OnSeS: A Novel Online Short Text Summarization based on BM25 and Neural Network

Jianwei Niu*, Qingjuan Zhao*, Lei Wang*, Huan Chen*, Mohammed Atiquzzaman†, Fei Peng‡
*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
†School of Computer Science, University of Oklahoma, Norman, OK 73019, USA
‡Shanghai Research Institute of Aerospace Computer Technology, Shanghai 200050, China
Email: niujianwei@buaa.edu.cn

***Abstract*—The last decade has witnessed a dramatic growth of social networks, such as Twitter, Sina Microblog, etc. Messages/short texts on these platforms are generally of limited length, causing difficulties for machines to understand. Moreover, it is rarely possible for users to read and understand all the content due to the large quantity. So it is imperative to cluster and extract the viewpoints of these short texts. To solve this, the representation of a word is enriched with additional features from external, but it is demanding in terms of computational and time resources. In this paper, we proposed OnSeS, a novel short text summarization method which makes full use of word2vec to represent a word and utilizes neural network model to generate each word of the summary. OnSeS consists of three phrases: 1) clustering short texts using the K-means algorithm; 2) ranking content of each cluster by building a graph-based ranking model using BM25; 3) generating main point of each cluster with the help of neural machine translation model on the top ranked sentence. The experimental results reveal that our proposed fully data-driven approach outperforms state-of-the-art method.**

***Index Terms*—short text clustering; text ranking; opinion extraction; short text summarization; neural machine translation**

I. Introduction

With the rapid development of social networks, increasing number of users are getting used to post their personal feelings or opinions towards current issues on their microblogs, which only allow less than 140 characters per piece. Massive amount of content is being generated and the task of clustering and summarizing it has drawn considerable attention. However, short texts, which tend to be sparser and noisier, are totally different from long paragraphs. Thus, traditional document summarization approaches perform badly when applied to short texts. Therefore it is necessary to find another approach to extract the main idea of short texts. The *objective* of this paper is to develop algorithms to cluster short texts and generate a summarization for each cluster. In this paper, we propose a novel short-text summarization method, called OnSeS, based on BM25 [1] and neural machine translation. It consists of three subtasks: 1) short text clustering, 2) short text ranking, and 3) short text summarizing.

Some algorithms have already been proposed to cluster texts, such as Affinity Propagation [2] and Spectral clustering [3]. The complexity of Affinity Propagation is quadratic with the increase in the number of documents; this makes the schemes unsuitable for coping with large datasets. Nayak et al. [4], proposed a method of clustering short texts by enriching their representation with additional features from Wikipedia, but it is demanding in terms of computational and time resources.

In order to obtain the top ranked text for a particular short-text cluster, we calculate the relevance between each piece of content based on BM25. BM25 is a bag-of-words ranking function applied in popular web search engines, ignoring the properties of the relationships between the neighboring words. Furthermore, text ranking model does not require deep linguistic knowledge or language specific annotated corpora, making it highly portable to other domains, genres and languages.

In the summarization stage, most of the previous studies just regard the top ranked sentence as the summarization, due to which the results are not satisfactory. In *contrast* to previous methods, we train a fully data-driven neural machine translation model on the Graphics Processing Unit (GPU) GTX TITAN Z, inspired by the recent popular recurrent neural network (RNN). The core of the summarization model is a neural probabilistic language model first proposed by Bengio et al. [5]. Additionally, an analyzing system to summarize the different opinions on a topic is proposed in this paper. With BM25 and neural machine translation based summarization, users can easily observe all kinds of opinions on a topic as well as the summary of each opinion.

In this paper, we *proposed* a novel approach called OnSeS which solved the sparse issue of the short text and generated a condensed representation. The major *contributions* of our work include:

- Clustering short texts using the K-means algorithm based on distributed representations of words and phrases.
- Ranking the short texts of each cluster by building a graph-based ranking model using BM25 and identifying the top ranked piece of content.
- Obtaining the main opinion of each cluster by using neural machine translation model on the top ranked sentence.

The rest of this paper is organized as follows. Section II gives an overview of the related work and background. The details of our proposed OnSeS approach are presented in Section III. Section IV describes the design of a series of experiments and evaluates their performances. Finally, we draw our conclusions in Section V.

978-1-5090-1328-9/16/$31.00 ©2016 IEEE

II. Related Work

In this section, we briefly present the previous approaches related to short text summarization. The methods of short text clustering usually involve the following three ways [3] [4] [6]: selecting effective clustering algorithms, incorporating the external corpus, and modifying weights of terms. Opinion summarizing technology is similar to multi-document summarizing technology, which has been studied for many years. Summarizing technologies have been mainly divided into two categories: summary based on statistics and summary based on understanding.

In order to effectively compute short text similarity, vectors are used to represent the text where every term is regarded as one dimension of the vector. Term Frequency-Inverse Document Frequency (TF-IDF) [6] is the most common representation strategy. However, Yin et al. [7] pointed out that it was not an efficient measure because of the sparse character of short text. To the best of our knowledge, researchers concentrated on overcoming the issue of sparse characteristic of shot text by adding external resources such as WordNet [8] and Wikipedia [4]. However, these methods increase the complexity in natural language processing.

There are many kinds of clustering algorithms, such as K-means and the agglomerative hierarchical clustering, as stated in [9] and [10] respectively. Some studies [11] [12] presented the clustering methods through extending the features, such as using statistical semantics, internal semantics and term-term similarity. While some studies estimated the similarity according to frequency of co-occurrence. Vectors of co-occurrence were used to express the original terms.

In recent years, some researchers have focused on mining and summarizing reviews on movies or products [13]. There are a lot of studies investigating the sentiment of the reviews. Feature-based methods are applied in specific domains. So their suitability largely depends on the training process. To sum up, these methods are not suitable for short texts due to domain limitations.

Nenkova et al. [14] extracted the top representative segments as the text summarization. Hu et al. [15] put forward a nice method to identify and extract specific product features and opinions. Unlike these studies, our summarization identifies the similarities, finds out the representative short text, and finally summarizes the opinion.

There are also some Information Extraction (IE) techniques such as traditional IE and broader Open IE [16]. Traditional IE focuses on interested specific relations, while Open IE focuses on the relation tuples which mediated by verbs such as REVERB [17] and WOE [18]. WOE required several steps, one of which is training a large number of examples, and REVERB further introduced syntactic and lexical constraints. However, these methods do not yield good performance in real world because of the limitation regarding to relation mediation and context.

III. Proposed Method

Our proposed approach, OnSeS, entails three stages: 1) clustering the short texts; 2) ranking all short texts of each opinion cluster; and 3) summarizing the description of each opinion cluster. Fig. 1 shows the overview of opinion summarization structure based on BM25 and neural network.

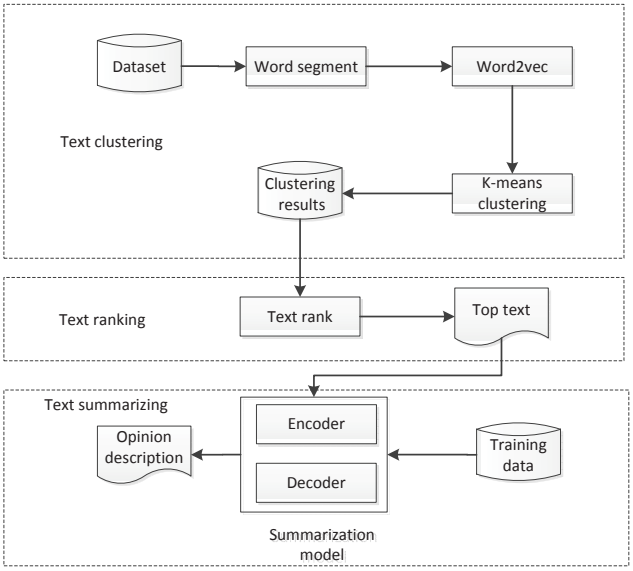


Fig. 1. The structure the OnSeS.

A. Short Texts Clustering

Analysis of Chinese texts is different from texts analysis in English. English words are separated by spaces. Therefore, English sentence segmentation is simpler than Chinese. Segmenting Chinese text accurately is really challenging. We cannot simply split the text into small characters directly for processing since it will lose the semantic relationship. In our system, we use the jieba[1] to parse the short texts to obtain word segmentation. Jieba is built to be the best Python Chinese word segmentation module.

It is a key step to convert the text into data that can be identified and calculated by computers. However, in our work, we use the deep learning algorithm rather than the TF-IDF method because the representation using TF-IDF is sparse. Deep Learning algorithm has made amazing achievements in the field of image and audio, but in the field of Natural Language Processing (NLP), it has not yet seen such exciting result. The reason is that the language (word, sentence, etc.) belongs to the high-level cognitive abstract entity which is produced in the process of human cognition, and the voice and the image belong to the original input signal. The word can be represented by word2vec means that Deep Learning is introduced into the field of NLP as a core technology. The first step is finding a way to make these symbols represented mathematically.

[1]https://pypi.python.org/pypi/jieba/

Deep learning in general uses a distributed representation of a low dimensional real vector often called word embedding. Word embedding makes the relevant or similar words close in distance. There are two main types of contributions of word embedding. The first one is improving the performance of the system through applying word embedding to existing systems. The second one is computing the similarity from the perspective of linguistics. The representation of this vector is not unique. The dimension of the vector is set to 50 or 100 generally.

Our representation of short text is based on the word2vec, with each word is represented by a vector. In order to represent texts by vectors, we train the word2vec to produce a vector with 100 dimensions. We choose a fixed number of words to represent the short text rather than all words of the short text so that the dimension of vectors had been effectively decreased. The value of each dimension of each short text is described by Eq (1):

$$w_{i,j} = \frac{\sum_{k=1}^{m} v_{i,k,j}}{m} \tag{1}$$

where $w_{i,j}$ denotes the weight of ith dimension of the short text d_j. m is the number of selected words and m is 20 in our work.

Therefore, the short text is represented by a vector, as shown in Eq (2), where n is the dimension of the vector. In this paper, we use the Euclidean distance given in Eq (3) to measure the similarity of two short texts.

$$V(d_j) = (w_{1,j}, w_{2,j}, ..., w_{n,j}) \tag{2}$$

$$dis(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (w_{ik} - w_{jk})^2} \tag{3}$$

As the clustering algorithm should be scalable with high-dimensional and large-volume data, we use the common K-means clustering algorithm. It is a simple and fast algorithm based on partition. The basic idea is to divide a dataset into several clusters based on the similarity among the data. We first randomly choose 2 short texts as the beginning cluster centers. Then all short texts will be assigned to the nearest cluster center. After that, the new center is formed by averaging all short texts in one cluster. Repeating the process until the centers of the clusters become constant. Thus each cluster represents one kind of opinions. In our system, we also set a parameter to limit the average Euclidian distance between observations and centroids in these clusters in case of the low efficiency. If the average Euclidian distance is greater than the threshold value, the number of the clusters will increase by 1 until the condition is met. The progress of K-means using word2vec is illustrated in Algorithm 1.

Algorithm 1 K-means using word2vec

Input: short texts $dataset$, the clusters number k_1
Output: the clusters number k_2, the clusters
Define: n_1:20; n_2:100; $threshold$:0.75
for all $d \in dataset$ **do**
  $Words \leftarrow$ cut d
  for all select n_1 words $\in d$ **do**
    $vector$ with n_2 dimensions $\leftarrow$ represent $word$
    $vectors = vectors + vector$
  end for
  $vectors = vectors/n_1$
end for
$k_1 \leftarrow 2$
$label \leftarrow true$
for all $label = true$ **do**
  $label \leftarrow false$
  for all $number \in k_1$ **do**
    $centroid \leftarrow$ random $vectors$
  end for
  clusters $\leftarrow vectors$
  average distance $avg \leftarrow$ Euclidian distance
  if $average < threshold$ **then**
    $k_1 \leftarrow k_1 + 1$
    $label \leftarrow true$
  end if
end for
$k_2 \leftarrow k_1$
Return k_2

B. Short Text Ranking

We now identify opinion clusters, which are used to express the different opinions on a particular topic. Clearly, identifying the representative opinion is related to the existing research on TextRank which is a graph-based ranking model [19]. Some other studies built a graph for a newspaper article, whereas our work is for short text with no structure or semi structure. Therefore the similarity of short texts based on bag-of-words is more effective than that based on co-occurrence relation.

The graph-based ranking model is aimed at deciding the importance of every vertex in a graph. We apply this ranking model to natural language processing for extracting the importance information. In this paper, we introduce the ranking model based on graph, and we investigate the processing of unsupervised extraction for short texts.

The main idea of ranking model is that one vertex is voting for another vertex if they are linked together. Meantime, a score will be assigned to a vertex according to its votes. Formally, we build an undirected graph $G = (V, E)$ where V represents the set of vertices and E represents the set of edges. The score of a vertex V_i is given in Eq (4):

$$S(V_i) = (1 - d) + d * \sum_{V_j \in V_i} \frac{\omega_{ij} * S(V_j)}{\sum_{V_k \in Out(V_j)} \omega_{kj}} \tag{4}$$

where d denotes a damping factor that is usually set to 0.85 [20]. The initial value of $S(V_i)$ is 1.0. In our system, the graph includes multiple links between vertices, consequently we add a weight to the edge that connects two vertices. It is important to notice that the final score of a vertex in a graph is obtained after iterating to convergence.

Such a relation of vertices is usually measured by the similarity. In our paper, we use the BM25 retrieval function to measure the similarity, which is based on the frequency, regardless of the relationship between the terms as well as the position of the term. The BM25 formula, is presented in Eq (5):

$$\omega(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\cdot(k+1)}{f(q_i,D)+k\cdot(1-b+b\cdot\frac{|D|}{avdl})} \quad (5)$$

where $avdl$ denotes the average doc length, $|D|$ is the length of doc D and $f(q_i, D)$ represents the frequency of q_i in D. In addition, b and k are tuning parameters. The weight of the edge between short texts D and Q is what we want to calculate, in other words, that is the similarity of D and Q.

After performing several iterations of the algorithm, scores are assigned to short texts. Finally, we use the final score as the importance degree of the short text. It presents each opinion cluster by top ranked short text with highest scores, then the text is selected for extracting summary.

C. Opinion Summarization

The opinion summarization facilitates quick reading, thus effective summarizing is meaningful for users to observe the topic comprehensively. Opinion summarization is a condensed representation of the top ranked short text, which is regarded as a highly difficult task and an important challenge. Our summarization is a data-driven approach to generate each word of the summary.

As we all know, statistical machine translation has already shown great success on natural language processing. Recently, neural machine translation proposed by cho et al. [21] is a novel approach, which reads a sentence and outputs a translation. The basic neural machine translation model is the encoder-decoder RNN (recurrent neural network) that has two RNNs for encoding the input text into a fixed-length vector and decoding to a correct translation respectively. The encoder RNN encodes the input sequence into a fixed-length vector. In contrast, the decoder RNN decodes the vector with a different sequence. The limitation of the basic encoder-decoder RNN model is that it has a subsequent effect with the increase length of the input text, because of the encoder RNN reads and encodes the input text to a fixed-length vector. Hence, there are many neural machine translation models which are variants of encoder-decoders.

In recent years, RNN has shown strong abilities on natural language processing, especially on speech recognition and machine translation. We use the word-based method which segments the text into words. The original order and relationships between neighboring words are considered in the encoder RNN, and the proved model gated recurrent unit (GRU) is adopted.

IV. Experimentss

We have conducted experiments to verify the effectiveness of our system, OnSeS. In this section, we present the corpus used in all experiments, and the experimental results and analysis. Our summarization system, called OnSeS, which is based on BM25 and neural machine translation, has been implemented with Python.

A. Experimental Setup

In order to verify the performance of OnSeS described in Section III, we carried out a series of performance studies on real datasets. Generally speaking, the datasets for text clustering always come from TREC, the medical literature (MEDILINE), etc. The public and standard available dataset DUC, TAC or TREC only has hundreds of English text for summarization. Furthermore, the public corpus of Chinese text for summarization is very rare and not yet very mature. In the experiment part, we use 2,400,591 short texts and the summary pairs provided by Hu et al. [22] to train the model for summarization. In addition, datasets 1 and 2 with 35,377 pairs in total have been used for testing the summarization.

The large-scale available dataset is crawled from formal and informative Sina Weibo posted by official organizations who verified their accounts. We also use the dataset 3 contains 18,126 short texts in total crawled from Twitter. In addition, the dataset 4 is used for testing the performance of short text clustering. All the datasets used are exposed in Table I.

TABLE I
DATASET DESCRIPTION USED IN ONSES

Dataset	Source	Number of short texts
Training dataset	Hu	2,400,591
Testing dataset 1	Hu	24,951
Testing dataset 2	Hu	10,426
Testing dataset 3	Twitter	18,126
Testing dataset 4	Sogou	572

To validate the efficiency of summarization based on BM25 and neural network, in this paper, we set up two experiments. In the first experiment, we applied the baseline method, information retrieval (IR) [23], which gives the sentence with the highest BM25 score as the summary of the short text. To summarize the online opinion, we select the top ranked short text as the input and split the input into sentences. After that, we rank the sentences in the same method and obtain the top ranked sentence. In the second experiment, we applied neural network used for machine translation, which was described in Section III (C). The summarization model was trained on the GPUs GTX TITAN Z for about 5 days with the training dataset provided by Hu et al. We set the vocabulary to 50,000. The next subsection showed the experimental results.

B. Results and Analysis

The standard metrics of evaluating short text clustering performance include Precision, Recall and F-measure. In Table II we show our short text clustering results on testing dataset 4 using the three evaluation metrics mentioned above.

In the ranking experiment, we process the short texts by building a graph-based ranking model and with similarity based on BM25. We evaluate the short texts ranking task based

TABLE II
CLUSTERING RESULTS ON DATASET 4

Clusters	Precision	Recall	F-measure
Cluster 0	0.896	0.986	0.939
Cluster 1	0.984	0.877	0.927
Average	0.94	0.932	0.933

on BM25, where b is set to 0.75. Text Ranking succeeds in identifying the most important short texts.

For the summarization task, we evaluate the performance of the system through several variants of the ROUGE metric proposed by Lin et al. [24], which have been shown to be successfully associated with human evaluations. Therefore, we adopt the ROUGE-1 (unigrams), ROUGE-2 (bigrams), AND ROUGE-L (longest-common substring) in this evaluation part. Formally, ROUGE-N is an n-gram recall between the generated summary and the reference.

Our experimental results are shown in Fig. 2. We run experiments using both datasets 1 and 2 on IR and OnSeS. The ROUGE-2 value is small than the values on ROUGE-1 and ROUGE-L due to the low probability of two neighbor words in the reference summary that appear in the generated summary. We first note that the value of OnSeS increases more than two times on ROUGE-1 and ROUGE-L, and more than three times on ROUGE-2 particularly. The main result of objective evaluation is that significantly better results can be obtained when using the neural network, in comparison with extractive method.

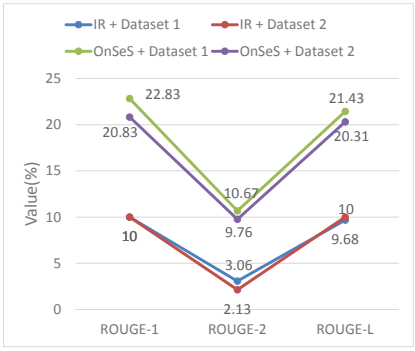


Fig. 2. Experimental results evaluated with objective various ROUGE metrics on IR and OnSeS.

As shown in Fig. 3, some example summaries produced by OnSeS are very close to human written summaries. For our training data, we limit the vocabulary to 50,000, and all the out-of-vocabulary words are replaced with UNK. Therefore, the summaries maybe contain many UNKs. In the summarizing experiment, we gave each summarization a score between 1 to 5: A score of 1 if the summarization could not represent the short text and a score of 5 if the summarization represented the short text correctly. A higher score means a stronger summarization. The human labeled score represents the comprehensiveness of the summarization. We show the percent of each score labeled by twelve volunteers on datasets in Fig. 4. As seen in Fig. 4(d), the percentage of sore of 1 by

I(1):11月25日，10岁女孩李某趁原原奶奶出电梯时，将原原抱起，在电梯里殴打了原原。据李某陈述，她将原原抱回家中，又在沙发上对原原实施殴打，后将原原抱至阳台栏杆上逗玩，致原原从25楼坠落。
On 25th November, 10-year-old girl Lee beat yuanyuan in an elevator when her grandma went out from the elevator. According to Lee, she took yuanyuan to home, and abused him on the sofa. Later, she took him to the balcony. While playing, yuanyuan fallen down from 25th floor.
R:重庆通报“女孩摔打男婴”案
The case that “a girl beat baby boy” was reported from Chongqing
G:重庆被摔打男婴案:女孩被女孩摔打后将其从25楼里摔打1至1将其从UNK从25楼坠落(高清组图)
The baby boy beaten case in Chongqing: a girl beat a baby girl at the 25th floor and caused him fallen down from the balcony.

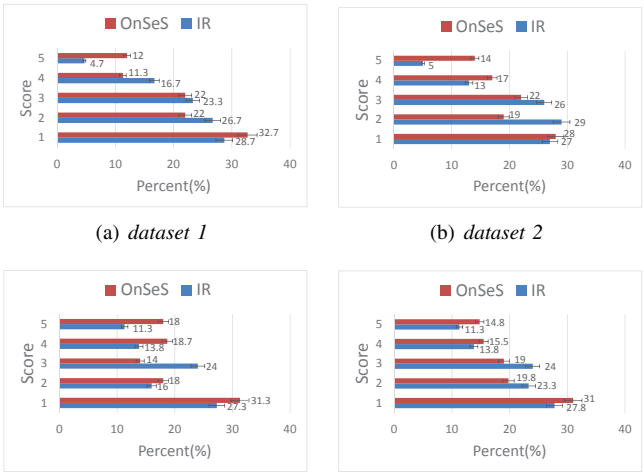
I(2):由于高尔基公园超负荷运转，该公园负责人称：想要打造“新概念”公园，吸引更多高素质人群和富裕群体来园游玩，未来很有可能只对高素质人群和富裕人群开放。
Due to the overload in the Gorky Park, the park manager says that they want to construct a “new concept” park which attracts more highly qualified and rich people. In the future, the park maybe will be specified only for highly qualified and rich people.
R:俄高尔基公园欲只对高素质和富裕人群开放
Russia Gorky Park attempts to open to only highly qualified and rich people.
G:""新概念""公园吸引更多富裕人群与富裕群体或在UNK游玩(图)
“New concept” park attracts more rich people to visit UNK.

I(3):春苗营养计划是为了改善贫困地区儿童营养状况。截止2014年9月，“春苗营养计划”覆盖全国20个省，受益儿童达150万人。安利公益基金会将在新疆维吾尔自治区建设总计80所春苗营养厨房。
Chunmiao nutrition plan is to improve the nutrirional status of children in poor areas. Until September 2014, the “chunmiao nutrition plan” covered 20 provinces over whole country and over 1.5 million children benefit from the plan. Amway Charity Foundation will build 80 Chunmiao nutrition kitchens in the Xinjiang Uygur Autonomous Region.
R:安利公益基金会第3000所春苗厨房落户新疆
The 3000th chunmiao kitchen by Amway charity foundation is located in Xinjiang.
G:安利公益基金会成员受益于今年9月“UNK营养计划”20个受益儿童达150万!
The number of the children who benefit from “the UNK nutrition plan” founded by Amway charitable foundation has reached 1.5 million until September.

I(4):从2005年2月调入足管中心担任主任，谢亚龙仅在足管中心待了3年多，但这3年却成了他的“发家史”。在其172万元的受贿款项中，包括2006年耐克公司市场部总监李彤向其行贿的人民币17.84万元。
From being promoted as the manager in February 2005, Xie Yalong has been working in the foot massage center for over 3 years. And he had built his fortune during these 3 years. He had accepted bribes in total 1,720,000 yuan, including 178,400 yuan that Li Tong, who is the marketing director of the Nike company bribed in 2006.
R:谢亚龙受贿172万详单：耐克公司行贿17.84万
The detailed list of bribes in total 1,720,000 yuan that Xie Yalong accepted: Nike company marketing director Li Tong bribed 178,400 yuan.
G:谢亚龙受贿千余万被曝曾在UNK行贿近10年曾给其行贿UNK万(图/简历)(图)--行贿UNK
Xie Yalong accepted thousand million bribes for 10 years in UNK, and once accepted UNK ten thousand yuan bribes (Figure).

I(5):今年第三季度，全新宝马X5xDrvie40e插电混合动力车型将上市；2016年3系插电混合动力版将会上市，以及未来的(即2系插电混合动力车)。（分享自@电动邦）
New BMW X5xDrvie40e plug-in hybrid model will be released during the third quarter of this year: plug-in hybrid third version will be released in 2016, as well as the future product (that is, plug-in hybrid car 2nd version).
R:宝马多款电动车上市计划公布将全部入华
The launch plan for several models of BMW electrical car has been introduced in China.
G:新一代宝马UNK混合动力车型将上市，2016年3系插电混合动力版将上市(图)
A new generation of BMW UNK hybrid models will be introduced to the market, and plug-in hybrid third version will come to the market in the 2016(Figure).

Fig. 3. Example sentence summaries produced by OnSeS. I is the input, R is the reference, and G is the generated summary.



(a) *dataset 1* (b) *dataset 2*

(c) *dataset 3* (d) *datasets 1-3*

Fig. 4. Summarization scores labeled by volunteers in each dataset.

OnSeS is higher than that by IR because there are many UNK in the summary generated by OnSeS. However, the percentage of sore of 5 and 4 by OnSeS is higher than that by IR.

Finally, we show the group statistics of human evaluation in Table III. It can be seen that the mean values for scores by OnSeS are higher than that by IR on datasets 1 and 2.

TABLE III
GROUP STATISTICS OF HUMAN EVALUATION

Methods	Dataset	Mean	Variance	Std. Deviation	Std. Error Mean
IR	dataset 1	2.420	1.440	1.200	0.098
	dataset 2	2.400	1.354	1.163	0.116
	dataset 3	2.850	2.238	1.496	0.122
OnSeS	dataset 1	2.480	1.862	1.365	0.111
	dataset 2	2.700	1.970	1.403	0.140
	dataset 3	2.740	2.288	1.512	0.123

V. CONCLUSION

In this paper, we proposed a novel method for online short text summarization based on BM25 and neural machine translation. In particular, the summary is generated with words which may not appear in the input rather than words which just are extracted from the input. We conducted experiments to verify the effectiveness of OnSeS, and the objective evaluation and human evaluation are both used for comparing the performances between OnSeS and IR. The experimental results show that our proposed a fully data-driven approach OnSeS outperforms state-of-the-art method.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61572060, 61190125, 61472024), and CERNET Innovation Project 2015 (NGII20151004).

REFERENCES

[1] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 232–241.

[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[3] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 2, pp. 849–856, 2002.

[4] R. Nayak, R. Mills, C. De-Vries, and S. Geva, "Clustering and labeling a web scale document collection using wikipedia clusters," in *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning*, 2014, pp. 23–30.

[5] 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.

[6] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[7] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 233–242.

[8] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using wordnet and lexical chains," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2264–2275, 2015.

[9] M. P. Parmar and M. G. S. Pandi, "Performance analysis and augmentation of k-means clustering, based approach for human detection in videos," 2015.

[10] S. S. A. Syed and T. S. Kumaran, "An energy efficiency distributed routing algorithm based on hac clustering method for wsns," *Indian Journal of Science and Technology*, vol. 7, no. S7, pp. 66–75, 2014.

[11] S. Seifzadeh, A. K. Farahat, M. S. Kamel, and F. Karray, "Short-text clustering using statistical semantics," in *Proceedings of the 24th International Conference on World Wide Web Companion*, 2015, pp. 805–810.

[12] A. K. Farahat and M. S. Kamel, "Statistical semantics for enhancing document clustering," *Knowledge and Information Systems*, vol. 28, no. 2, pp. 365–393, 2011.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[14] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining Text Data*, 2012, pp. 43–76.

[15] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, vol. 4, no. 4, 2004, pp. 755–760.

[16] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *AAAI*, vol. 5, 2010, pp. 529–573.

[17] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1535–1545.

[18] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 118–127.

[19] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," 2004.

[20] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] B. Hu, Q. Chen, and F. Zhu, "Lcsts: a large scale chinese short text summarization dataset," *arXiv preprint arXiv:1506.05865*, 2015.

[23] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[24] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the ACL-04 Workshop on Text Summarization*, vol. 8, 2004.