INCORPORATING PARAGRAPH EMBEDDINGS AND DENSITY PEAKS CLUSTERING FOR SPOKEN DOCUMENT SUMMARIZATION

Kuan-Yu Chen[*], *Kai-Wun Shih*[#], *Shih-Hung Liu*[*], *Berlin Chen*[#], *Hsin-Min Wang*[*]

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

E-mail: [*]{kychen, journey, whm}@iis.sinica.edu.tw, [#]{60247065S, berlin}@csie.ntnu.edu.tw

ABSTRACT

Representation learning has emerged as a newly active research subject in many machine learning applications because of its excellent performance. As an instantiation, word embedding has been widely used in the natural language processing area. However, as far as we are aware, there are relatively few studies investigating paragraph embedding methods in extractive text or speech summarization. Extractive summarization aims at selecting a set of indicative sentences from a source document to express the most important theme of the document. There is a general consensus that relevance and redundancy are both critical issues for users in a realistic summarization scenario. However, most of the existing methods focus on determining only the relevance degree between sentences and a given document, while the redundancy degree is calculated by a post-processing step. Based on these observations, three contributions are proposed in this paper. First, we comprehensively compare the word and paragraph embedding methods for spoken document summarization. Next, we propose a novel summarization framework which can take both relevance and redundancy information into account simultaneously. Consequently, a set of representative sentences can be automatically selected through a one-pass process. Third, we further plug in paragraph embedding methods into the proposed framework to enhance the summarization performance. Experimental results demonstrate the effectiveness of our proposed methods, compared to existing state-of-the-art methods.

Index Terms— Spoken document, summarization, embedding, relevance, redundancy

1. INTRODUCTION

Owing to the popularity of various Internet applications, rapidly growing multimedia content, such as music video, broadcast news programs, and lecture recordings, has been continuously filling our daily life [1-4]. Obviously, speech is one of the most important sources of information about multimedia. By virtue of spoken document summarization (SDS), one can efficiently digest multimedia content by listening to the associated speech summary. Extractive SDS manages to select a set of indicative sentences from a spoken document according to a target summarization ratio and concatenate them together to form a summary [5-8].

Representation learning has emerged as an attractive subject of research and experimentation in many machine learning applications because of its remarkable performance. When it comes to the field of natural language processing (NLP), word embedding methods can be viewed as pioneer studies [9-12]. The central idea of these methods is to learn continuously distributed vector representations of words using neural networks, which can probe latent semantic and/or syntactic cues that can in turn be used to induce similarity measures among words. A common thread of leveraging word embedding methods to NLP-related tasks is to represent the paragraph (or sentence and document) by averaging the word embeddings corresponding to the words occurring in the paragraph (or sentence and document). Then, intuitively, the cosine similarity measure can be applied to determine the relevance degree between a pair of representations. By doing so, this school of methods has recently demonstrated promising performance in many NLP-related tasks, such as relational analogy prediction, sentiment analysis, and sentence completion [13-17].

Although the utilities and abilities of word embedding methods have been proven recently, the composite representation for a paragraph (or sentence and document) is a bit queer especially in a manifold space. Theoretically, paragraph (or sentence and document)-based representation learning is more suitable/reasonable for some tasks, such as information retrieval and document summarization [18-21]. In this paper, we thus provide a thorough comparison of the word and paragraph embedding methods for extractive SDS. On the other hand, it is generally agreed upon that relevance and redundancy are two key aspects for generating a concise summary [5, 7, 8, 22, 23]. However, most of the existing methods only focus on determining the relevance degree between sentences and a given document, while the redundancy is tackled by an additional post-processing step. Beyond the continued and tremendous efforts made to measure the relevance between sentences and a given document, this paper proposes a novel and efficient summarization framework that can take both relevance and

redundancy information into account simultaneously for generating a concise extractive summary.

The remainder of this paper is organized as follows. We first briefly review some related work on extractive summarization in Section 2. Section 3 introduces the notion of leveraging the word and paragraph embedding methods for extractive SDS. After that, Section 4 sheds light on our proposed summarization framework to further improve the summarization performance. Finally, experimental setup, experimental results and conclusions are presented in Sections 5, 6, and 7, respectively.

2. RELATED WORK

The wide spectrum of extractive SDS methods developed so far spreads from methods simply based on the sentence position or structure information, methods based on unsupervised sentence ranking, to methods based on supervised sentence classification [5, 8].

For the first category, important sentences are selected from some salient parts of a spoken document [24], such as the introductory and/or concluding parts. However, such methods can be only applied to some specific domains with limited document structures. Unsupervised sentence ranking methods attempt to select important sentences based on statistical features of the sentences or of the words in the sentences without human annotations involved. Popular methods include, but are not limited to, the vector space model (VSM) [25], the latent semantic analysis (LSA) method [25], the Markov random walk (MRW) method [26], the maximum marginal relevance (MMR) method [23], the sentence significant score method [27], the language model-based method [28, 29], the LexRank method [30], the submodularity-based method [31], and the integer linear programming (ILP) method [32]. The statistical features may include, for example, the term (word) frequency, linguistic score, recognition confidence measure, and prosodic information. Among them, the ability of reducing redundant information has been aptly incorporated into the submodular-based method and the ILP method, in addition to the MMR method. In contrast, supervised sentence classification methods, such as the Gaussian mixture model (GMM) [25], the Bayesian classifier (BC) [33], the support vector machine (SVM) [34], and the conditional random fields (CRFs) [35], usually formulate sentence selection as a binary classification problem, i.e., a sentence can either be included in a summary or not. Interested readers may refer to [5-8] for comprehensive reviews and new insights into the major methods that have been developed and applied with good success to a wide range of text and spoken document summarization tasks.

3. WORD & PARAGRAPH EMBEDDINGS

3.1. Word Embedding Methods

Perhaps one of the most-known seminal studies on developing word embedding methods was presented in [9]. It estimated a statistical (*n*-gram) language model, formalized as a feed-forward neural network, for predicting future words in context while inducing word embeddings (or representations) as a by-product. Such an attempt has already motivated many follow-up extensions to develop similar methods for probing latent semantic and syntactic regularities in the representation of a word. Representative methods include, among others, the continuous bag-of-words (CBOW) model [11, 36], the skip-gram (SG) model [11, 37], and the global vector (GloVe) model [12].

3.1.1. The Continuous Bag-of-words (CBOW) Model

Rather than seeking to learn a statistical language model, the CBOW model manages to obtain a dense vector representation (embedding) of each word directly [11]. The structure of CBOW is similar to a feed-forward neural network, with the exception that the non-linear hidden layer in the former is removed. By getting around the heavy computational burden incurred by the non-linear hidden layer, the model can be trained on a large corpus efficiently, while still retains good performance. Formally, given a sequence of words, $w^1,w^2,\ldots,w^T$, the objective function of CBOW is to maximize the log-probability,

$$\Sigma_{t=1}^{T} \log P(w^t \mid w^{t-c},\ldots,w^{t-1},w^{t+1},\ldots,w^{t+c}), \quad (1)$$

where c is the window size of contextual words being considered for the central word w^t, T denotes the length of the training corpus, and

$$P(w^t \mid w^{t-c},\ldots,w^{t-1},w^{t+1},\ldots,w^{t+c}) = \frac{\exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w^t})}{\Sigma_{i=1}^{V} \exp(\mathbf{v}_{\overline{w}^t} \cdot \mathbf{v}_{w_i})}, \quad (2)$$

where $\mathbf{v}_{w^t}$ denotes the vector representation of the word w at position t; V is the size of the vocabulary; and $\mathbf{v}_{\overline{w}^t}$ denotes the (weighted) average of the vector representations of the contextual words of w^t [11, 16]. The concept of CBOW is motivated by the distributional hypothesis [36], which states that words with similar meanings often occur in similar contexts, and it is thus suggested to look for w^t whose word representation can capture the distributions of its context well.

3.1.2. The Skip-gram (SG) Model

In contrast to the CBOW model, the SG model employs an inverse training objective for learning word representations with a simplified feed-forward neural network [11, 37]. Given the word sequence, $w^1,w^2,\ldots,w^T$, the objective function of SG is to maximize the log-probability,

$$\Sigma_{t=1}^{T} \Sigma_{j=-c, j\neq 0}^{c} \log P(w^{t+j} \mid w^t), \quad (3)$$

where c is the window size of the contextual words for the central word w^t, and the conditional probability is computed by

$$P(w^{t+j} \mid w^t) = \frac{\exp(\mathbf{v}_{w^{t+j}} \cdot \mathbf{v}_{w^t})}{\Sigma_{i=1}^{V} \exp(\mathbf{v}_{w_i} \cdot \mathbf{v}_{w^t})}, \quad (4)$$

where $\mathbf{v}_{w^{t+j}}$ and $\mathbf{v}_{w^t}$ denote the word representations of the words at positions $t+j$ and t, respectively. In the

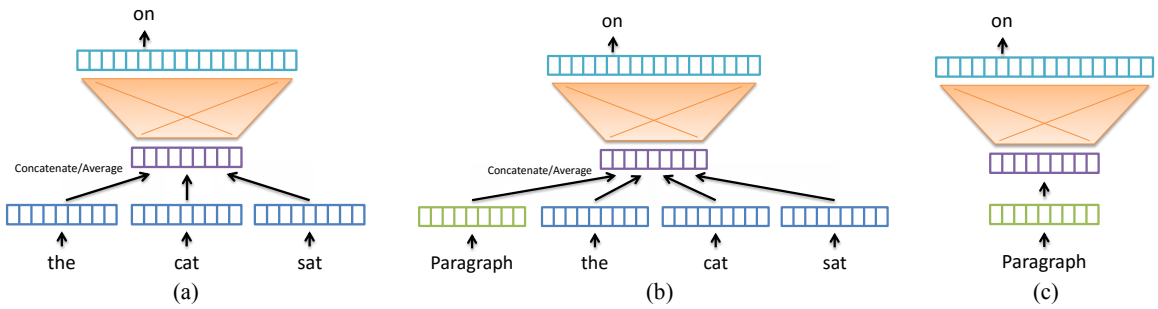

Fig. 1. Illustrations of (a) the feed-forward neural network language model (NNLM), (b) the distributed memory model (DM), and (c) the distributed bag-of-words model (DBOW).

implementations of CBOW and SG, the hierarchical soft-max algorithm [37, 38] and the negative sampling algorithm [37, 39] can make the training process more efficient and effective.

3.1.3. The Global Vector (GloVe) Model

The GloVe model suggests that an appropriate starting point for word representation learning should be associated with the ratios of co-occurrence probabilities rather than the prediction probabilities of words [12]. More precisely, GloVe makes use of weighted least squares regression, which aims at learning word representations by preserving the co-occurrence frequencies between each pair of words:

$$\Sigma_{i=1}^{V}\Sigma_{j=1}^{V} f(X_{w_i w_j})(\mathbf{v}_{w_i}\cdot\mathbf{v}_{w_j}+b_{w_i}+b_{w_j}-\log X_{w_i w_j})^2, \quad (5)$$

where $X_{w_i w_j}$ denotes the number of times words w_i and w_j co-occur in a pre-defined sliding context window; $f(\cdot)$ is a monotonic smoothing function used to modulate the impact of each pair of words involved in model training; and $\mathbf{v}_w$ and b_w denote the word representation and the bias term of word w, respectively.

3.2. Paragraph Embedding Methods

In contrast to the large body of work developing various word embedding methods, there are relatively few studies concentrating on learning paragraph representations [18-21]. Representative methods include the distributed memory (DM) model [19] and the distributed bag-of-words (DBOW) model [18, 19], to name a few. As far as we are aware, there is little work contextualizing these methods for use in speech summarization.

3.2.1. The Distributed Memory (DM) Model

The DM model is inspired and hybridized by the traditional feed-forward neural network language model (NNLM) [9] and the CBOW model. Formally, given a sequence of words, $w^1,w^2,\ldots,w^T$, the objective function of feed-forward NNLM is to maximize the total log-likelihood,

$$\Sigma_{t=1}^{T}\log P(w^t \mid w^{t-n+1},\ldots,w^{t-1}). \quad (6)$$

Obviously, NNLM is designed to predict the probability of the future word, given its n-1 previous words. The input of NNLM is a high-dimensional vector, which is constructed by concatenating (or averaging) the word representations of all words in the context (i.e., $w^{t-n+1},\ldots,w^{t-1}$), and the output can be viewed as that of a multi-class classifier. By doing so, the n-gram probability can be calculated through a softmax function at the output layer:

$$P(w^t \mid w^{t-n+1},\ldots,w^{t-1})=\frac{\exp(y_{w^t})}{\Sigma_{w_i\in V}\exp(y_{w_i})} \quad (7)$$

where y_{w_i} denotes the output value for word w_i. A simple example is shown in Fig. 1(a).

Based on the NNLM, the idea underlying the DM model is that a given paragraph also contributes to the prediction of the next word, given its previous words in the paragraph [19]. To make the idea to go, the training objective function is defined by

$$\Sigma_{i=1}^{\mathbf{T}}\Sigma_{t=1}^{T_i}\log P(w^t \mid w^{t-n+1},\ldots,w^{t-1},D_i). \quad (8)$$

where $\mathbf{T}$ denotes the number of paragraphs in the training corpus, D_i denotes the i-th paragraph, and T_i is the length of D_i. Since it acts as a memory unit that remembers what is missing from the current context, the model is named the distributed memory model. A simple example for the DM model is schematically depicted in Fig. 1(b).

3.2.2. The Distributed Bag-of-Words (DBOW) Model

Opposite to the DM model, a simplified version is to only leverage the paragraph representation to predict all of the words occurring in the paragraph [19]. The training objective function can then be defined by maximizing the predictive probabilities all over the words occurring in the paragraph,

$$\Sigma_{i=1}^{\mathbf{T}}\Sigma_{t=1}^{T_i}\log P(w^t \mid D_i). \quad (9)$$

Since the simplified model ignores the contextual words at the input layer, the model is named the distributed bag-of-words (DBOW) model. In addition to being conceptually simple, the DBOW model only needs to store the softmax weights in contrast to storing both softmax weights and word vectors for the DM model [19]. Fig. 1(c) is a running example to illustrate the architecture of the DBOW model.

3.3. Embedding Methods for extractive SDS

Inspired by the vector space model (VSM), a straightforward way to leverage the word embedding methods for extractive SDS is to represent a sentence S_i (and a document D to be summarized) by averaging the vector representations of words occurring in the sentence S_i (and the document D) [13, 15]:

$$\mathbf{v}_{S_i} = \Sigma_{w \in S_i} \frac{n(w, S_i)}{|S_i|} \mathbf{v}_w. \quad (10)$$

Alternatively, the sentence (and document) representations can be inferred directly by using the paragraph embedding methods introduced in Section 3.2. Consequently, the document D and each sentence S_i of D have a respective fixed-length dense vector representation, and their relevance degree can be evaluated by the cosine similarity measure.

4. A SUMMARIZATION FRAMEWORK

The most common belief in the document summarization community is that relevance and redundancy are two key issues for generating a concise summary. However, existing methods usually focus on determining only the relevance degree between a given document and its sentences, and the redundancy is considered in a post-processing step. Maximum margin relevance (MMR) is the most popularly used criterion for automatic summarization [23], based on which redundancy is computed by comparing a candidate sentence to the already selected sentences, and a greedy post-processing step is performed iteratively to select sentences. To avoid the time-consuming post-processing step, we propose a novel summarization framework, which can take both relevance and redundancy information into account at the same time. That is, a concise summary for a given document is automatically generated through a one-pass process instead of an iterative process.

The idea consists of two aspects: the representative sentences should have 1) a higher *density score* than other sentences and 2) a higher *divergence score* than other sentences that also have high density scores. In the SDS task, the density score for sentence S_i in a document D to be summarized can be defined by:

$$density(S_i) = \frac{1}{K-1} \sum_{j=1, j \neq i}^{K} \chi(sim(S_i, S_j) - \delta) \quad (11)$$

$$\chi(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & otherwise \end{cases} \quad (12)$$

where K is the number of sentences in D, $sim(S_i,S_j)$ is the similarity degree between sentences S_i and S_j, and δ denotes a pre-defined threshold, which is used to determine whether the pair of sentences is relevant to each other or not. After the density score for each sentence is obtained, the divergence scores of the sentences are calculated by

$$divergence(S_i) = 1 - \max_{\substack{\forall S_j \in D \\ density(S_j) > density(S_i)}} sim(S_i, S_j), \quad (13)$$

except that the sentence with the highest density score whose divergence score is set to 1 directly.

The concept behind the proposed framework is explained as follows. The sentences in a document can be classified into some classes, where each class represents a subtopic. Therefore, the cluster centers (i.e., the sentences with higher density scores) can be selected as the representative sentences. At the same time, the divergence score can be used to determine the importance of each subtopic. While the role of the divergence score here is similar to the MMR criterion, the former considers the redundancy information in a more general way than the latter. It is worthwhile to note that the proposed framework selects the representative sentences through a one-pass process. Since the framework is inspired from the density peaks clustering algorithm [40-42], we term it as DPC in short. This is the first time the density peaks clustering algorithm is introduced and evaluated in the SDS task, as far as we are aware.

Unfortunately, the threshold δ for the density score in Eq. (11) is hard to define or tune empirically [41]. In order to remedy the imperfection, a parameter-free variation can be obtained by modifying the density score as

$$density(S_i) = \frac{1}{K-1} \sum_{j=1, j \neq i}^{K} sim(S_i, S_j). \quad (14)$$

The model is then named DPC_sum hereafter.

In practice, the multiplication score (i.e., $density(S_i)*divergence(S_i)$) can be used alone or linearly combined with the conventional relevance score between the sentence and the document (i.e., $sim(S_i,D)$) to select sentences. We use the cosine measure as the similarity score $sim(\cdot,\cdot)$ throughout the paper. The vector representation for a sentence (or a document) is characterized by the conventional term frequency multiplied by the inverse document frequency (TF-IDF) or by inferring through the word or paragraph embedding methods (*c.f.* Section 3).

5. EXPERIMENTAL SETUP

The dataset used in this study is the MATBN broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [43]. The corpus 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 set. 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. Each document has three reference summaries annotated by three subjects. For the assessment of summarization performance, we adopted the widely-used ROUGE metrics [44]. All the experimental results reported hereafter are obtained by

Table 1. Summarization results achieved by the word and paragraph embedding methods.

Method	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
CBOW	0.382	0.249	0.322	0.362	0.214	0.314
SG	0.371	0.239	0.311	0.364	0.215	0.311
GloVe	0.366	0.244	0.310	0.363	0.214	0.310
DM	0.406	0.290	0.355	0.364	0.218	0.313
DBOW	0.418	0.293	0.364	0.375	0.232	0.323

Table 2. Summarization results achieved by the proposed summarization framework.

Method	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
VSM	0.347	0.228	0.290	0.342	0.189	0.287
MMR	0.362	0.238	0.312	0.369	0.218	0.317
DPC	0.409	0.285	0.356	0.352	0.200	0.297
DPC_sum	0.383	0.266	0.336	0.368	0.219	0.316

Table 3. Summarization results achieved by incorporating paragraph embedding with the proposed summarization framework.

Method		Text Documents (TD)			Spoken Documents (SD)		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
DM	+ DPC	0.441	0.339	0.409	0.387	0.242	0.337
	+ DPC_sum	0.445	0.332	0.400	0.376	0.236	0.329
DBOW	+ DPC	0.446	0.334	0.405	0.396	0.250	0.344
	+ DPC_sum	0.418	0.293	0.364	0.375	0.232	0.323

calculating the F-scores [34] of these ROUGE metrics. The summarization ratio was set to 10%. A corpus of 14,000 text news documents, compiled during the same period as the broadcast news documents, was used to estimate related models compared in this paper. A subset of 25-hour speech data from MATBN compiled from November 2001 to December 2002 was used to bootstrap the acoustic training with the minimum phone error rate (MPE) criterion and a training data selection scheme [45]. The vocabulary size is about 72 thousand words. The average word error rate of automatic transcription is about 40%.

6. EXPERIMENTAL RESULTS

To begin with, we investigate the utilities of three popular word embedding methods (i.e., CBOW, SG, and GloVe) and two paragraph embedding methods (i.e., DM and DBOW) for SDS (*c.f.* Section 3.3). The results are shown in Table 1, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the speech recognition transcripts that may contain recognition errors. From Table 1, we can see that the three word embedding methods, though with disparate model structures and learning strategies, achieve comparable results to each other in both the TD and SD cases. DBOW consistently outperforms DM in both the TD and SD cases, though the performance difference is mostly small. The results also indicate that the paragraph embedding methods outperform the word embedding methods as expected in the TD case, while they offer only a small performance gain in the SD case. The reason might be that the recognition errors disturb the paragraph embedding methods more severely than the word embedding methods.

In the next set of experiments, we evaluate the capability of the proposed summarization framework (i.e., DPC and DPC_sum) in improving the performance of SDS. Since DPC (and DPC_sum) is based on the cosine similarity measure, the vector space model (VSM) is treated as the first baseline system. MMR, which considers both relevance and redundancy information when generating a summary, has been widely used in summarization; therefore, the MMR model is treated as another baseline system. All the methods are based on the conventional TF-IDF vector representation, without applying any word and paragraph embedding methods. Moreover, for the DPC and DPC_sum methods, the multiplication score of a sentence is linearly combined with the conventional relevance score between the sentence and the document (*c.f.* Section 4). The results are shown in Table 2. From the table, two observations can be drawn. First, it is clear that DPC, DPC_sum, and MMR outperform VSM in all cases. The results indicate that redundancy is indeed an important issue to SDS. Second, the proposed framework (i.e., DPC and DPC_sum) outperforms MMR by a large margin in the TD case, but only gives comparable performance with MMR in the SD case.

In the last set of experiments, we further integrate the paragraph embedding methods into the proposed summarization framework (i.e., DPC and DPC_sum). The results are shown in Table 3. It is obvious that the results in Table 3 are better than all the results in Tables 1 and 2. Comparing the results in Table 3 to that of the paragraph embedding methods (*c.f.* DM and DBOW in Table 1), it is evident again that redundancy is an important issue to SDS. Comparing the results in Table 3 to that of the basic implementations of the proposed framework (*c.f.* DPC and

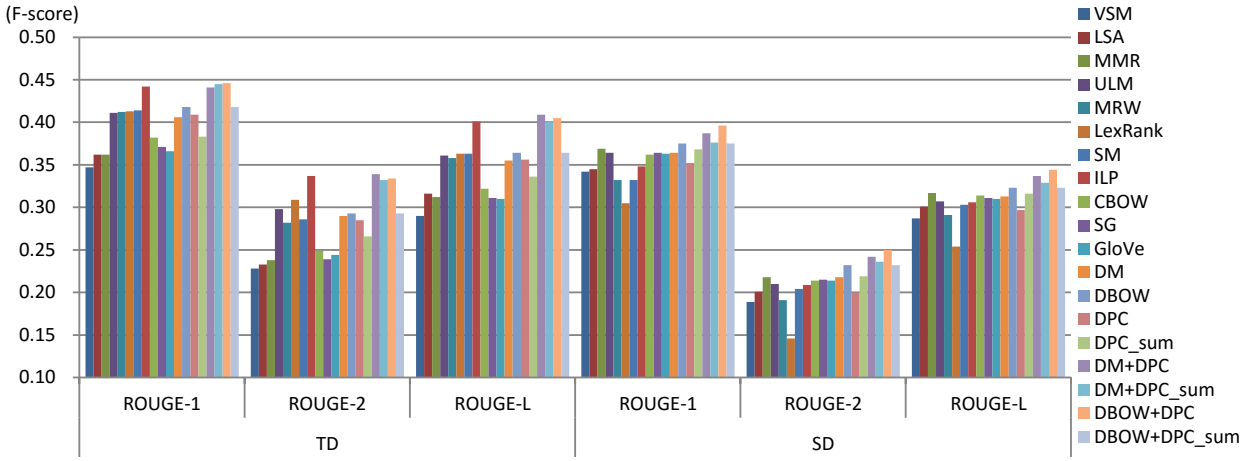

Fig. 2. Summarization results achieved by the proposed methods and well-studied or/and state-of-the-art unsupervised methods.

DPC_sum in Table 2), it is clear that, instead of only considering literal term matching for determining the similarity degree between a pair of sentence and document, incorporating concept (semantic) matching into the similarity measure leads to better performance. Comparing the results in Table 1 to the results of VSM in Table 2 also signals such evidence.

In the last set of experiments, we assess the performance levels of several well-practiced or/and state-of-the-art summarization methods for extractive SDS, including the vector space-based methods (i.e., VSM, LSA and MMR), the unigram language model method (i.e., ULM), the graph-based methods (i.e., MRW and LexRank), the submodularity method (SM), and the integer linear programming method (ILP). The results are illustrated in Fig. 2. Several noteworthy observations can be drawn from the results of various existing methods. First, the two graph-based methods (i.e., MRW and LexRank) are quite competitive with each other and perform better than the vector space-based methods (i.e., VSM, LSA, and MMR) for the TD case. However, for the SD case, the situation is reversed. It reveals that imperfect speech recognition may affect the graph-based methods more seriously than the vector space-based 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 degraded results. Second, LSA, which represents the sentences of a spoken document and the document itself in the latent semantic space instead of the index term (word) space, performs slightly better than VSM in both the TD and SD cases. Third, ILP, which also considers reducing the redundant information at the same time when producing a summary for a given document to be summarized, achieves the best results in the TD case, but only achieves comparable performance to other methods in the SD case. Fourth, ULM shows competitive results compared to other state-of-the-art methods. Comparing the results of existing methods to that of the proposed methods, it is clear that the proposed methods (in particular, DBOW+DPC and DBOW+DPC_sum) are the most robust among all the methods compared in the paper.

7. CONCLUSIONS & FUTURE WORK

In this paper, the paragraph embedding methods have been evaluated for spoken document summarization. In addition, a novel and efficient summarization framework (instantiated with DPC or its simplified version DPC_sum) has also been proposed by adopting the density peaks clustering algorithm in the selection of indicative sentences. Finally, these two techniques have been further integrated into a formal framework. Experimental results demonstrate that the proposed summarization methods are the most robust among all the methods (including several well-practiced or/and state-of-the-art methods) compared in the paper, thereby indicating the potential of the new spoken document summarization framework. For future work, we will explore other effective ways to enrich the representations of words and integrate extra cues, such as speaker identities or prosodic (emotional) information, into the proposed framework. We are also interested in investigating more robust indexing techniques to represent spoken documents in an elegant way.

8. REFERENCES

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

[2] M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, 25(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] L. S. Lee *et al.*, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1389–1420, 2015.

[5] 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.

[6] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. of ACL*, pp. 470–478, 2008.

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

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

[9] Y. Bengio *et al.*, "A neural probabilistic language model," *Journal of Machine Learning Research* (3), pp. 1137–1155, 2003.

[10] A. Mnih and G. Hinton, "Three new graphical models for statistical language modeling," in *Proc. of ICML*, pp. 641–648, 2007.

[11] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," in *Proc. of ICLR*, pp. 1–12, 2013.

[12] J. Pennington *et al.*, "GloVe: Global vector for word representation," in *Proc. of EMNLP*, pp. 1532–1543, 2014.

[13] D. Tang *et al.*, "Learning sentiment-specific word embedding for twitter sentiment classification" in *Proc. of ACL*, pp. 1555–1565, 2014.

[14] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proc. of ICML*, pp. 160–167, 2008

[15] M. Kageback *et al.*, "Extractive summarization using continuous vector space models," in *Proc. of CVSC*, pp. 31–39, 2014.

[16] L. Qiu *et al.*, "Learning word representation considering proximity and ambiguity," in *Proc. of AAAI*, pp. 1572–1578, 2014.

[17] K. Y. Chen *et al.*, "Leveraging word embeddings for spoken document summarization," in *Proc. of INTERSPEECH*, 2015.

[18] K. Y. Chen *et al.*, "I-vector based language modeling for spoken document retrieval," in *Proc. of ICASSP*, pp. 7083–7088, 2014

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of ICML*, pp. 1188–1196, 2014.

[20] P. S. Huang *et al.*, "Learning deep structured semantic models for web search using clickthrough data," *in Proc. of CIKM*, pp. 2333–2338, 2013.

[21] H. Palangi *et al.*, "Deep sentence embedding using the long short term memory network: analysis and application to information retrieval," in *Proc. of arXiv*, 2015.

[22] J.-M. Torres-Moreno (Eds.), *Automatic text summarization*, WILEY-ISTE, 2014.

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

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

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

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

[27] 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.

[28] K. Y. Chen *et al.*, "Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 8, pp. 1322–1334, 2015.

[29] S. H. Liu *et al.*, "Combining relevance language modeling and clarity measure for extractive speech summarization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 957–969, 2015.

[30] 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.

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

[32] 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.

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

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

[35] M. Galley, "Skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. of EMNLP*, pp. 364–372, 2006.

[36] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[37] T. Mikolov *et al.*, "Distributed representations of words and phrases and their compositionality," in *Proc. of ICLR*, pp. 1–9, 2013.

[38] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proc. of AISTATS*, pp. 246–252, 2005.

[39] A. Mnih and K. Kavukcuoglu, “Learning word embeddings efficiently with noise-contrastive estimation,” in *Proc. of NIPS*, pp. 2265–2273, 2013.

[40] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[41] S. Wang *et al*., “Comment on “Clustering by fast search and find of density peaks”,” in *Proc. of arXiv*, 2015.

[42] Y. Zhang *et al*., “Clustering sentences with density peaks for multi-document summarization,” in *Proc. of NAACL*, pp. 1262–1267, 2015.

[43] 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.

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

[45] 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.