Extractive Broadcast News Summarization Leveraging Recurrent Neural Network Language Modeling Techniques

Kuan-Yu Chen, *Student Member, IEEE*, Shih-Hung Liu, *Student Member, IEEE*,
Berlin Chen, *Member, IEEE*, Hsin-Min Wang, *Senior Member, IEEE*,
Ea-Ee Jan, Wen-Lian Hsu, *Fellow, IEEE*, and Hsin-Hsi Chen

***Abstract*—Extractive text or speech summarization manages to select a set of salient sentences from an original document and concatenate them to form a summary, enabling users to better browse through and understand the content of the document. A recent stream of research on extractive summarization 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 sentence in the document to be summarized. In view of this, our work in this paper explores a novel use of recurrent neural network language modeling (RNNLM) framework for extractive broadcast news summarization. On top of such a framework, 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 broadcast news documents, getting around the need for the strict bag-of-words assumption. Furthermore, different model complexities and combinations are extensively analyzed and compared. Experimental results demonstrate the performance merits of our summarization methods when compared to several well-studied state-of-the-art unsupervised methods.**

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

I. Introduction

Along with the growing popularity 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 made available to our everyday 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 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 with less (or even no) human labor involvement. 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 arguably the most popular methods for extractive speech 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

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 and NSC 101-2221-E-003-024-MY3.

Kuan-Yu Chen, Shih-Hung Liu, and Hsin-Hsi Chen are with the National Taiwan University, Taipei 106, Taiwan (e-mail: kychen@iis.sinica.edu.tw; journey@iis.sinica.edu.tw; hhchen@ntu.edt.tw).

Berlin Chen is with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan (e-mail: berlin@csie.ntnu.edu.tw).

Hsin-Min Wang and Wen-Lian Hsu are with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan (e-mail: whm@iis.sinica.edu.tw; hsu@iis.sinica.edu.tw).

Ea-Ee Jan is with the IBM, USA. (e-mail: ejan@us.ibm.com).

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 comprehensive and enjoyable discussions of major methods that have been successfully developed and applied to a wide variety of text and speech summarization tasks.

As a departure from the aforementioned methods, an emerging line of research is to employ the language modeling (LM) approach for important sentence selection, which has shown preliminary success for performing extractive speech summarization in an unsupervised fashion [19-22]. However, one of central challenges facing the LM approach is how to formulate the sentence models and accurately estimate their parameters for each sentence in the spoken document to be summarized. We recently introduced a new perspective on this problem of the existing LM-based methods [23], saying that it can be approached with a framework building on the notion of recurrent neural network language modeling (RNNLM), which shows promise 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 made by most of the existing LM-based methods. Our work in this paper continues this general line of research, including exploring different model complexities and combination strategies, as well as providing more in-depth elucidations on the modeling characteristics and the associated summarization performance of various instantiated methods. Further, the utility of our RNNLM-based methods is verified by extensive comparisons with several state-of-the-art unsupervised summarization methods [4].

The remainder of this paper is organized as follows. We start by reviewing previous studies on text or speech summarization using various kinds of unsupervised summarization methods in Section II. In Section III, we shed light on the basic mathematical formulations of the LM-based summarization approach and the recurrent neural network language modeling framework we explore in this paper. After that, the experimental settings and a series of speech summarization experiments are presented in Sections IV and V, respectively. Finally, Section VI concludes this paper and discusses potential avenues for future work.

II. 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.

A. *The Vector-Space Methods*

The vector space model (VSM) and the latent semantic analysis (LSA) are two best-known 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, for example, can be calculated by using the cosine similarity measure between the semantic vectors of the sentence and the document represented in the LSA space. In addition, the original maximum marginal relevance method (MMR) [11] can be viewed as an extension of VSM from the standpoint that it also transforms each sentence (and document) into a vector representation and the sentence selection is also based on the cosine similarity measure. The primary distinction between MMR and VSR lies in that MMR performs sentence selection iteratively by simultaneously considering the criteria of topic relevance and redundancy.

More recently, there has been a flurry of research on developing various word-to-vector (W2V) embedding methods [24, 25], which can serve as important building blocks of many interesting natural language processing applications, including extractive text and speech summarization [26]. The primary objective of these methods revolves around learning fixed-length continuous distributed vector representations of words from texts with neural networks, which not only can capture semantic or syntactic cues, but also can be used to induce similarity measures of words in context. In terms of extractive text summarization, a composite vector representation for each document (or each sentence of the document) is first obtained by, for example, averaging or concatenating the vector representations of words occurring in it. Following that, the relevance degree between any pair of the document to be summarized and one of its sentences for summary sentence selection can be determined by a relevance measure such as the cosine similarity between the vector representations of them.

B. *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.

C. 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 thereafter selects an optimal sentence combination by using integer linear programming [15]. By doing so, ILP intends to select a preferred set of summary sentences that can retain the most important theme of a given document. Although ILP is an NP-hard problem, some exact algorithms (such as branch-and-bound) can be exploited 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 [27, 28].

D. Issues Specifically for Speech Summarization

Most of the above-mentioned methods can be applied to both text and speech summarization; the latter, however, presents unique difficulties, such as speech recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries [4, 5, 12, 17]. Empirical evidence has suggested that speech recognition errors seem to be the predominant factor for the performance degradation of speech summarization when using speech recognition transcripts instead of manual transcripts, whereas erroneous sentence boundaries cause relatively minor problems. To mitigate this problem, a recent line of research on speech summarization has been to explore different ways for robustly representing the recognition hypotheses of spoken documents, such as the use of word lattices, confusion networks, *N*-best lists and subword-level indexing mechanisms [3], beyond the continued and tremendous efforts made to improve speech recognition accuracy [1]. Yet another school of thought has been dedicated to estimating the relevance between a spoken document to be summarized and its sentences, as well as the redundancy among these sentences, based on some repeatedly occurring speech patterns embedded in the acoustic signal of the document, without recourse to a speech recognition system for generating the corresponding speech recognition transcript [29, 30]. Furthermore, extra prosodic (acoustic) features, e.g., intonation, pitch, formant, energy, and pause duration, can provide important clues for speech summarization. Some recent work has revealed that exploring more non-lexical features such as the prosodic features would be likely to be beneficial for speech summarization especially when the speech recognition accuracy is not perfect [4, 5], albeit that reliable and efficient ways to use such features remain awaiting further investigation.

III. Language Modeling Based Methods

Intuitively, extractive speech summarization could be cast 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 regarded as a candidate information unit to be retrieved according to its relevance (or importance) to the information need. As such, the primary goal of extractive speech summarization could be stated as the selection of the most representative sentences that can succinctly describe the main topics 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 [31, 32]; this modeling approach has subsequently been applied to extractive speech summarization recently [19-22].

A. Unigram Language Modeling

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 [33]:

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

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 [19], *P*(*S*) is assumed to be uniformly distributed (or identical) for all sentences, unless otherwise stated. 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, $D{=}w_1,w_2,\ldots,w_L$, where words are further assumed to be conditionally independent given the sentence and their order is assumed to be of no importance (i.e.,

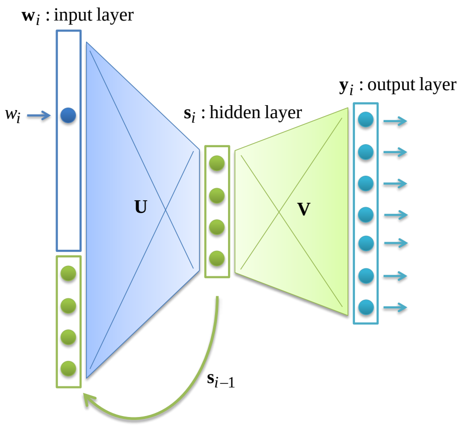
$\mathbf{w}_i$: input layer
w_i
$\mathbf{s}_i$: hidden layer
$\mathbf{y}_i$: output layer
U
V
$\mathbf{s}_{i-1}$

Fig 1. A schematic depiction of the fundamental network of RNNLM.

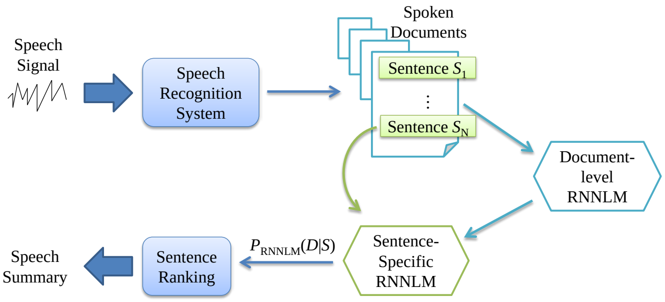

Fig 2. A sketch of the proposed RNNLM summarization framework.

the so-called "*bag-of-words*" assumption), then $P(D|S)$ can be approximated by

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

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 $P(w_i|S)$, 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 [31, 33]:

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

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.

B. *Recurrent Neural Network Language Modeling*

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 an additional indication of the semantic-relatedness of a given sentence with regard to the document to be summarized. 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]. 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 [34-36]. The fundamental network of RMMLM is schematically depicted in Fig. 1, which consists of three main ingredients: the input layer, the hidden layer and the output layer. 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., $\mathbf{s}_{i\text{-}1}$, 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 $\mathbf{y}_\text{i}$ are, respectively, represented or computed as follows [34-36]

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

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

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

where $f(\cdot)$ and $g(\cdot)$ are pre-defined sigmoid activation functions and softmax functions, respectively. Finally, the model parameters (i.e., **U** and **V**) of RNNLM can be derived by maximizing the likelihood of the training corpus using the back-propagation through time (BPTT) algorithm [37-39] that virtually unfolds the feedback loop of RNNLM making its model structure bear a close resemblance to the family of so-called deep neural networks [40] and thereby learn to

TABLE I
TRAINING OF RNNLM-BASED SENTENCE MODELS AND THE APPLICATION OF THEM FOR IMPORTANT SENTENCE RANKING.

```
Input:
$H$: Number of Hidden Layer Neurons
        $\mathbf{D} = \{D_1, \cdots, D_m, \cdots, D_M\}$
        $D_m = \{S_1^{D_m}, \cdots, S_j^{D_m}, \cdots, S_{|D_m|}^{D_m}\}$

Model Training & Important Sentence Ranking:
 1:  for $D_1$ to $D_M$ do
 2:    document-level RNNLM model training
 3:    $\mathcal{L}(\mathbf{U}_m, \mathbf{V}_m) = \sum_{i=1}^{|D_m|} \log(y_i)$
 4:    for $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ do
 5:      sentence-level RNNLM model training
 6:      $\mathcal{L}\left(\mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}} | \mathbf{U}_m, \mathbf{V}_m\right) = \sum_{i=1}^{|S_j^{D_m}|} \log(y_i)$
 7:    end for
 8:    for $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ do
 9:      calculate document likelihood
10:      $P\left(D_m \middle| S_j^{D_m}\right) = \prod_{i=1}^{|S_j^{D_m}|} P\left(w_i \middle| w_1, \ldots, w_{i-1}, S_j^{D_m}\right)$
11:                         $= \prod_{i=1}^{|S_j^{D_m}|} P\left(w_i \middle| \mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}}, S_j^{D_m}\right)$
12:    end for
13:    Sentence selection according to $P\left(D_m \middle| S_j^{D_m}\right)$
14:  end for
```

remember word usage information for several time steps that is packed into the hidden layer of RNNLM [34, 41].

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, which proceeds in three stages:

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 Stage 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.

TABLE II
THE STATISTICAL INFORMATION OF THE BROADCAST NEWS DOCUMENTS USED FOR THE SUMMARIZATION.

	Training Set	Evaluation Set
Recording Period	Nov. 7, 2001 – Jan. 22, 2002	Jan. 23, 2002 – Aug. 20, 2002
Number of Documents	185	20
Average Duration per Document (in sec.)	129.4	141.3
Avg. Number of words per Document	326.0	290.3
Avg. Number of Sentences per Document	20.0	23.3
Avg. Word Error Rate (WER)	38.0%	39.4%

TABLE III
THE AGREEMENT AMONG THE SUBJECTS FOR IMPORTANT SENTENCE RANKING FOR THE EVALUATION SET.

Kappa	ROUGE-1	ROUGE-2	ROUGE-L
0.544	0.600	0.532	0.527

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

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

A schematic illustration of the proposed RNNLM-based summarization framework is depicted in Fig. 2, while a highlight of the corresponding model training and important sentence ranking procedures is given in Table I. In the following, we elaborate on some important steps involved in Table I. 1) In the initial phase, a desired number of the hidden layer neurons H of each RNNLM and a set of documents $\mathbf{D}$ to be summarized, where each document D_m in $\mathbf{D}$ contains $|D_m|$ sentences (each of which is represented by $S_j^{D_m}$), are given. 2) Then, in the training phase, since the architecture of the prototype RNNLM model is a three-layer neural network, there are two sets of parameters (i.e., $\mathbf{U}_m$ and $\mathbf{V}_m$) for each document D_m to be summarized, which are estimated using the back-propagation through time (BPTT) algorithm (*cf.* Line 3 in Table I). Following that, the model parameters of the sentence-level RNNLM model (i.e., $\mathbf{U}_{S_j^{D_m}}$ and $\mathbf{V}_{S_j^{D_m}}$) for each sentence $S_j^{D_m}$ in D_m is estimated starting from the document-level model parameters (i.e., $\mathbf{U}_m$ and $\mathbf{V}_m$) of D_m obtained from previous step (*cf.* Line 6 in Table I). 3) Finally, in the important sentence ranking phase, we can calculate the document likelihood score offered by each sentence $S_j^{D_m}$ based on the corresponding RNNLM model of $S_j^{D_m}$ (*cf.* Lines 10 and 11 in Table I) and in turn select important sentences of D_m according their the document likelihood scores (*cf.* Line 13 in Table I). Interested readers may also refer to [42-44] for more in-depth discussions on a number of efficient training algorithms developed for RNNLM.

The training strategy described above can also be viewed as an instantiation of curriculum learning [45, 46], which seeks to apply a specific and well-planned ordering of the training data for estimating machine-learning models (such as neural networks) to be better suited for a target application. However, as far as we are aware, there is still not much research on leveraging RNNLM along with the aforementioned curriculum-learning strategy for extractive speech summarization. In this paper, we also make a step further by analyzing and comparing the effectiveness of the RNNLM-based summarization methods with other well-practiced state-of-the-art methods thoroughly.

Also worth mentioning is that there has been an alternative realization of the LM approach to extractive summarization that exploits the KL-divergence to measure, for example, the discrepancy of the word (unigram) distribution in a candidate sentence and that in the original document for important (summary) sentence ranking [20, 22]. With some algebraic manipulations, it is easy to show that the effect of the KL-divergence for important sentence ranking is negatively equivalent to the document likelihood (document unigram probability) generated by the sentence $P(D|S)$ (i.e., the ULM method), once the document model is estimated merely on the basis of the empirical frequency of words in the document. However, it seems to be more straightforward to extend ULM with higher-order language modeling strategies, such as leveraging RNNLM to measuring the relatedness between the document to be summarized and each of its sentences.

IV. Experiments Setup

A. 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 [47], 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. Since broadcast news stories often follow a relatively regular structure as compared to other speech materials like conversations, the positional information would play an important role in extractive summarization of broadcast news stories. We hence chose 20 documents, for which the generation of reference summaries is less correlated with the positional information (or the position of sentences), as the held-out test set to evaluate the general performance of the proposed summarization framework, while another subset of 100 documents the held-out development set for tuning the parameters of the various unsupervised summarization methods compared in the paper.

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 [48]. Table II shows some basic statistics about the spoken documents of the development and evaluation sets, where the average word error rate (WER) obtained for the spoken documents was about 38.1% [49]. A large number of text news documents collected by the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) were used. The documents collected in 2000 and 2001 were used to train *N*-gram language models for speech recognition with the SRI Language Modeling Toolkit [50].

B. Performance Evaluation

Three subjects were asked to create extractive 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 [51]. 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 [51].

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, unless otherwise stated. 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 III 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 extractive 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.

V. Experimental Results

A. Baseline Experiments

In the first place, we report on 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

TABLE IV

SUMMARIZATION RESULTS ACHIEVED BY A FEW WELL-STUDIED OR/AND STATE-OF-THE-ART UNSUPERVISED METHODS.

Method	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ULM	0.411	0.298	0.362	0.361	0.215	0.311
VSM	0.347	0.228	0.290	0.342	0.189	0.287
LSA	0.362	0.233	0.316	0.345	0.201	0.301
MMR	0.368	0.248	0.322	0.366	0.215	0.315
W2V	0.372	0.238	0.311	0.364	0.215	0.311
MRW	0.412	0.282	0.358	0.332	0.191	0.291
LexRank	0.413	0.309	0.363	0.305	0.146	0.254
Submodularity	0.414	0.286	0.363	0.332	0.204	0.303
ILP	0.442	0.337	0.401	0.348	0.209	0.306

TABLE V

SUMMARIZATION RESULTS ACHIEVED BY VARIOUS LM-BASED METHODS, INCLUDING ULM, BLM, PLSA, PLSA+ULM, RNNLM AND RNNLM+ULM.

Method	Text Documents (TD)				Spoken Documents (SD)			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
ULM	0.411	0.298	0.362	0.300	0.361	0.215	0.311	0.214
BLM	0.411	0.298	0.362	0.300	0.361	0.215	0.311	0.214
PLSA	0.382	0.260	0.350	0.266	0.327	0.188	0.284	0.189
PLSA+ULM	0.433	0.317	0.379	0.320	0.378	0.234	0.332	0.226
RNNLM	0.433	0.319	0.390	0.319	0.330	0.184	0.294	0.180
RNNLM+ULM	0.533	0.439	0.483	0.430	0.439	0.304	0.393	0.289

summarization methods, including the vector-space methods (i.e., VSM, LSA, MMR, and W2V), the graph-based methods (i.e., MRW and LexRank) and the combinational optimization methods (Submodularity and ILP). The corresponding summarization results of these unsupervised methods are illustrated in Table IV, 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 Table IV. 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, MMR, and W2V) 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 [32], in turn, will be performed based on these problematic measures, potentially leading to common results. Second, LSA and W2V, representing the sentences of a spoken document to be summarized and the document itself in a low-dimensional continuous 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 and ILP achieve the best results in the TD case, while the latter outperforms the former by a considerable margin. However, the superiority of these two methods seems to diminish for the SD case, again probably due to the effect of speech recognition errors. Fourth, the ULM method shows results that are competitive to those obtained by the other state-of-the-art unsupervised methods compared in this paper, which indeed justifies the viability of applying the language modeling approach for speech summarization. Lastly, there is a sizable performance gap between the TD and SD cases for all the above methods, indicating room for further improvements.

B. Experiments on Higher-order N-gram and Topic Language Modeling

In the second set of experiments, we first 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 a document to be summarized (denoted by BLM hereafter). As elaborated before (*cf.* Section III), 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 achieved by the BLM method are depicted in Table V. To our surprise, the integration of bigram and unigram cues together (i.e., BLM) for sentence modeling only arrives at the same performance level as that using the unigram information alone (i.e., ULM) for both the TD and SD cases. 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

TABLE VI
SUMMARIZATION RESULTS RESPECTIVELY ACHIEVED BY ULM AND RNNLM+ULM WITH RESPECT TO DIFFERENT SUMMARIZATION RATIOS.

Method	Summarization Ratio	Text Documents (TD)			Spoken Documents (SD)		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ULM	10%	0.411	0.298	0.361	0.364	0.210	0.307
	20%	0.483	0.368	0.420	0.428	0.255	0.355
	30%	0.551	0.432	0.481	0.471	0.304	0.399
RNNLM+ULM	10%	0.533	0.439	0.483	0.439	0.304	0.393
	20%	0.580	0.478	0.522	0.491	0.341	0.428
	30%	0.639	0.540	0.574	0.514	0.354	0.445

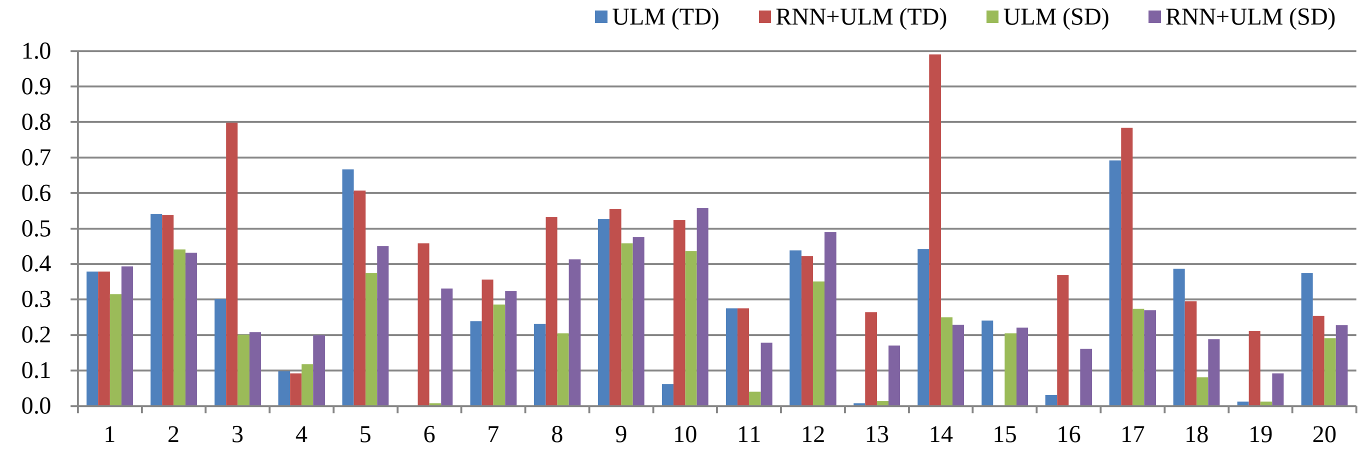

FIG 3. SUMMARIZATION RESULTS (IN ROUGE-2) FOR EACH INDIVIDUAL DOCUMENT (REPRESENTED WITH EITHER MANUAL OR SPEECH TRANSCRIPT) IN THE TEST SET, RESPECTIVELY, ACHIEVED BY ULM AND RNNLM+ULM.

of model parameters would be at most the square of that of the latter. As a side note, we have also experimented on using a trigram language model, smoothed with both unigram and bigram language models, to represent each spoken sentence; however, it delivered almost negligible improvements over the ULM and BLM methods.

Instead of constructing the sentence models based on literal term information (such as the statistics of word unigrams or bigrams), we also exploit probabilistic topic models to represent sentences through a latent topic space. For example, each sentence of a spoken document to be summarized is interpreted as a probabilistic latent semantic analysis (PLSA) model [31] consisting of a set of K shared latent topics $\{T_1,\ldots,T_k,\ldots,T_K\}$ with sentence-specific topic weights $P(T_k|S)$, while each topic offers a unigram (multinomial) distribution $P(w_i|T_k)$ for observing an arbitrary word w_i of the vocabulary:

$$P_{\text{PLSA}}(D\,|\,S)=\prod_{i=1}^{L}[\sum_{k=1}^{K}P(w_i\,|\,T_k)P(T_k\,|\,S)], \qquad (8)$$

where the probability $P(w_i|T_k)$ can be estimated beforehand based on a large set of text or speech documents, while the probability $P(T_k|S)$ of each sentence can be estimated on-the-fly during the summarization process using the expectation-maximization (EM) algorithm [31, 33]. The resulting sentence-specific PLSA model can be used in isolation (denoted by PLSA), or in linear combination with the unigram language model (denoted by PLSA+ULM), to compute the sentence generative probability for important sentence selection. As indicated in Table V, PLSA alone cannot match the performance of ULM, largely because PLSA only offers coarse-grained concept clues about the sentences at the expense of losing discriminative power among concept-related words in finer granularity. On the other hand, the combination of PLSA with ULM (PLSA+ULM) can lead to noticeable improvements as compared to that using either PLSA or ULM alone.

C. *Experiments on the Proposed RNNLM Summarizer*

In the third set of our experiments, we evaluate the effectiveness of the 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 Table V as well. In order to verify the utility of RNNLM and RNNLM+ULM in capturing long-distance word co-occurrence relationships (especially when compared to the other LM-based methods), we additionally include the summarization results evaluated with the ROUGE-SU4 (skip-bigram with maximum gap length of 4) metric in Table V [51]. ROUGE-SU4 is a frequently-used metric for summarization performance evaluation, which quantifies the degree of overlap between the

TABLE VII

Summarization results achieved by the proposed framework and a few well-studied or/and state-of-the-art unsupervised methods, which were measured by using the abstractive summaries written by the human subjects as the ground truth.

	Text Documents (TD)			Spoken Documents (SD)		
Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ULM	0.375	0.231	0.314	0.348	0.178	0.286
VSM	0.325	0.175	0.262	0.325	0.161	0.264
LSA	0.315	0.152	0.254	0.303	0.139	0.243
MMR	0.344	0.193	0.289	0.348	0.182	0.285
W2V	0.366	0.211	0.306	0.345	0.179	0.282
MRW	0.381	0.226	0.316	0.342	0.183	0.283
LexRank	0.312	0.173	0.262	0.281	0.120	0.227
Submodularity	0.394	0.235	0.334	0.336	0.188	0.295
ILP	0.368	0.234	0.317	0.313	0.158	0.268
PLSA+ULM	0.389	0.245	0.327	0.359	0.193	0.299
RNNLM	0.337	0.218	0.297	0.337	0.218	0.297
RNNLM+ULM	0.423	0.281	0.362	0.369	0.218	0.316

reference and automatically generated summaries in terms of not only unigrams but also distant skip-bigrams.

Comparing to the existing LM-based methods (i.e., ULM BLM, PLSA and PLSA+ULM) or the subcategories of unsupervised methods (*c.f.* Table IV), we can find that RNNLM+ULM consistently and significantly surpasses all the other models in both the 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 Table V. 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, ROUGE-L and ROUGE-SU4 metrics. On the other hand, RNNLM and ULM seem to be complementary of each other and indeed can conspire to obtain better sentence modeling. Furthermore, when we compare RNNLM (or RNNLM+ULM) with BLM, the experimental results demonstrate the obvious superiority of RNNLM that might be attributed to two causes. One is that RNNLM has the inherent advantage for capturing long-span structure information in a natural but systematic way. The other is that RNNLM can mitigate the data scarcity problem by implicitly performing clustering of words aside their histories (or preceding words) into a lower-dimensional continuous space, which makes the language model prediction (or probability calculation) based on such compact representations of words aside their histories become more robust [33, 52]. One thing to note is that we have also tried to combine ULM, PLSA and RNNLM together for achieving better summarization accuracy; however, such an attempt only leads to roughly comparable performance as RNNLM+ULM. It is thus believed that the way to systemically combine these models is still a challenging issue and needs further in-depth investigation and proper experimentation. Figure 3 depicts the summarization results (in ROUGE-2) for each individual document (represented with either manual or speech transcript) in the test set, achieved by ULM and RNNLM+ULM. A closer look at these results also reveals that RNNLM+ULM can indeed boost the performance of ULM significantly for most of the test documents that are more difficult to be summarized (for example, Documents 6, 13, 16 and 19 in the test set). In order to further assess the quality of the automatically generated summaries of our RNNLM-based methods and the other state-of-the-art methods compared in this paper, we also take an additional set of abstractive summaries written by the same three human subjects as the ground truth for performance evaluation. For this purpose, the human subjects were instructed to do human summarization, respectively, by writing an abstract for each document with a length (in words) being roughly 25% of the original broadcast news story. The corresponding results are shown in Table VII, which indicate that RNNLM+ULM can provide consistent and significant gains over the other methods as well, even though the reference summaries being used are the human-generated abstractive summaries instead of the human-generated extractive summaries.

D. More Empirical Analysis of the RNNLM Summarizer

To gain more insights into the merit of the RNNLM-based summarization framework, we additionally conduct empirical performance analysis on the RNNLM summarizer from three different aspects. First, we assess the statistical significance of the improvements that are delivered by RNNLM+ULM over ULM with the Student’s paired *t*-test, which confirms that RNNLM+ULM indeed significantly outperforms ULM (with the *p*-values smaller than 0.005 for both the TD and SD cases). Second, to further confirm such superiority of RNNLM+ULM over ULM, we also conduct speech summarization with different summarization ratios (i.e., 20% or 30%), in addition the default setting of 10%; the corresponding results are shown

TABLE VIII

SUMMARIZATION RESULTS ACHIEVED BY RNNLM+ULM WITH RESPECT TO DIFFERENT NUMBERS OF HIDDEN-LAYER NEURONS BEING USED.

Number of Neurons	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
25	0.526	0.436	0.474	0.439	0.304	0.393
50	0.533	0.439	0.483	0.432	0.296	0.385
100	0.465	0.359	0.474	0.426	0.289	0.373
150	0.492	0.386	0.439	0.407	0.263	0.358
200	0.428	0.310	0.376	0.425	0.281	0.374

TABLE IX

SUMMARIZATION RESULTS ACHIEVED BY RNNLM+ULM, MMR, ILP AND THEIR COMBINATIONS.

Method	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
RNNLM+ULM	0.533	0.439	0.483	0.439	0.304	0.393
MMR	0.368	0.248	0.322	0.366	0.215	0.315
ILP	0.442	0.337	0.401	0.348	0.209	0.306
RNNLM+ULM+MMR	0.538	0.450	0.489	0.445	0.312	0.395
RNNLM+ULM+ILP	0.554	0.465	0.505	0.444	0.312	0.399

TABLE X

SUMMARIZATION RESULTS ACHIEVED BY ULM, RNNLM AND RNNLM+ULM IN CONJUNCTION WITH SYLLABLE-LEVEL INDEX FEATURES.

Method	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ULM	0.444	0.327	0.393	0.370	0.217	0.310
RNNLM	0.497	0.398	0.454	0.396	0.248	0.345
RNNLM+ULM	0.558	0.468	0.514	0.474	0.337	0.426

in Table VI. It is evident that RNNLM+ULM consistently leads to marked improvements over ULM for summarization ratios of 20% and 30%, in terms of all the three ROUGE metrics; significance tests, again, indicate the statistical significance of such improvements. Third, we turn to investigate the impact of the model complexity of RNNLM (more specifically, the number of hidden neurons being used) on the ultimate summarization performance. As revealed by results shown in Table VIII, using a small number of hidden neurons (i.e., 25 or 50) seems to be adequate for the speech summarization task studied here. This can be attributed to the fact that since each sentence of a spoken document to be summarized usually consists of only a few words, the RNNLM model of each sentence, which has smaller complexity, tends to have more reliable estimation of its model parameters. Nevertheless, the way to systemically determine the optimal number of hidden-layer neurons of RNNLM for each spoken document to be summarized remains an open issue and needs further investigation. On the other hand, we have also experimented on deepening the architecture of our RNNLM model to be a four-layer network [42], which was in turn used to couple with our proposed training strategy for the modeling of each spoken sentence. Unfortunately, such a deeper RNNLM architecture only yielded mixed summarization results as compared to the three-layer RNNLM architecture we adopted in this paper.

E. Further Extensions on RNNLM Summarizer

A potential downside of our proposed RNNLM-based summarization framework is that the resulting summarizer performs important sentence ranking and selects the top-ranked sentences to form a summary simply based on (in decreasing order of) the relevance measure between a spoken document to be summarized and each sentence in the document (namely, the likelihood that the RNNLM+ULM (or RNNLM alone) model of each sentence generates the document; *cf.* Eq. (7)), without taking into account the relationships among sentences. However, it is generally expected that a desirable summary should not only include highly topic-relevant sentences as many as possible, but at the same time try to reduce the redundancy among these selected sentences as much as possible. To remedy this situation, we further explore to integrate the relevance measure provided by RNNLM+ULM into other state-of-the-art unsupervised summarizers that simultaneously consider the issues of topic coverage and redundancy removal during the summarization process. Here we take MMR [11] and ILP [15] as two examples for the purpose of exploration. For MMR, we use the RNNLM+ULM based measure to replace the original cosine similarity measure involved in the iterative selection process of MMR (denoted by RNNLM+ULM+MMR). On the other hand, for ILP, the RNNLM+ULM based measure is employed not only to

TABLE XI
FOUR TYPES OF ACOUSTIC FEATURES USED TO REPRESENT EACH SPOKEN SENTENCE.

1. Pitch Value (min, max, diff, avg.)
2. Peak Normalized Cross-correlation of Pitch Value (min, max, diff, avg.)
3. Energy Value (min, max, diff, avg.)
4. Duration Value (min, max, diff, avg.)

compute the importance (relevance) weights between any pair of the document to be summarized and one of its sentences, but also to estimate the redundancy degree involved in the constrained combinational optimization process of ILP (denoted by RNNLM+ULM+ILP). Their corresponding results are shown in Table IX. From these results, it is obvious that these two simple integrated methods can bring substantial gains to MMR and ILP, respectively, while they also considerably boost the summarization performance of using RNNLM+ULM in isolation, especially for the TD case. These results again corroborate the intuition that a good extractive summary should contain relevant and diverse sentences that cover the main topics or aspects of an original spoken document.

F. RNNLM with Syllable-level Index Units

In an attempt to mitigate the summarization performance degradation caused by imperfect speech recognition, we explore to make possible use of subword-level index units for the proposed RNNLM-based methods. To do this, syllable pairs are taken as the basic units for indexing instead of words. The recognition transcript of each spoken document, in form of a word stream, was automatically converted into a stream of overlapping syllable pairs. Then, all the distinct syllable pairs occurring in the spoken document collection were then identified to form a vocabulary of syllable pairs for indexing. We can thus use the syllable pairs (as a surrogate of words) to represent the spoken documents and sentences, and subsequently construct the associated summarization models of disparate methods based on such representations. The corresponding results for both the TD and SD cases, achieved by ULM, RNNLM and RNNLM+ULM in conjunction with syllable-level index units, are shown in Table X. We may draw attention to two observations here. First, the results, in general, have consistent trends with the previous sets of experiments where the documents are indexed with words (*c.f.* Table V). Second, the subword-level (syllable-level) index units seem to show comparable or even better performance than the word-level index units (*c.f.* Table V) when being used with the RNNLM-based methods for performing summarization with imperfect speech recognition transcripts (i.e., for the SD case). We conjecture this is because subword-level index units work more robustly against speech recognition errors and the out-of-vocabulary problem, thus likely leading to better summarization performance.

TABLE XII
SUMMARIZATION RESULTS ACHIEVED BY USING ACOUSTIC FEATURES IN ISOLATION AND ITS COMBINATION WITH ULM, RNNLM AND ULM+RNNLM BASED SENTENCE RANKING SCORES, RESPECTIVELY.

	Spoken Documents (SD)			
Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
SVM(AC)	0.373	0.235	0.332	0.220
SVM(AC+ULM)	0.378	0.236	0.335	0.224
SVM(AC+RNNLM)	0.387	0.250	0.344	0.239
SVM(AC+ULM+RNNLM)	0.407	0.268	0.363	0.255

G. Coupling RNNLM with Extra Acoustic Features

In the final set of experiments, we explore the potential of extracting extra acoustic features inherent in spoken sentences for use in summarization. To this end, we use a set of sixteen indicative features crafted based on four commonly-used types of acoustic values, as outlined in Table XI, to characterize a spoken sentence. In implementation, the acoustic features were extracted from the spoken sentences using the Praat toolkit [53]. Interested readers may refer to [54] for detailed accounts on the characteristics of these features and comparisons among them. Here SVM is chosen as the exemplar summarizer to integrate these derived acoustic features (i.e., taking them as the input that represents each sentence) for important spoken sentence ranking. The corresponding model was trained beforehand with the development set in a supervised manner, and the resulting SVM summarizer is denoted by SVM(AC) hereafter. Furthermore, we also study to take the ranking score of ULM, RNNLM and ULM+RNNLM implemented with syllable-level index units, respectively, as an additional indicative feature fed into SVM to represent each sentence (note that the score corresponds to the normalized document likelihood in the logarithmic domain, predicted by the respective sentence generative model), leading to an augmented set of seventeen features in total. The resulting SVM summarizers are denoted by SVM(AC+ULM), SVM(AC+RNNLM), and SVM(AC+ULM+RNNLM), respectively. Table XII shows the results of these summarizers for the SD case, from which at least two observations can be drawn. First, SVM(AC) exhibits superior performance over all the unsupervised summarizers compared in this paper, except for ULM+RNNLM and its variants (*cf.* Tables IV, V and IX). Unlike the unsupervised summarizers, SVM(AC), however, requires human annotation in the training phase. Second, SVM(AC+ULM), SVM(AC+RNNLM), and SVM(AC+ULM+RNNLM) all yield better performance than SVM(AC). Although SVM(AC+ULM+RNNLM) stands out in performance among these SVM-based summarizers, to our surprise, it does not in general operate as effectively as ULM+RNNLM and its variants (implemented with either word- or syllable-level index units). This means that the way to systemically combine the acoustic features with other indicative features (especially those seemingly superior-performing ones) for important spoken sentence selection remains a challenging issue and needs further in-depth investigation and proper experimentation.

VI. Conclusion and Outlook

In this paper, we have proposed a novel recurrent neural network language modeling (RNNLM) framework for performing speech summarization in an unsupervised manner. We have elaborated on how the notion of RNNLM can be crystallized so as to render both word usage cues and long-span word co-occurrence relationships that are deemed beneficial for speech summarization. Furthermore, the merits of the methods originated from our framework have also been validated by extensively comparisons with several state-of-the-art unsupervised summarization methods. Our future research directions include: 1) exploring the use of different kinds of prosodic, lexical and semantic information cues that can be incorporated into this framework so as to improve the empirical effectiveness of the RNNLM-based summarization methods, 2) developing robust indexing and confidence measuring techniques [55, 56] that can work in tandem with our summarization methods, and 3) integrating our summarization methods with other more sophisticated summary sentence selection criteria that are more closely coupled with the ultimate evaluation metrics of speech summarization.

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

[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] 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.

[20] 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.

[21] 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.

[22] 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.

[23] K.-Y. Chen *et al.*, “A recurrent neural network language modeling framework for extractive speech summarization,” in *Proc. the IEEE International Conference on Multimedia & Expo*, 2014.

[24] Y. Bengio *et al.*, “A neural probabilistic language model,” *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[25] T. Mikolov *et al.*, “Distributed representations of words and phrases and their compositionality,” in *Proc. Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

[26] M. Kageback *et al.*, “Extractive Summarization using Continuous Vector Space Models,” in *Proc. the Workshop on Continuous Vector Space Models and their Compositionality* (CVSC), pp. 31–39, 2014.

[27] D. Gillick and B. Favre, “A scalable global model for summarization,” in *Proc. the Workshop on Integer Linear*

Programming for Natural Language Processing, pp. 10–18, 2009.

[28] C. Li *et al.*, "Using supervised bigram-based ILP for extractive summarization," in *Proc. the Annual Conference of the International Speech Communication Association*, pages 1004–1013, August, 2013.

[29] A. Park and J. Glass, "Unsupervised word acquisition from speech using pattern discovery," in *Proc. the IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 409–412, 2006.

[30] X. Zhu *et al.*, "Summarizing multiple spoken documents: finding evidence from untranscribed audio," in *Proc. the Annual Meeting of the Association for Computational Linguistics*, pp. 549–557, 2009.

[31] 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.

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

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

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

[35] 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.

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

[37] M. Boden, "A guide to recurrent neural networks and backpropagation," in the Dallas Project, 2002.

[38] H. Jaeger, "A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach," GMD Report 159, German National Research Center for Information Technology, 2002.

[39] R. Pascanu *et al.*, "On the difficulty of training recurrent neural networks," in *Proc. JMLR:W&CP*, pp. 1310–1318, 2013.

[40] D. Li and D. Yu, *Deep Learning: Methods and Applications*, Foundations and Trends in Signal Processing, Now Publishers, June 2014.

[41] Y. Bengio *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[42] S. Kombrink *et al.*, "Recurrent neural network based language modeling in meeting recognition," in *Proc. Annual Conference of the International Speech Communication Association*, pp. 2877–2880, 2011.

[43] H.-S. Le *et al.*, "Large vocabulary SOUL neural network language models," in *Proc. Annual Conference of the International Speech Communication Association*, pp. 1469–1472, 2011.

[44] X. Liu *et al.*, "Efficient lattice rescoring using recurrent neural network language models," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4941–4945, 2014.

[45] Y. Bengio *et al.*, "Curriculum learning," in *Proc. Annual International Conference on Machine Learning*, pp. 41–48, 2009.

[46] J. L. Elman, "Learning and development in neural networks: the importance of starting small," *Cognition*, vol. 48, pp. 71–99, 1993.

[47] 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.

[48] 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.

[49] B. Chen, *et al.*, "Lightly supervised and data-driven approaches to Mandarin broadcast news transcription," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 777–780, 2004.

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

[51] C. Y. Lin, "ROUGE: Recall-oriented understudy for gisting evaluation." 2003 [Online]. Available: http://haydn.isi.edu/ROUGE/.T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," in *Proc. International Conference on Learning Representations*, pp. 1–12, 2013.

[52] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, Vol. 5, No. 9-10, pp. 341-345, 2001.

[53] S. H. Lin *et al.*, "A comparative study of probabilistic ranking models for Chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing*, Vol. 8, No. 1, pp. 3:1–3:23, 2009.

[54] S. Xie, and Y. Liu, "Using N-best lists and confusion networks for meeting summarization" *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1160–1169, 2011.

[55] C. Chelba *et al.*, "Soft indexing of speech content for search in spoken documents," *Computer Speech & Language*, vol. 21, no.3, pp. 458–478, 2007.

Kuan-Yu Chen (S'13) received the B.S. and M.S. degree in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2007 and 2010, respectively. He is currently pursuing his Ph.D. degree in the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. In November 2010, he joined the Speech, Language, and Music Processing Laboratory, Institute of Information Science, Academia Sinica, Taiwan, as a research assistant. His research interests are in the area of language modeling, large-vocabulary continuous speech recognition, information retrieval and natural language processing.

Shih-Hung Liu (S'13) received the B.S. and M.S. degree in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2003 and 2007, respectively. He is currently pursuing his Ph.D. degree in the Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan. In October 2007, he joined the Intelligent Agent Systems Laboratory, Institute of Information Science, Academia Sinica, Taiwan, as a research assistant. His research interests are in the area of speech summarization and natural language processing.

Berlin Chen (M'04) received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1994 and 1996, respectively, and the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, in 2001. He was with the Institute of Information Science, Academia Sinica, Taipei, from 1996 to 2001, and then with the Graduate Institute of Communication Engineering, National Taiwan University, from 2001 to 2002. In 2002, he joined the Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei. He is currently a Professor in the Department of Computer Science and Information Engineering of the same university. His research interests generally lie in the areas of speech and natural language processing, information retrieval, and artificial intelligence. He is the author/coauthor of over 100 academic publications.

Hsin-Min Wang (S'92–M'95–SM'04) received the BS and PhD degrees in electrical engineering from National Taiwan University in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, where he is currently a research fellow and deputy director. He also holds a joint appointment as professor in the Department of Computer Science and Information Engineering at National Cheng Kung University. He currently serves as the president of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), a managing editor of Journal of Information Science and Engineering, and an editorial board member of APSIPA Transactions on Signal and Information Processing and International Journal of Computational Linguistics and Chinese Language Processing. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, and pattern recognition. He received the Chinese Institute of Engineers (CIE) Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He is an APSIPA distinguished lecturer for 2014-2015. He is a member of the International Speech Communication Association (ISCA) and ACM. He is a senior member of the IEEE.

Ea-Ee Jan has been a Research Staff Member at IBM T. J. Watson Research Center since 1996. He worked on the area of speech recognition, statistical modeling, natural language processing, machine translation and multi-lingual information retrieval. Since 2011, he has joined the Service Innovation Lab in IBM Research. He has expanded his research areas in virtual machine image analytics for cloud computing, IT service delivery analytics and contracts risk analytics. He has published more than 45 technical papers, holds 11 US patents and won numerous IBM awards. Prior to joining IBM, Dr. Jan was a Research Assistant Professor at Rutgers University. He received the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Rutgers University, New Brunswick NJ, in 1992 and 1995, respectively.

Wen-Lian Hsu (F'06) is currently the Director and a Distinguished Research Fellow of the Institute of Information Science, Academia Sinica, Taiwan. He received a B.S. in Mathematics from National Taiwan University, and a Ph.D. in operations research from Cornell University in 1980. He was a tenured associate professor in Northwestern University before joining the Institute of Information Science in Academia Sinica as a research fellow in 1989. Dr. Hsu's earlier contribution was on graph algorithms and he has applied similar techniques to tackle computational problems in biology and natural language. In 1993, he developed a Chinese input software, GOING, which has since revolutionized Chinese input on computer. Dr. Hsu is particularly interested in applying natural language processing techniques to understanding DNA sequences as well as protein sequences, structures and functions and also to biological literature mining. Dr. Hsu received the Outstanding Research Award from the National Science Council in 1991, 1994, 1996, the first K. T. Li Research Breakthrough Award in 1999, the IEEE Fellow in 2006, and the Teco Award in 2008. He was the president of the Artificial Intelligence Society in Taiwan from

2001 to 2002 and the president of the Computational Linguistic Society of Taiwan from 2011 to 2012.”

Hsin-Hsi Chen is a professor in Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. He served as the chair of the department from 2002 to 2005. From 2006 to 2009, he was the chief director of Computer and Information Networking Center. His research interests are computational linguistics, Chinese language processing, information retrieval and extraction, and web mining. He will be the demo co-chair of ACL-IJCNLP 2015. Besides, he was the general chair of IJCNLP 2013 and the program co-chair of ACM SIGIR 2010. He also served as a senior PC member of ACM SIGIR (2006–2009), area/track chairs of ACL 2012, ACL–IJCNLP 2009 and ACM CIKM 2008, and PC members of many conferences (IJCAI, SIGIR, WSDM, AIRS, ACL, COLING, EMNLP, NAACL, EACL, IJCNLP, WWW, and so on). He has won Google research awards in 2007 and 2012.