Functional Structure Recognition of Scientific Documents in Information Science

Dayu Yan [1,2], Si Shen [1,2*] and Dongbo Wang [3]

[1] School of Economics & Management, Nanjing University of Science and Technology, Nanjing, China, 210000
[2] College of Information Management, Nanjing Agricultural University, Nanjing, China, 210000

Abstract

The recognition of the functional structure can help to understand the logical structure and content structure of the text, understand the text from a deep level, and further realize the academic big data analysis and data mining. In this study, the academic full text is taken as the research object, and the papers published in JASIST journals from 2010 to 2020 are selected to conduct a comparison experiment on the four models of Prompt, SciBERT, LSTM and TextCNN. The results showed that Prompt model had the best recognition effect.

Keywords

Functional structure, Text classification, Prompt, BERT

1. Introduction

Automatic recognition of the functional structure of academic texts is an important issue in the field of natural language processing (Lu, et al., 2018). On the one hand, it can make the logical structure of a paper clearer and present in a fine-grained way, which can enable researchers to retrieve the required literature information faster and save time. On the other hand, it helps to standardize the structure of the paper. At present, machine learning and deep learning have become the mainstream methods of paragraph structure recognition.

Traditional machine algorithms mainly include support vector machine, naive Bayes, Logistic regression and K- nearest neighbour algorithm. TUAROB et al. (2015) compared two models, naive Bayes and support vector machine, to divide chapter boundaries of academic literature. Since Hinton et al.(2006) put forward the concept of deep learning in 2006, in addition to these models based on traditional machine learning, the research on the use of deep models has gradually deepened. Tkaczyk et al. (2015) proposed a system based on modular open source flow to extract metadata in the paper. However, the proposal of pre-training language model makes the representation effect of word vector to a higher level. Ren et al. (2017) fully studied the structural features of academic texts, and combined with convolutional neural network, proposed an automatic functional structure detector to identify the structure of academic texts. The BERT model is excellent, but fine-tuning requires a lot of data and computational power, and not enough data is available for all scenarios. The prompt based downstream tasks have recently become a boon to small sample learning. Prompt allows the model input to be modified to bring downstream tasks closer to the pre-trained model.

Therefore, Prompt, SciBERT, LSTM and TextCNN models were selected in this research for comparative experiments. The optimal recognition model is sought and the influence of different models on structure function recognition is discussed.

2. Corpus & Method

2.1. Data Source & Data Annotation

This research obtained all the full texts of academic papers published in Journal of the Association for Information Science and Technology

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online
EMAIL: ydy@njust.edu.cn (D. Yan); shensi@njust.edu.cn (S. Shen); db.wang@njau.edu.cn (db. Wang)

(JASIST) from 2010 to 2020 by using self-made Python program.

As for data annotation, in this paper, the collected 2232 articles are divided into text according to paragraph as the basic unit, and the structure of these paragraphs is marked. Combined with previous research on the functional structure of academic texts, this paper divides the structure and function of academic texts into five parts: "introduction", "relevant research", "method", "experiment" and "conclusion", which are represented by "I", "R", "M", "E" and "C" respectively. The specific labeling process is as follows: BERT model is first invoked to complete automatic functional structure labeling of academic full-text data from 2010 to 2020. In order to ensure the accuracy of labeling, manual verification is required. After manual review and collation, preliminary text data is obtained. The number of marked paragraphs of each structure is shown in Table1 below .

Table 1 Basic Information of the Corpus

Num	Type	Count
1	introduction	62847
2	relevant research	90020
3	method	104061
4	experiment	152950
5	conclusion	84434

2.2. Method

Prompt reconstructs the template for different tasks, inputs human-made rules into the pre-training model, and makes the model better understand human instructions, bridging the gap between the training process and downstream tasks. SciBERT is a BERT pre-trained using a total of 1.14 million scientific papers in biomedical (82%) and computer science (12%) directions and may be more suitable for natural language processing tasks in the direction of scientific papers. LSTM is a variant of RNN. RNN can only have Short-Term Memory due to the disappearance of gradient. Compared with RNN network, LSTM network combines short-term memory with long-term memory by adding additional state c and using gate control. And to some extent, it solves the problem of disappearing gradient. TextCNN, as the name implies, is CNN for text tasks. Each word is mapped to a word vector by embedding, and then input to sofxmax layer through convolution layer and max-pooling layer to realize text classification.

3. Experiment

The main parameters of Prompt are as follows: epochs is 10, learning_rate is 1e-5, max_len is 512.The main parameters of SciBERT are as follows: max_len is 128, epochs is 3, learning_rate is 2e-5, num_attention_heads is 12, hidden_size is 300. LSTM parameters are as follows: max_len is 512, epochs is10, and learning_rate is 0.005. The main parameters of TextCNN are as follows :epochs is 5, filter_sizes is (3,4,5). In this paper, P (precision), R (recall), f1 (F1-score) and macro average of five paragraph functional structure recognition are used as evaluation indicators to measure the performance of these four deep learning models.

Table 2. Results of 10-Fold Cross-Validation

Model		Precision	Recall	F1-Value
Prompt	I	94.48%	96.07%	95.26%
	R	76.06%	86.40%	80.90%
	M	84.00%	85.71%	84.85%
	E	86.62%	79.87%	83.11%
	C	97.04%	91.11%	93.98%
	AVG	87.64%	87.83%	87.62%
SciBERT	I	92.09%	91.57%	91.83%
	R	72.67%	87.20%	79.27%
	M	83.33%	78.23%	80.70%
	E	81.88%	79.22%	80.53%
	C	94.12%	88.89%	91.43%
	AVG	84.82%	85.02%	84.75%
LSTM	I	89.86%	71.26%	79.49%
	R	55.10%	44.63%	49.32%
	M	59.90%	35.62%	44.44%
	E	51.49%	81.76%	63.19%
	C	75.12%	87.71%	80.93%
	AVG	66.13%	64.20%	63.47%
TextCNN	I	95.38%	69.66%	80.52%
	R	65.49%	74.40%	69.66%
	M	56.12%	90.48%	69.27%
	E	93.62%	57.14%	70.97%
	C	80.66%	81.11%	80.89%
	AVG	78.26%	74.56%	74.26%

As can be seen from Table 2, Prompt has the best effect in the functional structure recognition experiment of the paper, with F1 up to 87.62%, which is nearly 3 percentage higher than the SciBERT model. As a whole, Prompt model is superior to the other three models in the recognition of each part, showing that the fourth normal form of natural language has an unparalleled advantage .

From the perspective of various structural functions, the effect of the introduction is the best, and the average of the three indicators can reach 95%, followed by the conclusion and method, and the effect of related research is the worst. The reasons are as follows: (1) In the function of relevant research, the role of paragraphs is to summarize the current research status at home and abroad, sort out the research context, discover new research questions, and provide theoretical support for the following research. However, it overlaps with the following methods to a certain extent. (2) The experimental function partially overlaps with the method function to a certain extent, which leads to the lack of effect of experimental function. Introduction, conclusion and other functional structure repetition degree is low, so the effect is better.

4. Conclusion & Future Work

In the experiment of functional structure recognition of the full text, the overall performance of the functional structure Prompt on the test set is the best, because the model input the prompt information, can be more fully mining the semantic knowledge in the pre-training model. Secondly, from the view of the recognition effect of each structure, Prompt has the best performance in the introduction, related studies, methods, experiments, conclusions, showing the strong learning ability of the model, as well as a wide range of application prospects.

5. References

[1] Lu, W., Huang, Y., Bu, Y., et al.(2018). Functional structure identification of scientific documents in computer science. Scientometrics, 115(1), 463-486.

[2] Shen S, Jiang C, Hu H, et al. (2022). A model for the identification of the functional structures of unstructured abstracts in the social sciences. The Electronic Library, 40(6): 680-697.

[3] Tuarob S, Mitra P, Giles C L.(2015). A hybrid approach to discover semantic hierarchical sections in scholarly documents.2015 13th international conference on document analysis and recognition (ICDAR). IEEE, 1081-1085.

[4] Hinton, G. E., Salakhutdinov, R. R.(2006). Reducing the dimensionality of data with neural networks. science, 313(5786): 504-507.

[5] Tkaczy,k. D., Szostek, P, Fedoryszak, M., et al. (2015).CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4): 317-335.

[6] Ren, X., Zhou, Y., Huang, Z., et al.(2017). A novel functional structure feature extractor for Chinese scene text detection and recognition. IEEE Access, 5: 3193-3204.