Unsupervised Keyword Extraction Algorithm Based on Bert Model

Shouhao Zhang
School of International Economics and Management,
Beijing Technology and Business University, Beijing 100048, China

Keywords: Keyword Extraction, BERT Model, Unsupervised.

Abstract: Aiming at the problem that traditional word segmentation and unsupervised text keyword extraction methods ignore the context semantics, and the effect of candidate keyword extraction is limited, this paper proposes an algorithm based on the Bidirectional Encoder Representation from Transformers (BERT) model. In the word segmentation stage, the BERT model is used to segment the text to obtain candidate keywords, and then the text is input into BERT to extract the vector of candidate keywords. The word vector extracted in this paper is to reconstruct the hidden layer. According to the last four layers of the neural network sum, and average the word vectors obtained, then obtain the word vectors of candidate keywords, and finally score the similarity with sentence vectors combined with context semantics to obtain keyword ranking.

1 INTRODUCTION

The keyword extraction task can automatically summarize and extract the phrases of the core content of the text, and the phrases express the most critical information in the whole text, also known as keywords. Keyword extraction methods are mainly divided into unsupervised extraction methods and supervised extraction methods, each of which has its own advantages and disadvantages (Sterckx et al., 2016). The supervised keyword extraction task needs to manually label a large amount of corpus information, and then use a neural network to transform the keyword extraction task into a two-class task. The difficulty of supervised extraction method is that labeling corpus information takes a lot of resources, and there is a lack of large-scale keyword extraction data sets, so it is difficult to obtain corpus (Alzaidy et al., 2019). Unsupervised keyword extraction is more widely used in practical applications.

Extraction methods mainly include two aspects: firstly, obtaining candidate keywords, and then sorting the candidate keywords to select keywords (Wei et al., 2016). The disadvantage of unsupervised keyword extraction is that the difference between candidate keywords and text sequence length leads to the mismatch of representation, which affects the keyword extraction effect in long texts, and cannot make full use of the contextual semantic relevance of the pre-training model.

There are three subjects of unsupervised keyword extraction, and keyword extraction methods based on topic the model, statistical features and the graph model. Unsupervised algorithms based on the topic model (Campos et al., 2018) mainly include the LDA topic model, and on the basis of LDA, there is an algorithm model of LDA and TF-ID fusion. Topic models based on statistical features mainly include YAKE (Boudin, 2013), TF-IDF and improved algorithm with TF-IDF (Bordolin et al., 2020) as the core. In addition, in order to make the TF-IDF method suitable for corpus of different lengths Florescu and Caragea (Florescu and Caragea, 2017). Proposed to use the arithmetic mean of words instead of the logarithmic calculation method of IDF, and its effect was better than the traditional TF-IDF method. The keyword extraction method based on graph model mainly takes TextRank (Liu et al., 2009) as the main method, using the Word2Vec model to vectorize candidate words (Gagliard and Artese, 2020), or using FastText to train candidate word vectors, and using Sent2Vec and Doc2Vec models to obtain the semantic vectors of candidate words and full text.

These three unsupervised methods all have their own main problems in keyword extraction. TF-IDF algorithm can't reflect the position information of words, and the words with the front position and the words with the back position are regarded as the

Zhang, S.
Unsupervised Keyword Extraction Algorithm Based on Bert Model.
DOI: 10.5220/0011923800003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 306-309
ISBN: 978-989-758-622-4; ISSN: 2975-9463

same importance. The simple structure of TF-IDF can't effectively reflect the importance of words and the distribution of characteristic words, so it can't adjust the weights well. Although TextRank takes into account the relationship between words, it still tends to use frequent words as keywords. Compared with these models, BERT model has better effects, which can achieve better results according to the semantics of context, and can achieve better results according to the semantics of context.

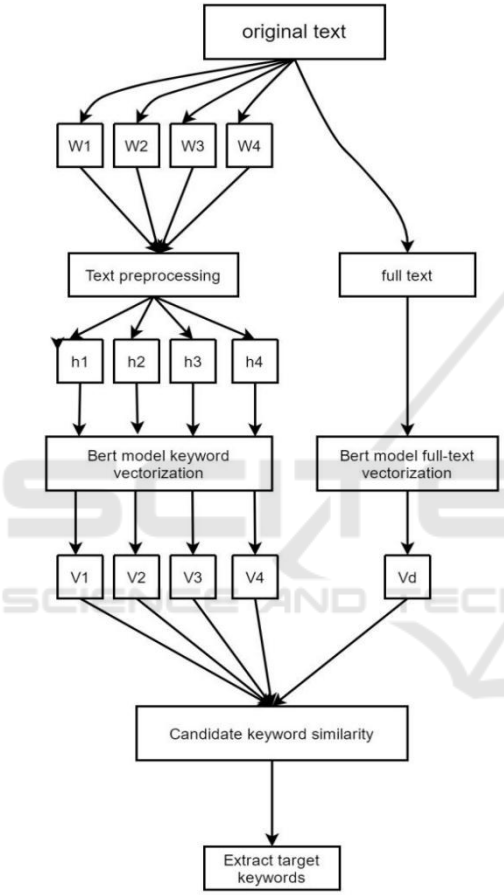

Figure 1: Schematic diagram of keyword extraction model.

2 MODEL DESIGN

The keyword extraction model proposed in this paper is shown in Figure 1, Where: Wi means the crown word separated from the original text by BERT toolkit; i=1, 2, 3, ..., indicating the position of each word; hi represents the candidate words generated after preprocessing Wi; Vi, a word vector representing candidate words; V_d full-text semantic vector; m is the total number of candidate words after preprocessing. The keyword extraction steps are as follows: firstly, the original text is cut into word sequences {Wi} of length n, and some words are filtered out by text preprocessing to obtain m candidate words h_i; Then, the candidate words are input into the BERT model to vectorize the words to obtain the corresponding word vector V_i; Finally, the word vectors of words are averaged to get the corresponding word vectors. At the same time, the full-text semantic vector is obtained by the full-text vectorization method of BERT model. Finally, the similarity between the candidate keyword vectors and the full-text semantic vectors is calculated, and the target keywords are sorted and extracted to complete the whole extraction process. The advantage of this model is that the candidate word vectors generated by the pre-training, model are dynamic, and the full-text semantic information is combined to screen keywords, at the same time, the problem of word meaning repetition caused by using similarity in keyword selection task is solved.

2.1 Experimental Environment and Data

There is a lot of information in the text that affects the experimental results, which needs to be preprocessed. Firstly, the text that needs to be preprocessed is segmented, then punctuation marks and noise are removed, and then candidate keywords of the text are extracted. The method of extracting candidate keywords from the text uses the unsupervised text segmentation method of BERT model. Based on the relevance between the context semantics of BERT model, the text is segmented, and Euclidean distance is used to calculate. After segmentation, candidate keywords are obtained.

The main process is to use BERT unsupervised word segmentation to divide the original text into word sequences: $\{W_1, W_2, \cdots, W_n\}$, the length of which is n, and then remove the stop words, punctuation and noise to get M candidate keywords: $\{h_1, h_2, \cdots, h_m\}$. Then, the original text sequence D is divided into T word sequences: $\{c_1, c_2, \cdots, c_T\}$, and the candidate words are set in the interval [a,b], so that the word sequence is $\{c_a, c_{a+1}, \cdots, c_b\}$. Then, according to the BERT model, the original text is processed to obtain the word vector. The process is to input the single word sequence obtained by BERT into the BERT model to obtain the hidden layer vector. Finally, the last four layers of the hidden layer vector are summed to obtain the vector of each

single word. The process of obtaining the hidden layer vector can be represented by (1):

$$\{d_1,d_2,...,d_T\}=\text{Bert\_encoder}(\{c_1,c_2,\ldots,c_T\}) \quad (1)$$

After the hidden layer vector is obtained, the hidden layer vectors of the last four layers are obtained respectively, and finally the sum is obtained. Get the word vector at the corresponding position of the word sequence, get the word vector sequence $\{d_a,d_{a+1},\ldots,d_b\}$ corresponding to the candidate words, and calculate the word vector V, the formula of which is:

$$v=\text{mean}(d_a,d_{a+1},\ldots,d_b) \quad (2)$$

The formula (2) is to calculate the average vector of the vector sequence to represent the composed word vectors, and to ensure the dimensional unity of word vectors with different lengths.

2.2 Full-Text Semantic Vector

First, add the "[CLS]" logo at the beginning of the original text. The purpose of adding this logo is to use this logo to obtain the full-text semantic vector. Then input the original text D into the BERT word breaker, that is:

$$\{c_{CLS},c_1,c_2\ldots,c_T\}=\text{Tokenizer(D)} \quad (3)$$

Then, the obtained single word sequence is input into the hidden layer of BERT model, and the hidden layer used is consistent with the hidden layer of the word vector acquisition task, namely:

$$\{d_{CLS},d_1,\ldots,d_T\}= \text{Bert\_encoder}(\{c_{CLS},c_1,c_2\ldots,c_T\}) \quad (4)$$

2.3 Similarity Calculation and Ranking of Candidate Words

Because they are all vectors output through the same hidden layer, this group of vectors is located in the same two-dimensional space. We can use cosine similarity and Euclidean distance and other distance algorithms to calculate the semantic distance and judge the similarity between them. In this paper, cosine similarity is selected to calculate the similarity between candidate keywords and full-text semantic vectors.

3 EXPERIMENT AND RESULT ANALYSIS

3.1 Experimental Environment and Model

This paper selects the laboratory data, including the business scope and content of logistics listed companies. The keywords are segmented by BERT model, then preprocessed, stop words are removed, to obtain candidate keywords, and then the keywords are extracted by Bert model according to the context semantic information.

3.1.1 Experimental Environment

The experimental computer configuration and environment are as follows: the CPU is Intel(R) Core(TM) i7-7700HQ CPU, the main frequency is 2.8GiHz, the memory is 4 GiB, the GPU is NVIDIA GeForce GTX 1050Ti, the system is Windows64-bit, python version number is 3.6, and Python version number is 1.7.0.

3.1.2 Pre-Training Model

The pre-training model adopts the BERT-base-Chinese model published by Google, which contains 12 layers of Transformer, and the vector of the fourth hidden layer from the bottom is used as the word vector.

4 DATA ANALYSIS AND CONCLUSIONS

Through keyword extraction of three logistics themes of China Railway Special Cargo Flow Co., Ltd., the corresponding keywords of each theme are obtained. As shown in Figure 1, the top three keyword phrases are extracted, and the top scores are 0.5737, 0.5395 and 0.6565, respectively. According to the length of the subject keyword, it is found that the subject keyword is not a word, but a phrase. By setting the length of different extracted keywords, the keyword phrases are extracted. Through the experimental comparison, Figure 2 shows that when the keyword phrase length is set to 2, the score is the highest, and the first scores are 0.6921, 0.7151 and 0.7797 respectively.

Through keyword extraction of three logistics themes of China Railway Special Cargo Flow Co., Ltd., the corresponding keywords of each theme are

obtained. As shown in Figure 1, the top three keyword phrases are extracted, and the top scores are 0.5737, 0.5395 and 0.6565, respectively. According to the length of the subject keyword, it is found that the subject keyword is not a word, but a phrase. By setting the length of different extracted keywords, the keyword phrases are extracted. Through the experimental comparison, Figure 2 shows that when the keyword phrase length is set to 2, the score is the highest, and the first scores are 0.6921, 0.7151 and 0.7797 respectively.

Table 1: Extraction result with keyword set to 1

introduction to logistics keywords	Keyword extraction
professional logistics service supply	(logistics, 0.5737) (railway transportation, 0.5546) (service provider, 0.549)
network advantage	(logistics, 0.5395) Railway, 0.5367) (transport, 0.4916)
modern logistics system	(logistics, 0.6565) (railway transportation, 0.6275) (railway, 0.5503)

Table 2: Extraction result with keyword set to 2

introduction to logistics keywords	Keyword extraction
professional logistics service supply	(logistics service provider, 0.6921) (enterprise logistics, 0.6807) (Railway special goods, 0.651)
Network advantage	(logistics network, 0.7151) (railway resources, 0.7147) (railway logistics, 0.7066)
modern logistics system	(logistics system, 0.7797) (logistics mode, 0.7319) (railway logistics, 0.73)

According to the keywords and contents, we can see that some key words are artificially summarized and not completely included in the contents. Therefore, if we process the data of the keywords and remove the manual summary, we can see that the experimental results are obviously better.

The unsupervised keyword extraction method used in this experiment uses single text data, which is limited. In the next step, we will learn to use supervised method to extract keywords.

REFERENCES

STERCKX L, CARAGEA C, Demeester T, et al. (2016) . Supervised keyphrase extraction as positive unlabeled learning[C]//Proceedings of the 2016 Conference on Empirical Methodsin Natural Language Processing, Austin, Nov 1- 4..Stroudsburg: ACL, 2016: 1924-1929.

Alzaidy R, Caragea C, Giles C L. (2019). Bi-LSTM-CRF sequence labeling for keyphrase extraction froms scholarly scholarly documents [C] //The World Wide Web Conferenceon-WWW'19.New York: ACM Press, 2019: 2551-2557.

WEI H X, GAO G L, SU X D. (2016). LDA-Based Word Image Representation for Keyword Spotting on Historical Mongolian Documents[C]//Neural Information Processing(ICONIP).Springer, 2016: 432-441.

Campos R, Vitor M, Pasouali A, et al. (2018). YAKE! Collection In dependent Automatic Keyword Extractor[C]//In Advances in Information Retrieval-40th European Conference on Information Retrieval. Springer ECIR 2018, Lecture Notes in Computer Science, Grenoble, France. Cham, 2018: 806-810.

Boudin F. (2013). A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction. [C]//Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya: Asian Federation of Natural Language Processing, 2013: 834-838.

Bordolin M, Chatterjee P C, Biswas S K, et al. (2020). Keyword extraction using supervised cumulative TextRank [J] . Multimedia Tools and Applications, 79 (41 / 42):31467-31496.

Florescu C, Caragea C. (2017). A New Scheme for Scoring Phrases in Unsupervised Keyphrase Extraction[C]//Proceedings of the Advances in Information Retrieval-39thEuropean Conference on Information Retrieval.ECIR 2017, Lecture Notes in Computer Science Aberdeen, UK.

Liu Z Y, Li P, Zheng Y B, et al. (2009). Clustering to find exemplar terms for keyphrase extraction[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Lan-guage Processing. Singapor: ACL Press: 257-266.

Gagliard I , Artese M T. (2020). Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering method [J].Multimoda Technologies and Interaction, 4 (2): 30.