Journal of Physics: Conference Series

**PAPER** • **OPEN ACCESS**

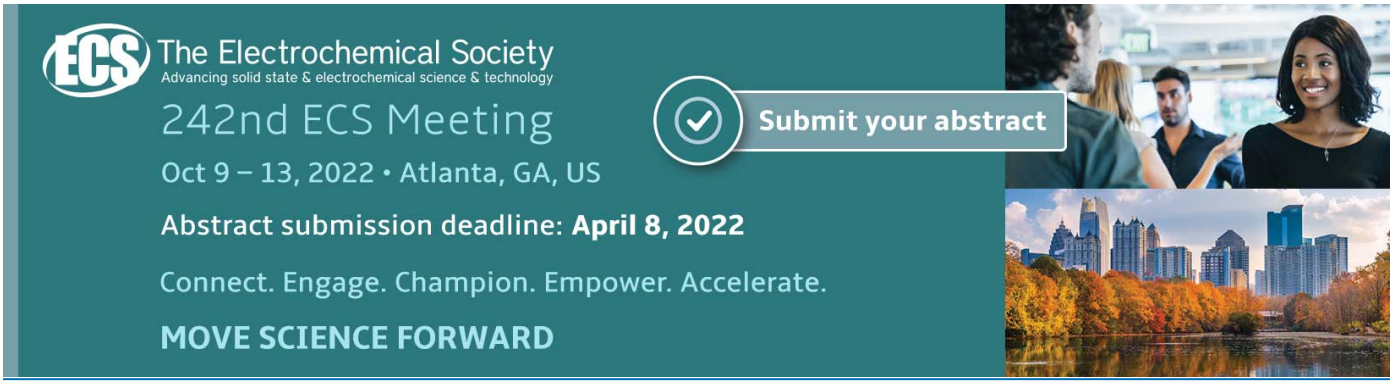
# Bert-Based Text Keyword Extraction

To cite this article: Yili Qian *et al* 2021 *J. Phys.: Conf. Ser.* **1992** 042077

View the article online for updates and enhancements.

You may also like

- 4D particle therapy PET simulation for moving targets irradiated with scanned ion beams
K Laube, S Menkel, C Bert et al.
- Chinese text multi-classification based on Sentences Order Prediction improved Bert model
Guanping Fu and Jianwei Sun
- Keyword extraction method for machine reading comprehension based on natural language processing
Ruiheng Li, Xuan Zhang, Chengdong Li et al.




This content was downloaded from IP address 103.57.85.28 on 18/03/2022 at 16:18

# Bert-Based Text Keyword Extraction

**Yili Qian, Chaochao Jia and Yimei Liu**

Shanxi University, Taiyuan 030006, China

*Corresponding author e-mail: qianyili@sxu.edu.cn

**Abstract.** With the explosive growth of network information, in order to obtain the information faster and more accurately, this paper proposes a text keyword extraction method based on Bert. Firstly, the key sentence set is extracted from the background material by Bert model as the information supplement to the text. Then, based on the extended text, TF-IDF, text rank and LDA are combined to extract keywords. The experimental results on real science and technology academic paper data sets show that the performance of the fusion multi type feature combination algorithm is better than that of the traditional single algorithm; and the F value of the algorithm is increased by 1.5% by extracting key sentences from background materials, which further improves the effect of key word extraction.

**Keywords:** Keyword Extraction, Feature Fusion, Bert

## 1. Introduction

In recent years, with the continuous progress of science and technology and the rapid development of computer technology, more and more information can be searched on the Internet, and our demand for information has been greatly met; but at the same time, the explosive growth of information also increases the difficulty of search. Facing the massive and long text, we cannot quickly and accurately obtain the key content [1].

Keywords are selected from the article and used to express the theme of the article. According to the keywords, we can quickly and accurately understand the central content of the long text, which brings great convenience for retrieval. Therefore, text keyword extraction has attracted more and more attention in recent years, and has been widely used in many natural language processing related fields, such as text similarity calculation [2], dialogue system generation [3], text topic mining [4], text classification [5].

To solve the problem of extracting keywords from the abstract text, this paper proposes a method of extracting key sentences from the main body of the paper by using the Bert model to supplement the information of the abstract, and using the combination algorithm of multi class features to extract keywords.

## 2. Related Work

At present, text keyword extraction algorithms can be divided into supervised method, semi supervised method and unsupervised method. Supervised keyword extraction algorithm regards keyword extraction as a binary classification problem to judge whether the words or phrases in the text

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

are keywords; semi supervised keyword extraction algorithm refers to the need for a small amount of training data, using these training data to build a keyword extraction model, and then using the constructed model to extract keywords from new text in the unsupervised method, the keywords are added to the training data, and then the model is retrained. The unsupervised method does not need to label the corpus, but directly uses the optional model to extract keywords. Although the accuracy of the former two methods is slightly higher than that of the unsupervised method, it needs a lot of manual annotation, which results in expensive labor cost, and the experimental process data is prone to over fitting, so it is not widely used. Unsupervised keyword extraction algorithm does not need training set. A large number of studies have improved the extraction method to improve the extraction accuracy, making it comparable to the supervised method.

Unsupervised keyword extraction algorithms can be divided into three categories: keyword extraction algorithm based on statistical features, keyword extraction algorithm based on word graph model and keyword extraction algorithm based on topic model. A large number of studies show that using a single method for keyword extraction is poor. People either improve the single algorithm or use combination methods to improve the accuracy of text keyword extraction. Xu Li [6-9] and others proposed an OPW text rank algorithm based on text rank. The accuracy rate of keyword extraction by this method is higher than that of traditional text rank algorithm. However, because the data set used by this algorithm is academic papers, and the papers are generally too long, the overall accuracy rate of extraction is not very high. On the basis of text rank, Fang Junwei [10] and others proposed a keyword extraction algorithm for academic texts based on prior knowledge text rank, which was evaluated on several literatures in the field of computer science, and the results were better than the traditional keyword extraction algorithm. Zeng Xi [11] and others proposed a short text keyword and extension based on topic model. In the keyword extraction work, the text topic classification information and word collocation relationship are introduced into the traditional TF-IDF algorithm, and the results are obviously better than the traditional LDA model for keyword extraction. Li Jiqi [12] and others improved the collaborative filtering algorithm by using TF-IDF to extract keywords, and the experimental results were significantly higher than those without keyword extraction. Chen Yiqun [13] and others transformed the keyword extraction algorithm into a classification problem in the patent data set. Based on the improved text rank algorithm to find the query words, SVM was used for patent keyword extraction. Xu Chen [14] and others studied an incremental patent semantic annotation based on keyword extraction algorithm. Saroj Kumar Biswas [15] and others proposed a weighted graph keyword extraction algorithm. The results of the algorithm on two public datasets are obviously better than the traditional algorithm. All of the above are related improvements on the three typical unsupervised keyword extraction algorithms mentioned before, and can be comparable with the supervised keyword extraction algorithm under specific conditions. However, the results obtained are not very ideal. Most of the data sets used in the study are relatively professional, and most of the keywords labeled are professional terms, which brings some difficulties for keyword extraction.

Therefore, this paper proposes a text keyword extraction algorithm based on Bert and multi class feature fusion. For scientific and technological academic papers, firstly, the text of the paper is taken as the background material, and the sentences with high semantic similarity to the abstract are extracted by using the Bert model to expand the abstract; then, the combination algorithm constructed by TF-IDF, text rank and LDA is used to extract keywords from abstract and key sentence sets based on multi category features [16-18].

## 3. Text Keyword Extraction Based On Bert and Multi Class Feature Fusion

### *3.1 Overall Framework of Keyword Extraction*

Abstract and text are included in scientific and technological academic papers. This paper takes the abstract as the main body and the main body as the background supplementary material. The unsupervised method based on Bert and multi class feature fusion is used to extract keywords. The overall framework is shown in Figure 1.

Figure 1 overall framework of keyword extraction Figure 1 overall framework of keyword extraction.

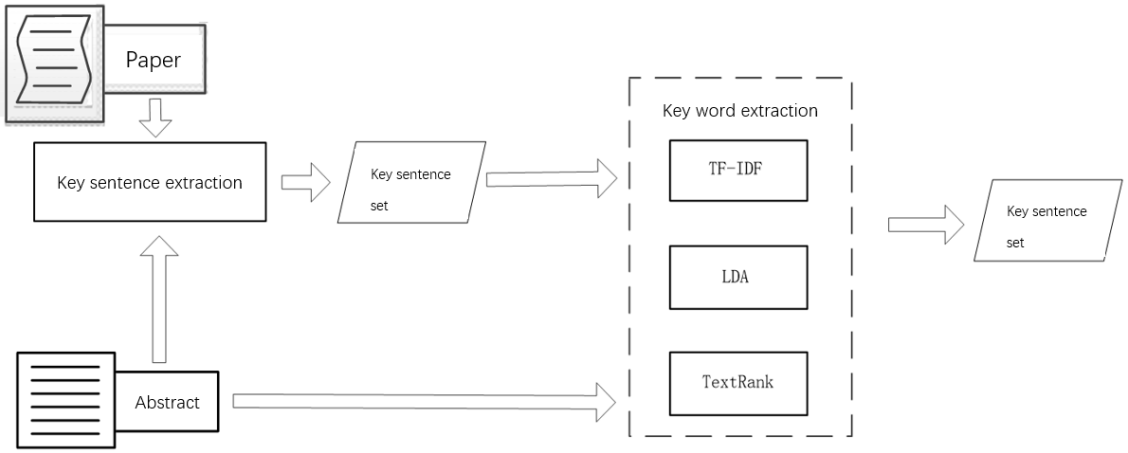



**Fig.1** Overall framework of keyword extraction

The input of the model is the abstract and body of the paper, and the output is the keyword set:

(1) Taking the text of the paper as the background material, the key sentences are extracted by using the Bert model, and N key sentences are extracted from each article to construct the key sentence set as the background supplement of the abstract;

(2) In this paper, TF-IDF, text rank and LDA combined algorithm are used to extract keywords based on summary and key sentence set. The first k words are selected as the extracted keywords in each algorithm, and then the final keyword set is obtained by using the idea of intersection.

*3.2 Key Sentence Extraction Based On Bert*

Abstract is short, but its content is limited, only based on the abstract keyword extraction, the effect is poor; the text of the paper is rich in information, but it is too long, so this paper extracts some key sentences from the text to expand and supplement the information of the abstract.

Because Bert has a good effect in training word vectors and capturing text semantics, this paper uses the Bert model to extract sentences with high semantic relevance from the text to construct the key sentence set. The specific process is as follows:

(1) The abstract and the main body of the article are divided into m and N sentences, respectively $S_{\text{abstract}}[s_1, s_2, s_3, \cdots\cdots s_m], T_{\text{passage}}[t_1, t_2, t_3, \cdots\cdots t_n,]$;

(2) For each $s_i$ sentence in $S_{\text{abstract}}$, the Bert model is used to find n sentences with the highest semantic similarity as candidate key sentences;

(3) The sentences in the candidate key sentence set are de duplicated and sorted, and the N sentences with the highest semantic similarity score are selected to form the final key sentence set.

*3.3 Text Keyword Extraction Based On Multi Class Feature Fusion*

Among the three algorithms used for keyword extraction in Figure 1, TF-IDF algorithm uses word frequency information, text rank algorithm uses graph related knowledge, and LDA algorithm uses text topic information. In this paper, the three methods are combined to extract text keywords. The specific processing flow is as follows:

(1) The first three keywords are extracted from the text by using the LDA, and then extracted by using the first three algorithms, respectively $K_1 = \{w_{11}, w_{12}, \ldots\ldots, w_{1k}\}$, $K_2 = \{w_{21}, w_{22}, \ldots\ldots, w_{2k}\}, K_3 = \{w_{31}, w_{32}, \ldots\ldots, w_{3k}\}$;

(2) Based on the union principle, the three groups of candidate keywords obtained in (1) are de duplicated and integrated as the final keyword set $K = K_1 \cup K_2 \cup K_3$.

## 4. Experiment and Analysis

### *4.1 Data set and Evaluation Index*

The data set used in the experiment is 300 scientific and technological papers downloaded from Wanfang database. There are 1520 keywords in total, and each paper has an average of 5.07 keywords.

The evaluation index selects the three most common indexes of information extraction, i.e. accuracy rate P, recall rate R and F value.

$$P = \frac{\text{extract}\,Word \cap \text{true}Word}{\text{extract}\,Word} \tag{1}$$

$$R = \frac{\text{true}Word \cap \text{extract}\,Word}{\text{true}Word} \tag{2}$$

$$F = \frac{2 * P * R}{P + R} \tag{3}$$

Among them, extract word is the number of keywords extracted by this algorithm, and true word is the number of keywords marked in the paper.

### *4.2 Parameter Selection*

Three classical algorithms are used to extract keywords from the abstract to determine the number of key sentences N and the number of keywords K. The influence of the number of key sentences and keywords on performance is shown in Figure 2 and figure 3.

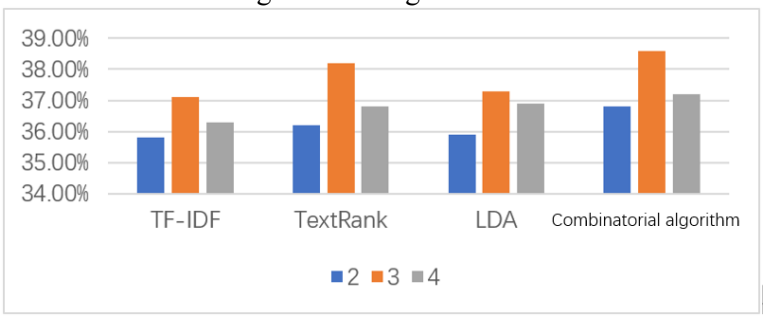



**Fig.2** Selection of the number of key sentences n

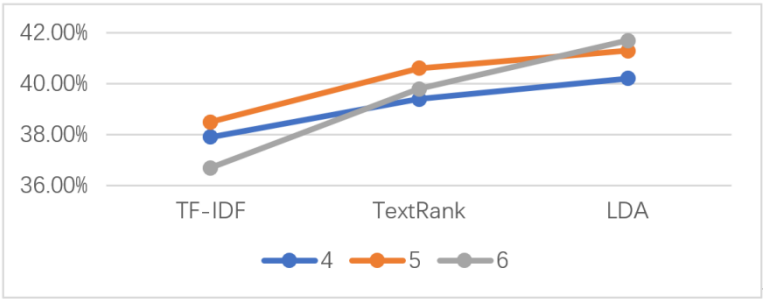



**Fig.3** Selection of the number of keywords K

It can be seen from Figure 2 and figure 3 that the overall performance of the keyword extraction model is the best when the number of key sentences n is set to 3 and the number of keywords K is set to 5.

*4.3 Experimental Results and Analysis*

In this paper, TF-IDF, text rank, LDA and the combination algorithm of this paper are used to test the key sentences extracted from the background of the article and not extracted. The results are shown in Table 1.

**Table 1.** Test results

| | Evaluating indicator | TF-IDF | Text rank | LDA | The algorithm in this paper |
|---|---|---|---|---|---|
| Key sentences are not extracted | P | 37.3% | 39.2% | 39.8% | 40.1% |
| | R | 39.1% | 42.2% | 43.0% | 44.2% |
| | F | 38.2% | 40.6% | 41.3% | 42.1% |
| Extract key sentences | P | 36.9% | 38.6% | 40.6% | *42.3%* |
| | R | 40.1% | 43.0% | 43.5% | *45.0%* |
| | F | 36.4% | 40.7% | 42.0% | *43.6%* |

It can be seen from table 1 that: (1) the recall rate, accuracy rate and F value of the combined algorithm based on multi type features are higher than that of single algorithm no matter whether the key sentences are extracted or not; (2) the performance of the algorithm is improved after the key sentence extraction based on the text of the paper, and its F value is improved by 1.5% compared with that without extracting key sentences, which proves that the key sentences extracted from background materials can be extracted to a certain extent On the summary of the content of the supplement, for the extraction of keywords made a certain contribution.

## 5. Conclusions

In this paper, a keyword extraction algorithm based on Bert and multi class feature fusion is proposed and verified on 300 scientific papers downloaded from Wanfang database. The experimental results show that the combination algorithm of multi class features is better than the single extraction algorithm; taking the text of the paper as the background material, using Bert to extract key sentences from it can solve the problem of missing information in the summary and improve the performance of the model. At the same time, although the model in this paper integrates the TF-IDF features, topic features and ranking features of the text, the fusion method is relatively simple, the semantic understanding of the model is not enough, and it needs to be further improved in the future work; in addition, the data set used in this paper is a professional scientific and technological academic paper, and the performance test can be carried out on other types of data sets in the next step Try.

## Acknowledgments

The 1331 Engineering Project of Shanxi Province, China

## References

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org, 2018.
[2] Chen Zelong. Patent text similarity assessment based on syntactic representation. 2019.
[3] Zhen Jiangjie. Research and implementation of multi-level semantic model in multi round dialogue system. 2019.
[4] Fang Junwei, Cui Haoran, he Guoxiu, et al. Keyword extraction from academic texts based on prior knowledge textrank. Information science, 2019, 37 (03): 77-82.
[5] Wang Tianshi, Zhang long, Liu huaiquan, et al. Text classification method based on keyword learning. Journal of Shandong Normal University: Natural Science Edition, 2019, 034 (001): 54-60.
[6] Xu Li. Text keyword extraction method based on weighted textrank. Computer science, 2019, 46 (z1).
[7] SALTONG, BUCKLEYC. Term-weighting approaches in auto-matic text retrieval. Information Processing&Management, 1987,24(5):513-523.

[8] MIHALCEAR, TARAUP. TextRank: Bringing-Order-into-Texts. Emnlp, 2004:404-411.
[9] BLEIDM, NGAY, JORDANMI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993-1022.
[10] Fang Junwei, Cui Haoran, he Guoxiu, et al. Keyword extraction from academic texts based on prior knowledge textrank. Information science, 2019, 37 (03): 77-82.
[11] Zeng Xi, Yang Hong, Chang Mingfang, et al. Short text keyword extraction and extension based on topic model . Journal of Shanxi University: Natural Science Edition, 2019, 42 (02): 37-45.
[12] Li Jiqi, Huang Gang. Research and application of improved collaborative filtering algorithm for keyword extraction. Computer technology and development, 2019, 029 (006): 154-158.
[13] Chen Yiqun, Zhou Ruqi, Zhu Weiheng, et al. Mining patent knowledge to realize automatic keyword extraction . Computer research and development, 2016, 053 (008): 1740-1752.
[14] Chen X, Zong W, Deng N, et al. Incremental Patent Semantic Annotation Based on Keyword Extraction and List Extraction. Conference on Complex. Springer, Cham, 2019.
[15] Bordoloi M, Biswas S K. Graph-Based Sentiment Analysis Model for E-Commerce Websites' Data. 2019.
[16] Csomai, Andras. Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing. unt theses & dissertations, 2008.
[17] Martin Dostál, Jezek K. Automatic Keyphrase Extraction based on NLP and Statistical Methods// Dateso: International Workshop on Databases. DBLP, 2011.
[18] TIMONEN M, TOIVANEN T, TENG Y, etal. Informative-ness-based Keyword Extraction from Short Documents KDIR. 2012: 411-42.