Research on Document Classification in the Field of Construction

Zhenyang Ren[1, 2, a], Fucheng Wan[1, 2, b, *], Hongzhi Yu [1, 2, c], Tiantian Wu [1, 2, d]

[1] Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

[2] Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu 730000, China

[a]1031439279@qq.com, [b, *]306261663@qq.com, [c] 3095911462@qq.com, [d] 1316170316@qq.com

Abstract. At present, the communication and preservation of project construction documents in most construction enterprises are still using traditional paper method, so the utilization of information is very low. This cannot guarantee the completeness, accuracy and systematic of architectural engineering records. At the same time, the environment for information management of construction documents is so complicated, which also increases the difficulty of document classification in the field of construction. This paper first constructs a corpus in the field of construction, then pre-treats and constructs word vector of documents to be categorized. Finally, finishing the task of classification with text classification model of Bi-LSTM combined with attention.

Keywords: Attention; Deep Learning; Document Classification; Construction.

1. Preface

With the continuous development of information technology, modern enterprises have put forward higher requirements for the systematic management of construction project documents. In the field of construction, a large number of documents will be produced. For a construction project, the entire life cycle usually includes: decision-making and preparation of the construction project, design and bidding of the project, construction of the project, completion of the project, and final demolition of the building. Along with the implementation and advancement of each specific stage, a large number of types of engineering project documents and information will be produced, most of which are in the form of various construction project documents, such as Drawings, photos, videos, electronic documents, etc. However, due to the weak awareness of many construction units, project files often fail to pay due attention in actual work, which often leads to the failure to archive documents such as project documents and project documents, which results in a large number of documents being disordered and mixed. The completeness, accuracy and system of the file cannot be guaranteed. Moreover, as the number of documents continues to increase, the difficulty of organizing and retrieving old documents is greatly increased.

On the other hand, based on the particularity of the construction industry, the classification of documents in this field is rather vague, whether it is based on the difference of the division of functions of the project or the difference of the project management organization, or the specific management activities or business processes from the project. From the perspective of document classification, they have their shortcomings. This paper studies the automatic text classification for the field of building construction. The research process intends to combine the characteristics of the documents in this field, and proposes a multi-level classification standard, and then uses the document classification technology to realize the automatic classification of documents in the field. Through the classification and sorting of documents in the field of building construction, the purpose of systematically and completely storing and managing documents in the field of building construction is achieved.

2. Construction of Corpus

2.1 Designing the Categories of Text Document

Based on the particularity of the construction industry, there are many related units such as general contracting, subcontracting and supervision, and the owner. Therefore, the classification of project documents and information is particularly complicated. Whether it is based on the difference of the job division function or the difference of the project management organization to classify the documents, or from the perspective of the specific management activities or business process of the project to classify the documents, they have their shortcomings. Aiming at the vague classification of construction documents in the field of building construction, based on the national standard (gbt50328-2014), a new category of document classification in the field of building construction was designed and proposed.

The classification category is divided into three levels. The first level is the classification category, including the engineering preparation stage documents (class A), supervision documents (class B), construction documents (class C), as-built drawings (class D), and project completion acceptance. For documents (class E), each primary category is subdivided into several secondary categories, and each secondary category also includes multiple tertiary categories.

2.2 Cataloging Design of Audio-visual Documents

For audiovisual files, this paper proposes to use the idea of cataloging to solve its classification. Through cataloging design, each audiovisual file corresponds to a text document, which contains all the information of the audiovisual document, and then classifies the text. Through this text, the corresponding audiovisual document can be found, thereby completing the classification of the audiovisual document.

The training corpus of the audiovisual document is all classified and cataloged according to the following figure. The classification standard uses the three-level classification category. The audiovisual document catalog is shown in Table 1:

Table 1. Catalog of audio-visual document

Attribute	Date
Drive letter number	
Project name	
Subordinate unit	
Start/Completion date	
Person in charge	
File name	
Number of files	
File description	
File type	
File storage address	
Keyword	

3. Vector Representation

3.1 Data Preprocessing

To complete the data pre-processing, the audio image document must be converted into text through the catalogue of the above design, so that all the text can be further processed. The content mainly includes the cleaning of invalid special characters and punctuation marks, and the cleaning of commonly used pause words in the language. Since there is no similar English interval between Chinese words and words, it is necessary to use the jieba word segmentation tool based on python to segment Chinese words.

3.2 Construction of Word Vectors

Words are often regarded as the basic unit of expressive ability in natural language models. Therefore, when performing text classification tasks, they are often first expressed in terms of words as basic units. The purpose is to use a real vector to represent a word, which is convenient for computers. Follow-up processing. It can be said that the expression of the word directly affects the effect of the text classification.

The representation of a word is divided into one-hot representation and distributed representation. The traditional one-hot representation tends to cause the dimension of the vector to be too large, nor can it represent the relevant information between words and words on the semantic level. This study uses a neural network-based distribution representation, ie word embedding, because documents in the construction field are generally dominated by long texts, and contexts have strong dependencies. Word embedding can not only control the dimensions of word vectors, but also Can express the complex context of the text. Here mainly consider using the Word2Vec tool to build word vectors. In Word2Vec, there are two model construction methods: CBOW and Skip-gram. Consider the CBOW model here because it is more suitable for small corpora.

The CBOW model predicts the central word through the context of the central word. Its network structure consists of the input layer, the projection layer and the output layer, as shown in Figure 1:

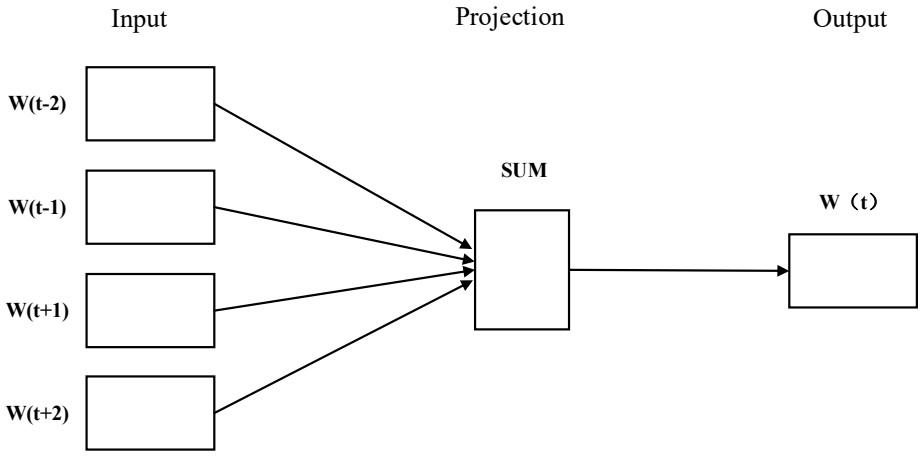

Figure 1. Schematic diagram of the cbow model structure

Since there is no hidden layer, a lot of computation time is saved. The output layer is responsible for calculating the probability of occurrence of the central word in context. The process of constructing a word vector in this study is:

(1) Generate a vocabulary.The word frequency of each word is counted, and the word frequency is sorted from high to low, so that there is an ont-hot vector for each word.

(2) Generate an ont-hot vector.Note here that the original location of each word is preserved to ensure contextual relevance.

(3) Determine the dimension n of the word vector and the window size of the cbow model.

(4) The neural network iteratively trains a certain number of times, and obtains the parameter matrix from the input layer to the hidden layer. The transposition of each row in the parameter matrix is the word vector of the corresponding word.

4. Feature Extraction and Classification Model Combined with Attention

The Attention used in natural language processing tasks first appeared in Encoder-Decoder, the core idea is to simulate the attention of people. For information to be processed, people tend to focus their attention on a few key points of information rather than distributing their attention evenly across all information. The introduction of the Attention into the feature extraction and classification model

has the advantage of enabling the data in the model to have different weights, so that the classification results can be improved based on the information values with higher weights when classifying.

This paper considers combining the Attention with Bi-LSTM. Based on the Attention model, the key is to calculate the probability distribution of attention.Different from the traditional Bi-LSTM, the two-layer state is directly combined as the final feature. The Bi-LSTM text classification model combined with the Attention utilizes the state at each moment and combines with the final state to calculate the attention of each moment to the final state. The probability distribution, which uses the attention distribution to further optimize the final state as the final text feature.

The objective function of the Bi-LSTM text classification model combined with the Attention is the same as that of Bi-LSTM, which uses softmax as the output layer normalization calculation and combines the cross entropy loss function.

The structure of the text classification model of Bi-LSTM combined with the Attention proposed in this study is shown in Figure 2:

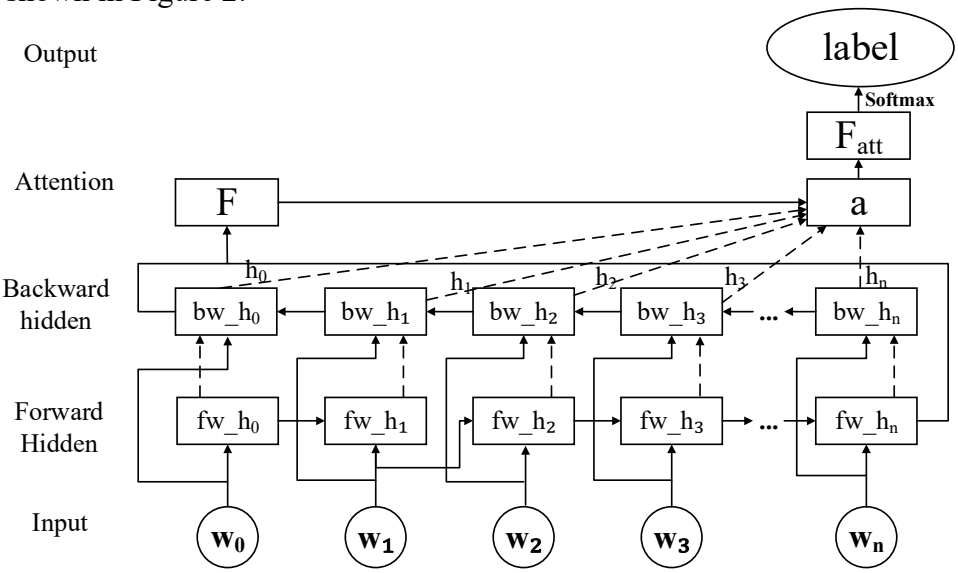

Figure 2. Schematic diagram of the text classification model of Bi-LSTM combined with attention

As shown in the figure, F represents the sum of the state values of the final hidden layer in the independent direction in Bi-LSTM, which is called the final state of Bi-LSTM, and a represents the attention probability distribution of the hidden layer unit state for the final state at a certain moment, the component a_n represents the attention probability of the Bi-LSTM state h_n for the final state at time n, and h_n is obtained by adding the states of the respective independent directions at that time, and F_{att} indicates that Attention-weighted text feature vector.

The model based on the Attention generally contains a two-part calculation process, one is the calculation process of the attention probability distribution, and the other is the final feature calculation process based on the attention distribution.

In this model, the attention probability of the output data at time n for the final state is calculated as follows:

$$a_n = \frac{e^{h'_n}}{\sum_{i=1}^{N} e^{h'_i}} \tag{1}$$
$$h'_n = h_n{}^T U F$$

The formula uses the softmax function as the way to calculate the attention probability distribution, where N is the number of input sequence elements. U is a weight matrix, F is the sum of the state values of the final hidden layers in the independent directions in Bi-LSTM, and h_n is the sum of the hidden layer state values in the n time. In this model, based on the final feature F_{att} of the attention distribution, the calculation process is expressed as:

$$F_{att} = \sum_{n=1}^{N} a_n h_n \tag{2}$$

N represents the number of input sequence elements, a_n represents the attention probability of the output data at the moment for the final state, h_n represents the sum of the hidden layer states in the two independent directions at time n.After obtaining the text feature vector F'_{att} based on the Attention, the probability distribution of the classification label is calculated by the softmax function of the output layer. The calculation process is expressed as:

$$y = \text{softmax}(F'_{att}) = \frac{e^{F'_{att(i)}}}{\sum_{j=1}^{T} e^{F'_{att(j)}}} \tag{3}$$
$$F'_{att} = VF_{att}$$

T is the number of category labels and V is the weight matrix of the model output layer. $F'_{att(i)}$ represents the i-th component value in the vector F'_{att}, and the vector length is equal to the number of classification labels.After the softmax function classification, the probability distribution y of the text category based on the Attention can be obtained, and the cross entropy loss is obtained from the real category distribution Y, which is expressed as:

$$E(Y, y) = -Y log(y) \tag{4}$$

Y represents the probability distribution of the real category and y represents the probability distribution of the category predicted by the model.

5. Conclusion

At present, most of the construction projects use the traditional paper method for communication and preservation, and the utilization rate of information is low, especially in the way of exchange and storage of Internet electronic documents. Paper-based documents will become aging, missing, and lost over time. If you want to organize these documents, you must convert them into electronic documents. Most of these documents are in unstructured form. Existence, the manual classification of these documents not only takes a lot of time, but also often has many subjective errors.By realizing the automatic classification of documents in the field, not only can a large amount of human resources be saved, the error of human factors can be reduced, the efficiency and accuracy of document sorting can be improved, and the reuse of knowledge can be expanded, and the interaction items and stability of building documents can be improved. Timeliness.On the other hand, aiming at the fuzzy phenomenon of document classification categories in the field of building construction, a multi-level classification management idea is proposed, which can improve the efficiency of document classification.Therefore, the research on automatic classification of documents in the field of building construction has certain positive significance for the field of building construction.

Acknowledgements

This research is supported by the National Natural Science Fund (NO. 61762076).

References

[1]. M. Ghiassi, M. Olschimke, B. Moon, P. Arnaudo. Automated text classification using a dynamic artificial neural network model[J]. Expert Systems with Applications,2012,39(12).

[2]. Fabrizio Sebastiani. Machine learning in automated text categorization[J]. ACM Computing Surveys (CSUR),2002,34(1).

[3]. Marina Sokolova,Guy Lapalme. A systematic analysis of performance measures for classification tasks [J]. Information Processing and Management,2009,45(4).

[4]. Yan Yan,Xu-Cheng Yin, Sujian Li, Mingyuan Yang, Hong-Wei Hao, Pasi A. Karjalainen. Learning Document Semantic Representation with Hybrid Deep Belief Network[J]. Computational Intelligence and Neuroscience,2015.

[5]. Collobert R, Weston J, Bottou L, et al, Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12(1): 2493-2537.

[6]. Zhang Chong. Research on text classification technology based on Attention-Based LSTM model [D]. Nanjing University, 2016.

[7]. Zhou Chao. Text classification based on deep learning hybrid model [d]. Lanzhou University, 2016.

[8]. Zhou Lian. The working principle and application of Word2vec [J]. Scientific and Technological Information Development and Economy, 2015, 25 (02): 145-148.

[9]. Yang Jieming. Text representation model and feature selection algorithm in text classification [d]. Jilin University, 2013.

[10]. Yan Wei. Research on text representation and classification based on deep learning [d]. Beijing University of Science and Technology, 2016.

[11]. Liang Bin, Liu Quan, Xu Jin, Zhou Qian, Zhang Peng. Specific target sentiment analysis based on multi-conflict convolutional neural network[J]. Computer Research and Development, 2017, 54(08): 1724-1735.

[12]. Wu Yuya, Wei Miao. Reviewing the method of natural language processing word embedding from deep learning [j]. Computer Knowledge and Technology, 2016, 12(36): 184-185.

[13]. Tang Ming, Zhu Lei, Zou Xianchun. A Document Vector Representation Based on Word2Vec [J]. Computer Science, 2016, 43(06): 214-217.