo-relevance feedback in conjunction with several effective proximity-based formulations of sentence modeling to improve the accuracy in the estimation of sentence models involved in the LM-based summarization framework. On the other hand, the utilities of our summarization methods and several widely-used methods are analyzed and compared extensively. The idea of utilizing word proximity for enhancing query modeling methods has recently attracted much attention and been applied with success to a few IR tasks [13]-[18] . However, to our best knowledge, this idea has never been extensively explored for probabilistic sentence modeling in extractive speech summarization.

The remainder of this paper is organized as follows. We first briefly review some related work on extractive summarization in Section II. Section III introduces the notion of leveraging the LM-based framework and various formulations of sentence modeling based on pseudo-relevance feedback for extractive speech summarization. After that, Section IV sheds light on our proposed methods to further improve sentence modeling. Finally, experiments and conclusions are presented in Sections V and VI, respectively.

II. RELATED WORK

The wide array of ASR-based extractive speech summarization methods developed so far may roughly fall into three main categories [3], [4], [7]: 1) methods simply

based on sentence structure or location cues, 2) methods based on unsupervised statistical mechanisms without the need of human-annotated ground truth while constructing the associated summarizers, and 3) methods based on supervised sentence classification.

For the first category, the important sentences are selected from specific parts of a spoken document, e.g., the introductory and/or concluding parts [8]. Such methods can only be applied to some limited domains or pre-known structured documents. The unsupervised methods attempt to extract salient sentences simply on the basis of our prior knowledge about the summarization process conducted by human, where some acoustic, phonetic and prosodic features of spoken words in the automatic transcript, or the statistical information resided in the transcript, such as word frequency, linguistic score and recognition confidence, are derived for measuring the importance of each sentence and/or the similarity among all sentences, in the spoken document. The associated methods based on these features have garnered much attention of research. Representative methods include, but are not limited to, the vector space model (VSM) [20], latent semantic analysis (LSA) [20], maximum marginal relevance (MMR) method [21], Markov random walk (MRW) [22], LexRank [23], submodularity-based method [24] and integer linear programming (ILP) method [25].

On the other hand, a number of supervised methods using various kinds of indicative features and explicit objectives for classification, such as Gaussian mixture models (GMM) [26], Bayesian classifiers (BC) [27], support vector machines (SVM) [28] and conditional random fields (CRFs) [29], to name just a few, also have been developed. In these methods, selection of important sentences is usually casted as a binary classification problem, i.e., to verify whether a given sentence should be included into the summary or not. However, these supervised methods require a set of training documents along with their corresponding hand-crafted summaries (or labeled data) for training the classifiers (or summarizers). In practice, manual annotation is expensive in terms of time and labor. Therefore, even if the performance of unsupervised summarizers is not always comparable to that of supervised ones, their easy-to-implement and portable property still makes them attractive for academic research or practical applications. Interested readers may also refer to [3], [4], [7], [30] for thorough and entertaining discussions of major methods that have been successfully developed and applied to a wide variety of text and speech summarization tasks.

III. LM-BASED FRAMEWORK FOR SUMMARIZATION

Intuitively, extractive speech summarization could be framed as an ad hoc information retrieval (IR) problem [31], where a spoken document to be summarized 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 (or importance) to the information need. In the past several years, the language modeling (LM)-based framework has been introduced to a wide spectrum of IR tasks and demonstrated with good empirical success [32]. This modeling paradigm also has been successfully adopted and applied to the context of speech summarization recently [9]-[12].

In the LM-based summarization framework, each sentence S of a spoken document D to be summarized is formulated as a probabilistic generative model for generating the document, and sentences are selected on the basis of their corresponding generative probability $P(D|S)$: the higher the probability $P(D|S)$, the more representative S is likely to be for D. With the "*bag-of-words*" assumption, the probability $P(D|S)$ can be approximated by:

$$P(D \mid S) \approx \sum_{w \in D} P(w \mid S)^{c(w,D)}, \tag{1}$$

where $c(w, D)$ is the occurrence count of a specific type of word (or term) w in D, reflecting that w will contribute more in the calculation of $P(D|S)$ if it occurs more frequently in D. The simplest way is to estimate the sentence model $P(w|S)$ on the basis of the frequency of words occurring in the sentence, with the maximum likelihood (ML) estimation [10]. In what follows, we will term (1) the document-likelihood measure (denoted by DLM for short). Due to that each sentence S of a spoken document D 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 this deficiency, in this paper, we explore several effective formulations for sentence modeling to enhance the sentence representation (or assign more accurate probability masses to words in the sentence) through leveraging the relevance cues gleaned from pseudo-relevance feedback (PRF) [31], [33]. A commonality among these formulations is that each sentence S is regarded as a query and in turn posted to an IR system to retrieve a set of top ranked text or spoken documents $D_S=\{D_1,\ldots,D_M\}$, counted as exemplars of pseudo-relevant documents, to be used for subsequent probabilistic sentence modeling.

A. Relevance Model (RM)

In speech summarization, it could be assumed that each sentence S of a spoken document D to be summarized is associated with an unknown relevance class R_S, and that words that are relevant to the semantic content expressed in S are samples drawn from R_S [34]. However, in reality, since there is no prior knowledge about R_S, we usually use the top-ranked contemporary text (or spoken) documents (denoted by D_S) in response to the sentence S, returned by a pseudo-relevance feedback process, to approximate the relevance class R_S. The corresponding relevance model (RM), on the grounds of a multinomial view of R_S, can be estimated by

$$P_{RM}(w \mid S) = \frac{\sum_{D_r \in D_S} P(D_r) P(w \mid D_r) \prod_{w' \in S} P(w' \mid D_r)}{\sum_{D_r \in D_S} P(D_r) \prod_{w'' \in S} P(w'' \mid D_r)}, \tag{2}$$

where the probability $P(D_r)$ can be simply set uniform or determined in accordance with the relevance of D_r to S, while $P(w|D_r)$ is estimated on the grounds of the occurrence count of w in D_r, with the ML estimation [12]. The RM model hypothesizes that words w, which co-occur with those words

of the sentence S, in the feedback documents will have higher probabilities. The resulting relevance model $P_{RM}(w|S)$ can be linearly combined with or used to replace the original sentence model $P(w|S)$.

B. Simple Mixture Model (SMM)

The simple mixture model (SMM) [35] is an alternative formulation to extract relevance cues from PRF for sentence modeling in extractive speech summarization. The basic idea is to assume that the set of top-ranked documents returned by PRF are relevant and the resulting model $P_{SMM}(w|S)$ estimated from these documents can potentially benefit sentence modeling. Specifically, SMM assumes that words in D_S are drawn from a two-component mixture model; one component is the SMM model $P_{SMM}(w|S)$, and the other component is a background model $P(w|BG)$. The SMM model $P_{SMM}(w|S)$ is estimated by maximizing the log-likelihood of the set of feedback documents D_S expressed as follows, using the expectation-maximization (EM) algorithm [36]:

$$LL_{D_S} = \sum_{D_r \in D_S} \sum_{w \in V} c(w, D_r) \log\big((1-\alpha) P_{SMM}(w|S) + \alpha P(w|BG)\big), \quad (3)$$

where α is a pre-defined weighting parameter used to control the degree of reliance between $P_{SMM}(w|S)$ and $P(w|BG)$.

The SMM estimation will enable more specific words (i.e., words in D_S that are not well-explained by the background model) to receive more probability mass, thereby leading to a more discriminative sentence model $P_{SMM}(w|S)$. Simply put, the SMM model $P_{SMM}(w|S)$ is anticipated to extract useful word usage cues from D_S, which are not only relevant to the sentence S, but also external to those already captured by the background model. Accordingly, the SMM model $P_{SMM}(w|S)$ can be combined with the language model, which is directly generated from the original sentence, through a simple linear interpolation.

IV. Sentence Modeling with Proximity Information

While the "*bag-of-words*" assumption can facilitate the derivation and estimation of the RM or SMM model, it seems to be an over-simplification for the problem of language modeling in extractive speech summarization. To mitigate such a drawback, one possible remedy is to incorporate the constraints of word order and adjacency relationships among previous words and the upcoming word into the formulation of the RM or SMM model (*cf.* Section III). For this idea to work, we explore three variants of such proximity-based counting method, including the window-based, Hyperspace Analogue to Language (HAL)-based and kernel-based methods, whose notions and formulations will be fleshed out, respectively, as follows.

A. Window-based Method

To consider the proximity effect within a fixed-length window while counting term frequencies, a given pseudo-relevant document is first segmented into a list of sliding windows, each of which has a fixed window size d. If the length (i.e., token counts) of a document is L, the document is segmented into (L-d+1) sliding windows, each of which contains d consecutive tokens. For example, if a document contains only five consecutive tokens {a, b, c, d, e}, and the window size is set to 3, there are three windows in this document, namely {a, b, c}, {b, c, d}, and {c, d, e}. The proximity frequency is then defined as the number of windows in which all n-gram terms co-occur and its corresponding counting and probability are expressed by

$$C_{window}(w, D_r) = \sum_{w' \in S} C_{D_r}(w, w') \cdot IDF(w'), \quad (4)$$

and

$$P_{window}(w \mid D_r) = \frac{C_{window}(w, D_r)}{\sum_{w''} C_{window}(w'', D_r)}. \quad (5)$$

In (4), $IDF(w')$ is used to designate the inverse document frequency of w', which reflects the importance of w' in the background collection; $C_{D_r}(w, w')$ denotes the number of times that w and w' co-occur within a fixed-length sliding window inside a pseudo-relevant document D_r, where the sliding window starts at each occurrence of w' with a span of d consecutive words. By substituting $P_{window}(w|D_r)$ in (5) into (2) in place of $P(w|D_r)$ and substituting $C_{window}(w, D_r)$ in (4) into (3) in place of $c(w, D_r)$, we can to some extent modulate the impact of the closeness of word proximity on relevance modeling in RM and SMM.

B. HAL-based Method

As a natural extension of the window-based method, we consider the relative strength between the expansion term and sentence term in a given distance within the window. This idea is motivated by the Hyperspace Analogue to Language (HAL) [37], which originates from a computational modeling of psychological theory of word meaning by considering only the context of words that immediately surround the given word. In HAL, all words within a window are considered co-occurring with each other with strengths inversely proportional to the distance between them, and the co-occurring measure is accumulated over the corpus. Then, a term w can be represented by a semantic vector, in which each dimension is the weight for this term w and another term w' as follows:

$$HAL(w, w') = \sum_{k=1}^{d} W(k) n(w, k, w'), \quad (6)$$

where k is the distance from term w to w', $n(w, k, w')$ is the co-occurrence frequency within a sliding window when the distance equals to k, and $W(k)$=d-k+1 denotes the strength. We adapt the original HAL model to our work. To measure the proximity between a candidate expansion term and the original sentence, the context is restricted to the sentence terms, not all the co-occurred terms in the feedback documents. With this setting, the resulting vector for each candidate term denotes a proximity relationship with the entire sentence. Then, the HAL-based proximity frequency counting is expressed as follows:

$$C_{HAL}(w, D_r) = \sum_{w' \in S} HAL(w, w') \cdot IDF(w'), \quad (7)$$

where $IDF(w')$ is the same as in (4). $C_{HAL}(w, D_r)$ can be readily used in RM and SMM like $C_{window}(w, D_r)$ in (4).

C. Kernel-based Method

More than the counts of two tokens co-occurring within a window, the kernel-based method also can give a weight to each count according to their relative distance. In this study, we use the Gaussian kernel to measure the proximity between a candidate expansion term w and a sentence term w', and the Gaussian kernel of proximity-based frequency counting is expressed as follows:

$$K_{D_r}(w, w') = \exp\left(\frac{-(Pos_w - Pos_{w'})^2}{2\sigma^2}\right), \quad (8)$$

where Pos_w and $Pos_{w'}$ are the positions of candidate term w and sentence term w' in a document, respectively, σ is a smoothing parameter called bandwidth, which controls the scale of Gaussian distribution. Notably, σ has a similar effect to capture regional information as the parameter d in window-based methods.

Different from the window-based method, the kernel-based method is a soft proximity measure. In particular, even if a candidate term and a sentence term do not co-occur in a window of d, its weight can still be slightly boosted. The kernel-based proximity frequency counting is expressed as follows:

$$C_{kernel}(w, D_r) = \sum_{w' \in S} K_{D_r}(w, w') \cdot IDF(w'), \quad (9)$$

where the definition of $IDF(w')$ is the same as that in (4). $C_{kernel}(w, D_r)$ can be readily used in RM and SMM like other proximity frequency counting methods.

The notion of leveraging word proximity for enhancing query modeling methods has recently attracted much attention and been applied with success to a few IR tasks [13]-[18]. However, to our best knowledge, this idea has never been extensively explored and well surveyed for probabilistic sentence modeling in extractive speech summarization.

V. EXPERIMENTS

A. Experimental Setup

The summarization dataset is extracted from a publicly available broadcast news corpus (MATBN) collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [38], which 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 205 broadcast news stories was selected for the summarization experiments. We chose 20 documents as the test set while the remaining 185 documents as the held-out development set. A set of about 100,000 text news documents, compiled during the same period as the broadcast news documents to be summarized, was employed to train the background language model and used as the collection for performing pseudo-relevance feedback. A subset of 25-hour speech data in MATBN was used to bootstrap the acoustic training with the minimum phone error rate (MPE) criterion and the training data selection scheme. The vocabulary size is about 72K words.

Three subjects were asked to create summaries of the 205 spoken documents for the summarization experiments as references (the gold standard) for evaluation. The reference summaries were generated by ranking the sentences in the manual transcript of a spoken document by importance without assigning a score to each sentence. For the assessment of summarization performance, we adopt the widely-used ROUGE metrics [39], including ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (the longest common subsequence). All the experimental results reported hereafter are obtained by calculating the F-scores [31] of these ROUGE metrics. 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.

Each news story consists of two kinds of transcripts, viz. TD and SD, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the speech recognition transcripts that may contain speech recognition errors. All the parameters mentioned above were tuned on the development set.

B. Experimental Results

To begin with, we assess the performance level of the baseline DLM method for extractive speech summarization, by comparing it with several well-practiced unsupervised summarization methods, including LEAD [8], VSM [20], MRW [22], LexRank [23], MMR [21], the submodularity-based method (denoted by Submodularity hereafter) [24] and the ILP method [25].

The corresponding summarization results of these unsupervised methods are illustrated in Table I. Several noteworthy observations can be drawn from Table I. First, the various graph-based methods (i.e., MRW, LexRank, and Submodularity) are quite competitive to each other and perform better than LEAD and VSM in both the TD and SD cases. Second, MMR that presents an extension of VSM, which takes the removal of redundant information as an additional criterion, can work as well as the various graph-based methods in the TD case, delivering even better performance than the latter ones for the SD case. Third, it is evident that DLM yields performance comparable to other unsupervised methods, confirming the applicability of the language modeling framework for speech summarization. Fourth, the ILP method turns out to be the best-performing one among all the unsupervised summarization methods compared here in the TD case, but it only offers mediocre performance in the SD case. Lastly, there is a sizable gap between the TD and SD cases, indicating room for further improvements. We may seek remedies, such as robust

TABLE I: SUMMARIZATION RESULTS ACHIEVED BY THE BASELINE DLM METHOD AND SEVERAL WIDELY-USED UNSUPERVISED METHODS.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	DLM	0.411	0.298	0.371
	LEAD	0.310	0.194	0.276
	VSM	0.347	0.228	0.290
	MMR	0.368	0.248	0.322
	MRW	0.412	0.282	0.358
	LexRank	0.413	0.309	0.363
	Submodularity	0.414	0.286	0.363
	ILP	**0.442**	**0.337**	**0.401**
SD	DLM	0.364	0.210	0.307
	LEAD	0.255	0.117	0.221
	VSM	0.342	0.189	0.287
	MMR	**0.366**	**0.215**	**0.315**
	MRW	0.332	0.191	0.291
	LexRank	0.305	0.146	0.254
	Submodularity	0.332	0.204	0.303
	ILP	0.348	0.209	0.306

TABLE II: SUMMARIZATION RESULTS ACHIEVED BY THE DLM METHOD INTEGRATED WITH VARIOUS SENTENCE MODELING FORMULATIONS.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	DLM	0.411	0.298	0.371
	RM	**0.453**	**0.335**	**0.403**
	SMM	0.439	0.320	0.388
SD	DLM	0.364	0.210	0.307
	RM	0.382	**0.239**	**0.331**
	SMM	**0.383**	0.229	0.327

indexing techniques, to compensate for imperfect speech recognition [40], [41].

In the second set of experiments, we evaluate the utilities of leveraging pseudo-relevance feedback in conjunction with two modeling formulations, i.e., RM and SMM, to enhance the sentence models involved in the DLM method. From the results shown in Table II, it is evident that both these two formulations can considerably improve the summarization performance of the DLM method, which corroborates the advantage of using PRF and the various formulations for enhanced sentence modeling.

In the third set of experiments, we compare the three proposed proximity-based frequency counting methods (viz. Window, HAL and Kernel) with the baseline unigram counting method in two kinds of relevance modeling, viz. RM and SMM. As can be seen from Tables III and IV, respectively for RM and SMM formulations, all the proximity-based counting methods deliver moderate improvements over simple unigram counting, in both the TD and SD cases. This confirms that incorporation of the word proximity information brings in more precise estimation of relevance modeling. There is no significant difference between the three proximity-based frequency counting methods, although HAL and Kernel slightly outperform Window. For RM, HAL performs the best in the TD case, while Kernel is the best-performing method in the SD case. On the other hand, for SMM, Kernel outperforms HAL in most cases. Generally speaking, kernel-based proximity counting seems to be more robust than the other two counting methods.

TABLE III: SUMMARIZATION RESULTS ACHIEVED BY PROXIMITY-BASED SENTENCE MODELING IN RM FORMULATION.

		Relevance Model (RM)		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	Unigram	0.453	0.335	0.403
	Window	0.456	0.345	0.406
	HAL	**0.460**	**0.349**	0.408
	Kernel	0.457	0.345	**0.409**
SD	Unigram	0.382	0.239	0.331
	Window	0.387	0.245	0.336
	HAL	0.387	0.246	0.335
	Kernel	**0.390**	**0.249**	**0.343**

TABLE IV: SUMMARIZATION RESULTS ACHIEVED BY PROXIMITY-BASED SENTENCE MODELING IN SMM FORMULATION.

		Simple Mixture Model (SMM)		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	Unigram	0.439	0.320	0.388
	Window	0.450	0.330	0.395
	HAL	**0.452**	0.343	0.403
	Kernel	0.451	**0.347**	**0.404**
SD	Unigram	0.383	0.229	0.327
	Window	**0.386**	0.237	0.332
	HAL	0.384	0.236	0.333
	Kernel	0.382	**0.240**	**0.334**

TABLE V: SUMMARIZATION RESULTS ACHIEVED BY SVM.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	SVM	0.484	0.383	0.437
SD	SVM	0.384	0.240	0.343

In the final set of experiments, we compare the RM method with SVM, one of the state-of-the-art supervised methods for extractive speech summarization [42]-[45]. SVM was trained with the documents along with their summaries in the development set, where each sentence of a spoken document was characterized with a set of 35 commonly-used lexical and prosodic features [3], [30], [43]-[45]. Comparing the results of SVM shown in Table V with that of the variants of the RM method shown in Table III, we notice that, although RM and its variants are, in essence, unsupervised methods that merely use word occurrence or co-occurrence statistics for important sentence selection, they can perform on par with or even better than SVM that utilizes handcrafted summaries and a rich set of features for model training.

VI. CONCLUSIONS

In this paper, we have presented a novel extension of the relevance modeling framework for use in extractive speech summarization. In particular, the so-called "*bag-of-words*" assumption of relevance modeling is relaxed by incorporating word proximity evidence (viz. the three variants explored in this work, including window-, HAL- and kernel-based

counting methods) into the several sentence modeling formulations. Experimental evidence supports that the various methods instantiated from our modeling framework outperform several existing state-of-the-art unsupervised methods for extractive speech summarization. In future work, we plan to integrate different kinds of proximity-based relevance modeling and more acoustic/prosodic information as well as lexical/semantic cues into the process of feedback document selection so as to improve the empirical effectiveness of sentence modeling. We are also interested in investigating more robust indexing techniques for representing spoken documents in order to bridge the performance gap between the TD and SD cases.

ACKNOWLEDGEMENT

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 104-2221-E-003-018-MY3, MOST 103-2221-E-003-016-MY2, MOST 104-2911-I-003-301, NSC 101-2221-E-003-024-MY3, and 101-2511-S-003-047-MY3.

REFERENCES

[1] S. Furui, *et al*., "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing,* 12(4), pp. 401–408, 2004.

[2] K. McKeown, *et al*., "From text to speech summarization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 997–1000, 2005.

[3] Y. Liu and D. Hakkani-Tur, "Speech summarization," *Chapter 13* in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, New York: Wiley, 2011.

[4] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, 5(2–3), pp. 103–233, 2011.

[5] S. Furui, *et al*., "Fundamental technologies in modern speech recognition," *IEEE Signal Processing Magazine*, 29(6), pp. 16–17, 2012

[6] D. O'Shaughnessy, *et al.*, "Speech information processing: Theory and applications," *Proceedings of the IEEE*, 101(5), pp. 1034–1037, 2013.

[7] I. Mani and M.T. Maybury (Eds.), *Advances in automatic text summarization*, Cambridge, MA: MIT Press, 1999.

[8] P. B. Baxendale, "Machine-made index for technical literature-an experiment," *IBM Journal*, October 1958.

[9] S.-H. Liu, *et al.*, "Effective pseudo-relevance feedback for language modeling in extractive speech summarization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3226-3230, 2014.

[10] S.-H. Lin, *et al*., "Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 19(4), pp. 871–882, 2011.

[11] A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in *Proc. Annual Meeting of the Association for Computational Linguistics*, pp. 815–824, 2010.

[12] B. Chen, *et al.*, "Sentence modeling for extractive speech summarization," in *Proc. IEEE International Conference on Multimedia & Expo*, pp. 1–6, 2013.

[13] J. Miao, *et al.*, "Proximity-based Rocchio's model for pseudo-relevance feedback," in *Proc. of SIGIR Conference*, pp. 535–544, 2012.

[14] Z. Ye, *et al.*, "A Bayesian network approach to context sensitive query expansion," in *SAC*, pp. 1138–1142, 2011.

[15] D. Metzler and W. B. Croft, "A markov random field model for term dependencies," in *Proc. SIGIR Conference*, pp. 472–479, 2005.

[16] Y. Lv and C.-X. Zhai, "Positional relevance model for pseudo-relevance feedback," in *Proc. SIGIR Conference*, pp. 579–586, 2010

[17] V. Plachouras and I. Ounis, "Multinomial randomness models for retrieval with document fields," in *Proc. European conference on IR research*, pp. 28–39, 2007

[18] J. Zhao, *et al.*, "CRTER: using cross terms to enhance probabilistic information retrieval," in *Proc. SIGIR conference*, pp. 155–164, 2011.

[19] X. Zhu G. Penn, and F. Rudzicz. "Summarizing multiple spoken documents: finding evidence from untranscribed audio." In *Proc. of Joint Conference of ACL and IJCNLP*, pp. 549–557, 2009.

[20] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. SIGIR Conference*, pp. 19–25, 2001.

[21] J. Carbonell and J. Goldstein, "The use of MMR, diversity based reranking for reordering documents and producing summaries," in *Proc. SIGIR Conference*, pp. 335–336, 1998.

[22] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in *Proc. SIGIR Conference*, pp. 299–306, 2008.

[23] G. Erkan and D.R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligent Research*, 22(1), pp. 457–479, 2004.

[24] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *Proc. NAACL HLT,* pp. 912–920, 2010.

[25] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. European conference on IR research*, pp. 557–564, 2007.

[26] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Computer Speech & Language*, 23(1), pp. 126–144, 2009.

[27] J. Kupiec, *et al.*, "A trainable document summarizer," in *Proc. SIGIR Conference*, pp. 68–73, 1995.

[28] A. Kolcz, *et al*., "Summarization as feature selection for text categorization," in *Proc. ACM Conference on Information and Knowledge Management*, pp. 365–370, 2001.

[29] M. Galley, "Skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. Empirical Methods in Natural Language Processing*, pp. 364–372, 2006.

[30] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. Annual Meeting of the Association for Computational Linguistics,* pp. 470–478, 2008.

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

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

[33] 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.

[34] V. Lavrenko and W.B. Croft, "Relevance-based language models," in *Proc. SIGIR Conference*, pp. 120–127, 2001.

[35] C.-X. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. SIGIR Conference*, pp. 403–410, 2001.

[36] A. P. Dempster, *et al*., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, 39(1), pp. 1–38, 1977.

[37] R.A.A.K. Lund and C. Burgess, "Semantic and associative priming in high-dimensional semantic space," in *Proc. Annual Conference of the Cognitive Science Society*, pp. 660–665, 1995.

[38] H.-M. Wang, *et al.*, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2), pp. 219–236, 2005.

[39] C.-Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation," 2003. Available: http://haydn.isi.edu/ROUGE/.

[40] S. Xie and Y. Liu, "Using *N*-best lists and confusion networks for meeting summarization" *IEEE Transactions on Audio, Speech and Language Processing*, 19(5), pp. 1160–1169, 2011.

[41] C. Chelba, *et al*., "Soft indexing of speech content for search in spoken documents," *Computer Speech & Language*, 21(3), pp. 458–478, 2007.

[42] S. Xie and Y. Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, 24(3), pp. 495–514, 2010.

[43] J. Zhang and P. Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Proc. NAACL HLT, Companion Volume*, pp. 213–216, 2007.

[44] B. Chen, *et al*., "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing & Managemen*t, 49(1), pp. 1–12, 2013.

[45] S.-H. Liu, *et al.*, "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