Abstract of News Text Based on Theme and Emotional Relationship

Hui Huang, Qing-tao Zeng

Beijing Institute of Graphic Communication, Beijing, China

Abstract

In the current rapidly developing society, people can easily and quickly get all kinds of news information from the Internet, such as browsing micro-blogs. Which contains a large amount of news text with emotion, the positive or negative emotional news subtly changing people's attitude to current events, affect social development situation of the news public opinion, the existing text based methods tend to consider factors such as the theme and sentence features, unable to get the text with emotional opinions in this paper. Therefore, this paper proposes a sentiment summarization method of Chinese news text which integrates sentence sentiment and topic similarity. In the in-depth analysis of TextRank graph model algorithm, the relationship between nodes in the model and the calculation of edge weight are improved through the integration of emotion information, and the sentence emotion weight with emotion is obtained. Then, by synthesizing the sentence emotion and topic relevance and other factors, the weight parameter is used to balance, and the sentence weight which integrates the sentence emotion and topic similarity is obtained. Finally, the emotion summary of news text is obtained.

Keywords

Abstract of text; Sentence emotion; TextRank;

1. Introduction

Text summarization is one of the Natural Language Processing (NLP) tasks, which aims to process text to generate concise and refined content, namely summary information. Text summary has the basic characteristics of simplicity, accuracy, clarity and so on. Text sentiment analysis is another important branch of NLP, which analyzes and processes subjective sentences containing emotional opinions in texts, inferences and induces information such as emotion, emotion and opinion in sentences [1]. Since Bo Pang [2] put forward emotion analysis in 2002, it has attracted extensive attention from scholars at home and abroad. With the rapid development of social media, micro blog emotion analysis has become the current research hotspot.Compared with the traditional text summarization, it is necessary to consider the emotional factors of the sentences in the text. The emotion summarization generated by it enables people to conveniently obtain the main content of the text and understand the views and attitudes of the public.

2. Related work

2.1 LDA model

The basic idea of LDA model is to describe documents as topic probability distribution and further describe topics as lexical item probability distribution [3]. The LDA model is a 3-layer Bayesian structure, and its LDA graph model is shown in Figure 1. Among them, the gray circle represents the observed

AHPCAI2022@2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence
EMAIL: e-mail: hh61650024@163.com (Huang Hui), e-mail: 954276545@qq.com (Zeng Qing-tao)

variable, the white circle represents the hidden variable, the arrow represents the association between the variables.

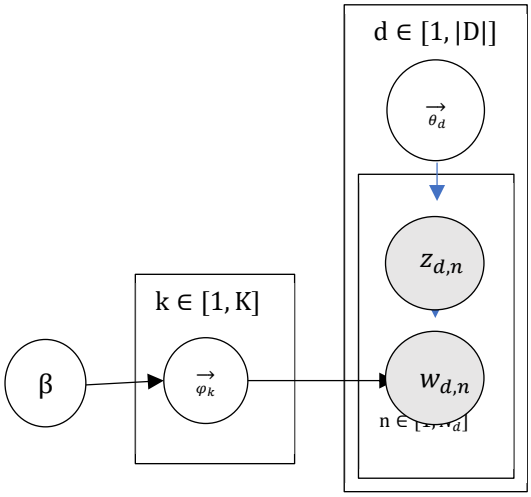

Figure 1. LDA graph model

The meanings of each symbol are shown in Table 1.

TABLE 1. SYMBOLIC MEANING DIAGRAM OF LDA MODEL

symbol	meaning
β	Hyperparameters of term distribution
α	The hyperparameters of the topic distribution
$\vec{\varphi}_k$	Word distribution under topic k
$z_{d,n}$	The first n subject word in document d
$\vec{\theta}_d$	The theme distribution of document d
$w_{d,n}$	The first n word item in document d
|D|	Total number of documents in the corpus
K	Number of topics
V	Number of words
N_d	The total number of words in the d document

In the above generation process, the joint probability of all observed variables and implied variables under a given hyperparameter is

$$p(w_i, z_i, \theta_i, \Phi | \alpha, \beta) = \prod_{i=1}^{N_i} p(w_{i,j}|\varphi_{z_{i,j}})$$
$$p(z_{i,j}|\theta_i).p(\theta_i|\alpha).p(\Phi|\beta) \quad (1)$$

Probability of obtaining document i by integrating θ_i, $z_{i,j}$ and Φ

$$p(w_i|\alpha,\beta) = \iint p(\theta_i|\alpha).p(\Phi|\beta).$$
$$\prod_{i=1}^{N_i}\sum_{Z_{i,j}} p\left(w_{i,j}\middle|\varphi_{z_{i,j}}\right) p(z_{i,j}|\theta_i) d\Phi d\theta_i =$$
$$\iint p(\theta_i|\alpha).p(\Phi|\beta).\prod_{i=1}^{N_i} p(w_{i,j}|\theta_i,\Phi) d\Phi d\theta_i \quad (2)$$

The whole corpus $D=\{w_i\}_{d=1}^{|D|}$ generate probability:

$$p(D|\alpha,\beta) = \prod_{i=1}^{|D|} p(w_i|\alpha,\beta) \quad (3)$$

2.2 Similarity calculation

Word similarity calculation has been widely used in various fields of automatic question answering system, word semantic disambiguation, information retrieval and so on [4]. The essence of Chinese sentence similarity is sentence semantic similarity, which involves semantic, syntactic, lexical and other factors.

In this paper, the semantic similarity calculation method of CNKI is used to calculate the semantic distance between words. The semenes in "Knowledge Network" have a hierarchical relationship. In this paper, we use the upper and lower relationship between the basic semenes and the basic semenes to calculate the similarity between semenes. The following formula is used to calculate the semantic similarity between Chinese words:

$$\mathrm{sim}(p_1, p_2) = \frac{a}{a+d} \tag{4}$$

Where, p_1 and p_2 are two basic semenes, a is a variable parameter, and d is the path length between semenes in the hierarchy. The process of calculating the semantic similarity of two words is the process of calculating the maximum sememe similarity between words:

$$wsim(w_1, w_2) = \max_{i=1\ldots2, j=1\ldots m} sim(m_{1i}, m_{2j}) \tag{5}$$

For the sentence $s_a = \{w_{a1}, w_{a2}, \ldots, w_{an}\}$ and $s_b = \{w_{b1}, w_{b2}, \ldots, w_{bn}\}$ to carry out pairwise similarity calculation for each word in sentence s_a and sentence s_b, and take the maximum value of each word in the calculation process as the similarity weight of this word. By summing and standardizing the lexical similarity between sentences, the sentence semantic similarity formula is finally obtained:

$$S(s_a, s_b) = \frac{\sum_{k=1\ldots n, l=1\ldots m} \max wsim(w_{ak}, w_{bl})}{log(|s_a|) + log(|s_b|)} \tag{6}$$

3. Emotional Summary

Sentiment summarization of news text is based on the common representation of text summarization and sentiment analysis. Abstract is the expression of the core content of the text, and the theme is the basis to represent the central idea of the text. Topic-based Chinese news text classification relies on the relationship between text sentences and topics. The text content can be classified effectively through the topic extraction of text content and the calculation of sentence topic orientation.

3.1 Sentiment summary process

The topic sentence association is divided into two parts: first, the topic model is used to extract the topic of the text content; The second is to calculate the similarity between the sentences in the text and each topic as the grouping basis. Firstly, the semantic similarity calculation method of CNKI is used to calculate the similarity between text sentences, and then the sentiment dictionary is constructed to extract the sentiment words in text sentences and identify the sentiment polarity of the sentiment words. This paper uses part-of-speech tagging and dependency parsing to extract content words and topic features from sentences in the text, and constructs an emotional sememe with the binary structure of "central word, affective word polarity" as the basic unit of sentence emotional semantics. The emotional similarity of sentences is calculated according to the co-occurrence rate of emotional sememes in sentences. Finally, a two-layer affective semantic graph model is constructed, which takes sentences as nodes, emotional similarity value and semantic similarity value as path weights between sentences, respectively. The affective semantic weights of the final sentences are obtained by graph calculation. By sorting the emotional semantic weight set of the obtained text sentences, the sentences with larger emotional semantic weight in the text are extracted proportionally as the final text sentiment summary result.

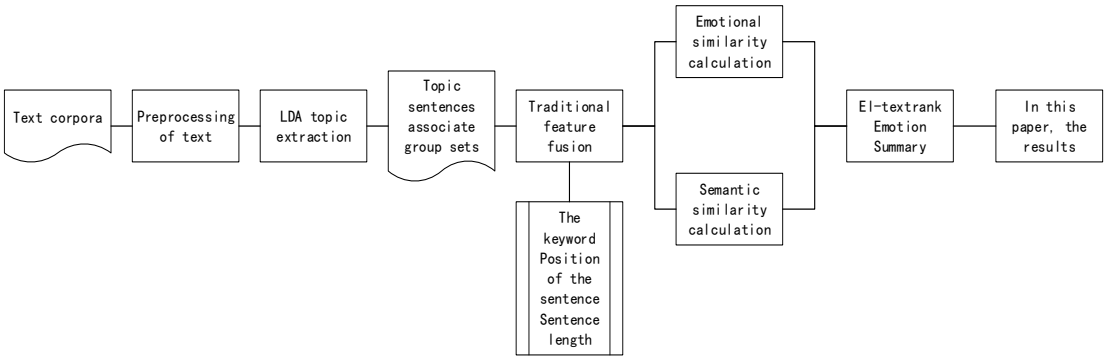

Figure 2. The flow char

3.2 Improved TextRank algorithm with semantic emotion

In an article, the sentences with rich emotion sometimes can not express the central idea of the text well, and the sentences with a high degree of agreement with the central idea of the text sometimes lack appropriate expression of emotion. Therefore, we need to take into account both the semantic importance and the emotional importance of sentences.

In this paper, the semantic similarity and emotional semantic similarity of sentences in the text are graphically calculated, and the two are statistically calculated

Finally, the emotion ranking result set of sentences is obtained. Figure 3 shows the El-Textrank graph model integrating semantic affective relationships.

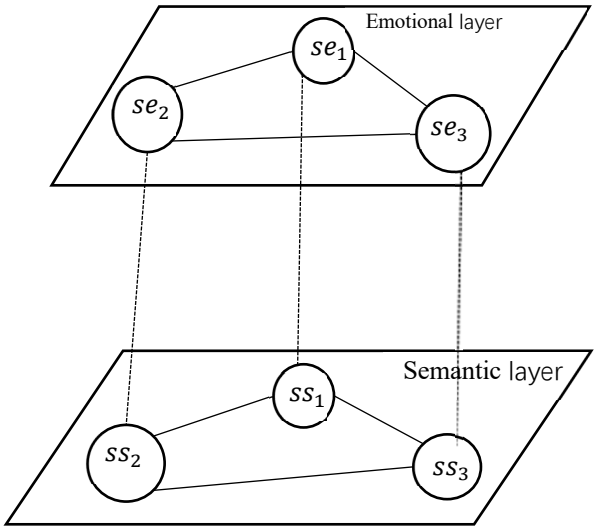

Figure 3. A graph model integrating semantic emotional relationships

The upper layer is the emotion model: $G' =< v, e, w >$, v is the set of sentences， e is the emotion relationship between sentences， w is the corresponding emotion similarity weight, se is the emotion weight of sentence node. The lower layer is the semantic layer: $G =< v, s, m >$, where s is the semantic relationship between sentences, m is the corresponding semantic similarity weight, and ss is the semantic weight of sentence nodes. The sentiment layer is associated with the semantic layer to calculate the sentiment ranking of text sentences.

In this paper, the sentence similarity calculation based on statistics in TextRank algorithm is replaced by the sentence similarity calculation method based on semantics, so as to improve the sentence similarity calculation effect in the field of news text. The method of obtaining sentence similarity weights

by word co-occurrence in text graph nodes is converted to the method of calculating sentence semantic similarity weights, and then the improved EL-Textrank formula is obtained:

$$Q(v_i) = (1-d) + d \cdot \sum_{v_j \in In(v_i)} \left(\frac{\omega \cdot E_{ij}}{\sum_{v_k \epsilon Out(v_j)} E_{jk}} + \frac{\lambda . L_{ij}}{\sum_{v_k \epsilon Out(v_j)} L_{jk}} \right) \cdot Q(v_j) \quad (7)$$

Where, d is the normalization factor (generally 0.85), ω and λ are the proportional parameters that bisect the semantic weight and the emotional weight, respectively. E is the semantic similarity weight of sentence, and L is the emotional similarity weight of sentence.

We use the El-Textrank algorithm to rank the sentence semantic emotional importance of the sentences in the topic sentence association group, and determine the number of important sentences extracted from each group according to the number of sentences in each group, and finally obtain the emotion summary of the text.

4. The experiment

4.1 The experimental data

The experimental corpus used in this paper is 250 news texts from NLPCC 2018 Weibo guided Chinese News Digest evaluation task. It includes topics such as society, politics, economy, culture, education and life, and also includes two kinds of standard abstracts to evaluate the system abstracts. According to the requirements of the evaluation, the artificial emotion summary is processed, and the consistency of the preprocessing results is checked, and the final text corpus is used as the evaluation data.

4.2 Evaluation standard

Chin-yew Lin[5] et al., inspired by BLEU's automatic evaluation method, developed an automatic ROUGE evaluation method applied in the field of automatic text summarization. The evaluation tool used in this paper is Rouge-1.5.5 provided by the government. This method evaluates the automatic digest by counting the total number of reproduction units between the automatic digest and the standard digest. Rouge-1, Rouge-2, Rouge-3, Rouge-W and ROUGE-SU* are the important indexes for evaluating the quality of digest. Rouge-n (based on N-gram co-occurrence statistics) is the recall rate of N-gram grammar, Rouge-L and Rouge-w are the longest common subsequence and the longest weighted common subsequence, respectively. The formula of Rouge-N is as follows：

$$\text{ROUGE} - \text{N} = \frac{\sum_{S \in Sum_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Sum_{ref}} \sum_{N-gram \in S} Count(N-gram)} \quad (8)$$

Where N stands for the length of N-gram, $Count_{match}(N-gram)$ is the total number of $N-gram$reproduced by automatic and standard abstracts under test, and $Count(N-gram)$ is the total number of $N-gram$ in standard abstracts.

Among the evaluation indexes, Rouge-1 reflects the ability of the abstract generated by the automatic summarization system to summarize the main content of the original document, that is, the extent of covering the important information of the original document. Rouge-2, Rouge-3, Rouge-4, Rouge-W reflect the readability and coherence of automatic summarization. Because the evaluation tool ROUGE is mainly used to measure the ability to contain the content of the standard summary by the recurrence number of the system summary and the standard summary, the Rouge-N method is the most commonly used and the most important evaluation index, and the Rouge-1 and Rouge-2 can best reflect that the system summary is close to the standard summary. Therefore, Rouge-1 and Rouge-2 are used in this paper to evaluate the content integrity, semantic readability and coherence of emotional summaries of news texts. Rouge-1 [6] is recognized as the evaluation parameter that can best reflect the system summary close to the standard summary.

4.3 The experimental setup

1) *Topic Extraction experiment*: In the topic extraction experiment, we set the proportion of sentences occupied by topics to obtain the number of topics extracted under the optimal text summarization result. We used 10% topic proportion interval as the experimental comparison distance, and ROUGE-SU* was used as the standard to evaluate the impact of the proportion of topic number in the text sentence set on the text summary. The change of ROUGE-SU* value under different topic number ratio is shown in Figure 4.

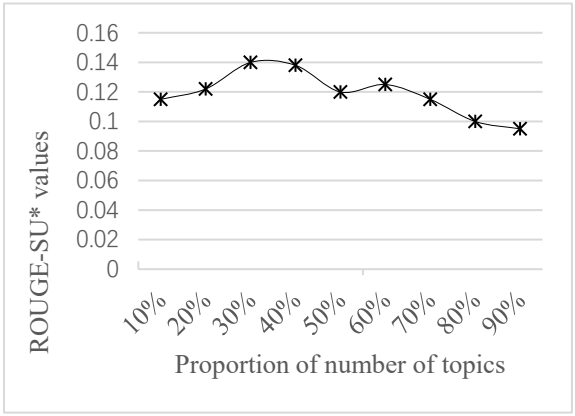

Figure 4. ROUGE-SU* values for different theme proportions

Through the analysis of the experimental results of topic extraction, the ROUGE-SU* value is the best when the number of topics accounts for 30% of the text sentences.

2) Keywords extraction experiment: In the process of the experiment, we conducted 8 groups of experiments on the test dataset, respectively selecting 5, 10, 15, 20, 25, 30, 35 and 40 keywords, and conducting text summarization experiments on different number of selected keywords. The final experimental results are shown in Figure 5.

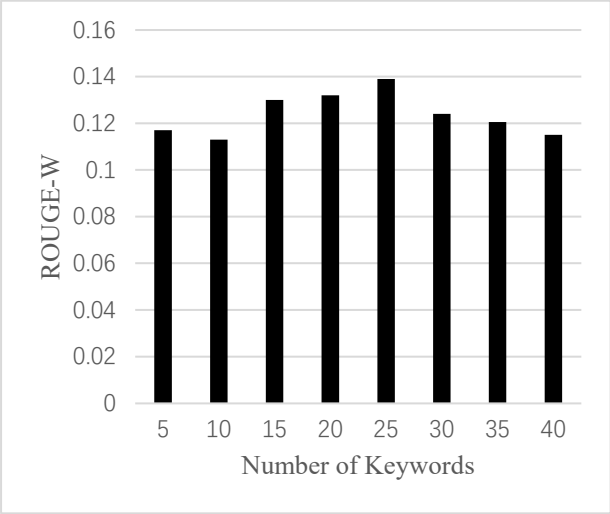

Figure 5. ROUGE-W value with different number of keywords

By analyzing the experimental results of the sample dataset, we find that the Rouge-W value of the text abstract is optimal when the number of keywords is 25 in the microblog oriented news articles.

3) Emotion fusion experiment: EL-textrank is used to train the similarity between emotion sememe and semantic, and the weighted parameters ω and λ of the optimal semantic feature value and the optimal emotion feature value are 0.6 and 0.4, respectively (ω+λ=1.0). The experimental process is shown in Table 2:

TABLE 2. COMPARISON OF SENTIMENT SUMMARY EVALUATION UNDER DIFFERENT EIGENVALUE PARAMETERS

ω	λ	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-N
0.1	0.9	0.29541	0.08121	0.01923	0.10149
0.2	0.8	0.32742	0.09382	0.02342	0.12415
0.3	0.7	0.34957	0.10207	0.03695	0.13342
0.4	0.6	0.38131	0.11828	0.05732	0.16592
0.5	0.5	0.41962	0.12031	0.05864	0.17139
0.6	0.4	0.42109	0.12765	0.06121	0.17781
0.7	0.3	0.41767	0.11471	0.05398	0.16485
0.8	0.2	0.40807	0.11056	0.04966	0.16037
0.9	0.1	0.36758	0.10466	0.03857	0.15854

4) Contrast experiment: We compare the traditional multi-feature text summarization methods: the multi-feature summarization method using the traditional TF-IDF method and the summarization method using PageRank algorithm with the topic-based El-Textrank sentiment summarization method in this paper. The comparison results are shown in Table 3:

TABLE 3. COMPARISON OF SENTIMENT SUMMARY EVALUATION UNDER DIFFERENT EIGENVALUE PARAMETERS

Method	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-N
TF-IDF	0.37541	0.10211	0.03522	0.13497
PageRank	0.40742	0.12458	0.05654	0.16415
EL-TextRank	0.42109	0.12765	0.06121	0.17781

The comparison of experimental evaluation results shows that the indexes of a topic-based El-Textrank sentiment summarization method adopted in this paper are close to the best values, which indicates that it is feasible to obtain sentences similarity features by grouping topic sentences with the new algorithm and combining the traditional multi-feature fusion to obtain summaries.

5. Conclusion

In this paper, a theme-based EL-Textrank sentiment summarization method is adopted, which fully considers the influence of multi-topic convergence in news text on the abstract content. Meanwhile, on the basis of the similarity between sentences and topics, semantic features and emotional features are fused under the grouping of topic sentences, which makes the summarization result more stable and efficient. The operability of the method is also relatively strong.

6. References

[1] Cardie C . Sentiment Analysis and Opinion Mining[J]. Synthesis Lectures on Human Language Technologies, 2014, 30(1):152-153.

[2] Chen Y , Pang B , Shao G , et al. Erratum to "DGA-Based Botnet Detection Toward Imbalanced Multiclass Learning" by Yijing Chen, Bo Pang, Guolin Shao, Guozhu Wen, and Xingshu Chen[J]. Tsinghua Science and Technology, 2021, 26(5):790-790.

[3] Figueroa G. Chen P C. Chen Y S.RankUp: Enhancing Graph-based Keyphrase Extraction Methods with Error-feedback Propagation[J]. Computer Speech & Language,2018,47:112-131.DOI:/10.1016/j.csl.2017.07.004.

[4] JI Wen-Qian Ji, LI Zhou-Jun Li, CHAO Wen-Han Chao, et al. Sentence similarity calculation and its application in automatic abstract system[J]. Intelligent Information Management, 2009, 1(1):38-45.

[5] Masuda K , Yukawa E , Nitta N , et al. The Evaluation of the Visual Function in Glaucoma Using the Newly Developed Letter Charts Under Fundus-Related[J]. software engineering ieee transactions on, 2009, 23(12):777-799.
[6] Lin C Y , Hovy E . Automatic evaluation of summaries using N-gram co-occurrence statistics[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003.