# A survey of research methods of automatic text summarization

Min Zhang *, Cuiju Luan

School of Shanghai Maritime University, Shanghai, China

* Corresponding Author Email: 1165990915@qq.com

**Abstract.** This template explains and demonstrates how to prepare your camera-ready paper for Trans Tech Publications. Automatic text summarization is an information compression technique that uses a computer to convert text or text collections into short summaries. Recently, studies on automatically summarizing texts using different methods have developed rapidly. By combing the relevant documents at home and at abroad, various techniques and methods involved in the existing automatic text summary task, as well as the commonly used evaluation indicators, the advantages and disadvantages of the current automatic text summary task are summarized and the future research trends are discussed.

**Keywords:** Automatic text summarization, extractive, abstractive, summarization approaches.

## 1. Introduction

Text summary refers to using a computer to automatically extract a short text that accurately reflects the content of the source text. It is a subset of natural language processing (NLP). The main aspects that should be considered when making a summary are: brief, reduce redundancy, and retain important information. The automated text summary process includes data preprocessing, model building, generating summaries, and evaluation of the automatically generated summaries using evaluation metrics.

The text summary method is classified in many ways. According to the different type of input text, it can be divided into single document text summary and multi-document text summary. Single document summaries are generated from a given document and multiple documents from a given set of subject-related documents. It can be divided into supervised summary and unsupervised summary. Supervised summary requires a large amount of annotated and label data to train models, and tests the trained model effects with unannotated data. Unsupervised summary does not require any training samples and directly models the data. According to the output type, it can be divided into extraction text summary method and abstractive text summary method. Extraction abstract is a summary of some important sentences and paragraphs extracted from the original text. Generative abstracts, which attempt to generate abstracts by understanding the meaning of the original text, allow for the generation of new words or phrases. Some scholars can generate abstracts, which can solve some problems encountered by abstractions and generative abstracts.

## 2. Common methods for text summary

### 2.1. Statistics-based approach

Statistical methods are based on statistical features. Commonly used statistical features include word frequency, TF-IDF, title, sentence position, subject word, sentence centrality and so on.

In 1958, Luhn extracted important sentences to obtain abstracts by ranking the sentences with high importance [1]. This was the first paper on automatic text summary technology, and the task of automatic text summary began. Word frequency, that is, in addition to stop words, the higher the frequency of words in the text, the higher the importance. Baxendale calculated the probability of the topic of the beginning and the end of the sentence position, and selected some sentences with high probability to generate the text summary [2]. Edmundson et al. combined multiple features such as word frequency, sentence location, cue word, and title word as comprehensive indicators to measure the importance of sentences [3]. The title, subheading, and words in the title of the text are generally

considered more important because it adds more meaning in determining the weight of each sentence [8]. Ko and Seo proposed a text summary method that integrates document information in summary extraction and extracts important sentences using contextual information and statistical methods [4]. First, combine each two adjacent sentences into a Bi-Gram pseudo sentence (BGPS), which contains more features than a single sentence. Then, BGPS is scored according to the statistical method, and finally the sentences corresponding to BGPS with high scores are selected as the summary sentences. Fattah through the position of the sentence, the sentence keyword data (according to the statistical characteristics such as word frequency, TF-IDF features, mutual information calculation), sentence length, the similarity of sentence and the title, the centrality of artificial features, the importance to the sentence, the sentence according to the scores, the important sentences, form a summary [5]. Specifically, the centrality of a sentence is determined based on words that overlap or appear more frequently in a given sentence in a document compared to other sentences in the document. Salon put forward the famous TF-IDF (word frequency reverse file frequency) method, the method is used to evaluate a file set or corpus a word in the importance of the importance of the word in the file, but also with the frequency in the corpus of inverse decline, its main idea is that if the article in a word in the frequency is higher, and in other articles in the frequency is less, it means the ability of the article is stronger [26].

The commonly used text summarization models based on statistical methods include Lead-3, text teaser [9], and text pronouns [10]. The Lead-3 algorithm extracts the first three sentences of an article as the abstract. In the text teaser paper, the optimal scoring sentence for the abstract is determined by calculating four factors: article title features, sentence length, sentence position, and keyword frequency. The features proposed in text pronouns include sentence position, presence of verbs and proper nouns, sentence length, n-grams features, word labels, word length, word weights, etc. Statistical methods are relatively traditional, feature extraction is relatively simple, and the methods are easy to implement. However, the use of words and sentences often remains superficial, so there are still some limitations.

### 2.2. Semantic-based approach

Semantics can simply be simply regarded as the meaning of the concepts represented by things in the real world, and the relationship between these meanings corresponding to the data. Semantics-based text summary method needs to use some tool to analyze the semantics, and then generate a final summary through a series of processes, the commonly used tools are HowNet, WordNet and so on.

Miller et al. proposed an automatic summary model based on the vocabulary chain, which evaluates the word chain together with the consistency of sentence length and vocabulary, and uses the heuristic algorithm to select the word chain with high score, and then generate the text summary [11]. Chen et al. applied the word chain method in the Chinese text abstract [12]. Take HowNet as the lexical chain, build the knowledge base, then identify the strong lexical chain, and finally select the summary sentences based on the heuristic rules. HowNet is a common-sense knowledge base with the concepts represented by Chinese and English words as the description object, and it reveals the relationship between concepts and the attributes of concepts as the basic content. Barzilay et al. collected the vocabulary related to a topic in the original document through the meaning of WordNet and part of speech marking tools to form a vocabulary chain [13]. WordNet is a broad semantic network of English vocabulary. Nouns, verbs, adjectives and adverbs are each organized into a network of synonyms, each set represents a basic semantic concept, and these sets are also connected by various relationships [14]. Liu et al. used the abstract semantic representation (abstract meaning representation, AMR) to parse the source text into a set of AMR graphs, then convert them into summary graphs, and then generate the text [30]. Cheng et al. proposed the method of automatic summarization of Chinese Web text. By analyzing the semantic associations between paragraphs, merging paragraphs with similar semantics, divides the theme level, and obtains the chapter structure. Combined with some heuristic rules, statistics statistical used to extract key words and key statements to generate the final summary [31].

Semantic-based methods can enhance the relevance of text entities with their related content and improve the accuracy of summaries. But it produces a summary that is not concise enough to describe the whole text very well.

### 2.3. The graph-based approach

The graph-based approach is to represent the relationship and importance between text units (words or sentences). The vertices of the graph represent the text units, and the edges represent the connection between the text units. The text is constructed into a topological graph, and the graph sorting algorithm is used to rank the word scores, and finally extracted from the graph. The commonly used algorithms are LexRank, TextRank, etc.

Mihalcea et al. proposed an algorithm based on graph sorting, and used PageRank algorithm and its improved algorithm TextRank to extract keywords and key sentences [15]. The relationship and importance between the sentences can be effectively illustrated by the graphs, and the sentences in the document are represented in the form of the nodes, and the edges represent the connection between the sentences. The connections between sentences are correlated based on similarity relationships. In addition, each sentence was scored and the highest scoring sentence was selected as the text summary. Erkan et al proposed LexRank, a graph sorting algorithm similar to TextRank. LexRank is an undirected powerless graph, with the method of matrix iterative convergence. If a sentence is similar to multiple sentences, then this sentence is more important and can be used as the content of the summary [16].

Chen et al. borrowed the PageRank algorithm and proposed WSRank, a single-text automatic summary algorithm for collaborative sorting of words and sentences [29]. It is an algorithm based on the cooperative ranking of words and sentences. It uses the interaction between words and sentences to obtain the iterative calculation of the weights between words and sentences to rank the weight of sentences and generate a summary. MM Raahman et al combined the TextRank algorithm and POSRank algorithm, and then proposed an optimization algorithm based on the fusion of three algorithms based on the factors based on whether the word appears in the title [17]. Huang Bo et al combined Word2vec with TextRank to improve the quality of abstracts [18]. Mehdad put forward the best path ranking strategy based on graph sorting algorithm, the method is based on graph sorting generation method, based on the original sentence extraction of the query phrases, using the vocabulary similarity for clustering of the selected sentences, in each sentence with words as the node set of the directed graph, and the adjacent words with directed edges [20]. Zheng et al. proposed the directed graph model PACSUM, which uses the pre-training model to generate the sentence vector form to capture the deeper semantic information of the sentence and make the abstract sentence more accurate [19]. The model adds the sentence position information to judge the pointing information between the sentences in the graph model.

The graph-based approach is unsupervised, does not require extensive corpus practice, and the language is independent and simple. But it is heavily calculated, slow, and it only depends on sentence similarity, so it is inevitable that the selected sentence similarity is extremely high.

### 2.4. Motif-based approach

Most text content in text information contains multiple topics, and the role of these topics cannot be ignored when extracting text summary. In order for computers to truly understand the text, some algorithms based on subject models gradually appear. Common methods include Latent Semantic Analysis (LSA), Latent Dirichlet Distribution model (LDA) and so on.

Gong and Liu initially proposed a method that uses LSA to select highly ranked sentences for single and multi-document summarization in journalism [21]. The core idea of LSA is to map words and documents into the latent semantic space, remove part of the noise in the original vector space by dimension reduction, and extract the concept of words in the document into the low-dimensional space, which is a data model. In order to make up for the shortcomings of LSA, Hofmann et al. proposed Probabilistic Latent Semantic Analysis (PLSA) [23]. PLSA is an unsupervised learning

algorithm that uses the probabilistic generation model to conduct topic analysis on text sets. The LDA proposed by Blei et al. is a subject model for processing documents. It is improved based on PLSA, adds hyperparameters, and follows the Dirichlet prior distribution in the word distribution and subject distribution part [22]. The LDA topic model is shown in Fig.1.

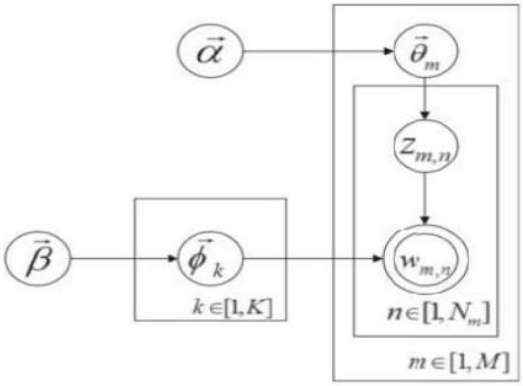


**Figure 1.** LDA topic model

With the emergence of clustering methods, based on topic models, researchers constantly abstrthem based on topic clustering. The method divides similar sentences or paragraphs into different clusters, from which a common theme or subtopic is found, and then text units are selected from these clusters for summary. The key factors to consider in clustering methods are the sentence scoring calculation of clustering sentence clustering ranking and the selection of representative sentences or the highest scoring sentences from each cluster, which is an unsupervised algorithm. Commonly used clustering algorithms are K-means [32], Hierarchical Clustering Algorithms [33], Expectation Maximum (EM) [34] and so on.

Teng et al proposed a single document abstract method combining word frequency and local topic identification [24]. This method first calculates the similarity of sentences, identifies them through sentence clustering, and then selects those sentences containing local topics according to the entry frequency. Liu et al proposed a rank-based sentence clustering framework that treats a word as an independent text object rather than the characteristics of the sentence [25]. Sentence clustering is a very important approach in topic-based summaries, which is able to discover various topics and cluster them according to them. Banerjee et al identified the most important documents from the collection of multiple documents, and each sentence in this document is initialized as a separate cluster, and then the sentences in the other documents are separately aggregated into the cluster with the highest similarity [27]. XU et al. proposed a Multi-Document Summarization Algorithm based on Topic Clustering (MDSTC) [28]. The step of text density sorting is first added to the typical clustering algorithm to determine the initial number of clustering centers, thus automatically discovering the number of subtopics hidden by the text collection. Then, the convolutional neural network algorithm is used to extract abstracts from different subtopic sets, supervised training of clustered topic text, score and mark all sentences, and finally select the statements that meet the central content as the text summary.

The topic-based method can better make the computer understand the text, and the extracted abstract is more suitable for the document itself, but its effect depends very much on the quality of the data set and the application field. Moreover, the number of initial clusters of the topic clustering algorithm usually needs to be given manually. The system cannot automatically generate the initial number of initial centers of the cluster, and the number of potential subtopics in the set.

### 2.5. Machine learning-based approach

With the development of artificial intelligence, more and more scholars further adopt the machine learning method to abstract the text. Kupiec et al applied the first statistical machine learning methods in the field of text summary [40]. The classical machine learning algorithm is used to identify whether the pre-processed generated sentences belong to the reference summary, that is, the currently learned

model is used to classify whether a given new sentence belongs to the reference summary. Through the idea of supervised learning, this technology uses the manually annotated data set to train and determine the parameters of the model, so that the computer can obtain the central meaning of the article and the characteristics of the text, build a reasonable model, and use the model to predict the text information in the unannotated data set.

At present, the widely used supervised learning algorithms include naive Bayes algorithm, decision tree algorithm, support vector machines (SVM), maximum entropy algorithm, logistic regression, etc. Naive Bayes algorithm is a classification algorithm based on Bayes theorem and feature independent assumption. Through this classification model, we can judge whether a sentence in a document should be selected as summary [40]. This method can be used for small-scale data sets, and the classification effect is good. Lin et al. proposed to use decision tree algorithm to score sentences, and selected several sentences with high scores to produce summaries [41]. This is a classification algorithm based on tree structure. The decision tree can show the process and results of classification very intuitively, and once the model is successfully built, it has a very high classification effect. The most classic decision tree algorithms are ID3, C4.5, and CART. Zhang et al. proposed to use Support Vector Machines (SVM) for extraction summary [37]. SVM is a kind of generalized linear classifier with binary classification of data. Its decision boundary is the maximum margin overplane for the learning sample, which can be solved into a convex quadratic programming problem [44]. The maximum entropy algorithm is to select the model with the maximum entropy from the set of models that meet the constraints to judge whether a sentence should be selected as a summary. It can use uncertain information for prediction and classification [42]. The Conditional Random Fields (CRF) is an undirected graph model that combines the maximum entropy model and the hidden Markov model [7]. Shen et al. was the first to use the conditional random airport to the extraction summary model, and added the common features of supervised learning to the conditional random airport, and achieved good results [6].

Common unsupervised learning algorithms include Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Hidden Markov Model (HMM) [43], clustering algorithm [45] and so on. Principal component analysis (PCA) is the conversion of some columns of possible related data into linear irrelevant data using orthogonal transformation to find the principal component. Gong Y using SVD to model text representation [21]. SVD was used to extract the most common and most relevant word combination patterns from the input document and represented them as singular vectors together with the sentences containing the pattern. Each singular vector implicitly represents the significant subject in the input document, and the sentences containing this word combination pattern will have the largest index value in the singular vector and are further arranged in descending order according to the highest index value included in the abstract. Conroy utilized the hidden Markov model in the summary algorithm to calculate the sentence score by using some features in the document, and then generate the summary based on the sentence score [43].

Machine learning methods take good advantage of the powerful computational performance of computers to efficiently and rationally model massive text information. But this approach relies heavily on the corpus, and the model is prone to overfitting. The subsequent research applied the neural network to the text summary, which played a very good effect.

### 2.6. Deep learning-based approach

In recent years, deep learning methods have been constantly studied and applied in generative text summary tasks, and have achieved good results. Compared with the extracted abstract, the generative abstract word is more flexible, which can avoid redundancy, and the generated abstract is more in line with the results of the manual abstract. At present, the mainstream model is the model based on the sequence-to-sequence (Seq2Seq) framework [46]. The basic structure is mainly composed of encoder and decoder. Its core idea is to map a sequence as input to one sequence as output through the deep neural network model. The Encoder end is responsible for encoding the input sequence into an intermediate context vector, and the Decoder end decodes the intermediate context vector into a

variable length vector. As shown in Fig.2. Encoders and decoders are usually implemented by either a recurrent neural network (RNN) or a convolutional neural network (CNN). At the same time, in order to further improve the effect of the model and solve the problems of generating the summary, such as not repeating words and sentences, the attention mechanism (attention mechanism) function to improve the training efficiency is generally added.

Rush et al. have applied the Seq2Seq model to the problem of generative text summary for the first time [48]. They used the Encoder-Decoder framework based on the Long Short-Term Memory (LSTM) [49] and combined the attention mechanism to build the model of generative text summary, which has proved that the end-to-end method for summary generation is very effective. Yu et al. used Seq2Seq to learn a sentence compression model on the basis of sequence annotation, use this model to measure the advantages and disadvantages of choosing sentences, and complete the model training combined with reinforcement learning [55]. Jadhav et al directly used the Seq2Seq model to alternately generate the index sequence of the statement to complete the decimation summary task [56].

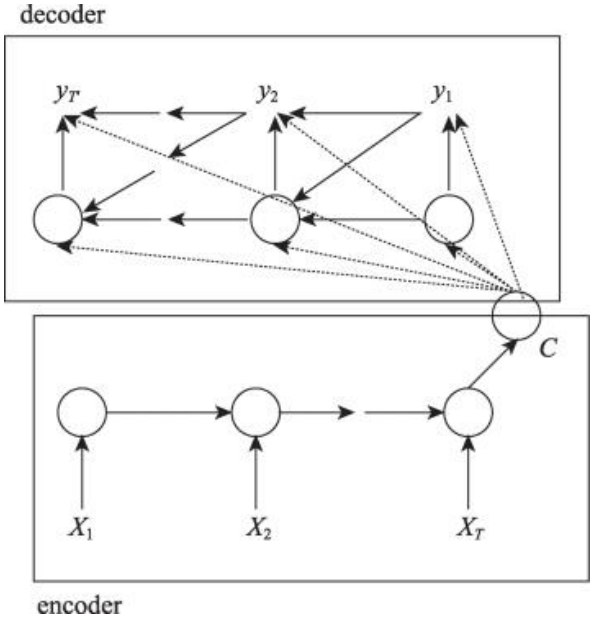


**Figure 2.** A sequential model of deep learning

Chopra et al introduced convolutional neural network to encode the original information, and used recurrent neural network to decoding and generate the summary, which achieved good results [51]. Nallapati et al performed experiments on dataset CNN / Daily Mail using a Gated recurrent unit (GRU) during the encoding phase [52]. Google used Beam-Search in its summary model, which uses RNN in the encoding and decoding stages to avoid wasting time and space [54]. Although the generative text summary method based on the Seq2Seq framework works very well, there are still some problems, such as the problem of generating continuous repeated words, and the problem of Out-of-Vocabulary (OOV). The Point-Generator Network (PGN) proposed by See et al. is widely used to generate summaries. The Copy and Coverage mechanism are added to the Seq2Seq based on attention mechanism, which effectively alleviates the problem of OOV and generated duplication [53].

In 2017, the Google team proposed a new network structure, Transformer. Transformer is a model based on the Encoder-Decoder framework, which is a deep learning model based entirely on the self-attention mechanism. It can solve the problem of long-term dependence, and it can also perform fast parallel computing. In addition, Transformer uses a multi-head self-attention (Multi-Head Self-Attention) mechanism to increase the diversity of Attention and capture effective information from different dimensions of the text [57].

Commonly used pre-training language models include ELMO [58], GPT [59], and BERT [60]. The bidirectional LSTM algorithm used in ELMo. The feature extraction algorithm used in the GPT is the Decoder model in the one-way Transformer. Compared with the LSTM model in ELMo, the

Transformer-based model has better feature extraction capabilities. BERT uses the Transformer-based feature extraction algorithm, the Encoder model of bidirectional Transformer. BERT uses the bidirectional Encoder part of Transformer to train the language model, which is then fine-tuned and applied to various downstream NLP tasks based on the pre-trained model. Zhang et al proposed a sequence-to-sequence model PEGASUS to solve the homomorphic encryption problem [61]. Song et al. proposed the mask sequence-to-sequence pre-trained model MASS based on the encoder-decoder-generated language [62]. Liu proposed BERTSUM, which is a simple variant of BERT, the first work of an extracted text summary using the BERT model [63]. Dong et al proposed the UNILM model, which can simultaneously handle natural language understanding and generation tasks, and can complete one-way, sequence-to-sequence and two-way prediction tasks [64]. Lewis et al proposed a denoised autoencoder BART for pre-trained sequence-to-sequence models [65].

Many scholars combine the generative abstraction method for summary, which alleviates the problems of redundancy and duplication. Chen et al. designed a model of extraction-generative structure to solve the problem of slow generative summary when facing long text [68]. The selected part selects the key sentences, and the generated part rewrite the selected sentences and generates the summary. Literature proposes an extraction-generative summary generation model based on the pointer network, first using the Text Rank algorithm and merging the subject similarity to extract important sentences in the article [67]. Then the summary generation task is realized by integrating the extraction information semantic and the generative framework based on Seq2Seq model introducing pointer network.

At present, the best summary is the generative text summary based on deep learning. In recent years, researchers have mostly adopted this method to conduct research. However, this method relies very much on labeled samples, and has problems such as repeated generation and early text termination. There are still many challenges and needs, which researchers should work hard to solve.

## 3. Evaluation Criterion

The current evaluation method is divided into manual evaluation method and automatic evaluation method according to whether the manual participation is made. The manual evaluation evaluates the candidate summary by experts, while the automatic evaluation method compares the similarity between the summary generated by the model and the reference abstract. Commonly used automatic evaluation methods are ROUGE [66], BLEU [39], and METEOR [38].

### 3.1. ROUGE

ROUGE is a recall-based similarity measure that works by comparing the abstracts automatically produced by the model to a set of reference abstracts usually produced manually, including four indicators: ROUGE-N, ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S, with the most commonly used indicators being the former two. The main calculation method of ROUGE-N is as follows:

$$ROUGE-N=\frac{\sum_{S\in\{\text{Re } feremce\ Summaries\}}\sum_{gram_n\in S}Count_{match}\left(gram_n\right)}{\sum_{S\in\{\text{Re } feremce\ Summaries\}}\sum_{gram_n\in S}Count\left(gram_n\right)} \tag{1}$$

Where the n-gram is a phrase composed of consecutive n words in a statement, $Count_{match}\left(gram_n\right)$ represents the n-gram numbers present in both the reference and candidate abstracts, $Count\left(gram_n\right)$ represents the number of n-grams appearing in the reference abstract.

ROUGE-L evaluates the summary based on the longest common subsequence (LCS). The calculation method of ROUGE-L is as follows:

$$ROUGE-L=\frac{Count\left(LCS\right)_{\max}}{Count\left(words\right)} \tag{2}$$

Among them, the molecule represents the number of words in the longest public sequence of the machine-generated summary and the manually annotated summary, and the denominator represents the number of words in the manually annotated summary.

ROUGE evaluation index is the most widely used at present, but this method can only evaluate the surface information of the reference summary and the system summary, which tends to investigate the importance and fluency of the summary, does not involve the semantic level, and cannot reflect whether a summary contains factual errors.

### 3.2. BLEU

BLEU evaluation index is often used in machine translation. The implementation method is to calculate the n-grams model of model generated sentence (candidate) and actual sentence (reference) respectively, and then calculate the number of matches. The values range from 0 to 1, and the closer the value is to 1, the better the machine translation result is. The calculation formula is as follows:

$$BLEU=BP\times\exp\left(\sum_{n=1}^{N}W_n\times\log P_n\right) \tag{3}$$

$$BP=\begin{cases}1 & lc>lr\\ \exp\left(1-lr/lc\right) & lc\le lr\end{cases} \tag{4}$$

Where, BP is the punishment factor, $W_n$ refers to the weight of the n-gram, generally set as 1 / N of the uniform weight, $P_n$ represents the accuracy of the n-gram, represents the detected text length, $lr$ represents the shortest reference text. The 1-gram of this evaluation index reflects the matching situation of the detected text and the reference text, while the other n-grams indicate the fluency of the detected text.

BLEU evaluation indicators are cheap and fast, easy to understand, and unrelated to language, and highly correlated to human evaluation results. However, it does not consider the accuracy of semantic, sentence structure and grammatical, and cannot deal with rich sentences well. The evaluation accuracy will be affected by common words, and the indicators are biased to short translation results, and it does not consider synonyms or similar expressions, which will have the possibility that reasonable translation will be denied.

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

### 3.3. METEOR

METEOR is proposed on the basis that the evaluation index based on recall rate is better than the evaluation index based on precision rate. METEOR Address the defect in BLEU of not using recall, while using higher-order n-grams. METEOR Put forward three methods to calculate the occurrence times of co-occurrence words, namely, exact method, root method and semantic method. After counting the number of co-appearing words in these three ways, and then calculated from the harmonic average of precision and recall $F_{mean}$ .The formula is as follows:

$$F_{mean}=\frac{\left(1+\beta^2\right)PR}{R+\beta P} \tag{5}$$

Candidate translations and reference translations can continuously and orderly match the words to form a block (chunks). METEOR Penalizes the shorter matches. The penalty factor Penalty is calculated by the total number of words unigrams_match and chunks matching by the summary text, as follows:

$$Penalty = \gamma \left( \frac{chunks}{unigrams\_match} \right)^{\theta} \tag{6}$$

Finally, the final evaluation of METEOR is based on a harmonic average of the block (chunk) decomposition matching and the characterization decomposition matching quality, using the following formula:

$$METEOR = (1 - Penalty) F_{mean} \tag{7}$$

Where α, γ, and θ are the default parameters used for the evaluation, as they were manually adjusted. METEOR The accuracy and recall on the whole corpus are considered, along with the effects of sentence fluency and synonyms on the semantics. But METEOR, like BLEU, is more sensitive to length. Therefore, in the future, researchers need to further improve the existing evaluation index or design a more perfect evaluation index.

## 4. Summary

Domestic and foreign research achievements on the direction of text abstract are relatively rich, but there are still some problems. The extracted text summary method is relatively easy, and it is not completely separate from the document itself, but it may cause redundancy and incoherent generated summaries. The generative method is more flexible and can avoid redundancy, but its corresponding technology is not very mature and can only deal with supervised scenarios and short sequence texts. So far, most of the research work is based on the abstraction summary method, because the generative summary model still has large problems in many aspects, such as the readability, compression rate and accuracy of the generated summary, and the model development is difficult. The generative text summary method based on deep learning is the best summary method. In recent years, researchers have mostly adopted this method for research. However, this method relies very much on labeled samples, and has problems such as repeated generation and text incoherence. In general, the existing text summary method still has many challenges and needs, which researchers should work hard to solve.

In the future, researchers will need to design more efficient algorithms to meet the increasingly complex text summary requirements. There are still many deficiencies in the two forms of abstraction and generation. The fusion of the two will make up for many defects. This generation-abstraction method will be an inevitable trend.

Since the text summary task relies highly on high-quality annotation data, it is costly to rely only on manual participation to build the dataset. In the absence of public datasets, building a semiautomatic dataset approach can be useful.

There are still many deficiencies in the existing evaluation index, so in the future, researchers need to further improve the existing evaluation index or design a more perfect evaluation index.

At present, the task of text summary is more concentrated, mostly in the field of news, and less used in other fields. In the future, it will be expanded to more new fields, such as commentary summary, conference summary, etc. Future text summary tasks can also be combined with other directions, such as sentiment analysis, text classification, etc.

## References

[1] LUHN H P. The automatic creation of literature abstract s[J]. IBM Journal of Research and Development, 1958, 2 (2): 159 - 165. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68 – 73.

[2] BAXENDALE P B. Machine-made index for technical literature—an experiment [J]. IBM Journal of Research and Development, 1958, 2 (4): 354 - 361.

[3] EDMUNDSON H P. New methods in automatic extracting [J]. Journal of the ACM, 1969, 16 (2): 264 - 285.

[4] Ko Y, Seo J. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Pattern Recognition Letters, 2008, 29 (9): 1366 - 1371.

[5] Fattah MA, Ren F (2009) GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput Speech Lang 23: 126 - 144.

[6] Shen D, Sun J-T, Li H, et al. Document Summarization Using Conditional Random Fields. [C]//IJCAI: Vol 7. 2007: 2862 – 2867.

[7] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. proceedings of icml, 2002.

[8] Abirami Rajasekaran, Dr. R. Varalakshmi, Review on automatic text summarization, International Journal of Engineering & Technology [J], 7 (2.33) (2018) 456 - 460.

[9] TARDAN P, ERWIN A, ENG K I, et al. Automatic text summarization based on semantic analysis approach for documents in Indonesian language[C]//2013 International Conference on Information Technology and Electrical Engineering (ICITEE). 2013: 47 - 52.

[10] JAGADEESH J, PINGALI P, VARMA V. Sentence extraction based single document summarization [R]. Work shop on Document Summarization,2005.

[11] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38 (11): 39 - 41, 1995.

[12] Chen Y, Wang X, Guan Y. Automatic text summarization based on lexical chains [C]//Proceedings of the1st International Conference on Natural Computation. Springer, 2005: 947 - 951.

[13] BARZILAY R, ELHADAD M. Using lexical chains for text summarization [J]. Advances in Automatic Text Summarization, 1999: 111 - 121.

[14] JAIN A, GAUR A. Summarizing long historical documents using significance and utility calculation using WordNet [J]. Imperial Journal of Interdisciplinary Research, 2017, 3 (3).

[15] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, page 20. Association for Computational Linguistics, 2004.

[16] Erkan G, Radev D R. Lex PageRank: Prestige in multi-document text summarization [C]//Proc of the 2004Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2004: 365 371.

[17] Rahman M, Roy C K. STRICT: Information retrieval-based search term identification for concept location [C]. International Conference on Software Analysis, Evolution and Reengineering. IEEE, 2017: 79 - 90.

[18] Huang Bo, Liu Chuancai. Chinese automatic text abstract based on weighted TextRank [J]. Application Research of Computers, 2020,37 (02): 407-410

[19] Zheng H, Lapata M. Sentence Centrality Revisited for Unsupervised Summarization [J]. ar Xiv preprint ar Xiv: 1906. 03508, 2019.

[20] Mehdad Y, Carenini G, Ng R T. Abstractive summarization of spoken and written conversations based on phrasal queries [C]//Proc of the 52nd Annual Meeting of the ACL. Stroudsburg: ACL, 2014: 1220 1230.

[21] Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis [C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001: 19 – 25.

[22] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993 1022.

[23] Hofmann T. Probabilistic latent semantic indexing [J]. ACM SIGIR Forum, 2017, 51 (2): 211 218.

[24] Teng Z, Liu Y, Ren F, et al. Single Document Summarization Based on Local Topic Identification and Word Frequency [J]. Seventh Mexican International Conference on Artificial Intelligence, 2008: 37 - 41.

[25] Liu S X, Yang L P. Method and apparatus for improving the readability of an automatically machine-generated summary: U. Spatent 8, 650, 483. 2014 - 2 - 11.

[26] Salton, G, Yu, et al. On the construction of effective vocabularies for information retrieval [J]. Acm Sigplan Notices, 1975.

[27] Banerjee S, Mitra P, Sugiyama K. Multi-document abstractive summarization using ILP based multi-sentence compression [C]//Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2015: 1208 1214.

[28] XU Xiaolong, YANG Chunchun, Multi-document summarization algorithm based on topic clustering.1673 - 5439 (2018) 05 - 0070 - 09.

[29] Chen Chen, Zhang Lu, Wu Zhiang. Automa-tic summarization algorithm for word sentence collaborative sorting [J]. Journal of Jiangsu University (Natural Science Edition), 2016, 37 (4): 443 - 449.

[30] Liu Fei, Flanigan J, Thomson S, et al. Toward abstractive summarization using semantic representations [C]//Proc of the 2015Conf of the NAACL. Stroudsburg: ACL, 2015: 1077 1086.

[31] Wang Jicheng, Wu Gangshan, Zhou Yuanyuan, et al. A text structure guided automatic summarization method for Chinese Web documents [J]. Journal of Computer Research and Development, 2003, 3: 398 - 405.

[32] J. MacQueen,Some methods for classification and analysis of multivariate observations, Berkeley Symposium on Mathematical Statistics and Probabil-ity1967.

[33] M. Steinbach, Ge. Karypis, V. Kumara. A Comparison of Document Clustering Techniques. In KDD Workshop on Text Mining, 2000. (also see TR 00-034, University of Minnesota, MN).

[34] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, 2001.

[35] Shi Mengjie A Survey of Text Clustering Algorithms [J] Modern Computer, 2014 (2): 5.

[36] Wang Sen, Liu Chen, Xing Shuaijie, Overview of K-mean-s Clustering Algorithm [J], Journal of East China Jiaotong University, 2022, 39 (05), 119 – 126.

[37] Zhang J, Fung P. Speech summarization without lexical features for Mandarin broad-cast news[C]//Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. 2007: 213 – 216.

[38] DENKOWSKI M, LAVIE A. Meteor universal: Language specific translation evaluation for any target language [C]// Proceedings of the 9th Workshop on Statistical Machine Translation. 2014: 376 - 380.

[39] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for Automatic Evaluation of Machine Translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 311 - 318.

[40] KUPIEC J, PEDERSEN J, CHEN F. A trainable document summarizer [C]//The 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1995: 68 - 73.

[41] LIN C Y. Training a selection function for extraction [C]// The Eighth International Conference on Information and Knowledge Management. ACM, 1999: 55 - 62.

[42] Chen Jianfei, Zhu Jun. Efficient learning algorithm of maximum entropy discriminant topic model [J] Pattern Recognition and Artificial Intelligence, 2019, 32 (08): 736 - 745.

[43] CONROY J M, O'LEARY D P. Text summarization via hidden markov models[C]//The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001: 406 - 407.

[44] V Vapnik, C Cortes, Support-vector networks, Machine leaning 20 (3), 273 - 297.

[45] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, 2009.

[46] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. Computer Science, 2014.

[47] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724 - 1734.

[48] Rush A M, Chopra S, Weston J. A Neural Attention Model for Abstractive Sentence Summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 379 – 389.

[49] Hochreiter S, Schmid Huber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735 - 1780.

[50] Zhao Hong An overview of deep learning methods for generative automatic summarization [J] Journal of Information Science, 2020, 39 (3): 15.

[51] Chopra S, Auli M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks [C]//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT, 2016: 93 - 98.

[52] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond [C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Co NLL, 2016: 280 - 290.

[53] A. See, P. J. Liu, D. C. Manning. Get to the point: Summarization with pointer-generator networks [J]. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, 1073 - 1083.

[54] Abadi M, Barham P, Chen J, et al. Tensor flow: a system for large-scale machine learning [C]//The Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. OSDI, 2016: 265 - 283.

[55] L. Yu, W. Zhang, J. Wang, et al. Seqgan: Sequence generative adversarial-nets with policy gradient [C]. Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[56] A. Jadhav, V. Rajan. Extractive summarization with swap-net: Sentences and words from alternating pointer networks [C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, 142 - 151.

[57] Vaswani A, Shazeer N, Parmar N et al. Attention is All You Need. Advances in Neural Information Processing Systems, 2017: 5998 - 6008.

[58] PETERSME, NEUMANN M, IYYERM, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North America-n Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 2227 - 2237.

[59] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pretraining [J]. 2018.

[60] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171 - 4186.

[61] ZHANG J Q, ZHAO Y, SALEH M, et al. Pegasus: Pretraining with extracted gap-sentences for abstractive summarization[C]//Proceedings of the 37th International Conference on Machine Learning. 2020: 11328 - 11339.

[62] SONG K T, TAN X, QIN T, et al. Mass: Masked sequence to sequence pretraining for language generation [J]. ar Xiv preprint ar Xiv: 1905. 02450, 2019.

[63] LIU Y. Fine-tune BERT for extractive summarization [J]. ar Xiv preprint ar Xiv: 1903. 10318, 2019.

[64] DONG L, YANG N, WANG W H, et al. Unified language model pretraining for natural language understanding and generation [J]. ar Xiv preprint ar Xiv: 1905. 03197, 2019.

[65] LEWIS M, LIU Y H, GOYAL N, et al. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2019.

[66] LIN C Y. Rouge: A package for automatic evaluation of summaries [C]// Text Summarization Branches Out. 2004: 74 - 81.

[67] Chen Wei, Yang Yan Pointer network-based extraction generative summary generation model [J] Computer Applications, 2021, 41 (12): 3527-3533.

[68] Chen Y C, Bansal M. Fast abstractive summarization with reinforce selected sentence rewriting [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 675 - 686.