Multi-Document News Summarization via Paragraph Embedding and Density Peak Clustering

Baoyan Wang[1,3] , Jian Zhang[2] , Fanggui Ding[4] , Yuexian Zou[1*]
[1*]ADSPLAB/ELIP, School of ECE, Peking University, China
[2]Dongguan University of Technology, China
[3]IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institute
[4]Shenzhen Press Group, China
*E-mail: zouyx@pkusz.edu.cn

***Abstract*—Multi-document news summarization (MDNS) aims to create a condensed summary while retaining the main characteristics of the original set of news documents. Research shows that the text representation is one of the keys for MDNS techniques. Without doubt, the bag-of-words (BOW) methods are most widely used. However, BOW methods generate high-dimensional representation vectors which ask for large storage and high computational complexity for MDNS. Besides, the generated representation vectors by BOW lack the semantic information and temporal information of the words, which limits the performance of MDNS. To tackle above issues, this paper introduces a word/paragraph embedding method via neural network modelling to generate lower dimensional word/paragraph representation vectors retaining word order and context information and semantic relationships between words/paragraphs. Besides, for MDNS, relevance and redundancy are both critical issues. Unlike the traditional MDNS methods quantifying the relevance among different sentences followed with a greedy post-processing module to ensure the diversity of summary, in this study, we concurrently take relevance, diversity and length constraint into account by employing density peak clustering (DPC) technique and the integrated sentence scoring method to select the more representative sentences and generate the summary with less redundancy. Experimental results on the DUC2003 and DUC2004 datasets demonstrate the effectiveness of our MDNS method, compared to the state-of-the-art methods.**

Keywords- Multi-Document News Summarization; Text Representation; Integrated Sentence Scoring Method; Density Peak Clustering

I. INTRODUCTION

With the explosively growing of information data overload over the Internet, consumers are flooded with all kinds of electronic documents i.e. news, tweets and blog. The end users of search engines and news providers have to read the same information over and over again conveyed by the presence of numerous documents. There are urgent demands for multi-document summarization now more than ever, as it aims at generating a concise and informative version for the large collection of original documents that facilitates readers quickly grasp the general information of them. Most existing studies are extractive based methods, which focus on extracting sentences directly from given materials and combining sentences together to form a summary. In this paper, we address the task of generic extractive-based summarization from multiple news documents (MDNS).

Text representation has occurred as an attractive subject of research in many applications of natural language processing (NLP) due to its remarkable performance i.e. text categorization [1], and named entity recognition [2]. Bag-of-words typically represent text as a fixed-length vector because of its concise, validness and often promising performance. Two main drawbacks are that the order and semantic of words are ignored. Word embedding [3] methods 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. And then the paragraph is represented by averaging/concatenating the word embeddings to measure the similarity among paragraphs. Despite considering the semantic, the order of words is lost. Paragraph embedding [4] learns continuous distributed vector representations for pieces of texts, anything from a sentence to a large document. In this paper, we investigate different text representation methods for MDNS thoroughly.

On the other hand, an effective summarization method always properly considers two key issues [5]: Relevance and Diversity. The extractive summarization methods can fall into two categories: supervised methods that rely on provided document-summary pairs, and unsupervised ones based upon properties derived from document clusters. The supervised methods generally treat the summarization task as a classification or regression problem [6]. For those methods, a huge amount of annotated data is required, which are costly and time-consuming. For another thing, unsupervised approaches are very enticing [7-15]. They tend to score and then rank sentences based on semantic, linguistic or statistic grouping extracted from the original documents. Whereas, most already existing methods tend to determine the relevance degree between sentences and documents firstly. And then an additional post-processing step is employed to remove redundancy and ensure the diversity of summary. They also tend to extract sentence based on greedy algorithm, which cannot guarantee the optimal summary. In this paper, we apply density peak clustering (DPC) to cluster sentences and an integrated scoring method to score sentences by considering relevance and diversity simultaneously. We extract sentences under length constraint based on dynamic programming (DP) strategy finally.

II. THE WORD AND PARAGRAPH EMBEDDIN METHODS

Text representation methods map words or paragraphs into a vector space, which facilitates learning algorithms to achieve better performance in NLP tasks i.e. text categorization. Bag-of-words methods bring about difficulty of data sparsity and always lack of the semantic information and words' order. Therefore, it's limited of

978-1-5386-1981-0/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Penn State University. Downloaded on March 30,2023 at 19:14:33 UTC from IEEE Xplore. Restrictions apply.

the ability to measure the similarity among sentences. In this section, we introduce the word embedding methods and paragraph embedding methods to address the problem, and compare them comprehensively.

A. Neural Network Language Model (NNLM)

In [16], the probabilistic feedforward NNLM is proposed to predict future words and generate word embedding as the by-product, which is a well-known pioneering research. Given a sequence of words, $w_1, w_2, ..., w_{\mathbb{C}}$, the objective function of NNLM is to maximize the log-probability:

$$\sum_{i=1}^{\mathbb{C}} \log P(w_i \mid w_{i-n},...,w_{i-1}) \tag{1}$$

where n denotes window size of previous words, $\mathbb{C}$ is the length of the corpus, and

$$P(w_i \mid w_{i-n},...,w_{i-1}) = \frac{\exp(y(w_i))}{\sum_{k=1}^{\mathbb{H}} \exp(y(w_k))} \tag{2}$$

where $\mathbb{H}$ denotes the size of vocabulary, $y(w_i)$ is the possibility which w_i is the next word.

B. Continuous Bag-of-words (CBOW) Model and Skip-gram (SG) Model

Similar to the feedforward NNLM, CBOW model [3] straightforward generates words embedding through context of target words instead of learning a statistical language model. For increasing efficiency, the CBOW model gets rid of hidden layers and changes from the neural network structure into the log linear structure directly. Besides, the CBOW model removes information of word order in context by using the average of word embedding, while still retains promising performance. Given a sequence of words, $w_1, w_2, ..., w_{\mathbb{C}}$, the objective function of CBOW are to maximize the log-probability:

$$\sum_{i=1}^{\mathbb{C}} \log P(w_i \mid w_{i-(n-1)/2},...,w_{i+(n-1)/2}) \tag{3}$$

Contrasted with the CBOW model, the SG model [3] uses target word to predict words of context rather than predicting the nearby word based on the context.

$$\sum_{i=1}^{\mathbb{C}} \sum_{j=i-(n-1)/2}^{+(n-1)/2} \log P(w_i \mid w_j, j \neq i) \tag{4}$$

C. Distributed Memory (DM) Model and Distributed Bag-of-Words (DBOW) Model

The DM model [4] is inspired by the method for learning word embedding. The DM model not only retains the semantics of the words, but considers the word order, at least in a small context. The DM model maps all of paragraphs into the unique vectors, and averages or concatenates the paragraph embeddings and word embeddings to predict the target word in the context. The paragraph embedding represents the missing information from the context and acts as a memory of the topic of the paragraph.

$$\sum_{i=1}^{M} \sum_{j=1}^{L_i} \log P(w_j \mid w_{j-n},...,w_{j-1}, S_i) \tag{5}$$

In contrast to DM model, the DBOW model [4] ignores the context words, but only predicts words randomly sampled from the paragraph.

$$\sum_{i=1}^{M} \sum_{j=1}^{L_i} \log P(w_j \mid S_i) \tag{6}$$

where M denotes the number of paragraphs in the corpus, S_i denotes the i-th paragraph, and L_i is the length of S_i. The DBOW model only needs to store the softmax weights, while the DM model has to store both softmax weights and word embeddings.

The sentence is represented by averaging the word embeddings of words appearing in that sentence or the paragraph embeddings directly in our method. Accordingly, each sentence S_i of corpus has a respective fixed-length dense vector representation.

III. Our Proposed MDNS Method

It's universally acknowledged that a good MDNS method should consider relevance and diversity properly. Whereas, most existing MDNS methods, such as FGB [9], LSA [13], BSTM [15], and LexRank [7], usually concentrate on investigating the relevance degree between a sentence with the others or documents. On the other hand, those methods tend to ensure the diversity of the sentences in summary and remove redundancy by a post-processing step. Therefore, we propose a new MNDS method termed as the integrated sentence scoring method, which use Density Peak Clustering (DPC) [17] to take relevance, diversity and length constraint into account simultaneously. Sentences are scored in the three aspects, and then the scores are log linearly combined. Finally, sentences are extracted to generate based on dynamic programming algorithm. Besides, our proposed method is a one-pass process and the details are given as follows.

Relevance Score. We show a relevance score to quantify the degree concerning how much a sentence is relevant to residual sentences in the documents. One of the underlying assumptions of DPC is that cluster centers are characterized by a higher density than their neighbors. Proceeding from it we consider that a sentence will be deemed to be more relevant and more representational when it possesses higher density meaning with more similar sentences. Thus, we define the function to compute the relevance scoring $SC_R(i)$ in sentences level for each sentence s_i as follows:

$$SC_R(i) = \sum_j^K f(Sim_{ij} - \omega), \quad f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases} \tag{7}$$

where Sim_{ij} is the similarity value between the i-th and j-th sentence and K denotes the number of sentences in the dataset. ω represents a predefined density threshold.

Diversity Score. Diversity scoring is presented to argue a good summary should not include analogical sentences. A document set usually contains one core topic and some subtopics. In addition to the most evident topic, it's also necessary to get the sub-topics most evident topic so as to better understand the whole corpus. Another hypothesis of DPC is that cluster centers also are characterized by a relatively large distance from points with higher densities, which ensure the similar sentences get larger difference scores. Therefore, by comparing with all the other sentences of the corpus, the sentence with a higher score could be extracted, which also can guarantee the diversity globally. The diversity score $SC_D(i)$ is defined in clusters level as the following function.

$$SC_D(i) = 1 - \max_{j: SC_R(j) > SC_R(i)} Sim_{ij} \tag{8}$$

Diversity score of the sentence with the highest density is assigned 1 conventionally.

Length Constraint. The longer sentence is, the more informativeness it owns. Therefore a fewer number of longer sentences tend to be extracted, which is contrary to

Authorized licensed use limited to: Penn State University. Downloaded on March 30,2023 at 19:14:33 UTC from IEEE Xplore. Restrictions apply.

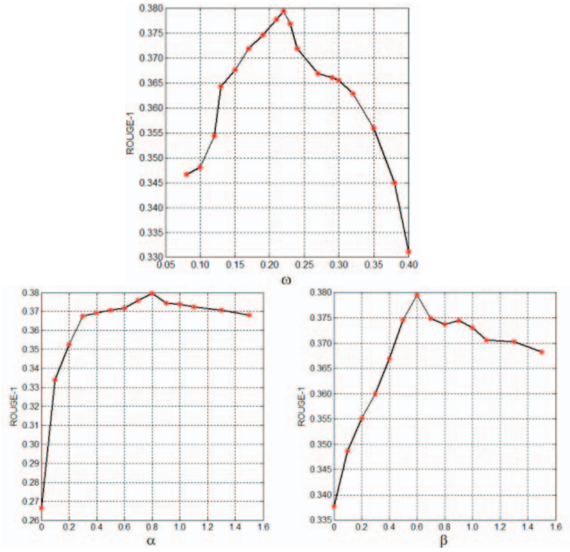

Figure 1. ROUGE-1 versus parameter α, β and ω on DUC2003

the human summarizers who tend to produce larger number of shorter sentences. The total number of words in the summary usually is limited. The longer sentences are, the fewer ones are selected. Therefore, it is requisite to provide a length constraint. Thus the length constrain is defined and normalized as follows.

$$SC_L(i) = \log((\max_j L_j) / L_i + 1) \qquad (9)$$

Integrated Sentence Scoring Method. The ultimate goal of our method is to select those sentences with higher relevance, and better diversity under the limitation of length. In order to adapt to the integrated scoring method, $SC_R(i)$, $SC_D(i)$ and $SC_L(i)$ should be normalized by divided their own highest values firstly. We define a function comprehensively considering the above purposes:

$$SC(i) = \alpha \log SC_R(i) + \beta \log SC_D(i) + \log SC_L(i) \qquad (10)$$

The parameters α, and β of our method is used to control the different weights of the three aspects.

Finally, we should generate a summary by extracting sentences under the limit of the exact length L. As every sentence is measured by an integrated score, the score sum of extracted sentences in summary should be as high as possible. The summary generation is regarded as the 0-1 knapsack problem:

$$\begin{aligned} &\arg\max \sum (SC(i) \times x_i) \\ &\mathit{Subject\ to} \sum L_i x_i \le L, x_i \in \{0,1\} \end{aligned} \qquad (11)$$

The problem is NP-hard. To alleviate this problem, we utilize the dynamic programming solution to select sentences until the expected length L of summaries is satisfied.

IV. Experimental Setup

A. Datasets and Evaluation Metrics

The open benchmark datasets DUC2003 and DUC2004, from Document Understanding Conference, are employed in our experiments. DUC2004 consists of 50 news document sets and 10 documents related to each set. Length Limit of summary is 665 bytes. DUC2003 consists of 60 news document sets and about 10 documents for each set. The structures of both datasets are similar. Therefore, we choose DUC2003 as the development dataset for parameters tuning and DUC2004 for evaluation. There are four human generated summaries provided as ground truth for each news set. We observe that the sentences of summaries are not strictly selected in their entirety, but changed considerably.

TABLE I. Validity of different text representation schemes in our method

Methods	ROUGE-1	ROUGE-2	ROUGE-SU
TF+DPC	0.38756	0.09278	0.13729
TF-ISF+DPC	0.38109	0.08934	0.13243
BOOL+DPC	0.39047	0.09559	0.13916
SG+DPC	0.37501	0.08972	0.13158
CBOW+DPC	0.38891	0.09396	0.13904
DBOW+DPC	0.39471	0.09689	0.14192
DM+DPC	**0.39947**	**0.09923**	**0.14631**

We apply widely used ROUGE version 1.5.5 toolkit [18] to evaluate summarization performance in our experiments. Among the evaluation methods implemented in Rouge, Rouge-1 focuses on the occurrence of the same words between generated summary and annotated summary, while Rouge-2 and Rouge-SU concerns more over the readability of the generated summary. We report the mean value over all topics of the F-measure scores of these three metrics in the experiment. Note that the higher ROUGE scores, the more similar between generated summary and annotated one.

B. Parameter Settings

We set word embeddings dimensionality 80 and context size 5 empirically. We find that significant improvements using pre-trained word embeddings over randomly initialized ones. Therefore, we use Wikipedia corpus to pre-train word embeddings and fine-tune them using our corpus. We investigate how parameters α, β and the density threshold ω relate to our method by a set of experiments on DUC2003. The parameters α, β and γ are set ranging from 0 to 1.5 respectively at the step size of 0.1. The best values of the parameters are selected by comparing all of the results. One parameter is tuned while set the others on their best values. The results of tuning parameters are shown in Figure 1. We find that α=0.8 and β=0.6 produce a better performance than α=1 and β=1, which indicates the effect of four scores do not equal each other for the integrated scoring method. Besides, the performance dropped a lot when α or β are set zero, which shows the three scores play an active role in our method. We observe that our method works best whenω is set 0.22. Using these settings, we apply our method on DUC2004.

V. Experimental Result

We compare the word and paragraph embedding methods with different bag-of-word methods firstly: 1) BOOL (presence or absence); 2) TF (term frequency); 3) TF-ISF (combine TF with ISF). The results of these experiments are listed in Table I. It is possible to see that BOOL term weighting achieves better results compared with TF, TF-ISF and the word embedding. It may due to the frequency of term repetition occur less in sentences.

Authorized licensed use limited to: Penn State University. Downloaded on March 30,2023 at 19:14:33 UTC from IEEE Xplore. Restrictions apply.

TABLE II . OVERALL PERFORMANCE COMPARISON ON DUC2004 DATASET USING ROUGE METRICS. REMARK: "–" INDICATES THAT THE METHOD DOES NOT OFFICIALLY REPORT THE RESULTS.

Methods	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU**
DUC best	0.38224	0.09216	0.13233
ClusterHITS [8]	0.36463	0.07632	–
FGB [9]	0.38724	0.08115	0.12957
RTC [10]	0.37475	0.08973	–
LSA [13]	0.34145	0.06538	0.11946
BSTM [11]	0.39065	0.09010	0.13218
MSFF [12]	–	0.09897	0.13951
LexRank [15]	0.37842	0.08572	0.13097
R2N2_ILP [6]	0.38780	0.09860	–
DM+DPC (OURS)	**0.39947**	**0.09923**	**0.14631**

The result also indicates that paragraph embedding methods (DM/DBOW) get better results than the word embedding methods (SG/CBOW) as expected. It may be because the paragraph embedding methods take the word order into consideration and capture the semantics of sentences preferably. Besides, the DM model and CBOW model outperform the SG and DBOW respectively. It may be because the context as the input better learns the semantics than as the target when trains the word and paragraph embeddings.

We work with the following widely used or recent published methods for general summarization as the baseline methods to compare with our proposed method. The results of these methods are listed in Table II.

From Table II, we can have the following observations: Our method clearly outperforms the DUC2004 best team work on the three metrics. It is obvious that our method outperforms other rivals significantly on the ROUGE-1 metric and the ROUGE-SU metric. It can be attributed to the integrated sentence scoring method to combine paragraph embedding method with DPC, which promotes the results mutually and ensure higher quality of the summaries. Compared with other cluster based method [8-10], our method removes redundancy when clustering and considers the semantics of sentences. Our method performs slightly better than MSSF and R2N2_ILP on ROUGE-2 score. Those methods are complex and even need multiple features and postprocessor. LSA, BSTM, and LexRank always need the maximum margin relevance method (MMR) [14] as the postprocessor to generate summary. The MMR quantifies the degree of dissimilarity between candidate sentences and already selected ones, and then select sentences based on greedy approach. The results indicate that diversity is indeed an important issue to MDNS. Besides, our method outperforms MMR based methods by a large margin.

VI. CONCLUSION

In this paper, we proposed an unsupervised method to handle the task of multi-document news summarization. We applied the paragraph embedding method to represent sentence and compared with other typical bag-of-words and word embedding methods comprehensively. For ranking sentences, we proposed an integrated sentence scoring method to take relevance, diversity and length constraint into consideration. DPC was employed to measure the relevance in sentences level and diversity of sentences in clusters level at the same time. We combined those scores with a length constraint and selected sentences based dynamic programming at last. Extensive experiments on the standard datasets showed that our method is quite effective for MDNS. In the future, we will apply our proposed method to topic-focused and updated summarization, to which the tasks of summarization have turned.

ACKNOWLEDGMENT

This work is partially supported by Key Technologies Research & Development Program of Shenzhen (No: JSGG20150512160434776) and Shenzhen Science & Research projects. (No: JCYJ20150331171116474).

REFERENCES

[1] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, pp. 649-657.

[2] A. Das, D. Ganguly, and U. Garain, "Named Entity Recognition with Word Embeddings and Wikipedia Categories for a Low-Resource Language," in TALLIP, vol. 16, p. 18, 2017.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv, preprint arXiv:1301.3781, 2013.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in NIPS, 2013, pp. 3111-3119.

[5] H. Liu, H. Yu, and Z.-H. Deng, "Multi-Document Summarization Based on Two-Level Sparse Representation Model," in AAAI, 2015, pp. 196-202.

[6] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization," in AAAI, 2015, pp. 2153-2159.

[7] K. Hong and A. Nenkova, "Improving the Estimation of Word Importance for News Multi-Document Summarization," in EACL, 2014, pp. 712-721.

[8] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in ACM SIGIR, 2008, pp. 299-306.

[9] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong, "Integrating document clustering and multidocument summarization," ACM Transactions on TKDD, vol. 5, p. 14, 2011.

[10] X. Cai and W. Li, "Ranking through clustering: An integrated approach to multi-document summarization," IEEE TALSP, vol. 21, pp. 1424-1433, 2013.

[11] D. Wang, T. Li, S. Zhu, C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in ACM SIGIR, 2008, pp. 307-314.

[12] J. Li, L. Li, and T. Li, "Multi-document summarization via submodularity," Applied Intelligence, vol. 37, pp. 420-430, 2012.

[13] S. Xiong and Y. Luo, "A new approach for multi-document summarization based on latent semantic analysis," in Computational Intelligence and Design, 2014, pp. 177-180.

[14] J. Goldstein, V.Mittal, J.Carbonell,M. Kantrowitz, Multi-document summarization by sentence extraction, in NAACL, 2000, pp. 40-48.

[15] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," Journal of machine learning research, vol. 3, pp. 1137-1155, 2003.

[17] A. Rodriguez and A. Laio, "Machine learning. Clustering by fast search and find of density peaks," Science (New York, NY), vol. 344, pp. 1492-1496, 2014.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out: Proceedings of the ACL-04 workshop, 2004.

Authorized licensed use limited to: Penn State University. Downloaded on March 30,2023 at 19:14:33 UTC from IEEE Xplore. Restrictions apply.