A Novel Paragraph Embedding Method for Spoken Document Summarization

Kuan-Yu Chen*, Shih-Hung Liu*, Berlin Chen† and Hsin-Min Wang*

*Academia Sinica, Taiwan
†National Taiwan Normal University, Taiwan
E-mail: {kychen, journey, whm}@iis.sinica.edu.tw, 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. In the context of natural language processing, paragraph (or sentence and document) embedding learning is more suitable/reasonable for some tasks, such as information retrieval and document summarization. However, as far as we are aware, there is only a dearth of research focusing on launching paragraph embedding methods. Extractive spoken document summarization, which can help us browse and digest multimedia data efficiently, aims at selecting a set of indicative sentences from a source document to express the most important theme of the document. A general consensus is that relevance and redundancy are both critical issues in a realistic summarization scenario. However, most of the existing methods focus on determining only the relevance degree between a pair of sentence and document. Motivated by these observations, three major contributions are proposed in this paper. First, we propose a novel unsupervised paragraph embedding method, named the essence vector model, which aims at not only distilling the most representative information from a paragraph but also getting rid of the general background information to produce a more informative low-dimensional vector representation. Second, we incorporate the deduced essence vectors with a density peaks clustering summarization method, which can take both relevance and redundancy information into account simultaneously, to enhance the spoken document summarization performance. Third, the effectiveness of our proposed methods over several well-practiced and state-of-the-art methods is confirmed by extensive spoken document summarization experiments.**

I. INTRODUCTION

With the popularity of the Internet and the increasing development of the digital storage capacity, unprecedented volumes of multimedia information, such as broadcast news, lecture recordings, voice mails and video streams, among others, have been quickly disseminated around the world and shared among people. 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 [1-3]. 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 concise summary [4-7].

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 pioneering studies [8-10]. The central idea of these methods is to learn continuously distributed vector representations of words using neural networks, which seek to 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 simply taking an average over the word embeddings corresponding to the words occurring within the paragraph (or sentence and document). By doing so, this thread of methods has recently demonstrated promising performance in many NLP-related tasks [11-14].

Although the empirical effectiveness of word embedding methods has been proven recently, the composite representation for a paragraph (or sentence and document) is a bit queer. Theoretically, paragraph-based representation learning is expected to be more suited for such tasks as information retrieval and document summarization [15-18]. However, to the best of our knowledge, there is only a dearth of research concentrating on proposing unsupervised paragraph embedding methods. Moreover, classic paragraph embedding methods infer a representation for a given paragraph by considering all of the words occurring in the paragraph. Consequently, those stop or function words may guide the embedding learning process to produce a misty paragraph representation. In order to complement such a flaw, we propose a novel unsupervised paragraph embedding method, named the essence vector model, which aims at not only distilling the most representative information from a paragraph but also getting rid of the general background information to produce a more discriminative low-dimensional vector representation for the paragraph.

In the context of extractive summarization, it is generally agreed upon that relevance and redundancy are two key aspects for generating a concise summary [19-21]. In this paper, we try to create a synergy of a density peaks clustering summarization method (which can take both relevance and redundancy information into account simultaneously) and the proposed essence vector model for generating a concise extractive summary for a document to be summarized.

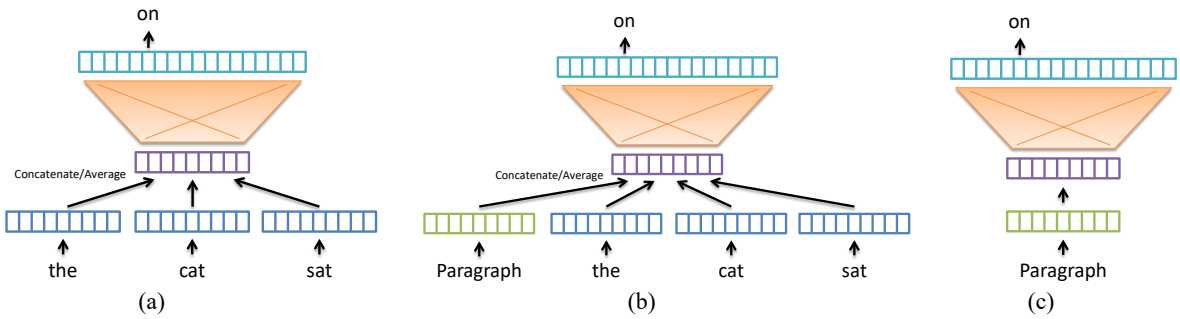

Fig. 1 Illustrations of (a) the feed-forward neural network language model (NNLM), (b) the distributed memory model (DM), and (c) the distributed bag-of-words model (DBOW).

The remainder of this paper is organized as follows. We first briefly review some classic paragraph embedding methods in Section II. Section III sheds light on our proposed essence vector model and the summarization framework. Then, experimental setup and results are presented in Sections IV and V, respectively. Finally, Section VI concludes the paper.

II. Classic Paragraph Embedding Methods

In contrast to the large body of work on developing various word embedding methods, there are relatively few studies concentrating on learning paragraph representations in an unsupervised manner [15-18]. Representative methods include the distributed memory model [15] and the distributed bag-of-words model [15, 16].

A. The Distributed Memory Model

The distributed memory (DM) model is inspired and hybridized from the traditional feed-forward neural network language model (NNLM) [8] and the recently proposed word embedding methods [9]. Formally, given a sequence of words, $\{w^1, w^2, \cdots, w^L\}$, the objective function of feed-forward NNLM is to maximize the total log-likelihood,

$$\sum_{l=1}^{L} \log P(w^l | w^{l-n+1}, \cdots, w^{l-1}). \quad (1)$$

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 taking an average over) the word representations of all words within the context (i.e., $w^{l-n+1}, \cdots, w^{l-1}$), 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:

$$P(w^l | w^{l-n+1}, \cdots, w^{l-1}) = \frac{\exp(y_{w^l})}{\sum_{w_i \in V} \exp(y_{w_i})}, \quad (2)$$

where y_{w_i} denotes the output value for word w_i, and V is the vocabulary. 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 [15]. To make the idea work, the training objective function is defined by

$$\sum_{t=1}^{T} \sum_{l=1}^{L_t} \log P(w^l | w^{l-n+1}, \cdots, w^{l-1}, D_t), \quad (3)$$

where T denotes the number of paragraphs in the training corpus, D_t denotes the t-th paragraph, and L_t is the length of D_t. 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).

B. The Distributed Bag-of-Words 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 [15, 16]. The training objective function can then be defined by maximizing the predictive probabilities all over the words occurring in the paragraph:

$$\sum_{t=1}^{T} \sum_{l=1}^{L_t} \log P(w^l | D_t). \quad (4)$$

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, whereas the DM model stores both softmax weights and word vectors [15]. Fig. 1(c) is a running example to illustrate the architecture of the DBOW model.

III. The Methodology

A. The Essence Vector Model

Classic paragraph embedding methods infer a representation for a given paragraph by considering all of the words occurring in the paragraph. However, we all agree upon that the number of content words in a paragraph is usually less than that of stop or function words. In other words, those stop or function words may guide the representation learning process to produce an ambiguous paragraph representation. Consequently, the associated performance gains will be limited. In order to complement such a flaw, we hence strive to develop a novel

unsupervised paragraph embedding method, which aims at not only distilling the most representative information from a given paragraph but also getting rid of the general background information (probably caused by stop or function words), so as to deduce an informative and discriminative low-dimensional vector representation for a given paragraph. We will henceforth term this novel unsupervised paragraph embedding method the essence vector (EV) model [22].

To make the idea to go, we begin with an assumption that each paragraph (or sentence and document) can be assembled by two components: the paragraph specific information and the general background information. The assumption also holds in the low-dimensional representation space. Accordingly, the proposed method consists of three modules: a paragraph encoder $f(\cdot)$, which can automatically infer the desired low-dimensional vector representation by considering only the paragraph-specific information; a background encoder $g(\cdot)$, which is used to map the general background information into a low-dimensional representation; and a decoder $h(\cdot)$ that can reconstruct the original paragraph by combining the paragraph representation and the background representation.

More formally, given a set of training paragraphs $\{D_1, \cdots, D_t, \cdots, D_T\}$, in order to modulate the effect of different lengths of paragraphs, each paragraph is first represented by a bag-of-words high-dimensional vector $P_{D_t} \in \mathbb{R}^{|V|}$, where each element corresponds to the frequency count of a word/term in the vocabulary V, and the vector is normalized to unit-sum. Then, a paragraph encoder is applied to extract the most specific information from the paragraph and encapsulate it into a low-dimensional vector representation:

$$f(P_{D_t}) = v_{D_t}. \tag{5}$$

At the same time, the general background is also represented by a high-dimensional vector with normalized word/term frequency counts, $P_{BG} \in \mathbb{R}^{|V|}$, and a background encoder is used to compress the general background information into a low-dimensional vector representation:

$$g(P_{BG}) = v_{BG}. \tag{6}$$

Both $f(\cdot)$ and $g(\cdot)$ are fully connected multilayer neural networks with different model parameters θ_f and θ_g, respectively. It is worthy to note that the model structures of $f(\cdot)$ and $g(\cdot)$ can be the same or different. Since each learned paragraph representation v_{D_t} only contains the most informative/discriminative part of P_{D_t}, we assume that the weighted combination of v_{D_t} and v_{BG} can be mapped back to P_{D_t} by a decoder $h(\cdot)$:

$$h(\alpha_{D_t} \cdot v_{D_t} + (1 - \alpha_{D_t}) \cdot v_{BG}) = P'_{D_t}, \tag{7}$$

where $h(\cdot)$ is also a fully connected multilayer neural network with parameters θ_h, and the interpolation weight can be determined by an attention function $q(\cdot,\cdot)$:

$$\alpha_{D_t} = q(v_{D_t}, v_{BG}). \tag{8}$$

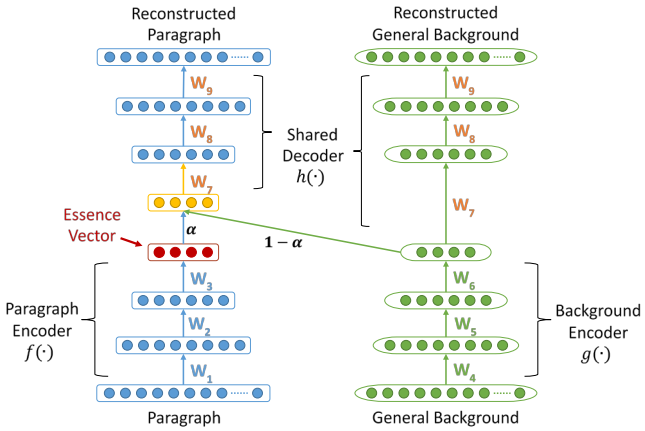

Fig. 2 A running example for the essence vector model.

The attention function can be realized by a trainable network or a simple linear/non-linear function only. Further, to ensure the quality of the learned background representation v_{BG}, it should also be mapped back to P_{BG} by $h(\cdot)$:

$$h(v_{BG}) = P'_{BG}. \tag{9}$$

In a nutshell, the objective function of the proposed essence vector model is to minimize the total KL-divergence measure:

$$\min_{\theta_f, \theta_g, \theta_h} \sum_{t=1}^{T} \left(P_{D_t} \log \frac{P_{D_t}}{P'_{D_t}} + P_{BG} \log \frac{P_{BG}}{P'_{BG}} \right). \tag{10}$$

The activation function used in the essence vector model is the hyperbolic tangent, except that the output layer in the decoder $h(\cdot)$ is the softmax [23], the cosine distance is used to calculate the attention coefficients, and the Adam [24] is employed to solve the optimization problem. Fig. 2 illustrates the architecture of the proposed paragraph embedding method.

B. The Enhanced Summarization Framework

The most common belief in the document summarization community is that relevance and redundancy are two key factors for generating a concise summary. Maximum margin relevance (MMR) is the most popularly used criterion for automatic summarization [20], 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, in this paper, we leverage a density peaks clustering summarization method [21, 25, 26], which can take both relevance and redundancy information into account at the same time. That is, a concise summary for a given document can be automatically generated through a one-pass process instead of an iterative process. Recently, the summarization method has proved its empirical effectiveness when being paired with classic paragraph embedding methods (cf. Section II) [21].

The underlying idea of the summarization framework consists of two aspects [21, 26]: 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. The density score for sentence S_i in a document D is defined as:

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

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

where K is the number of sentences in D, $\text{sim}(S_i,S_j)$ 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 score for sentence S_i is calculated by

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

except that, for the sentence with the highest density score, its divergence score is set to 1 directly.

In practice, the multiplication score (i.e., $\text{density}(S_i) \times \text{divergence}(S_i)$) can be used alone or linearly combined with the conventional relevance score between a sentence and a document (i.e., $\text{sim}(S_i,D)$) to select sentences. We use the cosine measure as the similarity score $\text{sim}(\cdot,\cdot)$ throughout the paper. The vector representation for a paragraph (i.e., sentence and document in this paper) is characterized by inferring through the classic paragraph embedding methods (cf. Section II) or the proposed essence vector model (cf. Section III-A).

IV. 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 [27]. 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 [28]. All the experimental results reported hereafter are obtained by calculating the F-scores [29] of these ROUGE metrics. The summarization ratio was set to 10%. 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 [30]. The vocabulary size is about 72 thousand words. The average word error rate of automatic transcription is about 38.1%. Furthermore, an external set of about 100,000 text news documents, which is assembled by the Central News Agency (CNA) during the same period as the broadcast news documents to be summarized (extracted from the Chinese Gigaword Corpus released by LDC), was used to obtain the background representation (cf. Section III-A).

TABLE I
SUMMARIZATION RESULTS ACHIEVED BY THE CLASSIC AND THE PROPOSED PARAGRAPH EMBEDDING METHODS WITH COSINE SIMILARITY MEASURE.

	Text Documents (TD)		Spoken Documents (SD)	
	ROUGE-2	ROUGE-L	ROUGE-2	ROUGE-L
DM	0.290	0.355	0.218	0.313
DBOW	0.293	0.364	0.232	0.323
EV	**0.338**	**0.404**	**0.266**	**0.357**

V. Experimental Results

To begin with, we assess the performance levels of the proposed essence vector (EV) model and two paragraph embedding methods (i.e., DM and DBOW) with conventional cosine similarity measure for SDS (cf. Sections II & III). The results are shown in Table I, 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 I, several observations can be drawn. First, DBOW consistently outperforms DM in both the TD and SD cases, though the performance difference is mostly small. Second, the proposed EV model outperforms DM and DBOW in both the TD and SD cases by a large margin, as expected. Third, the experimental results also confirm that EV can modulate the impact of those stop or function words when inferring representations for paragraphs. That is to say, the proposed paragraph embedding method can indeed distill the most important aspects of a given paragraph and get rid of the general background information to produce a more discriminative paragraph representation. Thus, the relevance degree between any pair of sentence and document representations can be estimated more accurately.

In the second set of experiments, we further integrate these paragraph embedding methods (i.e., DM, DBOW, and EV) with the density peaks clustering summarization method (cf. Section III-B). The results are shown in Table II. It is obvious that the results in Table II are better than almost all the results in Table I. The outcomes signal that redundancy is an important issue to text or spoken document summarization. A particular observation worthwhile to note is that while the combination of EV and the density peaks clustering method offers a quite promising performance gain in the TD case, it seems not to achieve a further performance gain as expected in the SD case. The reason should be explored further.

In the last set of experiments, we compare the results mentioned above with that of several well-practiced, state-of-the-art unsupervised summarization methods, including the graph-based methods (i.e., the Markov random walk (MRW) method [31] and the LexRank method [32]) and the combinatorial optimization methods (i.e., the submodularity-

TABLE II
SUMMARIZATION RESULTS ACHIEVED BY INCORPORATING DIFFERENT PARAGRAPH EMBEDDING METHODS WITH THE DENSITY PEAKS CLUSTERING SUMMARIZATION METHOD.

	Text Documents (TD)		Spoken Documents (SD)	
	ROUGE-2	ROUGE-L	ROUGE-2	ROUGE-L
DM	0.339	0.409	0.242	0.337
DBOW	0.334	0.405	0.250	0.344
EV	**0.405**	**0.453**	**0.264**	**0.361**

TABLE III
SUMMARIZATION RESULTS ACHIEVED BY SOME STATE-OF-THE-ART SUMMARIZATION METHODS.

	Text Documents (TD)		Spoken Documents (SD)	
	ROUGE-2	ROUGE-L	ROUGE-2	ROUGE-L
MRW	0.282	0.358	0.191	0.291
LexRank	0.309	0.363	0.146	0.254
SM	0.286	0.363	0.204	0.303
ILP	0.337	0.401	0.209	0.306
DM	0.339	0.409	0.242	0.337
DBOW	0.334	0.405	0.250	0.344
EV	**0.405**	**0.453**	**0.264**	**0.361**

based (SM) method [33] and the integer linear programming (ILP) method [34]). Among them, the ability of reducing redundant information has been aptly incorporated into the submodular-based method and the ILP method. Interested readers may refer to [4-7] 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. The corresponding results are listed in Table III. Several noteworthy observations can be drawn from the results of these methods. First, the two graph-based methods (i.e., MRW and LexRank) are quite competitive with each other in the TD case, while MRW outperforms LexRank in the SD case. Second, although both SM and ILP have the ability to reduce redundant information when selecting indicative sentences to form a summary for a given document, ILP consistently outperforms SM in both the TD and SD cases. The reason might be that ILP performs a global optimization process to select representative sentences, while SM chooses sentences with a recursive strategy. Comparing the results of these strong baseline systems to that of the paragraph embedding methods paired with the density peaks clustering summarization method, it is clear that all the paragraph embedding methods are better than the baseline methods. In particular, the proposed essence vector model is the most robust among all the methods compared in the paper. The results corroborate 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. Since the paragraph embedding methods can also be incorporated with the graph-based methods and the combinatorial optimization methods, we will study this in our future work.

VI. CONCLUSIONS & OUTLOOK

In this paper, we have proposed a novel paragraph embedding method, called the essence vector model, and made a step forward to plug in this new paragraph embedding method into the density peaks clustering summarization method to enhance the performance of SDS. Experimental results demonstrate that the proposed framework is 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 paragraph embedding method. For future work, we will first focus on pairing the essence vector model with other summarization methods. Moreover, we will explore other effective ways to 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.

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] Y. Liu and D. Hakkani-Tur, "Speech summarization," *Chapter 13* in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, G. Tur and R. D. Mori (Eds), New York: Wiley, 2011.

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

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

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

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

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

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

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

[12] 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

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

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

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

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

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

[18] 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.

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

[20] 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.

[21] K. Y. Chen *et al*., "Incorporating paragraph embeddings and density peaks clustering for spoken document summarization," in *Proc. of ASRU*, pp. 207–214, 2015.

[22] K.Y. Chen *et al*., "Learning to distill: the essence vector modeling framework" in *Proc. of Coling*, 2016.

[23] I. Goodfellow *et al*., *Deep Learning*, Cambridge, MA: MIT Press, 2016.

[24] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. of ICLR*, 2015.

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

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

[27] 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.

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

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.