ICEB 2021 Proceedings (Nanjing, China)

International Conference on Electronic Business (ICEB)

Winter 12-3-2021

Automatic Recognition of Knowledge Characteristics of Scientific and Technological Literature from the Perspective of Text Structure

Huan Ma
Hohai University, Nanjing, China, mahuanfit@163.com

Zhu Fu
Jiangsu University of Science and Technology, Zhenjiang, China, fuzhu886@163.com

Qiong Jia
Hohai University, Nanjing, China, jiaqiong@163.com

Recommended Citation

Ma, Huan; Fu, Zhu; and Jia, Qiong, "Automatic Recognition of Knowledge Characteristics of Scientific and Technological Literature from the Perspective of Text Structure" (2021). *ICEB 2021 Proceedings (Nanjing, China)*. 63.
https://aisel.aisnet.org/iceb2021/63

Ma, H., Fu, Z. & Jia, Q. (2021). Automatic recognition of knowledge characteristics of scientific and technological literature from the perspective of text structure. In *Proceedings of The International Conference on Electronic Business, Volume 21* (pp. 593-598). ICEB'21, Nanjing, China, December 3-7, 2021.

Automatic Recognition of Knowledge Characteristics of Scientific and Technological Literature from the Perspective of Text Structure

(Work in Progress)

Huan Ma [1]
Zhu Fu [2]
Qiong Jia [3,*]

*Corresponding author
[1] Master Student, Hohai University, Nanjing, China, mahuanfit@163.com
[2] Associate Professor, Jiangsu University of Science and Technology, Zhenjiang, China, fuzhu886@163.com
[3] Associate Professor, Hohai University, Nanjing, China, jiaqiong@163.com

ABSTRACT

This paper independently explores the chapter structure of scientific and technological literature in the field of shipbuilding in the natural sciences and the field of library and information in the social sciences. The chapter structure model of previous studies, namely 'background, purpose, method, result, conclusion, demonstration,' is quoted as the verification object of the document chapter structure in the field of exploration. In order to verify the rationality of the structure, this paper uses the deep learning models TextCNN, DPCNN, TextRCNN, and BiLSTM-Attention as experimental tools, and designs 5-fold cross-validation experiment and normal experiment, and finally verifies the rationality of the model structure, and It is concluded that the BiLSTM-Attention model can better identify the chapter structure in this field.

Keywords: Shipbuilding, library and information, chapter structure, deep learning; cross-validation.

INTRODUCTION

With the development of network information technology, scientific and technological resources are accelerating the pace of increase in quantity, variety, and distribution. As a result, during this period of time, a large amount of literature resources was not utilized in a timely and effective manner, which means that the value of literature has been weakened. Therefore, it is necessary to propose and improve the method of extracting the knowledge features of scientific and technological literature and automatically identifying the structure-function. Because it can help readers quickly understand the chapter structure and knowledge features of related documents in the field and improve the efficiency of the use of literature resources.

In this paper, the excavation of the text structure is actually based on the characteristics of the literature resources. The current feature recognition research of literature resources mainly uses methods and technologies such as natural language processing based on text classification and machine learning to automatically recognize the structure of the literature content.

Some scholars have the following opinions on what method is used to identify the text structure. First, Zhang *et al.* used characters as the data input unit and used deep learning of convolutional neural network to extract abstract semantics of character-level data. And achieved the purpose of the deep learning model for text understanding (Zhang & LeCun, 2015). Dasigi *et al.* used deep learning models such as RNN and LSTM to identify the structure of scientific and technological literature. Jung proposed to apply supervised learning technology (a technology of machine learning) to the semantic features of the main content of scientific and technological literature to identify. The first two researchers used deep learning methods, and the latter one chose one of the machine learning methods. Which of the two methods is better? Li Nan, Fang Li, etc., compared the performance of SVM, CRF, and NBC classification models in the field of machine learning and models represented by CNN and RNN structures in the field of deep learning. The results show that CNN, LSTM, LSTM+CRF, LSTM+SVM, CNN+CRF, CNN+SVM based on deep learning have better performance. In addition, Qin Chenglei *et al.* (2020) compared the three models of Bi-LSTM+CNN, Bi-LSTM+Attention, and Bi-LSTM+CNN+Attention, and proposed the performance of the deep learning model after adding hierarchical attention (Attention) better.

Then, this paper adopts the method of deep learning field as the experimental tool and adopts the TextCNN, DPCNN, TextRCNN, BiLSTM-Attention model, which contains CNN and RNN structure in this field as the experimental tool(Fesseha *et al.*, 2021; Wang *et al.*, 2019;).

Many scholars conduct research on text structure from different angles. For example, Lu Wei, Huang Yong, etc., studied the structural function recognition of academic texts from three different perspectives: chapter title, chapter content, and paragraph. However, Dasigi *et al.* only studied the composition structure of scientific and technological literature content from the sentence level. Then, Qin Chenglei and others carried out word-level, sentence-level, and paragraph-level structure recognition on scientific papers and concluded that the sentence-level level is more appropriate. Finally, this paper decides to explore the characteristics of the chapter structure of the literature in the shipbuilding field and the library and information field from the sentence level.

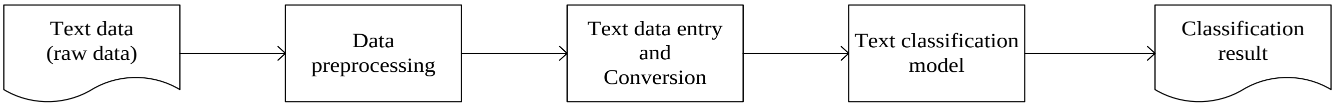
In summary, the general work of this article is as follows:
1. Construct a dataset in the field of shipbuilding and library and information.
2. Self-optimization of four models (parameter adjustment).
3. Comparison and analysis of experimental results of four optimized models.

RELATED THEORIES AND TOOLS

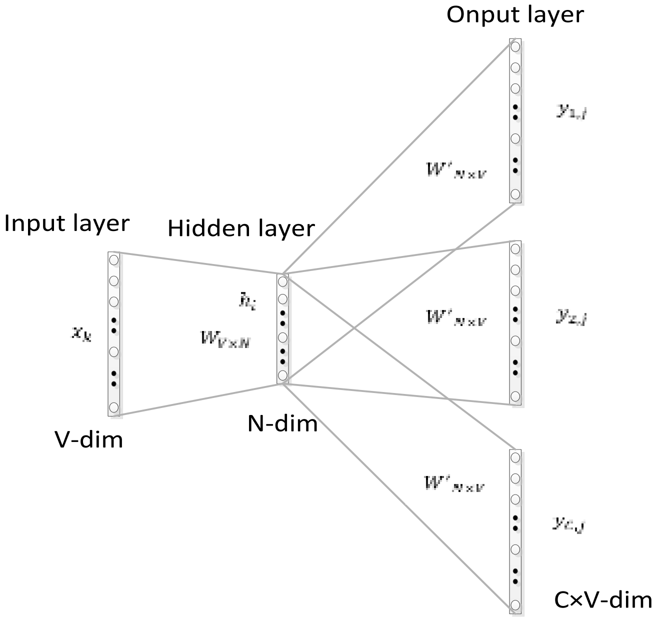
In this paper, the recognition of the chapter structure of scientific and technological literature in the field of shipbuilding and library information is based on the method of text classification. Wang Dongbo *et al.* proposed that the problem of sequence labeling of sentence units and the problem of academic text structure-function recognition can be converted to each other, and the realization of the two can achieve the same purpose. This means that the structure of the entire document can be identified by identifying the sequence to which the sentence unit belongs. As a result, the paper uses sentence-level sequence annotation to achieve the recognition of the document text structure, that is, the text classification method.

For text classification, first, the original data needs to be preprocessed so that the data obtained is sentence-level data with sequence tags. In the process of data preprocessing, it is necessary to remove the noise data and standardize the length of the text to ensure the quality of the data. Since the computer cannot directly recognize the text data automatically, and in order to reduce the burden of the computer, the text features of a whole sentence cannot be used as input. All, it is necessary to segment the sentence-type data so that important text features in a piece of data can be extracted. Then the divided words are converted into word vectors as the input of the computer.

For the word segmentation tool, this article uses Jieba word segmentation. The word vector conversion of the data afterword segmentation adopts the skip-gram model in the word2vec tool. Using word2vec is to express each word obtained by dividing text data into a unique word vector with a unified meaning and dimension. The essence of word2vec is a method of word clustering, which is suitable for sequence data and is also a simple (input layer-hidden layer-output layer) neural Network model. The skip-gram model takes the word vector of a word as input and predicts the context of the word. Figure 1 describes the general process of text classification, and Figure 2 shows the Skip-gram model.



Source: This study.
Figure 1: The general process of text classification



Source: https://blog.csdn.net/u010665216/article/details/78721354.
Figure 2: Skip-gram.

CONSTRUCTION OF SCIENTIFIC AND TECHNOLOGICAL LITERATURE DATASET

Data sources

As this article studies two scientific and technological literature fields, it is necessary to construct two datasets, namely the shipbuilding field dataset and the library and information field dataset. The scientific and technological literature in the field of ship construction comes from CNKI's literature resources in the past five years. Because the length of the dissertation is too long, academic papers are selected—a total of 173 scientific and technological literature resources in the field of shipbuilding selected by CNKI. The scientific and technological literature in the field of library and information comes from the literature resources of the library and information papers network in the past five years. In the field of library and information, 198 literature resources on library construction are selected.

Data preprocessing

Two datasets are obtained by removing non-critical information, sentence division, noise data removal, and short texting of long text data from the text data in the two fields. And the shipbuilding field dataset has a total of 10,680 records, the library, and the information field. The dataset has a total of 13,083 records.

Annotation result statistics

Manually label the processed data, and the computer learning rules can learn the rules only on the premise of the existing rules, so the process of data labeling is indispensable. The principle of data labeling is the chapter structure model adopted in this article. Table 1 is the definition of each part of the text structure model, and Table 2 is the statistics of the annotation results of the two datasets.

Table 1: The text structure.

Structure category	Definition
Background	Time background, theoretical background, including descriptions of past, present, and future situations.
Purpose	Put forward a clear goal that needs to be reached, including the improvement of technical level and feeling level, etc.
Method	All technologies, suggestions, etc., are adopted to achieve a certain purpose.
Result	The result is caused by the realization of the method or the unrealized phenomenon, which exists in itself
Conclusion	The author's summary of the above four parts
Demonstration	A method that can be used to prove an argument. The deduction from the argument to the topic in the proof is carried out in the form of reasoning, sometimes a series of reasoning methods

Source: This study.

Table 2: Annotation result statistics.

Category	Amount	
	Shipbuilding dataset	Library and information dataset
Background	1981	1821
Purpose	1289	1049
Method	865	1280
Result	2064	3348
Conclusion	2030	3262
Demonstration	2451	2323
Total	10680	13083

Source: This study.

EXPERIMENTAL DESIGN AND IMPLEMENTATION

Experimental standard

The experimental standards used in this article are accuracy rate Acc, recall rate R, and F1 value. In the following formula, TF represents the amount of text data whose actual and predicted categories are both positive. FP represents the amount of text data whose actual category is negative, but the predicted category is positive. FN is the amount of text data whose actual category is positive, but the predicted category is negative. TN is the amount of text data whose actual and predicted categories are both negative.

$$Acc=\frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$P=\frac{TP}{TP+FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{4}$$

Experimental model parameter adjustment

Before the TextCNN, DPCNN, TextRCNN, and BiLSTM-Attention models are used for text classification of experimental data, and each model needs to undergo a self-optimization process to ensure that the experimental results of each model are compared under the optimal conditions of the self-model. The self-optimization process of the model is actually the process of adjusting many parameters of the model. The specific process of parameter adjustment is displayed through four parameters. Table 3 below shows the final setting details of various parameters of each model.

Table 3: Experimental model parameter setting.

Model Parameter	TextCNN	DPCNN	TextRCNN	BiLSTM-Attention
sequence_length	300	200	200	200
vocab_size	5000	5000	5000	5000
hidden_size	-	-	224	128
hidden_size2	-	-	-	64
embedding_dim	300	300	300	300
num_filters	288	180	-	-
num_layers	-	2	1	2
filter_size	[2,3,4,5]	-	-	-
num_classes	6	6	6	6
learning_rate	1e-3	1e-3	1e-3	1e-3
batch_size	64	32	32	32
num_epochs	13	17	12	13
activation_func require_improvement	relu 1000	leakyrelu 1000	leakyrelu 1000	leakyrelu 1000

Source: This study.

EXPERIMENTAL RESULTS

Normal result

First, the experimental model is set according to the parameter values in the above table. The datasets of the shipbuilding field and library information field are respectively divided into a training set, val set, and test set according to 6:2:2. Among them, the data of the training set is used to train the model so that the model continuously extracts the corresponding rules from the data of the training set. The data in the validation set is used to test the quality of the model training with accuracy, recall, F1—the value as the standards. The test set is used to test the actual performance of the model. The specific results of the experiment are divided into two parts: the recognition result of the ship construction dataset and the recognition result of the library and information dataset. Not only need to obtain the specific results of text data recognition but also to obtain the model—the recognition result. The results are shown in Table 4.

Table 4: Normal experiment result.

Shipbuilding dataset				Library and information dataset			
Model	Acc	R	F1	Model	Acc	R	F1
TextCNN	0.9433	0.9447	0.9440	TextCNN	0.9746	0.9618	0.9682
DPCNN	0.8673	0.8297	0.8481	DPCNN	0.9109	0.9218	0.9163
TextRCNN	0.9562	0.9488	0.9525	TextRCNN	0.9831	0.9788	0.9810
BiLSTM-Attention	0.9641	0.9660	0.9651	BiLSTM-Attention	0.9863	0.9789	0.9826

Source: This study.

5-fold cross-validation experiment result

Using the 5-fold cross-validation method to evaluate the performance of the model can prevent over-fitting to a certain extent, and the cross-validation method can obtain as much information in the data as possible. The basic idea of the 5-fold cross-validation method is: First, divide the data into five equal parts that basically meet the average requirements, and each piece of data is mutually exclusive. Then, turn each piece of data as the test set in turn, and the remaining four. The dataset is merged into the training set. In the second experiment, there is no division of the validation set, only the training set, and the test set. Unlike Experiment 1, this experiment allows each piece of data to be used as a validation set, increasing the number of model training and taking the average of the model performance. The value evaluates the performance of the model and has certain feasibility. Since the application of the 5-fold cross-validation method is not the purpose of the experiment in this article, the second experiment is only a verification or supplementary explanation of the results of the first experiment. Therefore, take the training set in the first experiment as all the data in this experiment. The experimental results are shown in Table 5.

Table 5: 5-fold cross-validation experiment result.

Shipbuilding dataset				Library and information dataset			
Model	Acc	R	F1	Model	Acc	R	F1
TextCNN	0.9001	0.883	0.8914	TextCNN	0.8549	0.8876	0.8709
DPCNN	0.8045	0.7851	0.7947	DPCNN	0.795	0.8207	0.8076
TextRCNN	0.9000	0.8839	0.8919	TextRCNN	0.8577	0.8904	0.8737
BiLSTM-Attention	0.9021	0.8849	0.8934	BiLSTM-Attention	0.8584	0.8934	0.8756

Source: This study.

CONCLUSION AND FUTURE RESEARCH

Through the analysis of the two experimental results, the following conclusions are drawn.

1) Judging from the recognition accuracy rate of various models, the accuracy rate is about 90%, indicating that the six-term structure adopted is in line with the chapter structure of the literature in the research field.
2) The model based on the RNN structure is more suitable for the identification of the text structure of the documents in these two fields than the model based on the CNN structure.
3) A comparison between the TextRCNN and BiLSTM-Attention models based on the RNN structure shows that the BiLSTM-Attention model that introduces the attention mechanism has a better recognition effect.
4) The reference of the attention mechanism helps to improve the accuracy of classification and recognition under certain conditions.
5) In the process of text classification and recognition, the deep learning model can also improve the accuracy of classification and recognition by enhancing the long-distance capture of text information. That is, the bidirectional RNN or LSTM model can capture text information for a long time and long-distance.

In the end, this paper draws a text structure model for the scientific and technological literature in the field of shipbuilding and library information to be explored. The experimental results are reasonable and feasible, and the BiLSTN-Attention model based on two-way long-term and short-term memory and introducing the attention mechanism is more suitable. The recognition of text structure based on knowledge feature vector in these two fields.

This article only uses four models as experimental tools, the number is limited, and the experimental data may have human errors, so the experimental process needs to be further improved, here is mainly to provide a case for follow-up research, hope that follow-up researchers can check the omissions. Further, supplement it.

ACKNOWLEDGMENT

Generous financial support was provided by the National Natural Science Foundation of China [grant numbers71904043, 71872061, 71702045], Social Science Foundation of Jiangsu Province (18TQC002), the Key Program of NSFC-FRQSC Joint Project (NSFC No. 72061127002, FRQSC No. 295837), and Fundamental Research Funds for the Central Universities [grant numbers B210207008].

REFERENCES

Chenglei, Q., & Chengzhi, Z. (2020). Recognizing Structure Functions of Academic Articles with Hierarchical Attention Network. *Data Analysis and Knowledge Discovery,* 4(11), 26-42. https://doi.org/10.11925/infotech.2096-3467.2020.0364

Dasigi, P., Burns, G. A., Hovy, E., & de Waard, A. (2017). Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *arXiv preprint arXiv:1702.05398.*

Dongbo, W., Ruiqig, G., Wenhao, Y., Xin, Z., & Danhao, Z. (2018). Research on the structure recognition of academic texts under different characteristics. *Information Journal,* (10),997-1008. doi: CNKI:SUN:QBXB.0.2018-10-004 (in Chinese).

Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., & Dahou, A. (2021). Text Classification Based on Convolutional Neural Networks and Word Embedding for Low-Resource Languages: Tigr-inya. *Information,* 12(2), 52. https://doi.org/10.3390/info12020052

Huang, Y., Chen, J., Zheng, S., Xue, Y., & Hu, X. (2021). Hierarchical multi-attention ne-tworks for document classification. *International Journal of Machine Learning and Cybernetics,* 12(6), 1639-1647. https://doi.org/10.1007/s13042-020-01260-x

Jung, Y. (2017). A semantic annotation framework for scientific publications. *Quality & Qu-antity,* 51(3), 1009-1025. https://doi.org/10.1007/s11135-016-0369-3

Lu, W., Huang, Y., & Cheng, Q. (2014). The structure function of academic text and its classification. *Journal of the China Society for Scientific and Technical Information,* 33(09), 979-985.

Nan, L., Li, F., & Yifei, Z. (2019). Multi-disciplinary comparative study on methods of academic text structure function recognition based on deep learning model. *Modern Information,* 39(12), 55-63+87. https://doi.org/10.3969/j.issn.1008-0821.2019.12.007

Wang, R., Li, Z., Cao, J., Chen, T., & Wang, L. (2019, July). Convolutional recurrent neural networks for text classification. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE. https://doi.org/10.1109/IJCNN.2019.8852406

Yong, H., Wei, L., & Qikai, C. (2016). The structure function recognition of academic text—chapter content based recognition. *Information Science News,* (03), 293-300. doi:CNKI:SUN:QBXB.0.2016-03-008 (in Chinese).

Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710.*