A New Text Summarization Approach based on Relative Entropy and Document Decomposition

Nawaf Alharbe[1], Mohamed Ali Rakrouki[2], Abeer Aljohani[3]
Applied College
Taibah University
Saudi Arabia

Mashael Khayyat[4]
College of Computer Science and Engineering
University of Jeddah
Saudi Arabia

***Abstract*—In the era of the fourth industrial revolution, the rapid relay on using the Internet made online resources explosively grow. This revolution emphasized the demand for new approaches to utilize the use of online resources such as texts. Thus, the difficulty to compare unstructured resources (text) is urging the demand of proposing a new approach, which is the core of this paper. In fact, text summarization technology is a vital part of text processing, therefore. The focus is on the semantic information not just on the basic information. It requires mining topic features in order to obtain topic-words and topic-sentences relationships. This automatic text summarization is document decomposition according to relative entropy analysis; which means measuring the difference of the probability distribution to measure the correlation between sentences. This paper introduced a new method for document decomposition, which categorizes the sentences into three types of content. The performance demonstrated the efficiency of using the relative entropy of the topic probability distribution over sentences, which enriched the horizon of text processing and summarization research field.**

***Keywords*—Natural language processing; text summarization; extractive methods; relative entropy**

I. Introduction

At present, the rapid popularization of the Internet makes resources explosively growth. On the one hand, rich information resources bring great convenience. On the other hand, it also makes people difficult to select suitable resources. From the view of network information resources, the proportion of unstructured resources has been growing rapidly, and the processing of this type of data is more difficult compared to structured data. A text is a typical unstructured data, its effective analysis and processing has practical significance for Internet users.

Text summarization technology is a very important part of text processing. From a technical point of view, the techniques based on slight semantic features are different from those based on word features. The focus of this paper is not the basic information that can be observed in the composition of a document, such as words or sentences, but the deeper semantic information behind. By mining topic features, we can get topic-word features and topic-sentence features. Based on the relationship between sentences, it is possible to measure the ability to express the topic of a document, and then choose a sentence as a text summarization.

In this context, this paper proposes a new method of automatic text summarization based on relative entropy and document decomposition. The relative entropy is used to measure the difference of the probability distribution in order to measure the correlation between sentences. Also, a new document decomposition method is introduced for categorizing sentences as three types of content.

The remainder of this paper is organized as follows. In Section II, some important related works to text summarization are presented. A brief review of the LDA model is presented in Section III. Section IV is devoted to the presentation of probability distribution of topics over sentences. An improved sentence similarity calculation method is proposed in Section V. In Section VI, a candidate abstract sentence selection method based a greedy algorithm is proposed. The experimental results are presented in Section VII. Finally, Section VIII summarizes this research work.

II. Literature Review

In the 1950s, the rise of statistics prompted the germination of text summarization techniques, and statistical methods were limited to the surface features of documents. For example, according to the position of the sentence in the paragraph, the position of the paragraph in the article, the word frequency and the inverse text word frequency, the similarity between the sentence and the title and other characteristics to evaluate the importance of the sentence. Lunh [1] believed that words with a large number of occurrences are relatively closely related to the topic of the document, so the weight of words can be calculated according to the number of times they appear in the document, and sentence weights can be obtained based on the weight of words. Select sentences with higher weights as the abstract of the document. This idea has also become a cornerstone of the subsequent development of text summarization technology. Although the principle seems simple, the implementation results have a high accuracy rate, even surpassing many later more complex algorithms. Later, Baxendale [2] proposed that some summary words in the document also represent the topic of the document and should be given a higher weight. Edmundson [3] measures the importance of sentences according to three factors: clue words, keywords, and location, and selects sentences with greater weight as abstracts. In statistics, text is a linear sequence of sentences, and a sentence is a linear sequence of words. When analyzing text, it can finally be attributed to the analysis of words, and the weight of sentences can be obtained by analyzing the characteristics of words. In recent years, the academic community has further proposed methods based on integer linear programming [4]–[6] and methods of maximizing sub-

modular functions, which can consider sentence redundancy in the process of sentence selection [7], [8].

In the 1990s, with the rise of the Internet, the number of documents increased exponentially. At the same time, the rise of machine learning has made great progress in natural language processing, which has given new inspiration to text summarization technology. On the basis of statistics, Kupiec et al. [9] proposed a Naive Bayes classification model to select document summary sentences. With the development of machine learning, more advanced algorithms have been applied to text summarization techniques, such as decision tree model, hidden Markov model, conditional random field model, neural network, etc. Conroy and O'leary [10] calculated the correlation between words based on the hidden Markov model and on mutual dependencies, Goularte et al. [11] used linear regression model modeling, Svore et al. [12] proposed a neural network-based abstract method. Machine learning methods mainly focus on how to convert text summarization problems into machine learning problems. Although the text summarization obtained by the machine learning method has achieved good results, the lack of corpus in this aspect greatly restricts the training effect. In some recent work, the summarization is represented as a word or sentence level classification problem based on neural network architectures, and it is addressed by computing sentence representations [13]–[17]. Zhong et al. [18] reranked extractive summaries using document-level features.

A recent comprehensive and consistent review of text summarization for papers published between 2008 and 2019 can be found in Widyassari et al. [19]. Some in-depth investigation and analysis of automatic text summarization techniques have been provided by [20]–[22].

III. Text Summarization based on LDA Model

Since the LDA (Latent Dirichlet Analysis) [23] model was proposed, it has been widely used in the literature. It can be seen that the effect of the LDA model in the field of text topic extraction has been extremely recognized, and it has become a popular technology in the direction of text mining.

LDA is a hierarchical Bayesian model, in order to represent topics in each document in the form of a probability distribution. It is a "bag of words" model that treats the document as a set of words. There is no order of words, each one in the document is selected according to a certain probability from the thesaurus of the input document topic.

From a topology perspective, the LDA model assumes that the text consists of several randomly selected topics, and each topic is expressed by several randomly selected words in the corresponding thesaurus. This is an assumption that obeys objective reality. Based on this document composition method, the topic can be regarded as the probability distribution on the vocabulary (topic-word), and the document can be regarded as the probability distribution on the topic (doc-topic). This assumption can also be applied to large-scale data processing, that is, mapping documents to the subject space, so as to achieve the effect of dimensionality reduction.

A. Model Solving

The solution for the LDA model is a very complex optimization solution process, and it is very difficult to solve it optimally. For solving this model approximately, heuristic methods are used. There are roughly three types: One is an approach based on expectation advancement, one is based on variational EM solving, and one is based on Gibbs sampling [24]. Generally speaking, Gibbs sampling method is simpler than the other two types and works well. Therefore, for most computing tasks, this method is used to solve the LDA model.

B. Determination of the Number of Topics

A parameter that needs to be specified manually is the number of topics in the training corpus. The determination of the number of topics is a process of selecting models corresponding to different numbers of topics, which is a difficult problem to solve. There are generally two ways to determine the number of topics:

1) **Experience setting:** In the process of text mining, the corpus usually used as training needs to be relatively comprehensive, and the corpus can be basically determined in several aspects. For example, it is known in advance that these topics are about: culture, news, sports, politics, entertainment, then the number of topics can be clearly set to 5. However, for most of the training corpus, it is not known in advance which topics it contains, which requires repeated debugging or the use of enumeration methods. Since there is no model that can evaluate the results well, the debugging process needs to observe the correspondence between words and topics in the results to judge in the way of human understanding, and then determine a reasonable number of topics.
2) **Perplexity-based determination method:** This metric represents the uncertainty in predicting data. If a topic model obtains a low perplexity degree on the test corpus, then the model is considered to be very expressive, and the number of topics determined by the model is considered reasonable.

The characteristic of the LDA model is that the more accurate the model is, the narrower the scope of use is. Therefore, for different corpora, the size of the number of topics cannot be set completely by experience. Instead, the number of topics is determined by two methods: experience setting and perplexity calculation. The number of topics needs to be continuously set, and the number of topics with the lowest perplexity is taken as the training parameter.

IV. Probability Distribution of Topics Over Sentences

From the analysis in Section III, it can be seen that the LDA model represents the document in the form of a topic by representing the document as a certain probability distribution of the topic. Similarly, the topic is also represented as a certain probability distribution of words, thus forming a hierarchical structure: document-sentence-topic-word. Since we know the probability distribution of topic-words, we can use this hierarchical model to calculate the probability distribution of sentences-topics.

In Arora and Ravindran [25], three methods (generative, semi-derivative, and derivation) are proposed to estimate the probability distribution of sentences given a topic based on a hierarchical Bayesian model, and there is a strong assumption about the calculation: All sentences of a document express a topic, and each word in each sentence corresponds to only one topic. The performance of these methods has been verified, and in this paper, the derivation method with relatively good performance is selected for improvement.

Let assume that S_i ($i \leq length(D)$) are the sentences in a document D, W_j ($j \leq n$, where n is the number of words) is a word in the document, and T_k ($k \leq K$, where K is the number of all topics) are the topics contained in the document. We calculate the probability $P(T|S)$ of topic T given a sentence S, thereby calculating the probability that the topic T_k belongs to the sentence S_i.

To find the topic probability distribution over a sentence, it can be given by the Bayesian formula:

$$P(T|S) = \frac{P(S|T) \cdot P(T)}{P(S)} \tag{1}$$

From the output of the LDA model, the topic probability $P(T)$ of the document can be obtained, so that:

$$P(T) = \sum_{k=1}^{K} P(T_k|D) \tag{2}$$

For the probability $P(S)$ of a sentence, it can be calculated according to the words contained in the sentence. Similarly, it can be calculated by the known $P(W_i|S)$ as follows:

$$P(S) = \sum_{j=1}^{n} P(W_j|S) \tag{3}$$

where n is le length of the sentence S.

In order to calculate $P(S|T)$, we assume that each sentence in the document contains only one topic, and each word expresses only one topic. Then, in the case of known topics, we calculate the probability that the sentence belongs to each topic.

From Arora and Ravindran [25], three methods are pointed out for computing $P(S|T)$ for multiple documents. Since the text summarization of a single document is similar to the text summarization of multiple documents, we propose an explicit improvement using partially generated derivation as follows:

$$P(S_i|T_k) = P(D|T_k) \cdot \sum P(T_k|W_j) \tag{4}$$

where $P(S_i|T_k)$ represents the probability that sentence S_i expresses topic T_k; $P(D|T_k)$ represents the probability that document D generates topic T_k; $P(T_k|W_j)$ represents the probability that topic T_k generates word W_j.

From the Bayesian generation formula, the above formula can be rewritten as:

$$\begin{aligned} P(S_i|T_j) =& \frac{P(T_k|D)P(D)}{P(T_k)} \cdot \frac{P(\prod W_j|T_k)}{P(\prod_{W_j \in S_i} W_k)} \\ =& \prod_{W_j \in S_i} P(W_j|T_k) \cdot \frac{P(T_k|D)}{\prod_{W_j \in S_i} W_j} \end{aligned} \tag{5}$$

In order to calculate $P(W_j)$, it can be calculated according to the output data of LDA:

$$P(W_j) = \sum_{k=1}^{K} P(W_j|T_k) \cdot P(T_k) \tag{6}$$

Combining the above Eq. 1 - 6, we can get the representation of $P(T|S)$ as follows.

$$P(T_k|S_i) = \frac{P(T_k|D)^2 \prod_{W_j \in S_i} P(W_j|T_k)}{\sum_{j=1}^{n} P(W_j|S_i) \cdot \prod_{W_j \in S_i} \sum_{k=1}^{K} P(W_j|T_k) P(T_k|D)} \tag{7}$$

V. Improvement of Sentence Similarity Calculation Method

The goal of extraction-based text summarization is to use a good method to calculate the weight of each sentence as a basis for measuring their importance. Among several methods used for automatic text summarization, machine learning methods use sentences and words of documents as learning features. Statistical methods measure sentence weights according to word frequency, position of sentences in paragraphs, and similarity of sentences and topics on words. In graph model, the same words are used as the basis for establishing edges between nodes (sentences). The vector space model uses the words as the vector dimension, and each document forms a matrix to calculate keywords through singular value decomposition. All these methods build models based on the features of sentences or words in sentences.

A big drawback of modeling based on word features is that it cannot solve semantic associations well. Different words may express the same topic. In this case, how to determine the semantic association between words is a key point that needs to be improved. The characteristics of topic model can just make up for the shortcomings of the semantic relationship that cannot be mined in word-based modeling. This paper combines the characteristics of the topic model to calculate the similarity of two sentences in the semantic dimension.

Combined with the computational model presented in Section IV, the probability distribution of topic over sentence is transformed into sentence-to-sentence, sentence-to-document similarity through relative entropy, so the calculation method of sentence weight is obtained.

A. Relative Entropy Definition

Relative Entropy (also known as Kullback–Leibler Divergence, KLD for short) is a measure of the difference between two probability distributions P and Q. It can be expressed

in the form of $D_{KL}(P||Q)$, where Q is the probability distribution of theoretical data, as a measurement standard, and P is the probability distribution of real data, as the object of estimation. $D_{KL}(P||Q)$ represents the loss, or difference, when fitting the probability distribution P of real data with the probability distribution Q of theoretical data. The biggest feature of relative entropy is asymmetry, that is to say $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and relative entropy does not satisfy the triangle inequality relation.

In Shannon's information theory, if the probability distribution of the character set is given, then a way to encode the character set with the least number of bits can be designed according to this probability distribution. Assuming that the character set is X, and the probability of one character x is $P(x)$, then the average number of optimally encoded bits of character x is equal to the entropy set of the character set:

$$H(x) = \sum_{x \in X} P(x) \log \left(\frac{1}{P(x)} \right) \tag{8}$$

On the same character set, there is also another probability distribution $Q(x)$, if the optimal encoding based on $P(x)$ is used to encode characters conforming to $Q(x)$. Due to the difference in probability distribution, the number of bits required for encoding will be higher. In this case, the concept of relative entropy is proposed, which measures the average number of bits used to encode each character. According to this relationship, the divergence between two probability distributions P and Q is measured as follows.

$$\begin{aligned} D_{KL}(P||Q) &= \sum_{x \in X} P(x) \log \left(\frac{1}{Q(x)} \right) - \sum_{x \in X} P(x) \log \left(\frac{1}{P(x)} \right) \\ &= \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \end{aligned} \tag{9}$$

Relative entropy is greater than zero and has a value of 0 if and only if the two probability distributions are the same. From Eq. (9), it can be concluded that when the probability distributions P and Q are discrete random variables, the calculation method of relative entropy is as follows.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \ln \left(\frac{P(x)}{Q(x)} \right) \tag{10}$$

B. Application of Relative Entropy to Distance Metrics

Since relative entropy is based on a probability distribution, it is tested against another probability distribution, and the difference between test distribution and reference distribution is not an absolute method to measure the distance, because it does not have symmetry and transitivity.

It is generally believed that if the topic distribution on a sentence is closer to the topic distribution on the document than the topic distribution on other sentences is closer to the topic distribution on the document, then it can be considered that the ideas expressed in this sentence can more comprehensively include the theme of the entire document. For a sentence, what you want to get is the similarity of the sentence relative to the article on the topic, without calculating the similarity of the document relative to a sentence on the topic. Therefore, the asymmetric nature of relative entropy will not have any effect on the content that needs attention.

Based on the above ideas, the topic distribution on the document is regarded as a theoretical probability distribution, and the probability distribution on the sentence is regarded as the actual probability distribution. Then, with the help of relative entropy, a method to measure sentence similarity and sentence weight is obtained.

The similarity between the topic probability distribution on the sentence and the document can be expressed as:

$$D_{KL}(P(T|S_i)||P(T|D)) = \sum_{k=1}^{K} P(T_k|S_i) \log \left(\frac{P(T_k|S_i)}{P(T_k|D)} \right) \tag{11}$$

At the same time, the similarity of two sentences S_r and S_t can be calculated:

$$D_{KL}(P(T|S_r)||P(T|S_t)) = P(T|S_r) \log \left(\frac{P(T|S_r)}{P(T|S_t)} \right) \tag{12}$$

VI. Selection of Candidate Abstract Sentences

The strategy usually used for the selection of the center is to uniformly calculate the weight of the sentences in the document. After the weights of all sentences are obtained, several sentences with larger weights are selected as text summaries, and then subsequent semantic modification is performed.

Although this strategy has the ability to better express the main idea of the document, but through the understanding of the writing characteristics, we can know that an expository writing usually has some general statements to describe the summary of all central points described in the document. Then it will be expanded according to the specific content of each center point, and this will be clearly explained. So, a document can be seen as three levels of content:

1) The central sentence that describes all the ideas of the document, referred here as the general thesis.
2) The general situation described by each central idea of the document becomes a sub-thesis here.
3) Other auxiliary sentences describing each central idea become general descriptive sentences.

For a good abstract, it should include two levels of content: general thesis and sub-thesis, including sentences describing all the central ideas of the document, as well as sentences explaining what each central idea is, but not the content of general descriptive sentences. That is, it does not include examples or analytic sentences to demonstrate a central idea, which are redundant information for the abstract.

We believe that this paper is the first work considering the decomposition of sentences that constitute a document according to three levels, by observing their characteristics.

Three kinds of sentences make up a document: sentences that express the general thesis (first-level sentences), sentences that express the sub-thesis (second-level sentences), and general descriptive sentences (third-level sentences). The sentences describing the general thesis are concise and comprehensive, including the most important topics of the document, which appear most often in expository texts or news. It is the most needed content in the summarizing task of this document. The sub-thesis is used to illustrate or enrich the general thesis. The topics of the general thesis are scattered and included in different sub-theses, with relatively more length. It is usually argued from several different aspects and is the composition of an ideal summary. For general descriptive sentences, which are the lengthiest and contain the largest number of topics but are all irrelevant and don't appear many times throughout the document.

In order to select a sentence set that contains all the topics of the document, that is, when selecting the general thesis sentence and the sub-thesis sentence, it is only necessary to consider the number of topics contained in the sentence. The more topics are included, the more topics are included in the candidate set. Although feasible, this method ignores the importance of the narrative in the document itself.

For example, let a document containing only three central ideas. The document introduces these three central ideas in different proportions, respectively 20%, 30%, and 50%. If both sentences contain these three topics, but the topic probability distribution is different, one is 15%, 35%, 40%, and the other is 40%, 30%, 30%. Obviously the first sentence is more in line with the content of the document, and they state the same central idea in "similar" proportions. This coincides with the application of relative entropy to the distance measure mentioned above. Based on the concept of relative entropy, a summary sentence that is more representative of the document can be found.

A. Selection of Candidate Sentences for General Thesis

The first-level sentences are used in order to more accurately clarify all the points of view in the document. Thus, it is necessary to express more topics in a sentence as short as possible, i.e. playing the role of an outline. These sentences contain relatively few and important topics that will not only appear in the general thesis of the document, but also in the sub-thesis and general descriptive sentences. So, the number of occurrences in the entire document will be much more than other topics. The high frequency of such topics coincides with the relative entropy characteristics introduced in Section IV. Therefore, a strategy is proposed: For each sentence in the document, the relative entropy of the probability distribution between this sentence and the document topic is calculated. If a sentence has a high degree of coincidence with the topic of the document, it means that the content of its expression is closer to the overall document. In this way, using relative entropy, a sentence that is as similar as possible to the topic of the document is selected as the sentence that expresses the general thesis of the document. The implementation of this strategy is presented in Algorithm 1.

Algorithm 1 General thesis candidate sentences selection algorithm

Input: Document sentences $S = \{S_1, S_2, ..., S_m\}$; number of abstracts L; number of topics K
Output: General thesis candidate sentences Θ

Candidate sentences set $\psi = \emptyset$; General thesis sentences set $\Theta = \emptyset$; Topics $T = \{T_1, T_2, ..., T_K\}$; $L = K$
while $i < m$ **do**
  if $length(\psi) < L$ **then**
    $\psi = \psi \bigcup S_i$
  else if $D_{KL}(P(T|S_i)||P(D)) < \text{argmax}(\psi)$ **then**
    $\psi = [\psi - \text{argmax}(\psi)] \bigcup S_i$
  end if
end while
while $|\Theta| < K$ **do**
  $\Theta = \Theta \bigcup \max(\psi)$
end while
return Θ

The basic idea of Algorithm 1 is based on the greedy algorithm. The most similar sentences to the topic of the document are selected as candidate sentences. The process of the proposed algorithm can be stated as follows.

1) First of all, it is necessary to determine the size of the candidate sentence. Since the LDA model requires the user to input the number of topics, and the text summarization model requires the user to input the number of abstracts, the size of the candidate sentence can be jointly determined according to the required number of abstracts and the number of topics of the article. Here, it is set to be the same as the number of abstract sentences, and further selections will be made in the last step.
2) Based on the idea of a greedy algorithm, the topic similarity of each sentence to the document (relative entropy-based method) is calculated, and sentences with a high degree of similarity with the document topic are selected. If the number of current candidate sentences is less than the target number, the current sentence will be added. Otherwise, the sentence will be used as a candidate sentence only when the current sentence has a high degree of similarity with the topic of the document.
3) Again, based on the greedy algorithm, from the candidate sentences selected in the second step, the smallest set of sentences that can cover all topics is selected. We end up with the smallest set of sentences that cover the most topics.

B. Selection of Candidate Sentences for Sub-thesis

Sub-thesis candidate sentences refer to a certain length of description in a document in order to clearly describe a certain point of view, and select sentences that can summarize the sub-thesis. These sentences are sub-thesis candidate sentences. A notable feature of sub-thesis writing is that each sentence describing the relevant content expresses the same content to a large extent, and the topics contained in the sentences are basically the same. But it will also be mixed with some third-level content, that is, general construction sentences, which

will cause the deviation of the topic center. Therefore, how to eliminate the influence of the topic of the general construction sentence as much as possible, and choose the sentence that really expresses the point of view, needs further analysis.

For a paragraph in a document that expresses a sub-thesis as the object of analysis, in order to clarify an idea, it is usually necessary to use other sentences to explain and supplement, and usually the sub-thesis will appear in each sentence. For example, for the paragraph analysis of a document during the two sessions on "people's livelihood", the author will illustrate through examples of environmental protection, food, safety, etc. Through the topic analysis of LDA, three topics of environment, food and life will be obtained, and the most of the content is related to the topic of "life".

Based on this writing habit, this paper proposes a selection strategy of candidate sentences for sub-thesis based on topic selection: Taking the paragraph as the basic unit of analysis, after the LDA model analysis, the topic distribution of the paragraph is obtained, and several topics with the highest probability are selected as the target probability distribution. After that, the relative entropy of each sentence with this distribution is calculated. Then, the sentence with the smallest entropy value is selected as the candidate sentence for sub-thesis.

Algorithm 2 Sub-thesis candidate sentences selection algorithm

Input: Document $D = \{C_1, C_2, ..., C_m\}$
Output: The sub-thesis candidate sentences θ

```
Abstract sentences set θ = ∅; Topics T = {T_1, T_2, ..., T_K}; L = K; Paragraph C_i = {S_i1, S_i2, ..., S_im}
Divide sentences according to paragraphs, taking sentences of the same paragraph as a unit of analysis
while i < m do
    Count topics included in paragraph C_i
    Calculate the probability distribution P(T|C_i) of the topic on the paragraph
    while j < length(C_i) do
        if |θ_i| = ∅ or D_KL(P(T|S_ij)||P(T|C_i)) < value(θ_i) then
            θ_i = S_ij
        end if
    end while
end while
return θ
```

In Algorithm 2, the input text D consists of m paragraphs of text. For each paragraph C_i, first count the topics contained in the paragraph and the probability distribution of the topics. Then for each sentence S_{ij} in the paragraph C_i, we calculate the relative entropy of the topic probability distribution over the sentence and the paragraph. If the paragraph C_i has no candidate, that is, $|\theta_i| = \emptyset$ or the current relative entropy is greater than the relative entropy of the most similar sentence in the paragraph, that is, $D_{KL}(P(T|S_{ij})||P(T|C_i)) < \text{value}(\theta_i)$, set the current sentence as the candidate center sentence of the current paragraph, i.e. $\theta_i = S_{ij}$. In this way, until the calculation of all paragraphs is completed, the sentence saved in the θ_i array is the central sentence of each paragraph, that is, the sub-thesis of the document.

VII. Experimental Results and Analysis

The purpose of this experiment is to use the relative entropy method on the basis of the topic model to verify the correctness of the general thesis and sub-thesis summary methods, and how to set the parameters to maximize their accuracy.

A. Dataset and Evaluation Method

This paper uses the NLPCC 2015 [26] corpus as the experimental object. NLPCC2015 has three tasks. The third one is to perform news text summarization task for Weibo. The dataset consists of 250 news predictions that have been sentenced, completely sourced from Sina.com. The model training data uses the Internet corpus provided by Sogou Lab, involving 1761 articles in 9 aspects of recruitment, tourism, military, health, IT, sports, finance, market, and culture.

The evaluation method used for scoring is ROUGE [27] adapted to Chinese[1]. It's a recall-based calculation method adopted by most works as an automatic evaluation tool for the quality of text summaries. ROUGE measures the quality of automatic text summarization by calculating the degree of overlap between automatic summaries and expert summaries on various evaluation criteria. Usually, the degree of overlap between automatic summarization and expert summarization in N-grams or the length of the maximum co-occurrence subsequence is used. Among them, $N = \{1, 2, 3, 4\}$, indicating the coverage ability of automatic summarization on the content of expert summarization. It is generally believed that the 1-gram-based ROUGE score (ROUGE-1) reflects the closeness of automatic summarization and expert summarization [28], while ROUGE-2 reflects the smoothness of automatic summarization. The value range of the ROUGE score is $[0, 1]$. The closer to 1, the closer the automatic abstract to the expert abstract.

B. Experimental Results

The LDA model is used for model training on the training dataset. Since the main categories contained in the dataset are known, the number of topics of the LDA model is artificially set to 9. The output model of LDA is a two-dimensional array representing the corresponding probability of word-topic.Since the number of vocabulary is too large, the top ten words with the highest probability under each category are selected for display here. The output results are presented in Table I.

From Table I, we can see that the categories corresponding to the nine topics are: recruitment, tourism, military, health, IT, sports, finance, market, and culture.

1) Experiment of General thesis Abstract Method: Based on the output results of the LDA model, Algorithm 1 is used to extract the sentences expressing the general thesis in the text, and compare the accuracy with the manual summary. The accuracy of the result is only about 30%, and the effect is very poor. Although there are many topics that may be

[1]All experiments were re-run by ourselves since our algorithm did not participate in the NLPCC.

TABLE I. SOME OUTPUT RESULTS OF THE LDA MODEL

Topic 0:	Topic 1:	Topic 2:
Major=0.0083278164279716	Travel=0.013141023064702821	USA=0.007125444347199101
Student=0.007842661514385616	Visitors=0.004266203667474156	Japan=0.004809696664837551
Work=0.00733710875165395	Culture=0.004019739964102812	Plane=0.003791742025381582
School=0.0062748712404097975	City=0.0028504078737947554	Training=0.0037772489295390833
Education=0.005451361658338121	World=0.00240451233961009	System=0.003626238305766903
University=0.005270162961600818	Travel agency=0.0021930419407089076	Troop=0.003605801384330944
Candidate=0.005198484819478605	Chengdu=0.001931720410524329	Military=0.0035635128929766818
Exam=0.005015208574173687	Active=0.001658761327230168	Combat=0.0033607386032706472
Occupation=0.0035060068912367887	Park=0.0016147225498504244	Equipment=0.003248380941722047
Admissions=0.003493704451970437	Passenger=0.0015058548118690985	Progress=0.0032425202170355927
Topic 3:	**Topic 4:**	**Topic 5:**
Hospital=0.0059093508141070915	Computer=0.0020191139266127283	Match=0.00949530427474704
Sohu=0.005476764815671471	Internet=0.002007672796693706	Team member=0.0036720771812709914
Treatment=0.004703755393063685	iPhone=0.0017364146623097224	League=0.0032591843413982942
Live member=0.00459868303471546	Index=0.0015133078417762656	Team=0.0029232011897390537
Patient=0.0031455965425699617	Big data = 0.0013829300513194772	Reporter=0.002426053667587339
Health=0.0025964168608791265	System=0.0013096471915589481	Champion=0.0023930186697693075
Surgery=0.002488186684299358	Mobile=0.001085641175064798	Club=0.00237514767051908
Female=0.002363139272183045	BAT=0.0010612095334178121	Players=0.0022026031092450336
Found=0.0022240460467755595	AI=0.001029019868443202	Final=0.0021788234475517664
Occurrence=0.0021507363034214463	Understanding=9.899805996801138E-4	Season=0.002021560606839102
Topic 6:	**Topic 7:**	**Topic 8:**
Company=0.01909300611739068	China=0.009771496614544694	Interest=0.008119586724125997
Shareholder=0.008083219805765936	Company=0.007841044022748052	Life=0.004442234745400481
Shares=0.006859302052065136	Market=0.007116032951475622	Work=0.0037912371185839034
Ltd = 0.005513593668215	Enterprise=0.0067178444904189465	Child=0.0036027960884871127
Investment=0.004954318183600884	Development=0.006082827574448052	Know=0.0033194226653292082
Equity=0.004667721949358522	Current=0.0043594695781138115	Problem=0.0031881717400462665
Item=0.004523799009845025	Reporter=0.004117429660299634	AC=0.0030608909613718408
Securities=0.004491724613939593	Already=0.0037238429854752793	Man=0.0030123506377184076
Funding=0.004018111411489305	Product=0.003661836682905302	China=0.0027328446003346604
Reform=0.003952631004273419	Country=0.003587531849997642	Culture=0.0026000345127893025

expressed in a document, most of them have nothing to do with the main thesis expressed, so the topics in the document are ordered in non-ascending order of probability, and the topics with relatively small probability are gradually removed. The relative entropy of the topic distribution in the document and the topic distribution of each sentence is calculated, and the accuracy is calculated for different reduction rates (the interval between two reductions is 3%).

Observing Fig. 1, we can see that although there are some fluctuations (caused by the reduction of the interval), it can be observed that when the subject of a document is reduced to between 26% and 33%, the accuracy rate is the highest, that is, the calculation is performed at this time. The resulting automatic summary is the most likely to be a manual summary. Based on this conclusion, this paper sets the text topic reduction rate as 30% as the calculation parameter for the subsequent experiments.

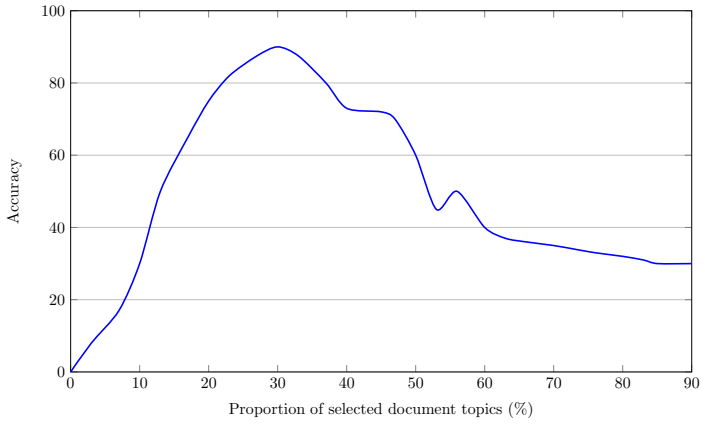

Fig. 1. The Impact of Document Topic Selection Ratio on Abstract.

The requirement of NLPCC2015 for text summarization is for Sina Weibo [29]. Given a document, it needs to be summarized into an abstract that can be used as a Weibo (up to 140 words), so the summary needs to be as short as possible. This is in line with the concept of the general thesis presented

TABLE II. ROUGE SCORE OF EACH ALGORITHM WHEN THE ABSTRACT LENGTH IS 80

Algorithm	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
WUST-1	0.563	0.322	0.234	0.189	0.624
WUST-2	0.572	0.331	0.239	0.191	0.631
Team-Best	0.542	0.312	0.229	0.178	0.610
FMAS	0.483	0.297	0.211	0.164	0.601
SentenceRank	0.470	0.308	0.226	0.175	0.598
GenSubE	**0.571**	**0.328**	**0.232**	**0.198**	**0.637**

TABLE III. ROUGE SCORE OF EACH ALGORITHM WHEN THE ABSTRACT LENGTH IS 140

Algorithm	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
WUST-1	0.586	0.395	0.275	0.206	0.650
WUST-2	0.599	0.409	0.290	0.218	0.653
Team-Best	0.587	0.413	0.264	0.212	0.639
FMAS	0.511	0.332	0.217	0.201	0.621
SentenceRank	0.535	0.386	0.283	0.207	0.582
GenSubE	0.532	0.351	0.254	0.203	0.631

in this paper. Our proposed algorithm (denoted GenSubE) is compared to the following algorithms:

- SentenceRank [30]: It evaluates the importance of sentences by measuring the relationship between them.
- Team-Best [25]: It's based on a super-edge sorting method, first find the subject word, calculate the sentence weight, and then use the edge-based random walk algorithm.
- WUST-1 and WUST-2 [26]: The best systems from the NLPCC 2015.
- FMAS as the baseline on the dataset. It's a pure statistical-based summarization method, considering TF-IDF (Term Frequency Inverse Document Frequency), sentence position, sentence length, and sentence similarity to calculate sentence weights.

The comparative results of the performance of the compared algorithms are presented in Tables II and III in terms of ROUGE evaluation indicators, when the target abstract lengths are 80 and 140 characters, respectively.

From Table II, it can be seen that for the 80-word abstract, the score of our proposed algorithm is almost similar as the best competition algorithms (WUST-1 and WUST-2) with the highest scores in the dataset, which are higher than the commonly used SentenceRank algorithm and the statistics-based FMAS algorithm. The ROUGE-1 evaluation index is usually regarded as the best criterion for judging automatic summaries and manual summaries. It can be seen that the recall rate achieved on ROUGE-1 is close to the best results. On the 140-word abstract, in Table III, the results of our proposed algorithm are not particularly outstanding. Only higher than the FMAS algorithm, and there is no advantage in the ROUGE score compared to the other text summarization methods. This is because the sentences extracted by the general thesis abstract method are usually not very long, usually only two sentences, so it has a good performance on the 80-word abstract. In the comparison of the 140-word abstract, it only maintains an above-average level.

2) Experiment of Sub-thesis Abstract Method: In this section, the general thesis and sub-thesis abstracting methods are used, and only the 140-word text abstract task is compared with other algorithms to observe the performance results. Algorithms 1 and 2 are combined to extract the sentences of the document's general thesis and sub-thesis, respectively. The performance of our method is compared with the other summary algorithms on the 140-word summary task. The performance results are presented in Fig. 2.

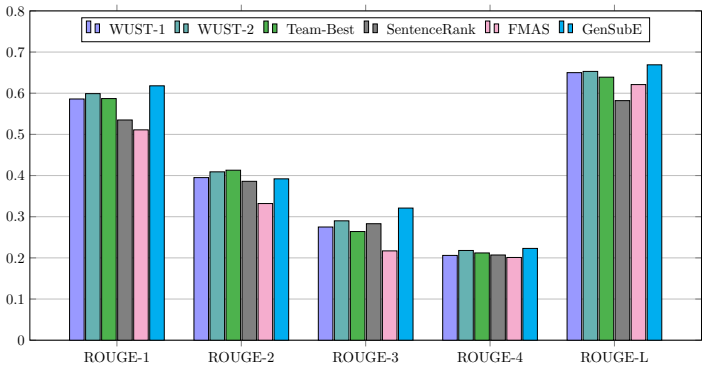

Fig. 2. ROUGE Value of Each Algorithm when the Abstract Length is 140 Words.

Fig. 2 shows that after extending the length of the abstract to 140 words, the score of our proposed method has been greatly improved compared with the method of simply extracting the general thesis. Compared with the SentenceRank algorithm and the statistics-based FMAS algorithm, it has great advantages in the all ROUGE evaluation indicators. The results show that our algorithm outperforms the best competition algorithms on all scores except on ROUGE-2 which is very close. Compared with the general thesis method, the text summaries obtained by the combined method have a great improvement in sentence fluency and comprehensive coverage. At the same time, compared with the other algorithms, it has obtained better than the best competition algorithms. The method proposed in this paper performs very well on the NLPCC2015 dataset.

VIII. CONCLUSION

In this paper, we propose a new method for document summarization. After processing the document through the LDA model, the probability distribution of the word-topic can be obtained. Firstly, we convert the probability distribution of word-topic into the probability distribution of topic-sentence to extract sentences in the document based on semantic analysis. After that, in order to measure the relationship between two sentences or sentences and document, relative entropy is introduced to measure the similarity of two probability distributions. The relative entropy of the topic probability distribution of the sentence over the document, and of the sentence over the paragraph are calculated, respectively. The smallest entropy value indicates that the difference is relatively small, and can be used as the central sentence of the paragraph. Also, this paper introduces a new document decomposition method based on relative entropy analysis. Through experiments and analysis on the NLPCC 2015 dataset, it can be known that when the

number of document topics is reduced to 30%, the probability distribution of abstract sentences and document topics are the most similar. At the same time, the results obtained on the 80-word abstract task and the 140-word abstract task are compared with other method. The performance results demonstrated the efficiency of using the relative entropy of the topic probability distribution over sentences to measure sentence relations.

REFERENCES

[1] H. P. Lunh, "The automatic creation of literature abstracts," *IBM Journal of Research Development*, vol. 2, pp. 159–165, 1958.

[2] P. B. Baxendale, "Machine-made index for technical literature—an experiment," *IBM Journal of Research and Development*, vol. 2, pp. 354–361, 1958.

[3] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, pp. 264–285, 1969.

[4] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4425 LNCS, 2007, pp. 557–564.

[5] D. Gillick and B. Favre, "A scalable global model for summarization," in *Association for Computational Linguistics (ACL)*, 2009, pp. 10–18.

[6] C. Li, X. Qian, and Y. Liu, "Using supervised bigram-based ilp for extractive summarization," in *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 1, 2013, pp. 1004–1013.

[7] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 2010, pp. 912–920.

[8] ——, "A class of submodular functions for document summarization," in *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2011, pp. 510–520.

[9] J. Kupiec, J. Pedersen, and F. Chen, "Trainable document summarizer," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 1995, pp. 68–73.

[10] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 2001, pp. 406–407.

[11] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, "A text summarization method based on fuzzy rules and applicable to automated assessment," *Expert Systems with Applications*, vol. 115, pp. 264–275, 2019.

[12] K. M. Svore, L. Vanderwende, and C. J. Burges, "Enhancing single-document summarization by combining ranknet and third-party sources," in *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 448–457.

[13] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 1, 2016, pp. 484–494.

[14] R. Nallapati, B. Zhou, C. dos Santos, Çaglar Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, 2016, pp. 280–290.

[15] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner : A recurrent neural network based sequence model for," *The Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3075–3081, 2017.

[16] J. Xu and G. Durrett, "Neural extractive text summarization with syntactic compression," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020, pp. 3292–3303.

[17] D. Anand and R. Wagh, "Effective deep learning approaches for summarization of legal texts," *Journal of King Saud University - Computer and Information Sciences*, 2019.

[18] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive summarization as text matching," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6197–6208.

[19] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D. R. I. M. Setiadi, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, 2020.

[20] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, pp. 1–66, 2017.

[21] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, 2021.

[22] J. Li, C. Zhang, X. Chen, Y. Hu, and P. Liao, "Survey on automatic text summarization," *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, vol. 58, pp. 1–21, 2021.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[24] G. Heinrich, "Parameter estimation for text analysis (based on gibbs sampling)," *Technical Report, Fraunhofer IGD*, 2009.

[25] R. Arora and B. Ravindran, "Latent dirichlet allocation based multi-document summarization," in *Proceedings of SIGIR 2008 Workshop on Analytics for Noisy Unstructured Text Data, AND'08*, 2008, pp. 91–97.

[26] X. Wan, J. Zhang, S. Wen, and J. Tan, "Overview of the nlpcc 2015 shared task: Weibo-oriented chinese news summarization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9362, 2015, pp. 557–561.

[27] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proceedings of the workshop on text summarization branches out (WAS 2004)*, 2004.

[28] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003*, 2003.

[29] C. SINA. (2009) Sina weibo. www.weibo.com. [Online]. Available: www.weibo.com

[30] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A meeting of SIGDAT, a Special Interest Group of the ACL held in conjunction with ACL 2004*, 2004, pp. 404–411.