International Core Journal of Engineering Volume 8 Issue 12, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

Research on Automatic Summarization based on the Fusion of TextRank and Multi-dimensional Semantic Features

Fei Xu[a], Man Yun[b], and Bo Yang[c]

School of Computer Science and Engineering, Xi'an Technological University, Xi'an, China

[a]29112462@qq.com, [b]133629626@qq.com, [c]16764496@qq.com

Abstract

In today's Internet age, a large amount of news appears in people's daily life, and the automatic text summarization can summarize the key information and subject content of news, and thus help people reduce reading time. However, the core information of the summary generated by the traditional graph model algorithm and missing text cannot fully reflect the content of the article, so this paper proposes the automatic summary model MD-TextRank (Multi-dimensional TextRank) based on the fusion of TextRank with multi-dimensional semantic features. The method introduces word2vec to represent the news text information on the traditional TextRank algorithm, and updates the weight of sentenes from subject similarity, sentence and title similarity, keyword coverage, and feature words; and taking the military field as an example, the domain dictionary has been designed to make the generated summary better reflects the concern of the field. Experiments demonstrate that the MD-TextRank text summarization model proposed in this paper has improved 8.9,6.8, and 5.9 percentage points compared to the traditional algorithm TextRank on ROUGE-1, ROUGE-2, ROUGE-L, significantly improving the quality of automatic text summarization.

Keywords

Text Summarization; TextRank; Word2Vec; MMR Redundant Processing.

1. Introduction

The text summarization [1] technology can express the originally complicated and varied text with a brief news summary without changing the meaning of the article and without losing its important information, thereby helping people to reduce the reading time.

At present, the implementation of automatic summarization are mainly divided into extractive method [2] and abstractive method [3]. Among them, the extractive summary is to directly extract sentences that can represent the key information of the article from the original text as a summary. The abstractive summarization is based on the understanding of the original text to form a summary. It is closer to the essence of the summary and has the potential to generate high-quality summary. However, this method have problems such as word order errors and length dependence. The quality of the abstractive summary has not yet reached the requirements of practical applications [4]. Therefore, it is of great significance to study how to improve the extractive method.

The classic algorithm TextRank was first proposed by Mihalcea in 2004 [5]. This method uses iteratively calculating the similarity between sentences to score sentences to filter out the key sentences in the text to form a summary. Subsequent studies are based on the improvement of this method. S.Karlsson [6] modified the calculation method of similarity between sentences by means of semantic folding on the basis of TextRank, and achieved good results. Zhang Lu et al. [7] believe that the higher the coverage of keywords in a sentence, the more important the sentence is. The ROUGE

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

score on the DUC2002 dataset is improved by 13%-30% compared to TextRank. Li Feng et al. [8] expanded the keywords into the TextRank score to improve the effect of the summary, but ignored other themes, semantic and other factors that affect the quality of the summary. Yu Shanshan [9] and others introduced the title, paragraph, special sentence, sentence position and length and other information into the construction of the TextRank network graph, and proposed an improved sentence similarity calculation method, thereby improving the effect of the TextRank algorithm .S. Sehgal [10] added the similarity between the article and the title to the sentence weight, which improved the accuracy of the text summary. Hilário [11] and others proposed a weighting method that combines the coverage and location of the sentence to judge the importance of the sentence, covering the important concepts in the summary to the greatest extent. Cao Yang et al. [12] compared the automatic summarization effects of different similarity calculation methods, selected the best similarity calculation method, and combined the sentence position, clue words and classic TextRank to calculate the weight of the sentence. Chuanming Yu et al. [13] proposed an extractive text summarization model based on maximum boundary relevance. This model combines Maximum Border Relevance (MMR) with deep learning, and comprehensively considers features such as sentence and full text similarity, keywords, and location information to extract summary. Liu Zhiming et al. [14] proposed a topic-based emotional summarization method, which uses the LDA model to obtain the article topic, and then integrates traditional multi-features to extract the summary.

All the above researches have promoted the development of extractive summarization. However, there are problems of incomplete consideration when scoring sentences based on TextRank. Literature [7-8] only considers one factor. Although literature [9-13] considers multiple factors, it ignores the theme of the article. Literature [14] uses the LDA model to extract the topic of the article, but ignores the influence of feature words on the sentence weight. At the same time, when focusing on a specific field, the summary generated by the above related research may not contain the content that the field is really concerned about.

To solve the above problems, this paper proposes to take the four dimensions of similarity between sentences and topics, similarity between sentences and titles, keyword coverage, whether there are characteristic words as the influencing factors of sentence weight, and apply them to automatic summarization with TextRank in an optimized combination to improve the quality of summarization.

2. TextRank

2.1 TextRank Algorithm

2.1.1 Sub-section Headings

The TextRank [15] is an unsupervised algorithm based on graph model that can extract key sentences in the text. The essence is based on the iterative calculation of sentence weights.

The similarity between sentences is calculated in the formula as follows:

$$Similarity\ (S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i \ \&\ w_k \in S_j\}|}{log(|S_i|) + log(S_j)} \tag{1}$$

Where $Similarity\ (S_i, S_j)$ represents the similarity between two sentences S_i and S_j.

The sentence score formula is as follows:

$$TextRank(S_i) = WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{2}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

Where $WS(V_i)$ represents the weight of sentence S_i, $In(V_i)$ represents the sentence pointing to node V_i ,and $Out(V_j)$ represents the set of sentences pointed to by node V_j .

2.2 MD-TextRank

It can be seen from the formula (1-2) that when the TextRank algorithm calculates the sentence score, it simply counts the words that co-occur between sentences. Therefore, this paper proposes an automatic summarization model MD-TextRank (Multi-dimensional TextRank) based on the fusion of TextRank and multi-dimensional semantic features. The steps are shown in Figure 1.

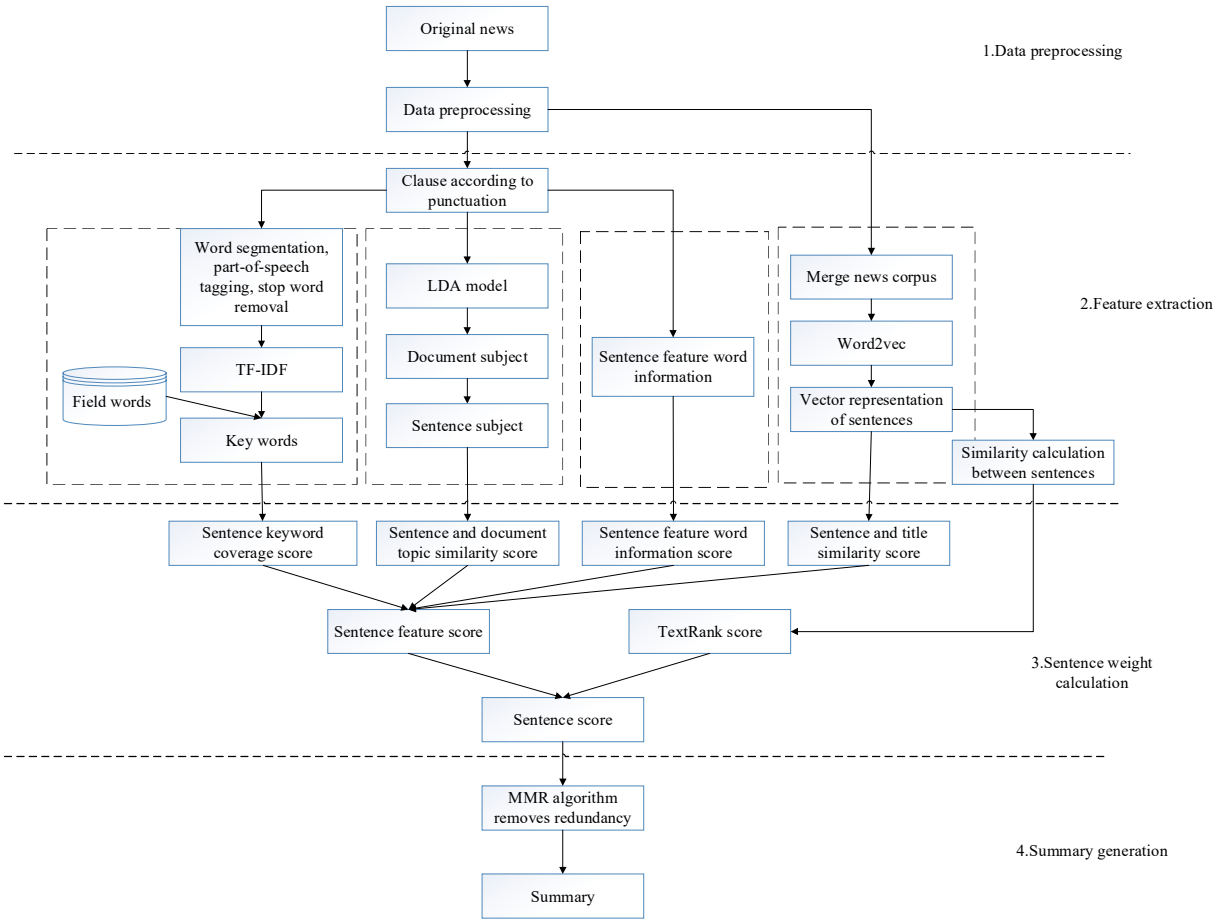

Fig. 1 MD-TextRank

2.2.1 Feature Extraction

(1) Keyword extraction

Generally speaking, the more keywords a sentence contains, the more likely it is to be a sentence in the abstract.The basic method adopted in this paper is the TF-IDF algorithm [19]. The specific formula is as follows[16]:

$$TF_w = \frac{N_w}{N} \tag{3}$$

$$IDF_w = log(\frac{Y}{Y_w + 1}) \tag{4}$$

$$TF - IDF_w = TF_w * IDF_w \tag{5}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

Where N_w refers to the number of occurrences of word w in a certain article, and N is the total number of words in the article. Y is the total number of documents in the corpus, and Y_w is the number of documents containing the word w.

Then, based on the candidate keywords obtained by the TF-IDF algorithm, only nouns and verbs that are more representative of the key information of the article are retained:

$$weight(i,M) = h \times count(i,M) \tag{6}$$

Where $weight(i,M)$ indicates the weight of the word i in text M; $count(i,M)$ indicates the frequency of the word i in text M, calculated by the TF-IDF algorithm; h is the part-of-speech impact factor.

(2) Acquisition of news topics

The topic of an article often contains the most important content in the article. Therefore, if the topic of the sentence is more similar to the topic of the document, the sentence is more likely to be the sentence in the summary of this article. This section uses the topic generation model based on Latent Dirichlet Allocation (LDA) [17-18] proposed in literature to obtain the topic distribution of the document and the topic distribution of the sentence. The topic distribution of the document can be obtained directly from the parameter of the LDA generation model, and the topic distribution of the sentence is as follows [17]:

$$P(T \mid S) = \frac{\sum_{W_I \in S} P(W_i \mid T) \times P(T \mid D)}{len(S)} \tag{7}$$

Where $P(W_i|T)$ represents the distribution probability of the word W_i under the topic, which is obtained from the LDA parameters θ.

(3) Sentence feature information

For a news article, if a sentence contains the feature words defined in Table 1, it is very likely that the sentence is the keynote sentence of the article.

Table 1. News Features Words (Part)

Numble	Feature word	Numble	Feature word
1	According to reports	4	indicate
2	It is said that	5	therefore
3	all in all	6	Obviously

(4) Sentence vector representation

Generally speaking, if a sentence is more similar to a news title, it means that the sentence can express the main point of the text, and the sentence is more likely to be a key sentence. This paper is based on the word2vec model [19] to calculate the word vector to enhance the expression of the semantic information of the word, the formula is shown in (8).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

$$\vec{s} = \frac{\sum_i^N w_i * weight(i)}{N}, w_i = (w_i^1, w_i^2, ..., w_i^n) \tag{8}$$

Where $\vec{s}$ represents the sentence vector of sentence S

2.2.2 Calculation of Sentence Weight

(1) Score of keyword coverage

We define the impact of keyword coverage on sentence weight as $Key(S)$. The formula is as follows:

$$Key(S) = \frac{\sum_{j=1}^m weight(j, M) + \sum_{i=1}^n W}{len(S)} \tag{9}$$

Among them, n represents the number of domain words contained in the sentence S. W represents the weight of the domain word. In order to balance the weight of the candidate keywords.

(2) Score of sentence and document similarity

This paper applies the Jensen-Shannon (JS) divergence formula [20] to the summarization task to calculate the correlation between two distributions as the similarity between the sentence and the topic. The JS divergence formula is based on the Kullback-Leibler (KL) divergence formula [21], and solves the problem of asymmetry in the distance expressed when the KL divergence formula is used to compare the correlation between distributions [22]. The formula is as follows:

$$KL(P \| M) = \sum_i P(i) \times log \frac{P(i)}{\frac{P(i) + Q(i)}{2}} \tag{10}$$

$$KL(Q \| M) = \sum_i Q(i) \times log \frac{Q(i)}{\frac{P(i) + Q(i)}{2}} \tag{11}$$

$$JS(P \| Q) = \frac{KL(P \| M)}{2} + \frac{KL(Q \| M)}{2} \tag{12}$$

Among them, P represents the topic distribution of the document, Q represents the topic distribution of the sentence. We define the similarity between the sentence and the topic of the document as $Theme(S)$. The formula is as follows:

$$Theme(S) = 1 - JS(P \| Q) \tag{13}$$

(3) Score of sentence feature information

We define the influence of the feature words defined in Section (3) in 2.2.1 on the weight of the sentence as $Feature(S)$. If the sentence contains the defined feature words, then the sentence will be weighted. The formula is as follows:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

$$Feature(S)=\begin{cases}1, sentence\ with\ feature\ words\\ 0, otheras\end{cases} \tag{14}$$

(4) Score of similarity between sentence and title

On the basis of obtaining the sentence vector, this paper proposes the following sentence similarity calculation formula:

$$Similarity(S_i,S_j)=\cos\theta=\frac{\overrightarrow{s_i}\times\overrightarrow{s_j}}{\|\overrightarrow{s_i}\|\times\|\overrightarrow{s_j}\|}=\frac{\sum_{k=1}^{n}\overrightarrow{s_{ik}}\times\overrightarrow{s_{jk}}}{\sqrt{\sum_{k=1}^{n}(\overrightarrow{s_{ik}})^2}\times\sqrt{\sum_{k=1}^{n}(\overrightarrow{s_{jk}})^2}} \tag{15}$$

We replace the calculation of sentence similarity in the TextRank algorithm of formula (1) with this formula.We define the similarity between the sentence S and the title on the weight of the sentence as $Headline(S)$. The formula is as follows:

$$Headline(S)=Similarity(S,S_{title}) \tag{16}$$

2.2.3 Comprehensive Score of Sentences

Based on the TextRank scoring of sentences,, this paper comprehensively considers the impact of keyword coverage, similarity between sentences and topics, similarity between sentences and titles, and sentence feature information on the sentence weight to obtain the final score of the sentence. The formula is as follows:

$$Score(S_i)=\lambda_1 Key(S_i)+\lambda_2 Theme(S_i)+\lambda_3 Feature(S_i)+\lambda_4 Headline(S_i) \tag{17}$$

$$W(S_i)=\frac{TextRank(S_i)+Score(S_i)}{2} \tag{18}$$

The sum of each λ is 1. $W(S_i)$ is the final score of the sentence S_i.

2.2.4 Summarization

After obtaining the weight of each sentence through section 2.2.3, we need to extract the first n most important sentences as the summary. However, if the first n sentences with the largest weight are directly extracted, there may be redundant components, that is, different sentences in summary express the same meaning, which violates the principle of conciseness and novelty of the summary. Therefore, this paper proposes the following formula based on the Maximal Marginal Relevance (MMR) algorithm [23] to remove redundant sentences and increase the readability of the summary.

$$MMR(S_i)=\lambda\times W(S_i)-(1-\lambda)\times max[Similarity(S_i,D)] \tag{19}$$

Among them, D is the set of sentences that have been selected. The first half of the formula indicates the score of the key information contained in the sentence, and the second half indicates the redundancy between the sentence and the selected into summary.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

3. Experiments and Analyses

3.1 Experimental Data

The data source used in this paper is the Sogou news datasets [24], which contains news on sports, finance, entertainment, military and other topics, with a total of nearly 100,000 news data, which is now commonly used for tasks such as document classification and automatic summarization.

3.2 Setting of Evaluation Indicators

This experiment uses three aspects of the standard evaluation tool ROUGE [25] (ROUGE-1, ROUGE-2 and ROUGE-L) to evaluate the summary results. The tool achieves the purpose of automatically evaluating the quality of the summary by comparing the overlap amount of the same evaluation unit (unary words, binary words, and longest substring) in the standard summary and the model-generated summary.

3.3 Analysis of the Experimental Results

3.3.1 Analysis of Model Comparison

In this section, the traditional TextRank model and other representative models improved based on TextRank are used as benchmark experiments on the Sogou public news dataset. And in order to verify the superiority of the model in this paper integrating the four factors, the models after adding the two and three factors with the greatest influence to TextRank are also used as benchmark experiments to compare with the summary results generated by MD-TextRank. The results are shown in Figure 2, in which the setting of experimental parameters are shown in Table 2 and the specific description of the model is shown in Table 3.

Table 2. Setting of experimental parameters

Parameter	Value	Parameter	Value
Number of LDA topics K	10	Word vector dimension	200
hyperparameter α of LDA	5	Number of windows of CBOW K	5
hyperparameter β of LDA	0.1	Hyperparameter λ of MMR	0.85
$\lambda_1 - \lambda_4$	0.2,0.4,0.3,0.2	Number of sentences in summary	3

Table 3. Description of benchmark model

Model	Description
TextRank	The basic model of extractive summarization. This method only considers the similarity between sentences and ignores the influence of other factors such as semantic on sentence weight.
+Feature&Headline	On the basis of TextRank, only feature word information, title and document similarity information are incorporated.
+Feature&Headline&key	On the basis of TextRank, only feature word information, the similarity of title and document , and keyword information are incorporated.
iTextRank[9]	On the basis of TextRank, it integrates title, paragraph, and sentence position information.
SW-TextRank[26]	It is the best model in extractive summarization, but it does not consider the influence of document topic and sentence topic correlation on sentence weight.
MD-TextRank	The weight of the sentence is updated from the four dimensions of similarity between sentences and topics, similarity between sentences and titles, keyword coverage,and feature words is included. And the redundancy is processed.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

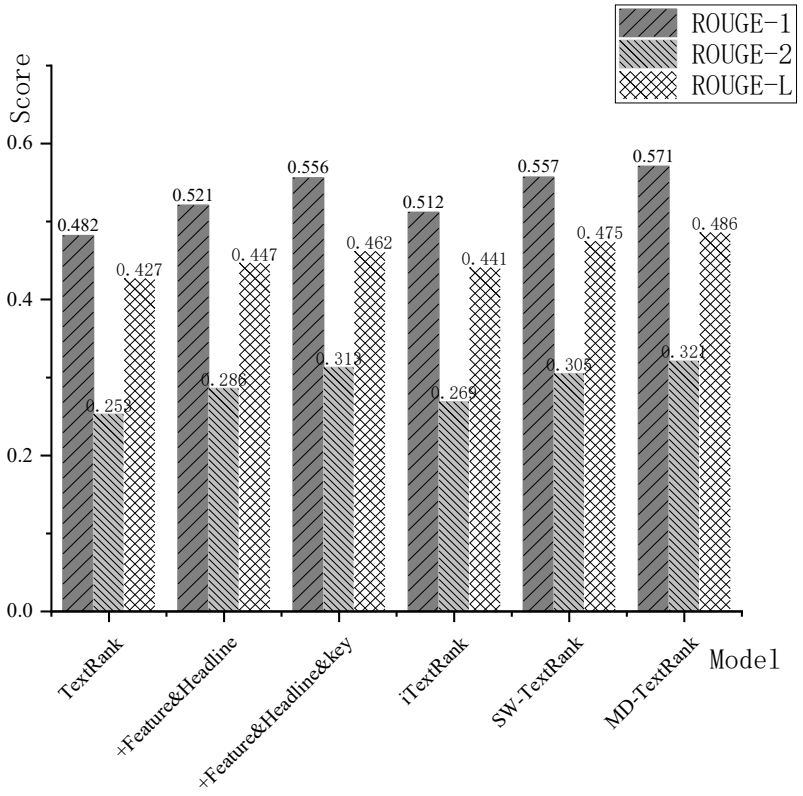

Fig. 2 The results of model comparison

As can be seen from the data in the figure, the summary generated by the model MD-TextRank in this paper performs best on ROUGE-1, ROUGE-2, and ROUGE-L compared to the traditional TextRank and other improved models. Among them, the highest scores of this model on ROUGE-1, ROUGE-2, and ROUGE-L are 0.571, 0.321, 0.486, respectively, while the highest scores of TextRank are only 0.482, 0.253, 0.427, which are increased by 8.9, 6.8, and 5.9 percentages respectively. Compared with the better-performing SW-TextRank model, the model in this paper has increased by 1.4, 1.6, and 1.1 percentages respectively. It can be seen that, compared with the traditional TextRank model and the improved models proposed by other scholars, the MD-TextRank model can effectively improve the quality of summary. On the basis of introducing word2vec to represent news text information, it enriches the calculation method of sentence weight by fusing the updated scores of the four dimensions of the sentence weight and the original TextRank score in a certain ratio.

3.3.2 Comparison of λ Coefficient Values

In section 2.2.3, we need to comprehensively consider the influence of different factors on sentence weight. Therefore, we set five sets of coefficient combinations for four different influencing factors, as shown in Table 4. And calculate the ROUGE score of the model for each set of coefficients, as shown in Figure 3.

It can be seen from the figure that when the group is 3, that is, λ_1, λ_2, λ_3, and λ_4 take 0.2, 0.2, 0.4, and 0.2 respectively, the ROUGE score is the highest, and the quality of the model generation is the best. It shows that the similarity between the sentence and the title has the greatest impact on the weight of the sentence, that is, if the sentence is more similar to the title, the more key information the sentence contains, the more likely it is the summary sentence of this article. Based on the results of this experiment, this paper sets the value of λ_{1-4} according to group 3.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

Table 4. Combination of influence factor coefficients

group	λ_1	λ_2	λ_3	λ_4
1	0.25	0.25	0.25	0.25
2	0.4	0.2	0.2	0.2
3	0.2	0.4	0.2	0.2
4	0.2	0.2	0.4	0.2
5	0.2	0.2	0.2	0.4

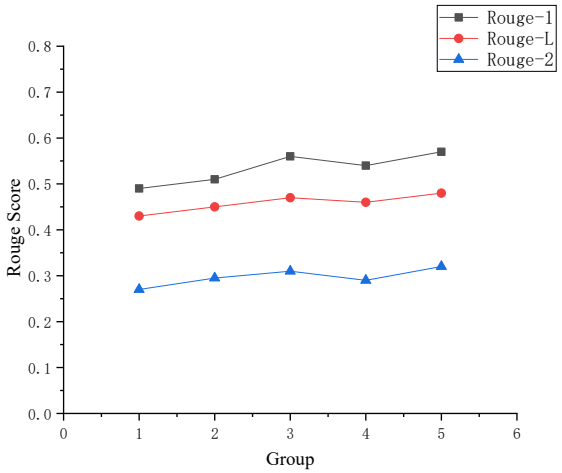

Fig. 3 Comparative evaluation of parameters λ

4. Conclusion

This paper proposes an automatic summarization method MD TextRank, which integrates TextRank scores with four semantic feature scores. This method comprehensively considers four influencing factors, namely, sentence and topic similarity, sentence and title similarity, keyword coverage, and whether there are feature words, and scores sentences from multiple dimensions and multiple perspectives. And the experiment proves that, compared with the traditional algorithm TextRank, which does not fuse factors and only fuses single factors, the score of the abstract extraction method that integrates four factors at the same time in this paper on ROUGE-1, ROUGE-2, and ROUGE-L has improved to varying degrees, indicating that the model MD TextRank in this paper can significantly improve the quality of news summary.

Acknowledgments

This work was supported by the State and Provincial Joint Engineering Lab. of Advanced Network, Monitoring and Control,China (grant number GSYSJ2018006); and Special scientific research project of Shaanxi Provincial Department of Education (grant number 18JK0399).

References

[1] Wang L. Several questions in the automated summary study [J].Book Intelligence Work, 2014,58 (20): 13-22.

[2] Le H T, Le T M. An approach to abstractive text summarization[C]//Proceedings of 2013 Soft Computing and Pattern Recognition (SoCPaR). Hanoi, Vietnam: IEEE,2013:371-376.

[3] Shi L, Ruan X, Wei Rui, Cheng Ying. Generative text summary study review based on the sequence-to-sequence model [J].Intelligence Journal, 2019,38 (10): 1102-1116.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0032

[4] Tang X B, Gu N, Tan M L. Automatic Summary of Chinese Based on Sentence Subject Discovery [J].Intelligence Science, 2020,38 (03): 11-16 + 28.

[5] Mihalcea R , Tarau P .TextRank: Bringing Order into Texts[C]// Proc Conference on Empirical Methods in Natural Language Processing.2004.

[6] Karlsson S. Using semantic folding with TextRank for automatic summarization[D]. Stockholm, Sweden: KTH royal institute of technology, School of computer science and communication ,2017.

[7] Zhang L, Cao J, Pu C Y, Wu Z. Single-document automatic summary algorithm based on the collaborative ordering of words and sentences [J].Computer application, 2017,37 (07): 2100-2105.

[8] Li F, Li Z J, Yang W M,et al. News Text automatic summary method using a keyword extension [J].Computer Science and Exploration, 2016,10 (03): 372-380.

[9] Yu S S, Su J D, Li P. Automatic summary extraction method based on the improved TextRank [J].Computer Science, 2016,43 (06): 240-247.

[10] Sehgal S , Kumar B , Maheshwar, et al. A Modification to Graph Based Approach for Extraction Based Automatic Text Summarization[J].2018.

[11] Oliveira H , Lima R , Lins R D , et al. A Concept-Based Integer Linear Programming Approach for Single-Document Summarization[C]// 2016 5th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2017.

[12] Cao Y.A Single-document Automatic Digest Study based on the TextRank algorithm [D].Nanjing University, 2016.

[13] Yu C M, Guo Y, Zhu X Y, Anlu.Extraction Text Summary Model Based on Maximum Border Correlation [J].Intelligence Science, 2021,39 (02): 34-43.

[14] Liu Z, Yu B, Yu Y, Yang X H, et al.Subject-based SE-TextRank emotion summary method [J].Intelligence Engineering, 2017,3 (03): 97-104.

[15] ASHARI A , RIASETIAWAN M .Document Summarization using TextRank and Semantic Network[J]. International Journal of Intelligent Systems & Applications, 2017, 9(11):26-33.

[16] GNH A , RS A , ACIA A , et al. Extractive Hotel Review Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends - ScienceDirect[J]. Procedia Computer Science, 2021, 179:558-565.

[17] ZHANG B F, LI C C. Research on automatic summarization of multiple documents based on the combination of LDA and TextRank [J]. Software Guide, 2018,17(04):13-15+18.Zhang Jin. Intelligence Keyword extraction method based on the improved TF-IDF algorithm [J].Intelligence Journal, 2014,33 (04): 153-155.

[18] WANG Y , XU W . Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud[J]. Decision Support Systems, 2018, 105(jan.):87-95.

[19] CHEN D H, WANG Y N, ZHOU Z L, et al. Research on Word Similarity Calculation Based on Word2Vec [J]. Computer Engineering and Applications, 2022, 58(3):8.

[20] BARZ B , RODNER E , GARCIA Y G , et al. Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,41(05):1088-1101.

[21] JI S , ZHANG Z , YING S , et al. Kullback-Leibler Divergence Metric Learning[J]. IEEE Transactions on Cybernetics, 2022, 52(04):2047-2058.

[22] Tang D. Research and Implementation of Multi-Characteristic Chinese Automatic Digest Based on LDA and Redundancy Control [D].Yunnan Normal University, 2021.

[23] CHENG K, LI C Y, JIA X X, et al. News text extraction summary method based on improved MMR algorithm [J]. Journal of Applied Science, 2021, 39(03): 443-455.

[24] Sogou Lab. Sogou News Data [EB/OL]. http://www.sogou.com/labs/resource/list_news.php. 2012-8-16.

[25] Lin C Y.Rouge: A Package for Automatic Evaluation of Summaries // Proc of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL. Barcelona, Spain, 2004: 74-81.

[26] Wang X, Han B, Gao R, et al. Automatic extraction of text summaries based on improved TextRank [J]. Computer Applications and Software, 2021, 38(06): 155-160.