A RECURRENT NEURAL NETWORK LANGUAGE MODELING FRAMEWORK FOR EXTRACTIVE SPEECH SUMMARIZATION

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

[†]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[*]National Taiwan Normal University, Taipei, Taiwan
[#]National Taiwan University, Taipei, Taiwan

[†]{kychen, journey, whm, hsu}@iis.sinica.edu.tw, [*]berlin@ntnu.edu.tw, [#]hhchen@ntu.edu.tw

ABSTRACT

Extractive speech summarization, with the purpose of automatically selecting a set of representative sentences from a spoken document so as to concisely express the most important theme of the document, has been an active area of research and development. A recent school of thought is to employ the language modeling (LM) approach for important sentence selection, which has proven to be effective for performing speech summarization in an unsupervised fashion. However, one of the major challenges facing the LM approach is how to formulate the sentence models and accurately estimate their parameters for each spoken document to be summarized. This paper presents a continuation of this general line of research and its contribution is two-fold. First, we propose a novel and effective recurrent neural network language modeling (RNNLM) framework for speech summarization, on top of which the deduced sentence models are able to render not only word usage cues but also long-span structural information of word co-occurrence relationships within spoken documents, getting around the need for the strict bag-of-words assumption. Second, the utilities of the method originated from our proposed framework and several widely-used unsupervised methods are analyzed and compared extensively. A series of experiments conducted on a broadcast news summarization task seem to demonstrate the performance merits of our summarization method when compared to several state-of-the-art existing unsupervised methods.

Index Terms— speech summarization, language modeling, recurrent neural network, long-span structural information

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, are continuously growing and filling our daily life [1-3]. Obviously, speech is one of the most important sources of information about multimedia. Users can listen to and digest multimedia associated with spoken documents efficiently by virtue 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 array of extractive speech summarization methods that have been developed so far may roughly fall into three main categories [4, 7]: 1) methods simply based on the sentence position or structure information, 2) methods based on unsupervised sentence ranking, and 3) methods based on supervised sentence classification.

For the first category, the important sentences can be selected from some salient parts of a spoken document [8]. For instance, sentences can be selected from the introductory and/or concluding parts of a spoken document. However, such methods can be only applied to some specific domains with limited document structures. On the other hand, unsupervised sentence ranking methods attempt to select important sentences based on statistical features of spoken sentences or of the words in the sentences without human labor involved. Statistical features, for example, can be the term (word) frequency, linguistic score and recognition confidence measure, as well as the prosodic information. The associated unsupervised methods based on these features have gained much attention of research. Among them, the vector space model (VSM) [9], the latent semantic analysis (LSA) method [9], 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 spoken document summarization. Apart from that, a number of classification-based methods using various kinds of representative features also have been investigated, such as the Gaussian mixture models (GMM) [9], the Bayesian classifier (BC) [16], the support vector machine (SVM) [17] and the conditional random fields (CRFs) [18], to name just a few. In these methods, important sentence selection is usually formulated as a binary classification problem. A sentence can either be included in a summary or not. These classification-based methods need a set of training documents along with their corresponding handcrafted summaries (or labeled data) for training the classifiers (or summarizers). However, manual annotation is expensive in terms of time and personnel. Even if the performance of unsupervised summarizers is not always comparable to that of supervised summarizers, their easy-to-implement and flexible property (i.e., they can be readily adapted and carried over to summarization tasks pertaining to different languages, genres or domains) still makes them attractive. Interested readers may also refer to [4-7] 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.

A recent line of research is to employ the language modeling (LM) approach in an unsupervised fashion, which has been applied to extractive speech summarization with preliminary success. However, one of the major challenges facing the LM approach is how to formulate the sentence models and accurately estimate their parameters for each spoken document to be summarized. This paper presents a continuation of this general line of research and its contribution is two-fold. First, we propose a novel and effective recurrent neural network language modeling (RNNLM) framework for speech summarization, on top of which the deduced sentence models are able to render not only word usage cues but also long-span structural information of word co-occurrence relationships within spoken documents, getting around the need for the strict bag-of-words assumption. Second, the utilities of the method originated from our proposed framework and several widely-used unsupervised methods are analyzed and compared extensively [4].

The remainder of this paper is organized as follows. Several unsupervised approaches are briefly introduced in Section 2. In Section 3, we shed light on the basic mathematical formulations of the LM-based summarization approach and the recurrent neural network language modeling framework proposed in this paper. After that, the experimental settings and a series of speech summarization experiments are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper and discusses avenues for future work.

2. POPULAR UNSUPERVISED METHODS

The wide spectrum of unsupervised summarization methods developed thus far may be further grouped into three subcategories: 1) the vector-space methods, 2) the graph-based methods, and 3) the combinatorial optimization methods.

2.1. The Vector-Space Methods

The vector space model (VSM), the latent semantic analysis (LSA), and the maximum marginal relevance method (MMR) are three representatives of the subcategory. VSM represents each sentence of a document and the whole document, respectively, in a vector form, where each dimension specifies the weighted statistics, for example the product of the term frequency (TF) and inverse document frequency (IDF), associated with an indexing term (or word) in the sentence or document. Sentences with the highest relevance scores (usually calculated by the cosine similarity of two vectors) to the whole document are included in the summary [9]. On the other hand, LSA projects the vector representation of the sentence (and document) into a latent semantic space, which is usually obtained by performing singular value decomposition (SVD) [9] on a word-by-sentence matrix of a given document. The ranking score of each sentence in the document to be summarized can be calculated by using the cosine similarity measure between the semantic vectors of the sentence and the document represented in the LSA space. Additionally, MMR can be viewed as an extension of VSM, because it also represents each sentence (and document) into a vector representation and the sentence selection is also based on the cosine similarity measure. The major difference is that MMR performs sentence selection iteratively by simultaneously considering the criteria of theme relevance and redundancy [11].

2.2. The Graph-Based Methods

The Markov random walk (MRW) method conceptualizes the document to be summarized as a graph of sentences, where each node represents a sentence and the associated weight of each link represents the lexical similarity relationship between a pair of nodes. Document summarization thus relies on the global structural information embedded in such conceptualized graph, rather than merely considering the similarity solely between each sentence of the document to be summarized and the document itself. Put simply, sentences that are more similar to others are deemed more salient to the main theme of the document [10]. In addition, LexRank bears a close resemblance to MRW by selecting salient sentences based on the notion of eigen-centrality of the sentence graph [13]. Both MRW and LexRank in essence are inspired from the well-known PageRank algorithm that is widely adopted by most of today's commercial search engines on the Internet.

2.3. The Combinatorial Optimization Methods

Among others, an interesting research direction is to frame the extractive speech summarization task as a combinatorial optimization problem, for which two widely studied and practiced methods are the submodularity-based method and the integer linear programming (ILP) method. The submodularity-based method views important sentence selection as a combinatorial optimization problem with a few objective functions defined on the sentence graph. A reasonable property of diminishing returns, stemming from the field of economics, is employed for important sentence selection. Several polynomial-time implementations have been proposed, with the intention to solve the summarization problem near-optimally [14]. In contrast, the ILP method leverages integer linear programming to deal with the constrained combinatorial optimization problem pertaining to extractive speech summarization. More specifically, ILP method reformulates the extractive summarization task as an optimization problem with a set of constrains, and then selects an optimal sentence combination by using integer linear programming. By doing so, ILP manages to select a preferred set of summary sentences that can retain the most important theme of a given document. Despite ILP is faced with an NP-hard problem, there exist some exact algorithms (such as branch-and-bound) for ILP. However, these algorithms are not readily suited for large-scale problems, since they almost invariably involve a rather time-consuming process for important sentence selection [15, 20, 21].

3. LANGUAGE MODELING BASED METHODS

Intuitively, extractive speech summarization could be cast as an ad-hoc information retrieval (IR) problem, where the spoken document is taken 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. As such, the ultimate goal of extractive speech summarization could be stated as the selection of the most representative sentences that can succinctly describe the main theme of the spoken document. In the recent past, the LM-based approach has been introduced to a wide spectrum of IR tasks with

good empirical success [22]; this modeling approach has been applied to extractive speech summarization recently [19, 23, 24].

3.1. Unigram Language Model

When applying the LM-based approach to extractive speech summarization, a principal realization is to use a probabilistic generative paradigm for ranking each sentence S of a spoken document D to be summarized, which can be expressed by $P(S|D)$. Instead of calculating this probability directly, we can apply the Bayes' rule and rewrite it as follows [24-26]:

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)}, \tag{1}$$

where $P(D|S)$ is the sentence generative probability, i.e., the likelihood of D being generated by S, $P(S)$ is the prior probability of the sentence S being relevant, and $P(D)$ is the prior probability of the document D. $P(D)$ in Eq. (1) can be eliminated because it is identical for all sentences and will not affect the ranking of the sentences. Furthermore, because the way to estimate the probability $P(S)$ is still under active study [25], we may simply assume that $P(S)$ is uniformly distributed, or identical for all sentences. In this way, the sentences of a spoken document to be summarized can be ranked by means of the probability $P(D|S)$ instead of using the probability $P(S|D)$: the higher the probability $P(D|S)$, the more representative S is likely to be for D. If the document D is expressed as a sequence of words, $D=w_1,w_2,...,w_L$, where words are further assumed to be conditionally independent given the sentence and their order is assumed to be of no importance (i.e., the so-called "*bag-of-words*" assumption), then $P(D|S)$ can be approximated by

$$P(D \mid S) \approx \prod_{i=1}^{L} P(w_i \mid S) \tag{2}$$

where L denotes the length of the document D. The sentence ranking problem has now been reduced to the problem of how to accurately infer the probability distribution $P(D|S)$, i.e., the corresponding sentence model for each sentence of the document. The simplest way is to estimate a unigram language model (ULM) on the basis of the frequency of each distinct word w occurring in the sentence, with the maximum likelihood (ML) criterion [25]:

$$P(w \mid S) = \frac{c(w,S)}{|S|}, \tag{3}$$

where $c(w,S)$ is the number of times that word w occurs in S and $|S|$ is the length of S. The ULM model can be further smoothed by a background unigram language model estimated from a large general collection to model the general properties of the language as well as to avoid the problem of zero probability. It turns out that a sentence S with more document words w occurring frequently in it would tend to have a higher probability of generating the document.

3.2. Recurrent Neural Network Language Model

While the bag-of-words assumption makes ULM a clean and efficient method for sentence ranking, it is an oversimplification of the problem of extractive speech summarization. Intuitively, long-span context dependence (or word proximity) cues might provide additional indication of the semantic-relatedness of a given sentence with regard to the document to be summarized [18, 19, 23]. Although a number of studies had been done on extending ULM to further capture local context dependence simply based on n-grams of various orders (e.g., bigrams or trigram), most of them resulted in leading to mild gains or mixed results [19, 23]. This is due in large part to the fact that a sentence usually consists of only a few words and the complexity of the n-gram model increases exponentially with the order n, making it difficult to obtain reliable probability estimates with the ML criterion.

In view of such phenomena, we explore in this paper a novel recurrent neural network language modeling (RNNLM) framework for the formulation of the sentence models involved in the LM-based summarization approach. RNNLM has recently emerged as a promising modeling framework that can effectively and efficiently render the long-span context relationships among words (or more precisely, the dependence between an upcoming word and its whole history) for use in speech recognition [27-29]. For each time index i, the input vector $\mathbf{w}_i$ is in one-of-V encoding, indicating the currently encountered word w_i, where the vector size V is set equal to the number of distinct vocabulary words; the hidden vector $\mathbf{s}_i$ represents the statistical cues encapsulated thus far in the network for the history (i.e., all preceding words) of w_i; and the output layer vector $\mathbf{y}_i$ stores the predicted likelihood values for each possible succeeding word (or word class) of w_i. An attractive aspect of RNNLM is that the statistical cues of previously encountered word retained in the hidden layer, i.e., $\mathbf{s}_{i-1}$, can be fed back to the input layer and work in combination with the currently encountered word w_i as an "augmented" input vector for predicting an arbitrary succeeding word w_{i+1}. By doing so, RNNLM can naturally take into account not only word usage cues but also long-span structural information of word co-occurrence relationships for language modeling. A bit of terminology: the augmented input vector $\mathbf{x}_i$, the hidden vector $\mathbf{s}_i$ and the output vector $\mathbf{y}_i$ are, respectively, represented or computed as follows [27-29]

$$\mathbf{x}_i = [(\mathbf{w}_i)^T, (\mathbf{s}_{i-1})^T]^T, \tag{4}$$

$$\mathbf{s}_i = f(\mathbf{U}\mathbf{x}_i), \tag{5}$$

$$\mathbf{y}_i = g(\mathbf{V}\mathbf{s}_i), \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ are pre-defined sigmoid activation functions and softmax functions, respectively. Finally, the model parameters (i.e., $\mathbf{U}$ and $\mathbf{V}$) of RNNLM can be derived by maximizing the likelihood of the training corpus using the back-propagation through time algorithm (BPTT).

As the notion of RNNLM is adopted and formalized for sentence modeling in extractive speech summarization, we devise a hierarchical training strategy to obtain the corresponding RNNLM model for each sentence of a spoken document to be summarized:

1) First of all, a document-level RNNLM model is trained for each document to be summarized by using the document itself as the training data. The resulting RNNLM model will memorize not only word usage but also long-span word dependence cues inherent in the document.
2) After that, for each sentence of the spoken document to be summarized, the corresponding sentence-specific RNNLM model is trained, starting from the document-level RNNLM

model obtained in Step 1 and using the sentence itself as the adaptation data for model training. That is, the parameters of RNNLM are optimized by maximize the likelihood of the sentence.

3) Consequently, the resulting sentence-specific RNNLM model can be used in place of, or to complement, the original sentence model (i.e., ULM). The RNNLM-based sentence generative probability for use in sentence ranking can be computed by

$$P_{\mathrm{RNNLM}}(D \mid S) = \prod_{i=1}^{L} P_{\mathrm{RNNLM}}(w_i \mid w_1, \ldots, w_{i-1}, S). \quad (7)$$

4. EXPERIMENTAL SETUP

4.1. Speech and Language Corpora

The summarization dataset employed in this study is a broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [30], 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 documents compiled between November 2001 and August 2002 was reserved for the summarization experiments. We chose 20 documents as the test set while the remaining 185 documents as the held-out development. Notice that the constants and weighting (interpolation) coefficients of all summarization methods compared in this paper were all tuned at optimum values.

On the other hand, twenty-five hours of gender-balanced speech from the remaining speech data were used to train the acoustic models for speech recognition. The data was first used to bootstrap the acoustic model training with the ML criterion. Then, the acoustic models were further optimized by the minimum phone error (MPE) discriminative training algorithm [31]. The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 30%. A large number of text news documents collected by the Central News Agency (CNA) between 2000 and 2001 (the Chinese Gigaword Corpus released by LDC) were used to train trigram language models for speech recognition with the SRI Language Modeling Toolkit [32].

4.2. Performance Evaluation

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 adopted the widely-used ROUGE metrics [33]. It evaluates the quality of the summarization by counting the number of overlapping units, such as *N*-grams, longest common subsequences or skip-bigram, between the automatic summary and a set of reference summaries. Three variants of the ROUGE metrics were used to quantify the utility of the proposed methods. They are, respectively, the ROUGE-1 (unigram) metric, the ROUGE-2 (bigram) metric and the ROUGE-L (longest common subsequence) metric [33].

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. Since increasing the summary length tends to increase the chance of getting higher scores in the recall rate of the various ROUGE metrics and might not always select the right number of informative words in the automatic summary as compared to the reference summary, all the experimental results reported hereafter are obtained by calculating the F-scores of these ROUGE metrics. Table 1 shows the levels of agreement (the Kappa statistic and ROUGE metrics) between the three subjects for important sentence ranking. Each of these values was obtained by using the summary created by one of the three subjects as the reference summary, in turn for each subject, while those of the other two subjects as the test summaries, and then taking their average. These observations seem to reflect the fact that people may not always agree with each other in selecting the summary sentences for a given document.

Table 1. The agreement among the subjects for important sentence ranking for the evaluation set.

Kappa	ROUGE-1	ROUGE-2	ROUGE-L
0.432	0.600	0.532	0.527

5. EXPERIMENTAL RESULTS

At the outset, we assess the performance level of the baseline LM-based summarization method (i.e., ULM) for extractive speech summarization by comparing it with several well-practiced or/and state-of-the-art unsupervised summarization methods, including the vector-space methods (i.e., VSM, LSA and MMR), the graph-based methods (i.e., MRW and LexRank) and the submodularity-based method. The corresponding summarization results of these unsupervised methods are graphically illustrated in Fig. 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 that may contain speech recognition errors. Several noteworthy observations can be drawn from Fig. 1. First, the two graph-based methods (i.e., MRW and LexRank) are quite competitive with each other and perform better than the various vector-space methods (i.e., VSM, LSA, and MMR) for the TD case. However, for the results of the SD case, the situation is reversed. It reveals that imperfect speech recognition may adversely affect the performance of the graph-based methods as compared to vector-space methods; a possible reason for such a phenomenon is that the speech recognition errors may lead to inaccurate similarity measures between each pair of sentences. The PageRank-like procedure of the graph-based methods, in turn, will be performed based on these problematic measures, potentially leading to disastrous results. Second, LSA, representing the sentences of a spoken document to be summarized and the document itself in the latent semantic space instead of the index term (word) space, can perform slightly better than VSM in both of the TD and SD cases. Third, the submodularity-based method achieves the best results in the TD case, but only offers mediocre performance as compared to the other methods in the SD case. Finally, it is evident that ULM shows competitive results when compared to the other state-of-the-art unsupervised methods, confirming the applicability of the language modeling approach for speech summarization.

Going one step further, we investigate a simple extension of the ULM method by using a bigram language model smoothed with a

unigram language model to represent each sentence involved in the document (denoted by BLM hereafter). As elaborated before (*cf.* Section 3.1), the weakness of the ULM method lies in that it follows the strict bag-of-words assumption (an oversimplification) without considering the word regularity or proximity information within spoken documents. The corresponding summarization results of BLM are depicted in Fig 2. To our surprise, the incorporation of bigram and unigram cues together (i.e., BLM) for sentence modeling only arrives at the same performance level as that using unigram cues alone (i.e., ULM) in the SD case, but performs even worse than the latter in the TD case. A reasonable explanation is that the estimation of the bigram language model for each sentence inevitably suffers from a more serious data sparseness problem than the unigram model, since its number of model parameters would be at most the square of that of the latter.

In the third set of our experiments, we evaluate our proposed RNNLM method for extractive speech summarization. The deduced sentence-specific RNNLM model can be used in isolation (denoted by RNNLM) or linearly combined with the unigram language model (denoted by RNNLM+ULM) to compute the sentence generative probability; the corresponding results are shown in Fig. 2 as well. Comparing to the other LM-based methods (i.e., ULM and BLM) or the other subcategories of unsupervised methods, we can find that RNNLM+ULM outperforms all the other models by a large margin in both TD and SD cases; however, using RNNLM in isolation only leads to improved results in the TD case. Furthermore, two more particularities can be made when we look into the results of Fig. 2. On one hand, because RNNLM+ULM manages to encapsulate not only word usage cues but also long-distance word co-occurrence relationships for sentence modeling, it seems to perform particularly well when the evaluation metrics are based on counting the number of matched high-order word co-occurrence counts between the reference and automatically generated summaries, such as the ROUGE-2 and ROUGE-L metrics. On the other hand, RNNLM and ULM seem to be complementary of each other and indeed can conspire to obtain better sentence modeling.

In the last set of experiments, we report on the detailed results of RNNLM+ULM with respect to the number of hidden-layer neurons being used in RNNLM. As can be seen from Fig. 3, it would be preferable to set the number of hidden-layer neurons lower than 100, probably due to that each sentence usually consists of only a few words that can be used for model estimation. It is intuitively clear that the more complex the sentence model, the more training data is needed to obtain a reliable estimation. Nevertheless, the way to systemically determine the optimal number of hidden-layer neurons for RNNLM remains an open issue and needs further investigation.

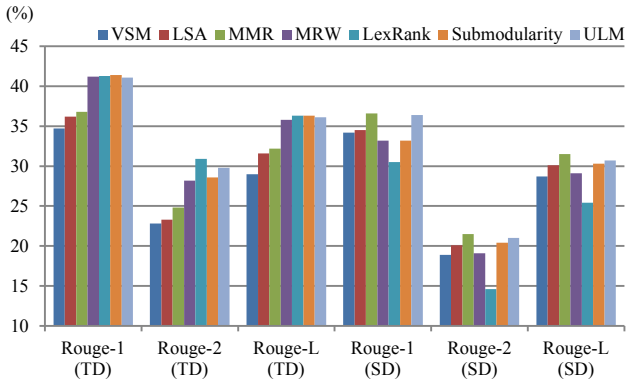

Fig. 1. Summarization results achieved by a few state-of-the-art unsupervised methods.

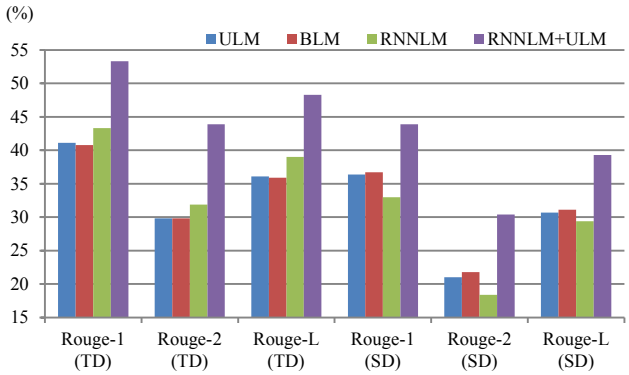

Fig. 2. Summarization results achieved by various LM-based methods, including ULM, BLM, RNNLM and RNNLM+ULM.

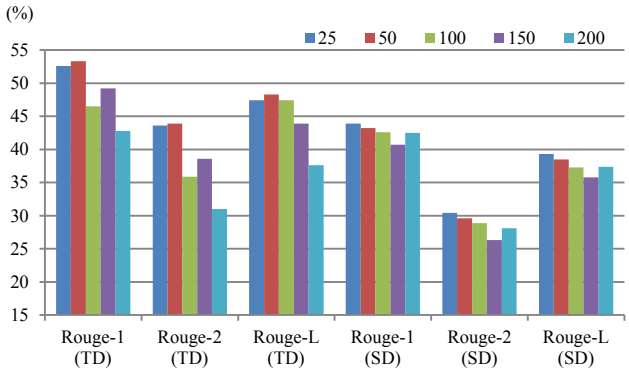

Fig. 3. Summarization results achieved by RNNLM+ULM with respect to different numbers of hidden-layer neurons used in RNNLM.

6. CONCLUSIONS

In this paper, we have proposed a novel and effective recurrent neural network language modeling (RNNLM) framework for extractive speech summarization. The deduced RNNLM sentence models are able to render both word usage cues and long-span structural information of word co-occurrence relationships that are expected to benefit speech summarization. Experimental evidence supports that the summarization method originated from such an RNNLM framework is quite comparable to several state-of-the-art unsupervised methods. As to the future work, we plan to investigate jointly integrating relevance-feedback, proximity, and other different kinds of lexical/semantic information cues into this framework so as to improve the empirical effectiveness of sentence modeling. We are also interested in investigating more robust indexing techniques for representing the spoken documents. In addition, we intend to further adopt and formalize this RNNLM framework for other multimedia-related tasks such as spoken document retrieval and music retrieval.

7. 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 NSC 101-2221-E-003-024-MY3, NSC 102-2221-E-003-014-, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 and NSC 103-2911-I-003-301.

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

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

[10] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in *Proc. 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. 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. 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. NAACL HLT, Companion Volume*, pp. 213–216, 2007.

[18] 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.

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

[20] D. Gillick and B. Favre. "A scalable global model for summarization," in *Proc. of the Workshop on Integer Linear Programming for Natural Language Processing*, pp. 10–18, 2009.

[21] C. Li, X. Qian, and Y. Liu. "Using supervised bigram-based ILP for extractive summarization," in *Proc. Annual Conference of the International Speech Communication Association*,, pages 1004–1013, August, 2013.

[22] 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.

[23] 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.

[24] 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.

[25] Y. T. Chen *et al.*, "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 95–106, 2009.

[26] J. Frederick, *Statistical Methods for Speech Recognition*, The MIT Press 1999.

[27] T. Mikolov et al., "Recurrent neural network based language model," in *Proc. Annual Conference of the International Speech Communication Association*, pp. 1045–1048, 2010.

[28] T. Mikolov *et al.*, "Extension of recurrent neural network language model," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 5528–5531, 2011.

[29] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *Proc. Spoken Language Technology*, pp. 234–239, 2012.

[30] 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.

[31] G. Heigold *et al.*, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 58–69, 2012.

[32] A. Stolcke, "SRILM–An extensible language modeling toolkit," in *Proc. Annual Conference of the International Speech Communication Association*, pp. 901–904, 2005.

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

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