Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 17(4): 3582–3600
DOI: 10.3934/mbe.2020202
Received: 23 March 2020
Accepted: 5 May 2020
Published: 12 May 2020

Research article

Topic-based automatic summarization algorithm for Chinese short text

Tinghuai Ma[1,*], Hongmei Wang[1], Yuwei Zhao[1], Yuan Tian[2] and Najla Al-Nabhan[3]

[1] Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] Nanjing Institute of Technology, Nanjing 211167, China

[3] King Saud University, Riyadh 11362, Saudi Arabia

* **Correspondence:** Email: thma@nuist.edu.cn.

Abstract: Most current automatic summarization methods are for English texts. The distinction between words in Chinese text is large, the types of parts of speech are many and complex, and polysemy or ambiguous words appear frequently. Therefore, compared with English text, Chinese text is more difficult to extract useful feature words. Due to the complex syntax of Chinese, there are currently relatively few automatic summarization methods for Chinese text. In the past, only the important sentences in the original text can be selected and simply arranged to obtain a summary with chaotic sentences and insufficient coherence. Meanwhile, because Chinese short text usually contains more redundant information and the sentence structure is not neat, we propose a topic-based automatic summary method for Chinese short text. Firstly, a key sentence selection method is proposed combining topic words and TF-IDF to obtain the score of each text corresponding to the topic in the original text data. Then the sentence with the highest score as the topic sentence of the topic is selected. Considering that the short text of Weibo may contain a lot of irrelevant information and sometimes even lack some important components of topic, three retouching mechanisms are proposed to improve the conciseness, richness and readability of topic sentence extraction results. We validate our approach on natural disaster and social hot event datasets from Sina Weibo. The experimental results show that the polished topic summary not only reflects the exact relationship between topic sentences and natural disasters or social hot events, but also has rich semantic information. More importantly, we can almost grasp the basic elements of natural disaster or social hot event from the topic sentence, so as to help the government guide disaster relief or meet the needs of users for quickly obtaining information of social hot events.

Keywords: Chinese short text; automatic summarization; topic sentence; natural disaster; social hot event; Sina Weibo

1. Introduction

With the development of Internet and the growing popularity of various social network platforms, millions of new messages are generated every day. They are spreading rapidly among the internet and people within several seconds. The social network platform has not only become the important way for most users to elaborate news and express their views, but also an important place for hot topics to be generated and spread. Taking Sina Weibo (China) as an example, the active users has reached 100 million and the daily number of micro-blogging posts is up to more than 400 million as of June 2018 [1, 2]. In China, people are accustomed to follow current events on Weibo. And news and events spread very quickly on Weibo, which can reach millions of retweets within a few minutes. Therefore, collecting Weibo texts and extracting topic sentences from them can understand the hot events that people pay attention to and grasp the trend of public opinion. In addition, sudden natural disasters can be discovered in time and detailed information about the disasters can be obtained, thereby providing assistance to the government in disaster relief. Above all, it will help the government to keep abreast of the network public opinion and guide the public opinion correctly.

The previous topic extraction method is normally relying on the scattered list of several keywords to achieve the effect of topic representation [3]. However, we can hardly get the core meaning of the topics and obtain what we want accurately in this way. Therefore, we hope to achieve a one-sentence summary of each topic. Through research, we found that the automatic summary algorithm can meet our demand which can extract important statements from long text and form a short summary [4]. Therefore, based on the analysis of automatic summary methods, a topic sentence extraction method is proposed to obtain a topic sentence containing elements such as time, place and things. Through this topic sentence, the main information of natural disasters or social hot events can be understood.

We mainly face the following challenges. On the one hand, most of the automatic summarization algorithms are often applied to long texts or novels in the previous research [5–8]. However, the Weibo text studied in this article is the short text with only 140 words or less. In addition, there is a lot of noise data in the microblog texts [1]. There are millions of blog posts every day, even if only the texts related to natural disasters or social hot events are captured, it will also contain a lot of interference information. Therefore, if the original dataset is utilized, a lot of irrelevant information will affect experimental results. In this paper, we propose an automatic summarization method based on topics for Chinese short text, which is combined topic probability model and feature words. We use the topic words obtained by topic detection based on graph analysis algorithm [9, 10] and TF-IDF [11] to sort the short text of Micro-blogging. Each topic word is assigned, and the score of each Micro-blogging post is counted. We considere that the larger the score of the post is, the richer the information it contains and more complete the topic is. Therefore, the text with the highest score is selected to prepare for the next step. After getting the key sentence, we hope to simplify the key sentence and add the missing information in the topic sentence extraction step. It mainly contains three steps: Sub-sentence filtering, information supplement and words-order adjustment [12]. Through the above processing, we have completed one sentence summary of the topic which is also the final result of topic sentence extraction. We use several evaluation indicators on the real Sina Weibo dataset, which is Chinese short text, to evaluate the proposed topic-based automatic summarization algorithm(TASA).

The main contributions of this article are as follows:

(1) We propose a key sentence selection method based on the topic word weight and TF-IDF value

to calculate the score of each blog post. According to the blog post score corresponding to each topic, representative and complete information sentence of each topic can be obtained from the complex microblog data. This method does not need to model the topic distribution of sentences in advance, but uses topic words to rank the importance of sentences while adding the classification and weight of topic information. In other words, this method simultaneously completes the topic filtering and sentence sorting process.

(2) We propose a novel topic sentence extraction method, which is to polish the topic sentence, including sub-sentence filtering, information supplement and words-order adjustment to improve the quality of topic sentence. The sub-sentence filtering makes the topic sentence refined and concise, which is more conducive to quick reading and grabbing of the topic content for people. Information supplement can add some important but non-appearing components to the topic sentence, which can improve the richness of the topic sentence. Words-order adjustment makes topic sentences more fluent and increase sentence readability.

(3) The necessity and advantages of touch-up mechanisms were confirmed in experiments. We compare our method with other five topic sentence extraction methods. However, the performance of TASA is superior to other baseline methods in ROUGE-1, ROUGE-2, ROUGE-SU4 and manual evaluation, which greatly proves the superiority of proposed method.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed topic-based automatic summarization algorithm for Chinese short text, followed by experimental setup and experimental results in Section 4. Finally, the conclusion and future work are discussed in Section 5.

2. Related works

The classic automatic summarization methods are based on feature words. These approaches used statistical techniques to extract surface-level features such as keywords, titles, topic words and prompt words [13–15]. Then they calculated the weight of sentence to get several key sentences. Our main idea is coming from this aspect. Milad and Nasser [16] proposed a Bayesian summarization method to map the input text to the Unified Medical Language System concepts and then selected the important ones to be used as classification features. In addition, there are many special features applied to the process of summarization extraction, such as hashtags, timestamps, and emotional vocabulary [17]. However, these methods may get feature words that do not match the topic, which will reduce their accuracy. Therefore, the researchers added the lexicon of a specific domain to enhance their semantics. Liu et al. [18] proposed a core semantic extraction model (CSEM), which considers the structural features of the core semantics of news corpus to reduce semantic redundancy and improve the novelty of abstracts. In order to enhance the richness of semantics, the semantic unit, that is, the association relationship of a group of keywords, is used to express the semantics of the text. A decay function is also introduced to adjust the importance of semantic units according to the time when the semantic units appear in the summary sentence candidate set to form a summary. Finally, CSEM extracts the minimum number of sentences to cover the core semantics of the corpus as a summary.

Probability topic model is used in automatic summary generation because of its excellent topic characteristics which can solve the topic-related problem to a certain extent. It aggregates documents into a long document with no boundaries and flexibly utilizes a scoring mechanism to rank sentences

such as K-L divergence [18]. LDA is the first topic model which can solve the problem of multi-document summarization and the researchers continuously improve this algorithm. Zhou and Zhong [19] used a semi-supervised learning framework based on LDA model to extract biomedical events. The sentences in the unannotated corpus are elaborately and automatically assigned with event annotations based on the calculated distance, which is described by the hidden topic distribution and the structure of the sentences in the annotated corpus and the sentences in the unannotated corpus. Xiong and Diane [20] combined LDA and review metadata, which is called review helpfulness ratings, to facilitate review summarization and form the helpfulness-guided review summarizers. Due to the proposed method is metadata driven, it does not require manual annotation and can be generalized to different types of online reviews. Wu et al. [21] proposed a topic modeling based approach to extractive automatic summarization. It extracted the candidate sentences associated with topic words from a preprocessed novel document and designed an importance evaluation function to select the most important sentences from the candidate sentences to generate an novel summary.

Moreover, the graph-based method is widely used in the field of automatic summarization algorithms [22]. Due to the emerging application of graph technology, graph-based approaches have been used to generate the summary of corpus. These methods consider the internal structural relationships between words or sentences in the document, rather than using words and other features only to rank sentences [23]. Fang et al. [24] proposed a word-sentence co-ranking model named CoRank, which combines the word-sentence relationship with the graph-based unsupervised ranking model to realize the automatic summarization. The extractive method proceeded by creating an intermediate representation of the original text, scoring sentences and finally selecting sentences with high scores as the summary. Manos et al. [25] proposed a multi-pattern time digest algorithm MGraph. MGraph is a graphics-based framework that creates a visual summary of real-world events through post-mortem analysis of event-related post flows. It calculates the importance of each piece of information according to the received social attention, and uses topic modeling techniques to capture the relevance of posts to topics. Then, it uses the graph-based ranking algorithm DivRank to get a set of relevant and important posts while maximizing the coverage of the event and minimizing visual redundancy. Ye et al. [26] proposed a two-layer graph algorithm combining sentences and words for multiple documents, and generated multiple document abstracts, namely charts and keywords. It extracted the correct keywords in each document to correct the sentence graph and improve the richness and meaning of the abstract.

In summary, these methods have their own strengths and considering our goal of extracting topic content in one sentence, we decide to combine feature words and topic model to form the summary. Therefore, we intend to start from the selection of feature words by NLPIR technology and TF-IDF value. Then, in order to improve the semantic relevance, we utilize the topic words obtained in the topic model to make the sentence marked with topics. Then we calculate the score of each sentence and pick out the highest one to form the key sentence. Finally, the key sentence is polished to get the final topic sentence.

3. Topic-based automatic summarization method for Chinese short text

In this section, we will detail our automatic sammarization model which can be divided into two steps. Firstly, due to the original data is noisy and sparse, we utilize a novel sentence selection method

based on topic words and TF-IDF to sort the data and filter most of useless messages. According to the research in our previous article [9], we can get topics from the initial data through the community detection method and select the top 10 topic words for each topic according to the topic word extraction algorithm. Therefore, we can use the topic words to sort the text of each topic without building topic model at first. In other words, it completes topic filtering and sentence sorting at the same time. Then, we choose the top one sentence to be the key sentence. Secondly, the initial key sentences usually contain some redundant information, useless information or missing some important factors, so we have to modify the key sentences further. We will correct the key sentence from three aspects: Sub-sentence filtering, information supplement and words-order adjustment. Through the above process, we can improve the readability and richness of the key sentences to form the final topic sentences.

Algorithm 1: Topic-based automatic summarization algorithm for Chinese short text

Input: Original blogs $(b_1, b_2, ..., b_k)$
Topic words set of topic $z_r(z_{r1}, z_{r2}, ..., z_{ri})$.
Output: Key sentences $(b_{z1}, b_{z2}, , b_{zr})$
Topic sentences $(B_{z1}, B_{z2}, , B_{zr})$

```
data preprocessing with NLPIR technology;
for z_1 to z_r do
    for b_1 to b_k do
        if w_i ∈ z_r then
            Score_1 += S(w_i), Score_2 += TF − IDF_{w_i};
        else
            Score_2 += TF − IDF_{w_i}
        end
        Score = Score1 + Score2/N;
    end
    get the Score of each blogs;
    sort blogs and return the highest blog b_{z_i};
end
get key sentence of each topic (b_{z_1}, b_{z_2}, , b_{z_r});
for b_{z_1} to b_{z_r} do
    sub-sentence filtering;
    information supplement;
    words-order adjustment;
end
get topic sentences (B_{z1}, B_{z2}, , B_{zr});
```

The pseudo code of the topic-based automatic summarization algorithm for Chinese short text(TASA) is shown in Algorithm 1. We will describe the details in the following two subsections.

3.1. Key sentence selection

In traditional automatic summary approach, it usually selects the sentence directly depending on the word frequency or the total weight of TF-IDF value. However, the method based on word

frequency will result in redundant information and the method of TF-IDF will decrease the accuracy of key sentence filtering due to the short text of Micro-blogging. In this paper, we propose a novel key sentence selection method combining topic words and traditional TF-IDF values based on feature words. This can make up for the shortcoming of missing important information, without the need to model the topic distribution in advance. In other words, the method completes both topic filtering and sentence sorting. Each different word is considered as the feature word. The following are the specific steps of the key sentence selecting algorithm.

Firstly, we utilize Eq (3.1) to calculate the weight of each feature word. A word is considered more representative if it appears more frequently in a post and appears less frequently in the remaining posts. In order to improve the efficiency of the algorithm, we remove the stop words and noise words by text preprocessing and NLPIR technology. Then, we calculate the weights of the remaining feature words.

$$TF - IDF_{w_i} = TF_{w_i} * IDF_{w_i} = TF_{w_i} * logN/N_w \tag{3.1}$$

Where TF_{w_i} is number of times the word w appears in Micro-blogging i while IDF_{w_i} measures the percentage of all blogs in the Micro-blogging text collection that contains this word w. N represents the number of blog posts in the Micro-blogging text data set, and N_w shows the number of blogs containing the word w.

After the above steps, we can get the weight of each feature word using TF-IDF. However, the topic information alone cannot distinguish the topic information or each topic. Therefore, drawing on the idea of topic word [9] from our previous research, the contribution of topic words is added, and the weight of each topic word is calculated by the topic word extraction algorithm. The method in reference [9] is based on graph analysis, which uses the community structure detection algorithm to detect topics in the feature word graph of the original micro-blogging data, that is, the feature words describing the same topic are divided into the same community. Therefore, each feature word corresponds to a topic number (ie, a certain community). In other words, each topic has a set of feature words, which can describe the content of the topic in detail. Then, since each word has its own topic label and weight, the most relevant topic words are extracted for each topic based on the score of the feature word in each topic. The Eq (3.2) below defines S to represent the weight score of each feature word in each topic. First, we calculate the weight $S(z_{ri})$ of each feature word in each topic according to Eq (3.2), and then arrange the feature words in descending order of weight score. We found that the feature words ranked in the top 10 are the most influential [27]. When the number exceeds 10, the topic relevance has begun to decline. So we utilize the top 10 feature words of each topic as the final topic words. The weight of these topic words is calculated by Eq (3.2) which can maximize the role of topic words [9].

$$S(z_{ri}) = \frac{\#\sum_{l}^{N}(w \in l_i)}{\#\sum_{l}^{N} l_i} \times \frac{\#l_w}{\#N} \tag{3.2}$$

Where z_{ri} represents the word w_i belongs to the topic z_r. l_i is the unordered list of words of blog post i. $\#\sum_{l}^{N}(w \in l_i)$ represents the number of times that the word w appears in the blog post i. $\#\sum_{l}^{N} l_i$ is the number of all words contained in the blog post i. $\#l_w$ is the number of users that have used word w. And $\#N$ is the total number of all users.

After the above analysis, we not only get the TF-IDF value of each feature word, but also get the weight of each topic word. Due to the topic words are belongs to feature words, the topic word and feature word must overlap when calculating the sentence score which is a considerable problem. In summary, this paper will use these feature words and topic words to calculate the score of each blog post. Since it is not scientific and prudent to artificially increase the weight of topic word, when encountering a topic word in a sentence, not only do we need to consider its score as a feature word, but also need to add its weight as a topic word. To simplify the complexity of the algorithm, we make judgments for each word in each blog. If it only appears in the feature word set, its TF-IDF value is accumulated. If it belongs to a topic word, accumulate its TF-IDF value first and accumulate its topic word weight for this sentence additionally. However, this will result in a problem that if a sentence has a long content with a large number of feature words, then the score of the sentence will be inevitably higher. In order to avoid the longest sentence always being in the summary generation result, this paper uses the average value of the TF-IDF values in each sentence as the final feature word weight. Due to the number of topic words is scarce, we utilize the sum of the topic word weights to enhanced the topic information. Therefore, the score of each blog is calculated by Eq (3.3). The higher the sentence score is, the richer content and the more complete representation the topic has.

$$Score(b_n) = \frac{\sum_{i=0}^{m} W_i}{m} + \sum_{i=0}^{n} S(z_{ri})(w_i \in Z_r) \tag{3.3}$$

Where b_n is the n-th blog post. $\#\sum_{i=0}^{n} z_{ri}$ represents the number of how many topic words the blog contains. W_i represents the TF-IDF weight of the feature word w_i in the blog post and m is the number of feature words in each blog post. $(w_i \in Z_r)$ indicates whether the feature word belongs to the topic word, and $\sum_{i=0}^{n} S(z_{ri})$ represents the sum of the topic word weights included in each blog post.

Then we sort these posts by their scores and pick out the highest one as the key sentence of topic r.

3.2. Topic sentence extraction

The traditional automatic summarization algorithm can only select important sentences in the original text and simply arrange them to obtain a summary result that is not coherent in context. However, Chinese text usually contains more redundant components or noise information and the sentence structure is not neat, so The generated topic sentence results need to be polished to enhance the conciseness and fluency of the summary results.

At this point, we have gotten the key sentences $(b_{z1}, b_{z2}, , b_{zr})$ for each topic. Since the microblog has much extra or irrelevant information, and even sometimes is lack of some important factors, we can not use the original microblogging post as the result of topic sentence extraction directly. It is necessary to polish the key sentence results generated by the automatic summary to enhance the conciseness and smoothness of the results. To solve this problem, we add three adjustment mechanisms: 1) Sub-sentence filtering, 2) information supplement and 3) words-order adjustment. The sub-sentence filtering part is aiming to prevent the content of some clauses from being empty or containing too less information. It can improve the pithiness of topic sentence extraction. The information supplement part can supplement some important missing components in key sentences. Although the possibility of this situation is small, it is also essential for the confidentiality of the

algorithm and improving the integrity of sentence information. The last part is words-order adjustment which is in order to make the whole sentence more fluent and the related information more concentrated. At the same time, the words-order adjustment also deletes the spaces, the carriage return and other such symbols between the sentences, and then indents the first line. In this way, the readability of the sentence is improved.

(1) sub-sentence filtering

Although the text in Sina-microblog is short, it can not be fully explained in one sentence. Usually, a post contains several sentences and each sentence contains several clauses. In order to select the important components of the sentence, we divide the posts into sub-sentences with semicolons, periods, question marks, etc. The normal sentences will contain three components with subject, predicate and object which have at least three words. When the number of clauses is less than or equal to three, the filtering has no significant meaning. Therefore, if the summary result contains more than three clauses, we firstly delete the clause containing less than four words. Secondly, the score of the remaining clause is calculated using formulate (3.3). If the value of the clause is bigger then the deleted one, then it is retained, otherwise it is deleted. There is no need to worry about the filtering, because the following step will ensure the sentence integrity. Through the filtering of this step, the summary results become refined and concise.

(2)information supplement

Due to the filtering effect, some essential element information may be removed, or the necessary information is not included in the original key sentence. In order to increase the information integrity of the topic sentence, we choose time, place and event as three essential factors. Therefore, we have to judge whether these three factors are included in this key sentence. That is to say whether this key sentence contains time words, position words and nouns which can be represented by three parts of speech $/t$,$/f$, $/n$. If a part is missing in the topic sentence, find the corresponding information in the topic word (if the information is included in the topic word) and add it to the topic sentence. According to the arrangement habit of time, place, person and event in Chinese expression, the principle of information supplement is to add the time word to the beginning of the sentence, add the place word after the time word and add the event word after the place. In this way, the basic element information in the topic sentence is complete and comprehensive.

(3)words-order adjustment

The content of the blog post is usually composed of a few sentences and the format is not uniform. Therefore, the spaces and newlines between sentences are deleted before the words-order adjustment, and the first line is indented. The sentence after the above processing, it seems to be a little incoherent. According to the habit of Chinese expression, when the sentence contains words such as, because, due, etc., these prepositions are usually at the beginning of the sentence. Therefore, we define that if the clause contains the preposition or the positional word, it will be unified in advance. It will increase the readability and consistency of the sentence to a certain extent. Through the above three steps, we will get the final topic sentence which is also the final topic sentence extraction result.

4. Experiments

The prototype of the proposed method is implemented mainly in JAVA SE 1.8. The data in this experiment is stored in mySQL. We are concerned with the Chinese short text data on Sina Weibo.

People can publish any opinions, events and feelings about recent news or events, personal encounters, emotional expressions and so on on the Sina platform. All we have to do is to crab the Weibo short text and analyze those data. So it can be imagined that the noise data in these data is very large which increased the difficulty of topic detection. In this paper, we focus on our experimental data on natural disasters and social hot events. Due to the wide variety of natural disasters, we only focus on earthquakes, typhoons, floods, debris flow and heavy rain. The content of the data is unordered and contains many extraneous data. The data is downloaded by crawler application written in python through the application of Sina microblogging API interface. We follow the microblogging account like the global weather, news, government, other authoritative bloggers and ordinary users and crawl the blog posts considering the microblogging user id and the content from June 2017 to August 2017. Due to the number of data collection restrictions of microblogging API interface and the strict review of the user application for the developer platform, we finally collected 3595 records related to natural disaster data which includes earthquakes, typhoons, floods, debris flow and heavy rain. For social hot event data sets, short text data of Weibo events is manually collected from the Sina Weibo platform by querying by entering event keywords. The data set contains 3 real social hot events, the time span is from June 2015 to May 2016, which are the sinking of the Yangtze River passenger ship, the Tianjin Tanggu blast and girl attacked at Yitel. All the data we collected is in Chinese. In the next step, we will use these data independently to test the experimental performance. The overall experimental data is shown in Table 1.

Table 1. Dataset.

Dataset name	Blog posts	Words	Distinct feature words
natural disaster	3595	53960	9984
social hot event	7697	115530	13704

To deal with the complicated data, we utilize NLPIR technology to preprocess data. NLPIR [28] is a powerful word segmentation tool, it can not only divided text content into the semantic feature words and phrases, but also be able to mark part of speech of each feature word which is contributed to the selection of characteristic words. With the support of NLPIR, we transfer each micro-blogging data into the words form and treat each word as a feature word. Remove the duplicate words and we get distinct feature words.

To evaluate the proposed topic sentence extraction method, we compared our proposed method with the method only rely on TF-IDF [29] and topic words [9], respectively, two mature algorithms: Extractor, and AutoSummarize [4] and a topic based method: Topic based Summary [21]. The following is a brief introduction to the principles of these comparison methods.

TF-IDF [29]: It only calculates the average weight of the words which is a basic method. While the value is bigger, the sentence is more important. We also choose the top one sentence to be the result of topic sentence.

Topic words [9]: It only relies on the number of occurrences of topic words obtained by topic detection method based on graph analysis in each blog to measure the importance of the sentence. The more times a topic word appears, the more relevant it is to the topic. When there are multiple blog posts that get the same score, their importance depends on the type of topic words included in the blog

post and the length of the text. After sorting, we get the highest one as the experimental result.

Extractor [4]: For summarizing documents, this system uses a genetic programming approach which itself provides an automatic learning process, which allows the summarizer to work on different domains without re-training it. This approach is the result of the research carried out in Turney [4], where several learning algorithms were analyzed and evaluated for determining the best for the key phrase extraction task. Currently, it is also a commercial system which has an online demo for testing it. Since it usually extracts 5 pieces of sentences, we only select the first one as a comparison here.

Auto Summarize [4]: This summarizer is integrated into Microsoft Word and it also generates summaries in several languages. It is a commercial system and the method mainly uses location information and statistical features to achieve summary generation based on the experimental analysis. It can select the proportion of the original text that needs to be generated according to the user's needs. In this paper, we select the results with the lowest proportion to be the comparison experiment results.

Topic based Summary [21]: It is a topic modeling based approach to extract automatic summaries, so as to achieve a good balance among compression ratio, summarization quality and machine readability. It first uses LDA topic model to get topic words and then get sentences related to these topic words. Moreover, it utilizes topic diversity to pick out several high score sentences and compose a short summary. This method is similar to ours to some extent, but the selection of topic words and the ordering of sentences are completely different.

In order to compare the experimental performance of these methods, we use the N-gram co-occurrence statistics called ROUGE-N to evaluate the generated summaries at the word level [30]. Without loss of generality, ROUGE-N is defined as follows:

$$ROUGE - N = \frac{\sum_{\text{gram}_N \in S} Count_{match}(gram_N)}{\sum_{\text{gram}_N \in S} Count(gram_N)} \tag{4.1}$$

where S is one of the standard summaries, N stands for the length of the N-gram, and Count match (gram N) is the maximum number of N-grams co-occurring in the generated summary and the standard summary. From Eq.(4.1), ROUGE-N is in fact an N-gram recall measure. In the literature [31], there are a series of ROUGE-N measures according to different values of N. ROUGE-1 refers to the overlap of 1-gram (each word) between generated summary and the benchmark summary. ROUGE-2 refers to the overlap of bigrams between generated summary and the benchmark summary. Since the Chinese text can be divided into two words in most cases, this paper uses ROUGE-1 and ROUGE-2 to evaluate the excellentness of the summary results in our experiments. In addition, ROUGE-SU4 is used to evaluate the generated summary results, where S represents Skip-bigram, U represents unigram, and 4 represents the maximum interval allowed for two words. This indicator not only calculates the number of Skip-bigram co-occurrences between the generated summary and the standard summary, but also calculates the unigram between the generated summary and the standard summary.

Because ROUGE evaluation only involves the surface information of the text, manual evaluation is used as a supplement to make the evaluation of the quality of the generated summary more accurate and complete. In the experiment, five students were invited to score the generated summaries by our proposed method and the comparison methods. And the summaries are scored from three aspects: The grammatical quality of the summary (that is, the readability of the summary), information richness and consistency. The score is 1–5. The higher the score is, the better the quality of the summary is.

Table 2. The comparison results of topic sentence extraction.

Method	Result
Key sentence	【洪灾中，大学生村官的坚守】6 月下旬以来，南方多省市遭遇连续强降雨，其中湖南、广西、江西三省灾情最为严重。据初步统计，截至 7 月 4 日，三省受灾群众达 1726.92 万人，因洪涝灾害直接导致死亡、失踪 78 人，农作物受灾面积 1302.29 千公顷，直接经济损失达 425.63 亿元。无情的洪灾面前，大学生村官们为了群 ?...展开全文 c,6
Topic sentence	无情的洪灾面前，截至 7 月 4 日，6 月下旬以来，南方多省市遭遇连续强降雨，其中湖南、广西、江西三省灾情最为严重。因洪涝灾害直接导致死亡、失踪 78 人。

The benchmark summary results used for comparison in all experiments are manually generated topic sentence extraction results.

4.1. Topic sentence extraction results

The results of our topic sentence extraction method are shown in the Table 4 using the natural disaster data set and the Table 2 is a description of topic sentence extraction process using the topic of flood. The English translations corresponding to Tables 2 and 4 are shown in Tables 3 and 5, respectively.

Table 3. Corresponding English translation of Table 2.

Method	Result
Key sentence	【Stubbornness of College Student Village Officials in Flood Disasters】Since late June, many provinces and cities in the south have experienced continuous heavy rainfall, of which Hunan, Guangxi, and Jiangxi provinces suffering the most. According to preliminary statistics, as of July 4, the number of disaster-affected people in the three provinces reached 17.2692 million, 78 people were directly killed or missing due to floods, the crop area affected was 130.229 hectares, and the direct economic loss reached 42.563 billion yuan. In the face of the ruthless floods, the college student village officials ?... Expand the full text c,6
Topic sentence	In the face of the ruthless floods, as of July 4, since late June, many provinces and cities in the south have experienced continuous heavy rainfall, of which Hunan, Guangxi, and Jiangxi provinces suffered the most. 78 people were directly killed or missing due to floods.

First, we pay attention to the key sentence of Table 2, which is only based on the key sentence selection algorithm. The result is more verbose and has more noisy data. The format of the result is not uniform enough. It contains many meaningless symbols and irrelevant content which reflects the topic

Table 4. The results of topic sentence extraction by TASA.

Topic	result
1	无情的洪灾面前，截至7月4日，6月下旬以来，南方多省市遭遇连续强降雨，其中湖南、广西、江西三省灾情最为严重。因洪涝灾害直接导致死亡、失踪78人。
2	中国地震台网正式测定：08月08日21时19分在四川阿坝州九寨沟县（北纬33.20度，东经103.82度）发生7.0级地震，震源深度20千米。
3	预计24日晚到25日白天州内大部地区阴有中到大雨，局部地区有暴雨，州内其余地区多云有阵雨，局地有中到大雨。25日晚到26日白天州内大部阵雨转多云，局地有中到大雨。请注意防范短时强降水可能引发的山洪、泥石流。
4	昨夜的威海暴雨给新的一天送来了清凉，但暴雨导致多个路段出现了不同程度的积水，另我市暴雨仍将持续。
5	台风天鸽在珠海市金湾区沿海地区登陆,今年第13号台风“天鸽”（强台风）已于23日12时50分在珠海市金湾区沿海地区登陆，登陆时中心最大风力14级（45米/秒）

representation is not clear enough. Although the summary results of the floods can also be obtained, the emphasis is not clear and the sentences are redundancy. After the topic sentence extraction process, the summary results (topic sentence) are concise and clear at a glance. The important information becomes concentrated, irrelevant content is excluded, and the sentence is more neat. We can quickly grasp the main information of this sentence with time, place, and things which are concentrated into two short sentences. Therefore, this process plays an important role in the summary generation and is an indispensable step.

Next, the performance of the summary results is further explored from an intuitive perspective through Table 4. It is not difficult to find the meaning of each topic. For example, topic 1 describes the natural disaster floods which happened in the south of China like Hunan, Guangxi and Jiangxi. We can also know the time of this disaster which is in the late June. Whats more, the cause of the disaster is continuous rain for many days and the flood has caused tremendous damage to people like death and missing. Topic 2 represents the earthquake which happened in Sichuan on 08^{th} August. We can also know the focal depth, earthquake intensity and other useful information from this topic sentence extraction result. Moreover, we can samely summarize the other three topics which is related to heavy rain and typhoon. Above all, we can see that our topic sentence extraction result is quite excellent. It can summarize thousands of texts into several short sentences which can explain each corresponding topic clearly. Therefore, the government can quickly make judgments on these disasters and minimize disasters losses.

In addition, on the social hot event data set, taking the event of “girl attacked at Yitel” as an example, the summary result generated by the method proposed in this paper are shown in Table 6. From Table 6, we can know the more comprehensive information of the event, which can save people time and effort in obtaining event information. The summary of the “girl attacked at Yitel” generated by our proposed method is clear and well readable. Therefore, it can meet the user’s demand for information acquisition of events.

4.2. Performance comparison

In order to further verify the performance of the proposed algorithm, we compared our method with five methods using ROUGE-1, ROUGE-2 and ROUGE-SU4 on natural disaster data set. The higher

Table 5. Corresponding English translation of Table 4.

Topic	Result
1	In the face of the ruthless floods, as of July 4, since late June, many provinces and cities in the south have experienced continuous heavy rainfall, of which Hunan, Guangxi, and Jiangxi provinces suffered the most. 78 people were directly killed or missing due to floods.
2	China Seismograph Network officially determined that a magnitude 7.0 earthquake occurred in Jiuzhaigou County, Aba Prefecture, Sichuan (at 33.20 degrees north latitude and 103.82 degrees east longitude) at 21:19 on August 08, with a focal depth of 20 kilometers.
3	It is estimated that 24 to 25 during the day and night, there will be cloudy with moderate to heavy rain in most parts, local heavy rainfall, cloudy with showers in the rest of the state and moderate to heavy rain in the local area. 25 to 26 during the day and night, most of the showers in the state turned cloudy, local with moderate to heavy rain. Please pay attention to prevent torrents and debris flows that may be caused by short-term heavy precipitation.
4	The heavy rainfall in Weihai last night brought coolness to the new day, but it led to the accumulation of water in varying degrees on multiple road sections, and the heavy rainfall in our city will continue.
5	Typhoon hato made landfall in the coastal area of Jinwan District, Zhuhai City. This year's No.13 typhoon hato (strong typhoon) has landed in the coastal area of Jinwan District, Zhuhai City at 12:50 on 23. The maximum wind force of the center when landing was 14(45m / s).

Table 6. The summary result of the "girl attacked at Yitel" by TASA.

Result
4月5日，网友弯弯_2016入住北京和颐酒店遇袭逃脱。北京警方回应彻查事件。据平安北京，4月7日涉案男子已抓获。和颐酒店经理称遇袭女子炒作。
The corresponding English translation of the above summary is: On April 5, a netizen named Wanwan_2016 was attacked and escaped while staying in Beijing Yitel. The Beijing police responded to the investigation. According to Pingan Beijing, the man involved was arrested on April 7. The hotel manager described the attacked woman as a publicity stunt.

the values of these three indicators are, the better the performance of the algorithm is. Table 7 shows the values of the three indicators of the automatic summary results of different methods on the natural disaster data set.

It can be seen from the values of ROUGE-1 in Table 7 that TASA is superior in the summary generation results of the five topics and its performance is relatively stable in terms of the matching degree of 1-gram. It is worth noting that TASA achieves a higher degree of matching in the summary generation results of Topic 3 and Topic 5. The highest value even reached 0.867 which shows the high accuracy of the topic summary. The methods Topic Words and Topic based Summary are also relatively stable and their performance is only next to TASA. They achieve the best results in topics 1 and 5. However, the value of ROUGE-1 was reduced by about 12% compared to the TASA. The experimental performance of Extractor and TF-IDF are similar on average, ranked fourth and fifth. Finally, the ROUGE-1 value of the Auto Summarize is generally low which gets the lowest value in the summary results of the five topics. This indicates that the Auto Summarize method has fewer effective words in the summary. Through the above analysis, we can find that the topic-related algorithms like TASA, Topic Words and Topic based Summary rank in the top three under the criterion of ROUGE-1, which shows that using topic information will improve the performance of automatic summary. Therefore, combining the automatic summary algorithm with the topic model can further improve the accuracy and topic relevance of the summary result. Although the TF-IDF method can also achieve the purpose of automatic summarization to a certain extent, the performance is limited. Therefore, TASA is better than other five methods and achieves the best performance in terms of topic correlation and semantic integrity.

The ROUGE-2 indicator allows us to learn more about the semantic information of topic summaries, rather than simply checking the number of single words contained in the results. It reflects the accuracy of the two-word results, and the semantics of the words are generally fixed, which further determines the semantic relevance of the summary content. It can be seen from Table 7 that the value of ROUGE-2 is slightly lower than ROUGE-1 in each topic. Because the single word constitutes the two-word form, which adds the word order and semantic choice, the value of ROUGE-2 will inevitably decrease. In Table 7, the ROUGE-2 of TASA is higher in each topic which indicates that the word order and semantics in the summary results are more reasonable and accurate. The topic representation has rich semantics and more complete content. At the same time, the result is more stable which indicates the wider applicability. Topic Words, Topic based Summary and Extractor are close behind. The ROUGE-2 of Topic based Summary method is almost equal to TASA in topic 2, but there is a small decrease in topic 4. The possible reason for this phenomenon is that the topic model LDA has defects in topic word extraction and cannot be well applied to detect each topic, so its experimental performance has great differences. Based on this, we can conclude again that the topic related summarization method has certain advantages. The methods of TF-IDF and Auto Summarize also have large fluctuations, and their values of ROUGE-2 are lower which indicates that the topic summary is not accurate enough. The possible reason is that the summary result is not polished, which results in more irrelevant information in the summary. Therefore, the summary results of TASA on each topic are better and have beter stability compared with other five methods.

In addition, by observing the value of ROUGE-SU4 in Table 7, we can find that the performance of each algorithm under this indicator is basically consistent with the trend under the first two indicators. And the value of ROUGE-SU4 for each method is slightly higher than the value of ROUGE-2 and

Table 7. Different ROUGE-1, ROUGE-2 and ROUGE-SU4 of TASA, TF-IDF, Topic words, Extractor, AutoSummarize and Topic based Summary on the natural disaster dataset.

	Method	Topic1	Topic2	Topic3	Topic4	Topic5
ROUGE-1	TASA	0.82	0.806	0.826	0.813	0.867
	TF-IDF	0.72	0.742	0.707	0.674	0.738
	Topic Words	0.77	0.742	0.76	0.744	0.803
	Extractor	0.68	0.71	0.747	0.721	0.77
	Auto Summarize	0.64	0.645	0.707	0.698	0.754
	Topic based Summary	0.78	0.774	0.747	0.721	0.82
ROUGE-2	TASA	0.714	0.7	0.744	0.714	0.75
	TF-IDF	0.53	0.6	0.5	0.452	0.517
	Topic Words	0.632	0.667	0.662	0.667	0.717
	Extractor	0.551	0.633	0.649	0.643	0.7
	Auto Summarize	0.51	0.633	0.5	0.5	0.7
	Topic based Summary	0.591	0.7	0.676	0.571	0.667
ROUGE-SU4	TASA	0.746	0.723	0.765	0.739	0.79
	TF-IDF	0.56	0.628	0.531	0.48	0.56
	Topic Words	0.665	0.697	0.695	0.698	0.758
	Extractor	0.579	0.665	0.67	0.647	0.718
	Auto Summarize	0.546	0.638	0.534	0.529	0.712
	Topic based Summary	0.62	0.73	0.68	0.613	0.72

lower than the value of ROUGE-1 in each topic. Our proposed method is higher than the ROUGE-SU4 of other comparison methods, which further proves the advantages of the topic-based automatic summarization algorithm in this paper.

Table 8 shows the average scores of the manual evaluation of the summaries generated by each method for topic 1 on natural disaster dataset in terms of the grammar, informativeness and coherence. We can see from the data shown in Table 8 that the method proposed in this paper has achieved the highest score in three aspects. Compared with other methods, the summary generated by the method in this paper has good readability, rich related information and good consistency because it is polished.

It can be seen from the above that by comprehensively comparing the six methods, three methods based on the topic, namely TASA, Topic words and Topic-based Summary, have better performance. Therefore, we only compare these three methods on the social hot event dataset and use ROUGE-1 to verify the performance of the proposed model. Table 9 shows the ROUGE-1 values of the automatic summary results of different methods on the social hot event dataset. It can be seen from Table 9 that compared with the other two methods, the summary result generated by our proposed method on each topic is the best.

In summary, we can get the following information. First, whether it is based on feature words, topic models or other automatic summarization algorithms, each method can obtain the topic-related summary results to a certain extent. Second, the summary generation algorithm based on the topic model can improve the performance of the automatic summary algorithm. Third, the TASA proposed

Table 8. The average scores of the manual evaluation of the summaries generated by each method for topic 1 on natural disaster dataset.

Method	Grammar	Informativeness	Coherence
TASA	4.2	3.85	4.2
TF-IDF	3.5	3	3
Topic Words	3.8	3.4	3.6
Extractor	3.5	3.1	3.1
Auto Summarize	3	2.5	2.5
Topic based Summary	3.9	3.45	3.6

Table 9. The ROUGE-1 values of TASA, Topic words and Topic based Summary on the social hot event dataset.

	Method	Topic1	Topic2	Topic3
ROUGE-1	TASA	0.68	0.687	0.715
	Topic Words	0.59	0.612	0.65
	Topic based Summary	0.62	0.605	0.663

in this paper has achieved excellent and stable results on ROUGE-1, ROUGE-2 and ROUGE-SU4 for each topic.

5. Conclusions

This paper proposes a topic-based automatic summarization algorithm for Chinese short text(TASA). The goal is to represent the content of each topic in the form of a sentence. Firstly, combining the topic words and TF-IDF, a key sentence selection mothod is proposed to obtain a ranked set of key sentences related to topics in Weibo and select the sentence with the highest score as the topic sentence. Considering the format irregularity and sparse content of Weibo texts, three retouching mechanisms are added to further improve the simplicity, richness and readability of topic sentence extraction results. Experimental results of polished topic summaries show that the TASA proposed in this paper can not only automatically identify each topic, but also obtain a concise summary of each topic in a sentence more accurately and contain rich information. Therefore, when natural disasters occur, the government can accurately detect the occurrence of natural disasters by collecting relevant text information from Weibo and guide the public to the related activities of disaster relief. We compare TASA with five advanced methods for automatic summarization. The experimental results on the real datasets show that three methods of fusing topic information have achieved good experimental results. However, the performance of TASA is superior to other baseline methods in ROUGE-1, ROUGE-2, ROUGE-SU4 and manual evaluation, which greatly proves the superiority of proposed method. TASA is also superior to the methods of using TF-IDF or topic words alone. Therefore, combining IF-IDF and topic words is the correct choice. In addition,

compared with other methods of adding topic information, TASA is more accurate, and is excellent in readability, richness and consistency.

In recent years, information grows and spreads faster and faster. The natural disaster usually happens frequently and suddenly. How to seize the time node to discover such topics as early as possible or achieve predictions will be the focus of our future research. In addition, although the topic-based automatic summarization algorithm proposed in this paper integrates topic information in Weibo text collection to a certain extent, it is limited. Therefore, we consider how to combine topic-related text, use machine learning to analyze the grammatical structure, and automatically generate a summary for a short content text set in future work, so that the topic information can be extracted at the same time as automatic summaries are generated. Finally, since it is very time-consuming to generate reference summaries manually, we will consider evaluating the generated summaries without reference summaries in subsequent research work.

Acknowledgments

This work was supported in part by the National Science Foundation of China (No.U1736105). The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through Research Group no. RG- 1441-331.

Conflict of Interests

The author declares no conflict of interest.

References

1. S. L. Lo, R. Chiong, D. Cornforth, An unsupervised multilingual approach for online social media topic identification, *Expert Syst. Appl.*, **81** (2017), 282–298.
2. J. F. Yeh, Y. S. Tan, C. H. Lee, Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation, *Neurocomputing*, **216** (2016), 310–318.
3. J. Christensen, Mausam, S. Soderland, O. Etzioni, *Towards coherent multi-document summarization*, Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2013, 1163–1173. Available from: https://www.aclweb.org/anthology/N13-1136/.
4. E. Lloret, M. Palomar, Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre, *Expert Syst. Appl.*, **40** (2013), 6624–6630.
5. G. Yang, D. Wen, Kinshuk, N. S. Chen, E. Sutinen, A novel contextual topic model for multi-document summarization, *Expert Syst. Appl.*, **42** (2015), 1340–1352.
6. I. Mani, M. T. Maybury, *Advances in Automatic Text Summarization*, (MITRE Corporation) Cambridge, The MIT Press, (1999).
7. J. M. Torres-Moreno, *Automatic Text Summarization*, John Wiley and Sons, 2014.
8. A. Nenkova, K. McKeown, A survey of text summarization techniques, *Min. Text Data*, **2012** (2012), 43–76.

9. T. Ma, Y. Zhao, H. Zhou, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan, Natural disaster topic extraction in sina microblogging based on graph analysis, *Expert Syst. Appl.*, **115** (2019), 346–355.
10. T. Ma, Q. Liu, J. Cao, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan, LGIEM: Global and local node influence based community detection, *Future Gener. Comput. Syst.*, **105** (2020), 533–546.
11. T. Ma, H. Rong, Y. Hao, J. Cao, Y. Tian, M. A. Al-Rodhaan, A Novel Sentiment Polarity Detection Framework for Chinese, *IEEE Trans. Affective Comput.*, 2019.
12. A. Kazantseva, S. Szpakowicz, Summarizing short stories, *Comput. Linguist.*, **36** (2010), 71–109.
13. M. T. Khan, M. Durrani, S. Khalid, F. Aziz, Online knowledge-based model for big data topic extraction, *Comput. Intell. Neurosci.*, **2016** (2016), 1–10.
14. Indra, E. Winarko, R. Pulungan, Trending topics detection of Indonesian tweets using BN-grams and Doc-p, *J. King Saud Univ. Comput. Inf. Sci.*, **31** (2019), 266–274.
15. W. M. Wang, Z. Li, J. W. Wang, Z. H. Zheng, How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds, *Expert Syst. Appl.*, **90** (2017), 439–463.
16. M. Moradi, N. Ghadiri, Different approaches for identifying important concepts in probabilistic biomedical text summarization, *Artif. Intell. Med.*, **84** (2018), 101–116.
17. R. Yan, L. Kong, C. Huang, X. Wan, X. Li, Y. Zhang, *Timeline generation through evolutionary trans-temporal summarization*, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, 433–443. Available from: https://www.aclweb.org/anthology/D11-1040/.
18. W. Liu, X. Luo, J. Zhang, R. Xue, R. Xu, Semantic summary automatic generation in news event, *Concurrency Comput. Pract. Exp.*, **29** (2017), e4287.
19. D. Zhou, D. Zhong, A semi-supervised learning framework for biomedical event extraction based on hidden topics, *Artif. Intell. Med.*, **64** (2015), 51–58.
20. W. Xiong, D. Litman, *Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews*, In Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers, 2014, 1985–1995. Available from: https://www.aclweb.org/anthology/C14-1187/.
21. Z. Wu, L. Lei, G. Li, H. Huang, C. Zheng, E. Chen, et al., A topic modeling based approach to novel document automatic summarization, *Expert Syst. Appl.*, **84** (2017), 12–23.
22. A. Barrera, R. Verma, *Combining syntax and semantics for automatic extractive single-document summarization*, In International Conference on Intelligent Text Processing and Computational Linguistics, 2012, 366–377. Available from: https://link.springer.com/chapter/10.1007/978-3-642-28601-8_31.
23. F. Barrios, F. López, L. Argerich, R. Wachenchauzer, Variations of the similarity function of textrank for automated summarization, preprint, arXiv1602.03606, 2016.
24. C. Fang, D. Mu, Z. Deng, Z. Wu, Word-sentence co-ranking for automatic extractive text summarization, *Expert Syst. Appl.*, **72** (2017), 189–195.

25. M. Schinas, S. Papadopoulos, Y. Kompatsiaris, P. A. Mitkas, Mgraph: Multimodal event summarization in social media using topic models and graph-based ranking, *Int. J. Multimedia Inf. Retr.*, **5** (2016), 51–69.

26. F. Ye, X. Xu, Automatic multi-document summarization based on keyword density and sentence-word graphs, *J. Shanghai Jiaotong Univ. Sci.*, **23** (2018), 584–592.

27. W. Xie, F. Zhu, J. Jiang, E. P. Lim, K. Wang, Topicsketch: Real-time bursty topic detection from twitter, *IEEE Trans. Knowl. Data Eng.*, **28** (2016), 2216–2229.

28. X. Yang, P. Jin, X. Chen, *The construction of a kind of chat corpus in chinese word segmentation*, In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, 168–172. Available from: https://ieeexplore.ieee.org/document/7397448.

29. D. Yan, E. Hua, B. Hu, *An improved single-pass algorithm for chinese microblog topic detection and tracking*, In 2016 IEEE International Congress on Big Data (BigData Congress), 2016, 251–258. Available from: https://ieeexplore.ieee.org/abstract/document/7584945.

30. C. C. Birant, O. Aktas, Rule-based turkish text summarizer (RB-TTS), *Adv. Electr. Comput. Eng.*, **18** (2018), 113–119.

31. A. Abdi, N. Idris, R. M. Alguliev, R. M. Aliguliyev, Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems, *Inf. Process. Manage.*, **51** (2015), 340–358.

32. H. Rong, T. Ma, J. Cao, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan, Deep Rolling: A Novel Emotion Prediction Model for a Multi-Participant Communication Context, *Inf. Sci.*, **488** (2019), 158–180.

AIMS Press