Extraction of Thesis Research Conclusion Sentences in Academic Literature

Litao Lin†
College of Information Management
Nanjing Agricultural University
Nanjing Jiangsu China
2020114016@njau.edu.cn

Dongbo Wang
College of Information Management
Nanjing Agricultural University
Nanjing Jiangsu China
db.wang@njau.edu.cn

Si Shen
School of Economics & Management
Nanjing University of Science and Technology
Nanjing Jiangsu China
shensi@njust.edu.cn

ABSTRACT

The extraction of sentences with specific meaning in academic literature is an important work in academic full-text bibliometrics. This research attempts to establish a practical model of extracting conclusion sentences from academic literature. In this research, SVM and SciBERT models were trained and tested using academic papers published in JASIST from 2017 to 2020. The experimental results show that SciBERT is more suitable for extracting thesis conclusion sentences and the optimal F1-value is 77.51%.

CCS CONCEPTS

Theory of computation~Theory and algorithms for application domains~Machine learning theory~Models of learning

KEYWORDS

SVM, BERT, Academic full text, Thesis research conclusion, Text mining, Deep learning

1 INTRODUCTION

Full-text data of academic literature mainly contains external characteristics and content characteristics. Since the creation of citation index by Garfield, citation analysis based on external characteristics of literature has been widely applied in various fields. However, due to the limitations of data and technology, the previous bibliometric have many defects, including rough statistical method and single indication ability [1]. Today, increasingly rich full-text data and evolving machine learning and deep learning techniques allow researchers to investigate the content characteristics of academic literature in depth.

Entity extraction and sentence extraction are two important basic works of full-text bibliometric analysis. At the entity level, the relevant research mainly includes theory method entity extraction [2], algorithm entity extraction [3] and software entity extraction [4]. At the sentence level，there are mainly four research directions, including extraction of experimental result sentences, extraction of research question sentences, extraction of research conclusion sentences and extraction of future work sentences. At present, there are more researches related to entity extraction and less researches on sentence extraction.

The research conclusion sentence refers to the sentence that contains the research conclusion. In the academic full text, research conclusion sentences are divided into citation research conclusion sentences and thesis research conclusion sentences. Citation research conclusion sentences refer to experimental results and conclusions in quotation sentences, such as 'Taylor's work shows that the special purpose syntactic parsers perform well on morphological descriptions.'. Thesis research conclusion is the author's statement of his own research results, such as 'In this way, we extended earlier work to the case that the impact factor can have a value lower than one.'.

Automatically extracting thesis research conclusion sentences can promote the development of automatic summarization and originality evaluation of academic papers. Therefore, this research attempts to construct an automatic recognition model of the thesis research conclusion sentence based on the deep learning techniques.

2 CORPUS AND METHOD

2.1 Data Source and Data Annotation

This research obtained all the full texts of academic papers published in JASIST (Journal of the Association for Information Science and Technology) from 2017 to 2020 by using self-made Python program.

As for data annotation, first, we use Python's NLTK module to segment the full text of the paper in sentence units. Then, 7 postgraduates majoring in information science manually annotated the sentences. For sentences that are not sure how to label, the decision will be made after group discussion and the experimenter completes the final review. The discriminant criteria of the thesis research conclusion sentence are as follows: (1) Semantically speaking, the sentence content is a summary of the author's own work experience, observations or actual research results. (2) The content of the sentence can be a reasoning and qualitative interpretation of the experimental

results, but it cannot be a straightforward description of the data of the experimental results.

Data imbalance, that is, the gap between the number of positive and negative samples used to train the model is too large, which is one of the most widespread problems in contemporary machine learning [5]. After the annotation is completed, the thesis research conclusion sentences only account for 3% of the total corpus (more than 130 thousand sentences in total). In order to alleviate the problem of data imbalance, we negatively sampled non-research conclusion sentences to increase the proportion of thesis research conclusion sentences to 8.9%. The basic information of the final corpus is shown in Table 1.

Table 1. Basic Information of the Corpus

Num.	Type	Count
1	Total article	502
2	Total Sentences	54,479
3	Thesis research conclusion sentences	4,870
4	Average number of marked sentences in each article	9.7
5	Average words number in each sentences	27.99
6	The longest sentence words number	255

2.2 Method

SVM and SciBERT are used in this research. SVM is called support vector machine and it is a classic model for text classification. In its simplest form, an SVM is able to perform a binary classification finding the 'best' separating hyperplane between two linearly separable classes. SciBERT [6] is a deep learning model based on the BERT architecture [7], which is trained on the full text corpus of 1.14 million scientific and technological documents. SciBERT uses the same configuration and size as BERT-base [7] in the construction process, and it performs better than BERT-Base on natural language processing tasks in scientific literature.

3 EXPERIMENT

Before the start of the formal experiment, we tested different hyper parameters combinations on a small part of the experimental corpus to explore the optimal settings for SVM and SciBERT. At the same time, considering the performance of the computer hardware used in the experiment, the final hyper parameters are set as follows. SciBERT (scibert-scivocab-uncased): 256 for Maximum sequence length, 64 for batch size, 2e-5 for learning rate, 3 for training epoch, case insensitive. The penalty function of SVM is set to 2, the kernel function is RBF, and TF-IDF is used to vectorize the text. The research uses a ten-fold cross-validation strategy, and the operating effect of the model is measured by Precision, Recall and F1-Value. Table 2 shows the results of the experiment.

Table 2. Results of 10-Fold Cross-Validation

Model		Precision	Recall	F1-Value
SciBERT	MAX	85.86%	78.61%	77.51%
	MIN	73.41%	44.13%	58.22%
	AVG	79.86%	64.51%	70.74%
SVM	MAX	98.19%	64.51%	77.03%
	MIN	90.37%	37.08%	53.80%
	AVG	95.97%	52.14%	67.24%

Table 2 shows that the SVM has a high precision rate for extracting thesis research conclusion sentences and a low recall rate. The SciBERT 's precision rate and recall rate are more balanced. From the perspective of the average F1-Value, SciBERT reached 70%, which is more than three percentage points higher than SVM. In summary, SciBERT performance is relatively better.

Compared to the sentences extracted by the SciBERT model with the manually annotated sentences, recognition errors of the SciBERT that have been discovered are as follows: (1) Recognizing the sentence describing the graph as the thesis research conclusion sentence. The possible reason for this problem is that the sentence describing the graph normally has phrases such as "as shown in" at the beginning, and these words are also important features of the thesis research conclusion sentence. (2) Recognizing research hypothesis sentences as thesis research conclusion sentences. According to observations, the thesis conclusion sentence is similar to the hypothesis sentence in terms of grammar and semantics. (3) Recognizing citation conclusion sentences without quotation mark as thesis research conclusion sentence. It indicates that some special words or symbols may affect the judgment of the model.

4 CONCLUSION & FUTURE WORK

This research provides a practical method for extracting conclusion sentences of thesis research from academic literature. This research shows that SciBERT is relatively superior than SVM for automatically extracting thesis conclusion sentences. This research uses a negative sample strategy to alleviate the problem of data imbalance and to enable faster model optimization, which may reduce the complexity of negative samples. Therefore, data augmentation needs to be achieved by adding more positive samples in the future. In addition, the position of the sentence in the article also needs to be considered to optimize the performance of the model. Finally, some research conclusion sentences extracted contain pronouns and do not have perfect semantics when read alone. Therefore, research on Co-Reference Resolution should be carried out.

ACKNOWLEDGMENTS

The authors acknowledge the National Natural Science Foundation of China (Grant Numbers:71974094) for financial support.

REFERENCES

[1] C. Lu, Y. Ding and C. Zhang, Understanding the impact change of a highly cited article: a content-based citation analysis, *SCIENTOMETRICS*, vol. 112, pp. 927-945, 2017.

[2] H. Zhang and C. Zhang, Using Full-text Content of Academic Articles to Build a Methodology Taxonomy of Information Science in China, *ArXiv*, vol. abs/2101.07924, 2021.

[3] Y. Wang and C. Zhang, Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language

processing, *J INFORMETR*, vol. 14, pp. 101091 - 101091, 2020.
[4] X. Pan, E. Yan, Q. Wang, and W. Hua, Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers, *J. Informetrics*, vol. 9, pp. 860-871, 2015.
[5] K. Micha, Radial-Based Undersampling for imbalanced data classification, *PATTERN RECOGN*, vol. 102, 2020-06-23 2020.
[6] I. Beltagy, A. Cohan and K. Lo, SciBERT: Pretrained Contextualized Embeddings for Scientific Text, *ArXiv*, vol. abs/1903.10676, 2019.
[7] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,, 2018.