Positional Language Modeling for Extractive Broadcast News Speech Summarization

Shih-Hung Liu[1, 2], *Kuan-Yu Chen*[1, 2], *Berlin Chen*[3], *Hsin-Min Wang*[1], *Hsu-Chun Yen*[2], *Wen-Lian Hsu*[1]

[1]Institute of Information Science, Academia Sinica, Taiwan
[2]National Taiwan University, Taiwan
[3]National Taiwan Normal University, Taiwan

E-mail: [1]{journey, kychen, whm, hsu}@iis.sinica.edu.tw, [2] yen@cc.ee.ntu.edu.tw, [3] berlin@ntnu.edu.tw

Abstract

Extractive summarization, with the intention of automatically selecting a set of representative sentences from a text (or spoken) document so as to concisely express the most important theme of the document, has been an active area of experimentation and development. A recent trend of research is to employ the language modeling (LM) approach for important sentence selection, which has proven to be effective for performing extractive summarization in an unsupervised fashion. However, one of the major challenges facing the LM approach is how to formulate the sentence models and estimate their parameters more accurately for each text (or spoken) document to be summarized. This paper extends this line of research and its contributions are three-fold. First, we propose a positional language modeling framework using different granularities of position-specific information to better estimate the sentence models involved in summarization. Second, we also explore to integrate the positional cues into relevance modeling through a pseudo-relevance feedback procedure. Third, the utilities of the various methods originated from our proposed framework and several well-established unsupervised methods are analyzed and compared extensively. Empirical evaluations conducted on a broadcast news summarization task seem to demonstrate the performance merits of our summarization methods.

Index Terms: extractive broadcast news summarization, positional language modeling, relevance modeling

1. Introduction

Following the rapid proliferation of Internet applications, ever-increasing volumes of multimedia, such as broadcast radio and television programs, lecture recordings, digital archives, among others, have been made available and become an integral part of our everyday life [1]. It is generally agreed upon that speech is one of the most important sources of information about multimedia [2, 3]. People can listen to and digest multimedia associated with spoken documents efficiently with the aid of extractive speech summarization, which selects a set of indicative sentences from an original spoken document according to a target summarization ratio and concatenates them together to form a summary accordingly [4-7]. The wide variety of extractive speech summarization methods that have been developed so far could be categorized into two broad groups, namely, unsupervised and supervised methods.

A common practice of most unsupervised methods is to select important sentences based on statistical features of sentences or of the words in the sentences, where the extraction of features and the training of corresponding models for sentence selection are typically conducted in the absence of human supervision. Statistical features, for example, can be the term (word) frequency, linguistic score and recognition confidence measure, as well as the prosodic information. Numerous unsupervised methods based on these features have been proposed and has sparked much interest recently. Among them, the vector space model (VSM) [8, 9], the latent semantic analysis (LSA) method [8], the Markov random walk (MRW) method [10], the maximum marginal relevance (MMR) method [11], the sentence significant score method [12], the LexRank [13], the submodularity-based method [14], and the integer linear programming (ILP) method [15] are the most popular approaches for speech summarization. On the other hand, a number of classification-based methods using various kinds of representative features also have been investigated, such as the Gaussian mixture models (GMM) [8], the Bayesian classifier (BC) [16], the support vector machine (SVM) [17], and the conditional random fields (CRFs) [18], to name just a few. These supervised methods need a set of training documents along with the corresponding handcrafted summaries for training their component models, whereas manual annotation would be expensive in terms of time and personnel.

A recent line of research is to employ the language modeling (LM) approach [19-22] in an unsupervised fashion for extractive speech summarization, showing some preliminary success. However, one of the major challenges facing the LM approach is how to formulate the sentence models involved in summarization and estimate their parameters more accurately for each spoken document to be summarized. This paper presents a continuation of this general line of research and has at least three main contributions. First, we propose a positional LM framework leveraging different granularities of positional-specific information to better estimate each individual sentence model. Second, we endeavor to further integrate the positional cues into relevance modeling via a pseudo-relevance feedback procedure. Third, the utilities of the various methods originated from our proposed framework and several widely-used unsupervised methods are analyzed and compared thoroughly.

2. Language Modeling for Summarization

In this work, we frame extractive speech summarization as an ad-hoc information retrieval (IR) problem, where a spoken document to be summarized is taken as an information need and each sentence of the document is thought of as a candidate unit to be retrieved according to its relevance (or importance) degree to the information need. This way, the primary goal of extractive speech summarization could be stated as selecting a set of representative sentences that can succinctly describe the main theme of the spoken document. Over the years, the language modeling (LM) approach has been introduced to a wide array of IR tasks, enjoying good empirical success [19]; this realm of research has been recently extended to speech summarization [20-23].

2.1. Probabilistic Generative Paradigm

A principal realization of leveraging the LM approach to extractive speech summarization is to adopt a probabilistic generative paradigm, which determines the importance of each sentence S in a document D to be summarized with regard to the likelihood of D being generated by the sentence model of S, i.e., the sentence generative probability $P(D|S)$. As such, sentences can be ranked in decreasing order of $P(D|S)$: the higher the probability $P(D|S)$, the more representative S is likely to be for D. If D is expressed as a sequence of words, $D=w_1,w_2,\ldots,w_{|D|}$ ($|D|$ is the length of D), where words are further assumed to be conditionally independent given S (i.e., the so-called "*bag-of-words*" assumption), then $P(D|S)$ can be factored as

$$P(D \mid S) \approx \prod_{i=1}^{|D|} P(w_i \mid S), \tag{1}$$

The sentence ranking problem has now boiled down to the problem of how to accurately infer the sentence model $P(w|S)$ for each S. Intuitively, the simplest way is to estimate $P(w|S)$ solely based on the frequency of each distinct word w occurring in S, with the maximum likelihood estimation (MLE) [24]. In what follow, we will term Eq. (1) the unigram language model method (denoted by ULM for short).

2.2. Relevance Model (RM)

For better estimation of the sentence models for use in summarization, each sentence S of a spoken document D to be summarized can be assumed to be associated with an unknown relevance class R_S and words that are relevant to the semantic content expressed in S are also samples drawn from R_S [25-27]. However, since there is no prior knowledge about R_S in reality, a pseudo-relevance feedback (PRF) procedure can be employed to probe R_S. More specifically, each sentence S is taken as a query and posed to an IR system to retrieve a set of top-ranked documents $\mathbf{D}_S$ from an external collection to approximate the relevance class R_S. The corresponding relevance model (RM), with a multinomial view of R_S, can be constructed with the following equation [26, 28]:

$$P_{\mathrm{RM}}(w \mid S) = \frac{\sum_{D' \in \mathbf{D}_S} P(D')P(w \mid D')\prod_{w' \in S} P(w' \mid D')}{\sum_{D'' \in \mathbf{D}_S} P(D'')\prod_{w' \in S} P(w' \mid D'')}, \tag{2}$$

where the document prior probability $P(D')$ can be determined in accordance with the relevance of D' to S (or simply assumed to be uniform), while $P(w \mid D')$ is estimated on the basis of the occurrence counts of w in D' with the MLE criterion. The resulting relevance model $P_{\mathrm{RM}}(w|S)$ can be linearly combined with or used to replace the original sentence model $P(w|S)$.

3. Positional Language Modeling

The existing variants of the LM approach for extractive summarization mostly build on the predominant "*bag-of-words*" assumption. A common first thought to mitigate this deficiency would be the seizing of high-order n-gram (e.g., bigram or trigram) information for sentence modeling. However, such a remedy is still too restrictive in rendering long-span dependency of non-adjacent words within a document to be summarized. In view of this, we alternatively explore a novel use of various position-specific LM methods that manage to additionally incorporate proximity or longer contextual evidence of words inside the document into the estimation of each individual sentence model. The key notion of employing the position-specific LM methods is based on the conjecture that a candidate sentence residing at a specific segment of the document comprising more content-carrying (important) words, or by itself containing more words that are located more close to other content-carrying words inside the document, are more qualified to be included in the final summary. Below we shed light on three instantiated position-specific LM methods we will explore in this paper.

3.1. Passage-based Language Model (PaLM)

As a first attempt, we devise a passage-based language model (denoted by PaLM) to explore the positional information inherent in a spoken document to be summarized. Ideally, the spoken document (like a broadcast news story) can be divided into several paragraphs according to its syntactic/semantic structure, such as the introductory remarks, related studies or events, elucidations of methodology or affairs, conclusions of articles, and references or footnotes of reporters. As such, a passage-based language model $P(w|L_m)$ can be constructed for each paragraph L_m, respectively. After that, the sentence generative probability can be determined by referring to not only the original sentence model $P(w|S)$ but also one of the passage-based language models (e.g., $P(w|L_m)$) that corresponds to the position of the sentence in the spoken document:

$$P(D \mid S) = \prod_{i=1}^{|D|} \left(\alpha \cdot P(w_i \mid S) + (1-\alpha) \cdot P(w_i \mid L_m) \right) \tag{3}$$

where α is a tunable weight used to control the balance between the original sentence model and the passage-based model. However, in reality, the syntactic/semantic structure of a document is hard to be determined correctly. In this paper, we hence split each spoken document into a predefined number of equal-length segments as the resulting passages.

3.2. Position-Specific Language Model (PoLM)

Not content with capturing the coarse-grained passage-based positional information, as done by PaLM (*c.f.* Section 3.1), in this paper we make a step forward to incorporate more fine-grained positional information into sentence modeling. For each word position in the document, the semantic/syntactic cues carried by it can be discovered by considering the context words around this specific position. Hence, a position-specific language model can be estimated for each word position, with the aid of a proximity-based word occurrence count discounting strategy that accounts for the proximity relationships among the word of interest at each position and words in its surrounding context. More specifically, a "virtual" passage is composed for each position in the document by accumulating the propagated occurrence counts from all other words in the spoken document (as illustrated in Figure 1). In mathematical terms, the position-specific language model $P(w|k)$ (denoted by PoLM) of the k-th word position inside the document can be calculated by

$$P(w \mid k) = \frac{\hat{c}(w,k)}{\sum_{w' \in V} \hat{c}(w',k)}, \tag{4}$$

$$\hat{c}(w,k) = \sum_{j=1}^{|D|} c(w,j) \times f(k,j), \tag{5}$$

$$f(k,j) = e^{\frac{-(k-j)^2}{2\sigma^2}}, \tag{6}$$

where V designates the vocabulary size, $c(w, j)$ corresponds to the original occurrence count of word w at position j (either 0 or 1), and $\hat{c}(w,k)$ is the propagated occurrence count of word w at position k. In addition, $f(k,j)$ is the proximity-based propagation function from position j to position k; here we simply use the Gaussian kernel as the default propagation function (other alternatives are also feasible). Such a PoLM framework intuitively offers a position-specific perspective on the content of the document and thus can provide more fine-grained cues for possible use in sentence modeling. In this paper, we explore two disparate ways to harness the power of PoLM for speech summarization: the best-position strategy and the multi-position strategy. For the best-position strategy, we determine the relevance degree between the document to be summarized and one of its sentences based on the highest generative probability predicted by one of the PoLM models involved in the sentence:

$$P(D \mid S) = \max_{k \in S} \prod_{i=1}^{|D|} P(w_i \mid k), \tag{7}$$

One the other hand, for the multi-position strategy, we average out the L most highest generative probabilities computed by the PoLM models belonging the sentence to construct its sentence model:

$$P(D \mid S) = \frac{1}{L} \sum_{k \in S \,\&\, k \in \mathrm{Top}\, L} \prod_{i=i}^{|D|} P(w_i \mid k), \tag{8}$$

3.3. Position-Specific Relevance Model (PRM)

In relevance modeling (*cf*. Section 3.2), the probability of a word is computed by sweeping over all of the pseudo-relevant documents. On top of concept of relevance modeling, we explore a novel position-specific relevance model (denoted by PRM) that further incorporates the position-specific proximity cues. To this end, for each sentence, the construction of its corresponding PRM model will take into account all potential position-specific language models (*c.f.* Section 3.2) of each pseudo-relevant document:

$$P_{\mathrm{PRM}}(w \mid S) \propto P_{\mathrm{PRM}}(w, S) = \sum_{D' \in \mathbf{D}_S} \sum_{k=1}^{|D'|} P(w, S, k, D'), \tag{9}$$

where k indicates a specific position in a pseudo-relevant document D'. The challenge now lies in estimating the joint probability $P(w, S, k, D')$ for each pseudo-relevant document. Here we may assume that based on the probabilistic generative paradigm, we can first pick a document D' according to $P(D')$, then choose a position k in D' with the probability $P(k|D')$, and lastly generate word w and sentence S conditioned on D' and k, with the probability $P(w, S|D', k)$:

$$P(w, S, D', k) = P(D')P(k \mid D')P(w, S \mid D', k), \tag{10}$$

where $P(D')$ can be interpreted as a document prior and may be simply set to be uniformly distributed. Although it is possible to estimate $P(k|D')$ based on the specialty of the document structure, we assume here that every position in a document is equally important, i.e., $P(k|D') = 1/D'$. If we further makes the hypothesis that the generation of w and S are conditionally independent given D' and k, then we have

$$P_{\mathrm{PRM}}(w \mid S) = \frac{\sum_{D' \in \mathbf{D}_S} \frac{1}{|D'|} \sum_{k=1}^{|D'|} P(w \mid D', k) P(S \mid D', k)}{\sum_{D'' \in \mathbf{D}_S} \frac{1}{|D''|} \sum_{k'=1}^{|D''|} P(S \mid D'', k')}, \tag{11}$$

where $P(w|D', k)$ is estimated on the grounds of the propagated occurrence counts of w in D' at position k (*cf*. Eq.(5)). As such, the resulting position-specific relevance model $P_{\mathrm{PRM}}(w|S)$ can be linearly combined with or used to replace the original sentence model $P(w|S)$.

The notion of leveraging positional information for enhancing query modeling has recently attracted much attention and been applied with some success to a few IR and speech recognition tasks [29-31]. However, to our knowledge, this notion has never been sufficiently explored for sentence modeling in the context of extractive text and speech summarization.

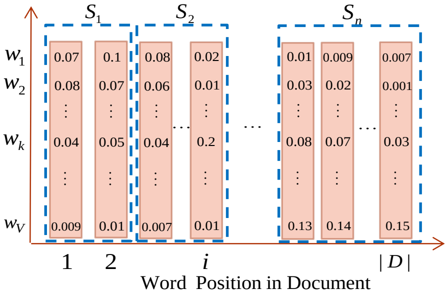

Fig 1. Illustration of PoLM, where each position has its own virtual passage characterized with a position-specific unigram model estimated based on the propagated occurrence counts gathered from its surrounding context. In this example, the sentence S_1 contains two words located at positions 1 and 2 of the spoken document.

4. Experimental Setup

The summarization dataset employed in this study is the broadcast news (MATBN) corpus assembled by the Academia Sinica and the Public Television Service Foundation of Taiwan [32]. A subset of 205 broadcast news documents was prepared for the summarization experiments. Three subjects were asked to create summaries of the 205 spoken documents for the summarization experiments as references (the gold standard) for evaluation. Further, 20 documents were reserved as the held-out test set, while 100 documents randomly selected from the rest were taken as the development set. An external set of about 100,000 text news documents, compiled during the same period as the broadcast news documents to be summarized, was also used to assist the estimation of RM and PRM.

For the assessment of summarization performance, we adopt the widely-used ROUGE metrics [33]. Three variants of the ROUGE metrics are used to quantify the performance of the summarization methods compared in this paper. They are, respectively, the ROUGE-1 (denoted by R-1, unigram) metric, the ROUGE-2 (denoted by R-2, bigram) metric, and the ROUGE-L (denoted by R-L, longest common subsequence) metric. 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, is set to 10% in this research.

5. Experimental Results

We first assess the performance level of the baseline LM-based summarized method (i.e., ULM) by comparing it with several well-established unsupervised methods, including VSM, LSA, MRW, LexRank, submodularity, ILP, CBOW and SG (the last two methods are based on distributed word representations derived based on the local proximity information among words [9][9]). In addition, a bigram language model method (denoted by BLM, which is regarded as a straightforward extension of ULM) is also investigated here. The corresponding summarization results of these unsupervised methods are illustrated in Table 1, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the speech recognition transcripts (the average word error rate for the speech recognition transcripts was about 40%). Several noteworthy observations can be drawn from Table 1. First, ULM works on par with or even better than the graph-based methods (viz. MRW, LexRank, and Submodularity) and surpasses VSM, LSA, CBOW, and SG for both the TD and SD cases. Meanwhile, BLM brings almost no additional gain over ULM. Second, ILP appears to be the best-performing one

Table 1. Summarization results achieved by a few well-established unsupervised methods.

	TD			SD		
	R-1	R-2	R-L	R-1	R-2	R-L
VSM	0.347	0.228	0.290	0.34	0.189	0.287
LSA	0.362	0.233	0.316	0.34	0.201	0.301
MRW	0.412	0.282	0.358	0.33	0.191	0.291
LexRank	0.413	0.309	0.363	0.30	0.146	0.254
Submodularity	0.414	0.286	0.363	0.33	0.204	0.303
ILP	0.442	0.337	0.401	0.34	0.209	0.306
ULM	0.408	0.298	0.359	0.36	0.218	0.306
BLM	0.408	0.298	0.359	0.36	0.218	0.311
CBOW	0.369	0.224	0.308	0.36	0.206	0.313
SG	0.367	0.230	0.306	0.35	0.205	0.303

Table 2. Summarization results achieved by PaLM with different numbers of document segments.

PaLM	TD			SD		
	R-1	R-2	R-L	R-1	R-2	R-L
2	0.427	0.320	0.383	0.380	0.234	0.335
4	0.412	0.298	0.364	0.381	0.231	0.324
8	0.415	0.321	0.372	0.369	0.225	0.317

Table 3. Summarization results achieved by PoLM with the best-position and multi-position strategies.

PoLM	TD			SD		
	R-1	R-2	R-L	R-1	R-2	R-L
Best-Position	0.443	0.326	0.387	0.381	0.237	0.332
Multi-Position	0.448	0.340	0.396	0.384	0.247	0.338

Table 4. Summarization results achieved by RM and PRM.

	TD			SD		
	R-1	R-2	R-L	R-1	R-2	R-L
RM	0.450	0.336	0.400	0.374	0.226	0.321
PRM	0.475	0.366	0.428	0.391	0.251	0.339

Table 5. Summarization results achieved by SVM and SVM+PRM.

	TD			SD		
	R-1	R-2	R-L	R-1	R-2	R-L
SVM	0.484	0.383	0.437	0.384	0.240	0.343
SVM+PRM	0.498	0.401	0.452	0.394	0.257	0.357

among all the methods compared here. However, the superiority of ILP seems to be less pronounced for the SD case, probably due to the effect of speech recognition errors. Lastly, it is evident that ULM shows competitive results when compared to the other well-practiced unsupervised methods, confirming the applicability of the LM approach for speech summarization.

In the second set of experiments, we evaluate the effectiveness of PaLM as a function of different numbers (2, 4, and 8) of equal-length passages (segments) being used to represent a spoken document; the corresponding results are shown in Table 2. As can be seen, PaLM arrives at almost the same performance level as or provides only slight improvements over ULM. Further, increasing the number of segments does not enhance the performance of PaLM. A possible explanation is that the estimation of PaLM inevitably suffers from the data sparseness problem, since increasing the number of segments will result in much less words being binned into each individual segment.

We then continue to examine the performance of PoLM. As elaborated earlier in Section 3.2, PoLM are formulated with two modeling strategies, i.e., the best-position strategy (denoted by Best-Position) and the multi-position strategy (denoted by Multi-Position). A closer look at the corresponding results illustrated in Table 3 reveals two noteworthy points. First, PoLM stands out in performance for both of the TD and SD cases in comparison to the aforementioned LM-based methods (i.e., ULM BLM, and PaLM) and the existing state-of-the-art unsupervised methods. Second, for PoLM, the multi-position strategy (L=3, *cf.* Eq.(8)) seems to perform better than the best-position strategy, which suggests that averaging out the several most highest generative probabilities computed by the PoLM models of a sentence presents a feasible and effective means for sentence modeling.

In the fourth set of experiments, we report on the results of PRM and its precursor, namely RM, which are shown in Table 4. Inspection of Table 4 we notice two particularities. On one hand, RM, which revolves around a goal of relevance modeling (disparate from proximity modeling pursued by PoLM) for each sentence, also can yield substantial performance improvements over the various LM-based methods (i.e., ULM BLM, and PaLM) and the existing unsupervised methods. On the other hand, PRM (representing a tight integration of the concepts of RM and PoLM) can further boost the performance as compared to that using either RM or PoLM in isolation. As a final note, we additionally compare PRM with SVM; SVM is arguably one of the state-of-the-art supervised methods for speech summarization. In this paper, SVM was trained with the documents of the development set along with their summaries, where each sentence of a spoken document was characterized with a set of 35 commonly-used prosodic and lexical features [34, 35]. Furthermore, we also attempt to combine SVM and PRM by taking the ranking score of PRM for each sentence as an additional indicative feature, leading to an augmented set of 36 features in total, for use in the model of SVM (denoted by SVM+PRM). Comparing the results of SVM shown in Table 5 with that of PRM method shown in Table 4, we observe that, although PRM in essence is an unsupervised method that merely uses word occurrence and proximity statistics for important sentence selection, it achieves performance almost comparable to SVM that utilizes handcrafted summaries and a rich set of features for model training. In addition, the integration of SVM and PRM (i.e., SVM+PRM) can yield consistent improvements over that using either one of them individually, with respect to all the three ROUGE metrics. This again manifests the utility of PRM.

6. Conclusions

In this paper, we have presented an effective positional language modeling framework for extractive speech summarization. The deduced various position-specific sentence models are able to render both word occurrence and proximity cues inherent a spoken document, which are anticipated to benefit speech summarization. Experimental evidence supports the performance merits of the summarization methods originated from such a framework in comparison to a few state-of-the-art unsupervised methods. As to future work, we envisage to leverage more sophisticated language models, such as the long short-term memory (LSTM) neural network and its variants [36, 37], to jointly integrating more proximity and different kinds of acoustic and lexical information, as well as discourse-related cues, for use in speech summarization.

7. 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 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.

8. References

[1] S. Furui *et al.*, "Fundamental technologies in modern speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 16–17, 2012.

[2] M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 150–152, 2008.

[3] L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42–60, 2005.

[4] Y. Liu and D. Hakkani-Tur, "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, 2011.

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

[6] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011.

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

[8] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 19–25, 2001.

[9] M. Kageback *et al.*, "Extractive summarization using continuous vector space models," in *Proc. of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 31–39, 2014.

[10] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299–306, 2008.

[11] J. Carbonell and J. Goldstein, "The use of MMR, diversity based reranking for reordering documents and producing summaries," in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[12] S. Furui *et al.*, "Speech-to-text and speech-to-speech summarization of spontaneous speech", *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[13] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligent Research*, vol. 22, no. 1, pp. 457–479, 2004.

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

[15] K. Riedhammer *et al.*, "Long story short - Global unsupervised models for keyphrase based meeting summarization," *Speech Communication*, vol. 52, no. 10, pp. 801–815, 2010.

[16] J. Kupiec *et al.*, "A trainable document summarizer," in *Proc. ACM SIGIR Conf. on R&D in Information Retrieval*, pp. 68–73, 1995.

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

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

[19] C. X. Zhai, "Statistical language models for information retrieval: A critical review," *Foundations and Trends in Information Retrieval*, vol. 2, no. 3, pp. 137–213, 2008.

[20] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proc. of Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting*, pp. 362–370, 2009.

[21] 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*, vol. 19, no. 4, pp.871-882, 2011.

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

[23] 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*, vol. 23, no. 6, pp. 957-969, 2015.

[24] J. Frederick. "Statistical methods for speech recognition," The MIT Press 1999.

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

[26] V. Lavrenko and B. Croft, "Relevance-based language models," in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127, 2001.

[27] B. Chen et al., "Spoken document retrieval with unsupervised query modeling techniques," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2602-2612, 2012.

[28] B. Chen and K.-Y. Chen, "Leveraging relevance cues for language modeling in speech recognition," *Information Processing & Management*, vol. 49, no. 4, pp. 807-816, 2013.

[29] H. S. Chiu *et al.*, "Leveraging topical and positional cues for language modeling in speech recognition", *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1465-1481, 2014.

[30] Y. Lv and C. X. Zhai, "Positional language models for information retrieval", in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299–306, 2009.

[31] Y. Lv and C. X. Zhai, "Positional relevance model for pseudo-relevance feedback", in *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 579–586, 2010.

[32] H. M. Wang *et al.*, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, 2005.

[33] C. Y. Lin, "ROUGE: Recall-oriented understudy for gisting evaluation." 2003 [Online]. Available: http://haydn.isi.edu/ROUGE/.

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

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

[36] M. Sundermeyer *et al.*, "From feedforward to recurrent LSTM Neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.

[37] H. Palangi *et al.*, "Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval," in *Proc. of the International Conference on Machine Learning*, 2015.