Research on Organization Name Matching Based on Word Vector

KaiLi Song¹,a, YunLing Li², LuLu Yao³, Yuan Wang⁴

¹Shandong University of Science and Technology, Qingdao, ShanDong, China
²Shandong University of Science and Technology, Qingdao, ShanDong, China
³Shandong University of Science and Technology, Qingdao, ShanDong, China
⁴Shandong University of Science and Technology, Qingdao, ShanDong, China

anetnews@sdust.edu.cn
1515138537@qq.com

Abstract—Organization names usually exist in the form of abbreviations, aliases, etc. in daily use, which brings great difficulties to online search and matching in the Internet era. Through extensive analysis of Chinese organization name data and research on word embedding technology based on neural network, this paper proposes a BERT model based on organization name field data training to achieve organization name matching. The experimental results show that the accuracy is improved by 17% compared with the method based on edit distance.

1. INTRODUCTION

Natural language processing, as a sub-field of artificial intelligence and linguistics, is mainly to study how to use computers to process human language, namely the sound, shape and meaning of natural language, and realize information exchange between humans and computers. Its related technologies have been applied in machine translation, text classification, text summarization, question answering systems and other fields. The prerequisite for the development of natural language processing is to understand human language, which makes natural language processing one of the most difficult areas of artificial intelligence, because compared with images and speech, human language is more abstract, allowing computers to understand. Before processing natural language, we need to let the computer learn the social culture, human habits and other knowledge behind different languages, and the learning of these knowledge has certain difficulties and challenges.

Organizations, as a general term for agencies, groups, or other enterprises and institutions, include many categories, including government agencies, such as "Qingdao Meteorological Bureau" and "Qingdao Municipal People's Government"; schools, such as "Shandong University of Science and Technology", "East China Normal University"; corporate enterprises, such as "China UnionPay Headquarters", "Qingdao Zhengdong Construction Group"; medical institutions, such as "Qingdao University Hospital", "Huangdao District Traditional Chinese Medicine Hospital"; non-governmental organizations, such as "Pangjia Town Committee of Boxing County of China Education Union", "Red Cross Society of Zhonggong Town, Jinan City" and so on. In our daily life, the use of organization names is very common. Whether it is the establishment of the country’s government agencies, the
organizational structure of the government, or the business transactions in the market economy, the daily operation of various enterprises and institutions requires the cooperation of various institutions. Even our most daily necessities, food, housing, and transportation are also separated. Do not open related agencies. Therefore, the use of organization names has become a necessity.

Organization names are widely used in people's daily life, with the characteristics of large number, wide coverage, complex composition, etc., and with the continuous expansion of social and economic development, new terms have been produced [1]. The names of some organizations with a large number of characters, such as "The First Institute of Oceanography of the Ministry of Natural Resources", often exist in the form of abbreviations such as "One Institute of Oceans" and "One Institute" or common names in different contexts. When a user uses a retrieval system to retrieve the name of an organization, he often enters its abbreviation or other types of common names. These irregular inputs make natural language processing difficult. In addition, the unique grammatical structure of Chinese also greatly increases the complexity of the organization's abbreviations. The latest research is an improved matching algorithm based on edit distance proposed by Huang Linsheng [2], which uses "words" as the basic unit of editing operations. The matching accuracy of this algorithm is basically stable between 60% and 70%. Therefore, how to quickly and accurately retrieve the matching target institution in the results, this research has a wide range of application value. Based on the superiority of the current deep learning model, this paper converts the organization name matching into the similarity calculation of word vectors. After the model training of professional field data, the organization name is converted into the representation form of word embedding, and the training model is continuously optimized to improve the search and matching accuracy of the Chinese organization name.

2. BERT MODEL

2.1. BERT model principle

Traditional text matching methods include algorithms such as VSM, TF-IDF, BM25, etc. These methods mainly solve matching problems at the literal level. For example, the BM25 algorithm calculates the matching score between the two through the coverage of the network field to the query field. The higher the score, the better the match between the webpage and the query. However, this kind of literal matching cannot explain that "taxi" and "taxi" mean the same thing. Although the words "machine learning" and "learning machine" contain the same text, they are different. Meaning. This shows that the task of text matching should not only stay at the level of literal meaning, but also consider semantic issues.

With the continuous development of deep learning and its successful application in the fields of computer vision, speech recognition and recommendation systems, in recent years, many scholars have also applied deep neural network models to various tasks in the field of NLP to reduce artificial features. influences. The text matching calculation is based on the word vector representation trained by the neural network. Its training method is more concise, and the semantic computability of the word vector representation obtained is further enhanced. In 2018, Google released a deep language model based on a large amount of unsupervised data training-the BERT model [3], which uses a two-way Transformer as an encoder to extract deep-level text semantic information. Since its inception, BERT has made new progress in multiple tasks in natural language processing, and has shown excellent information extraction capabilities, including text matching, text classification, sentiment analysis and other fields. As a deep learning model, the BERT model has shown amazing results in the top level test of machine reading comprehension SQuAD1.1: It surpassed humans on all two metrics, and also achieved the best results in 11 different NLP tests, including pushing the GLUE benchmark to 80.4% (absolute improvement of 7.6%) and MultiNLI accuracy of 86.7% (absolute improvement rate) 5.6%) etc.

The advantage of the BERT model comes from the fact that it is based on the encoder part of Transformer [4], which is a model constructed using only the multi-head self-attention mechanism. Researchers believe that this mechanism can more directly capture the relationship between words, so
that the coding of the sequence is more integrated, and it can better represent the meaning of the entire sequence. It is a deep model suitable for NLP research tasks.

![Figure 1 BERT model](image)

Attention [5] has been widely used in various deep learning tasks in recent years, including computer vision, natural language processing, etc. The Attention mechanism is essentially a problem-solving method proposed by imitating human attention. The core goal is to quickly filter out high-value information from a large amount of information. It assigns a weight coefficient to each element in the sequence. If each element in the sequence is stored in the form of (K, V), the attention is addressed by calculating the similarity between Q and K. The similarity calculated by Q and K reflects the importance of the extracted V value, that is the weight, and then the weighted summation obtains the attention value.

Self-Attention is one of the attention mechanisms and a special form of it. Different from the attention mechanism, it does attention inside the sequence, and uses the sequence itself and itself for attention processing. The special point in the KQV model is that Q=K=V. For example, input a sentence, then each word in it must be calculated with all the words in the sentence, the purpose is to learn the word dependency relationship within the sentence and capture the internal structure of the sentence. In short, the Self-Attention mechanism obtains a new representation of each word considering the context information. Multi-headed self-attention mechanism (Multi-headed self-attention) is a derivative mechanism of the self-attention mechanism, which is multiple independent self-attention calculations as an integrated function to prevent overfitting.

2.2. Fine-tuning of the BERT model

The full name of BERT is Bidirectional Encoder Representations from Transformer, which is a two-way encoder representation based on Transformer. As the name suggests, BERT uses Transformer, and when processing a word, it can also consider the words before and after the word to get its meaning in the context. We know that Transformer's attention mechanism has a very good effect on feature extraction of words in the context, and intuitively, considering the two-way encoding of the context is more than considering only the one-way effect of the above (or below) it is good.

In the BERT model, the input of a word or sentence is the superposition of token embeddings, segmentation embeddings, and position embeddings, that is, the embedding of each word is the superposition of three embeddings. Token embeddings use WordPiece embedding, which is realized through the vocab.txt word list; while positional embeddings are used to represent the position information of words in a sentence; segment embeddings are for the whole sentence, and different sentences have different items. It uses To achieve sentence-level tasks.
BERT is essentially a two-stage NLP model. The first stage is called: Pre-training, which is similar to WordEmbedding, using the existing unlabeled corpus to train a language model. The main innovation of the BERT model is the pre-training method, that is, the Masked LM and Next Sentence Prediction methods are used to capture the word and sentence-level representation respectively[6].

2.2.1. Masked LM
When training BERT with the Masked Language Model method, randomly mask 15% of the words in the corpus. The mask operation for these 15% of the words is divided into three situations: 80% of the words are directly replaced with [Mask], 10% of the words are directly replaced with another new word, and 10% of the words remain unchanged. In the training process, 15% of the labels in each sequence are randomly masked. Unlike CBOW in Word2Vec, which predicts every word, Masked LM masks some words randomly from the input, and its goal is to predict based on context. The original vocabulary of the masked word. Unlike the pre-training of the language model from left to right, the representation learned by Masked LM can fuse the context of the left and right sides. The two-way Transformer in the model does not know which words it will be asked to predict, or which words have been replaced by random words, so it must maintain a distributed contextual representation for each input word. In addition, random replacement only occurs in 1.5% of all words, which does not affect the model's understanding of the language.

2.2.2. Next Sentence Prediction
Many sentence-level tasks in natural language, such as automatic question answering (QA) and natural language inference (NLI), require understanding of the relationship between two sentences. For example, in the above Masked LM task, after the first step of processing, 1.5% of the vocabulary is covered, then in this task, the data needs to be randomly divided into two parts of the same size, and the two sentence pairs in one part of the data are The context is continuous, and the two sentence pairs in the other part of the data are discontinuous. Then let the Transformer model identify these sentence pairs and determine whether the next sentence is continuous with the current sentence.

The second stage is called: Fine-tuning, using pre-trained language models to complete specific NLP downstream tasks, such as: text classification, similarity judgment, reading comprehension, etc. This article applies the bert model to institution name matching. Specific tasks.

3. EXPERIMENT AND RESULT ANALYSIS

3.1. Experiment procedure

3.1.1. Full name data download
A total of 422,128 institution names in Shandong Province were obtained through AutoNavi API, and 100,000 of them were randomly selected to establish a database of institution names.

3.1.2. Abbreviation data acquisition
Part of the organization name abbreviation data is crawled through web pages, mainly using Baidu Baike information. Then analyze the structure characteristics of the full name of the organization and the characteristics of the abbreviation of the organization name, summarize the rules for generating the abbreviation of the organization name, and obtain the organization name abbreviation data.

3.1.3. Model training
The processed full name data and abbreviated data are formed according to the data requirements of the BERT model to form corresponding train.csv, test.csv, dev.csv and other data. Set training parameters such as Batch size, Learning rate, and epochs, Use the base version released by Google as a pre-training model, leaving the number of network layers, hidden layers, and self attention heads unchanged.
Constantly optimize the parameters during the training process to complete the fine-tuning on the task of matching organization names.

3.1.4. **Organization name matching**

Using the trained BERT model in the field of organization names, define a BertVector class, generate the word vector representation of the organization name, and then use the cosine similarity calculation method. Use the python programming language to define the cosine_similarity function, get the cosine value of the input word and the full name of the institution in the database, and then do a result sorting, and return the matching result with the highest degree of similarity.

3.2. **Result analysis**

When using the pre-training model released by Google for fine-tuning, the accuracy and loss value are selected as the evaluation indicators for model training, which can simply and intuitively reflect the quality of the model. After continuous optimization of parameter settings, the final accuracy is 0.98, auc is 0.947, and the loss value is 0.083. The values of these evaluation indicators can ensure the reliability of the model. After that, different numbers of input words are sequentially matched with the organization name data. After statistics, the accuracy of the model for organization name matching is basically stable at about 87%. Compared with the method based on improved edit distance, it has been greatly improved.

4. **CONCLUSION**

This paper uses the pre-training model chinese_L-12_H-768_A-12 publicly released by Google and the data of Shandong province institution name to train the BERT model, and continuously optimizes the accuracy of the model training to be 0.98. Then we use the trained BERT model to perform the word embedding representation of the institution name and the calculation of the cos value, and traverse all the data for sorting, and return the optimal result. The statistical accuracy can be maintained at about 87%. The experimental results show that the BERT model is very advantageous for organization name matching.

**REFERENCES**

[1] Lian Yushun. Design and Implementation of the Chinese Organization Name Retrieval System [D]. Shanghai Jiaotong University, 2013.

[2] Huang Linsheng, Deng Zhihong, Tang Shiwei, Wang Wenqing, Chen Ling. Chinese organization name abbreviation-full name matching algorithm based on edit distance [J]. Journal of Shandong University (Science Edition), 2012, 47(05): 43-48.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. ArXiv: Computation and Language, 2018,23 (2):3-19.

[4] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[5] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[6] Wang Ziniu, Jiang Meng, Gao Jianling, Chen Yaxian. Chinese named entity recognition method based on BERT[J]. Computer Science, 2019, 46(S2): 138-142.