Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 1
AIEA 2023

Text Processing Method based on Natural Language Processing

-- Taking Two Model as Examples

Shiyi Yu and Jialin Ji *

School of law, Shenzhen university, Shenzhen, China

* Corresponding author Email: 2020031082@email.szu.edu.cn

Abstract. Today, natural language processing is booming and is widely used in different fields, especially for text processing. In many Chinese natural language processing tasks such as information retrieval, named entity recognition, syntactic analysis, etc., word segmentation and part-of-speech tagging are often the first steps. This paper mainly selects sequence labeling model and BERT pre-training model, specifically studies how to use natural language processing technology for text processing, and analyzes and summarizes the advantages and disadvantages of the two models.

Keywords: Chinese Natural Language Processing Tasks; Sequence Labeling Model; BERT Pre-training Model; Text Processing.

1. Development and Use of Natural Language Processing

Natural language processing (NLP) is a subfield of artificial intelligence that aims to enable machines and humans to communicate using natural languages, such as English, Chinese, Spanish, etc. NLP involves the analysis, understanding and generation of natural language, as well as the interaction with humans in a human-like manner. This technology is a significant direction in the field of artificial intelligence, and has various applications and impacts on different domains.

NLP has a long history, tracing back to the 1950s. Since then, NLP has undergone three stages, namely rule-based methods, statistical methods and deep learning methods. Rule-based methods rely on manually crafted rules and dictionaries to process natural language, which are effective for specific domains but have limited scalability and robustness. Statistical methods use probabilistic models and machine learning algorithms to learn from large-scale corpora, which can handle more complex and diverse natural language phenomena but require a lot of annotated data and computational resources. Deep learning methods use neural networks and representation learning to automatically learn from raw data, which can achieve state-of-the-art performance on many NLP tasks but are often difficult to interpret and explain.

NLP also encompasses a variety of tasks, such as machine translation, text classification and proofreading, information extraction, speech synthesis and recognition, etc. These tasks are all for the purpose of achieving effective communication between humans and machines or between humans. Machine translation is the task of automatically translating text or speech from one language to another, which can facilitate cross-language communication and information exchange. Text classification and proofreading are the tasks of assigning labels or correcting errors for text documents, which can help users organize and improve their texts. Information extraction is the task of extracting structured information from unstructured text, such as entities, relations, events, etc., which can help users access and analyze relevant information. Speech synthesis and recognition are the tasks of converting text to speech or speech to text, which can enable users to interact with machines using voice.

Currently, NLP is applied to many fields, such as search engines, social media, e-commerce, education, health care, finance, etc. The development and application of NLP brings many benefits. First, it improves the efficiency and quality of information acquisition and retrieval, allowing users to find the information they need faster and more accurately. For example, search engines use NLP to understand user queries and provide relevant results. Second, it realizes cross-language and multilingual

communication and cooperation, promoting understanding and communication between different cultures and countries. For example, social media platforms use NLP to translate user posts and comments into different languages. Third, it promotes the development of artificial intelligence, making computers have higher levels of intelligence and cognitive abilities. For example, e-commerce platforms use NLP to generate product descriptions and recommendations based on user preferences.

NLP is a fascinating and challenging field that has great potential for the future. However, it also faces some limitations and risks that need to be addressed. For instance, NLP may encounter ethical issues such as privacy protection, bias detection and fairness enhancement. Moreover, NLP may require more human involvement and collaboration to ensure its quality and reliability. Therefore, it is important to develop NLP in a responsible and sustainable way that can benefit both humans and machines.

2. Text Processing Models based on Natural Language Processing

2.1. Sequence Labelling Model

Sequence labeling is an important research topic in natural language processing tasks, whose goal is to predict the corresponding labels based on a given text sequence. [1]

The basic principle of sequence labeling models is to predict the label at each position in the output sequence, which is achieved by annotating the input sequence. Models usually use annotated data for training, where each input sequence is paired with the corresponding output sequence label. The model learns to recognize patterns in the input text and predicts the appropriate label at each position in the output sequence. In sequence labeling tasks, common algorithms include hidden Markov models (HMMs), conditional random fields (CRFs), and recurrent neural networks (RNNs), etc. Sequence labeling models usually need to consider the context information of the input text, in order to more accurately predict the label at each position. Some models will add additional context information, such as surrounding words or sentence structure, to improve the accuracy of prediction. When training sequence labeling models, a large amount of annotated data is needed for training, and these annotated data need to accurately reflect the situation in the real world.

The advantages of sequence labeling models are that they can usually process a large amount of text data in a short time, and can predict the label at each position according to the context information of the input text, thereby improving the accuracy of annotation. In addition, some sequence labeling models can provide interpretable results, allowing users to understand how the model makes predictions. However, the disadvantages of sequence labeling models are that they have high requirements for annotated data, and the model needs a large amount of annotated data for training, and these annotated data need to accurately reflect the situation in the real world, and need to consider the context information of the input text, so they may be affected by incomplete or erroneous context information. Secondly, they may encounter difficulties when dealing with long sequences, because they need to consider the context information of the entire sequence. Moreover, sequence labeling models can usually only handle known words or phrases, and cannot handle unknown words or phrases.

2.2. B.BERT Pre-training Model

BERT models are a family of language models introduced in 2018 by researchers at Google, which stand for Bidirectional Encoder Representations from Transformers [2]. BERT models are based on the transformer architecture, which uses attention mechanisms to encode and decode natural language. BERT models are pre-trained on large-scale corpora using two objectives: masked language modeling and next sentence prediction. BERT models have achieved state-of-the-art results on many natural language processing tasks, such as question answering, natural language inference, sentiment analysis, etc. However, BERT models also have some advantages and disadvantages that need to be considered.

One advantage of BERT models is that they can capture the context and dependencies of natural language from both left and right directions, which can improve the accuracy and quality of natural language understanding. For example, BERT models can disambiguate the meaning of words based

on their surrounding words, such as “bank” in “He went to the bank” versus “He sat on the bank”. Another advantage of BERT models is that they can leverage the large amount of unlabeled data available on the web, which can enhance the generalization and robustness of natural language processing. For example, BERT models can learn from diverse domains and genres of text, such as books, Wikipedia, news, social media, etc. A third advantage of BERT models is that they can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, without substantial task-specific architecture modifications. For example, BERT models can be fine-tuned for question answering by adding a span prediction layer on top of the pre-trained model.

One disadvantage of BERT models is that they require a lot of computational resources and time to pre-train and fine-tune, which can be costly and inaccessible for some researchers and practitioners. For example, BERT-Base has 110 million parameters and BERT-Large has 340 million parameters, which require 16 TPU chips to pre-train for four days on a corpus of 3.3 billion words (Devlin et al., 2019). Another disadvantage of BERT models is that they may encounter ethical issues such as privacy protection, bias detection and fairness enhancement, especially when dealing with sensitive or personal data. For example, BERT models may inadvertently reveal private information or discriminate against certain groups based on their representations or predictions [3]. A third disadvantage of BERT models is that they may be difficult to interpret and explain, especially when using complex or black-box models, which can affect the trust and reliability of natural language processing. For example, BERT models may produce unexpected or erroneous outputs that are hard to justify or debug [4].

3. Analysis and Conclusion

3.1. Lexical Analysis based on Sequence Annotation Model for Chinese

Sequence annotation refers to the process of putting an input sequence of observations $x_1x_2x_3...x_n$ into a string of marker sequences $y_1y_2y_3...y_n$. To solve the sequence labeling problem, it is actually to find a mapping of observation sequences to labeled sequences $f(x_i)\rightarrow y_i$ (i=1,2,3...n) [5]

In Chinese lexical analysis, since there is no explicit word boundary in Chinese, word segmentation is required before part-of-speech labeling. Word segmentation and part-of-speech tagging are the first step in many Chinese natural language processing tasks such as information retrieval and syntactic analysis. There are two main types of statistical Chinese word segmentation, one is word-based Chinese word segmentation, which marks each word in the sentence with a category that identifies the word boundary, and Chinese the word segmentation is converted into a sequence labeling problem of word sequence. Another type of method is word-based Chinese word segmentation, and the part-of-speech labeling problem is also a typical sequence labeling problem.

In this paper, the traditional serial model of the sequence annotation model is used to carry out the Chinese word segmentation analysis experiment, and the word segmentation and part-of-speech labeling are divided into two stages. In general, the process of the system performing word segmentation-part-of-speech labeling of sentences can be summarized into two main parts, namely feature extraction (FE) and prediction (PD).

First, set up a word breaker and part of speech tagger, and use them to form a word segment-part of speech tag serial model. The tokenizer delineates word boundaries for each word in Chinese sentence, such as A (indicating the beginning of the word), B (indicating the middle of the word), C (indicating the ending), and D (indicating the self-contained word), Then a word sequence can be expressed as {A, B, C, D}. Then the part-of-speech annotator delineates the part of speech for each word, including but not limited to pronouns (P), Prepositional (Prep), nouns (N), verbs (V), adjectives (J), etc.

Next, take the phrase "I want to go home" as an example to conduct simple participle and part-of-speech labeling experiments.

a. Word segmentation stage:

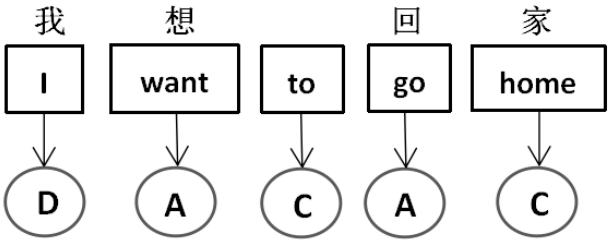

Figure 1. Word segmentation stage

b. Part of speech tagging stage:

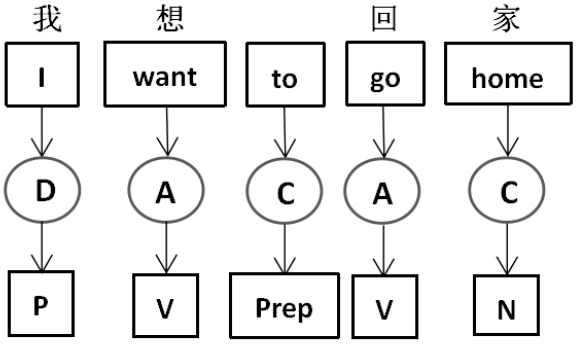

Figure 2. Part of speech tagging stage

In the two-stage sequence annotation serial model described above, the first stage divides the given sentence into a sequence of words with reasonable semantics. In the word segmentation problem, the "word" of the sequence node corresponds to each word in the sentence, and the label space of the node is {A, B, C, D}. This article takes the Chinese sentence "I want to go home" as an example, and the label space of the sequence node is {D, A, C, A, C,}, which indicates the beginning, middle, ending, or separate words, respectively. The second stage is to label parts of speech. In the sentence "I want to go home", there are pronouns, verbs, prepositions, nouns and other parts of speech.

Since the part-of-speech stage and the part-of-speech labeling stage are carried out independently, the two stages are separated. Errors in the previous stage affect the accuracy of the later stage, which greatly reduces the performance of Chinese lexical analysis.

3.2. Application Analysis of BERT Pre-training Model

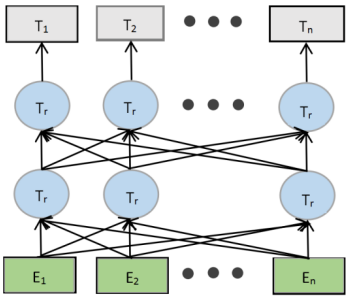

Figure 3. The neural network architecture diagram of the model

BERT (Bidirectional Encoder Representations from Transformers) is a model proposed by Google AI language researchers in 2018, it is "a deep learning model built on a bidirectional transformer basis for masked language model and next sentence prediction tasks." [6] In the BERT Chinese pre-training model, Chinese characters are processed in a characterized way and pre-trained using an open-domain corpus, the neural network architecture diagram of the model is shown in Figure 3. The arrows in the

figure represent the flow of information from one layer to the next, and the sequences T_1, T_2,⋯, T_n represent the final context representation of each input word.

From a macro point of view, Transformer is a Seq2seq model (sequence model) based on the Self-Attention mechanism, it consists of an encoder and a decoder. As shown in Figure 4, x_1, x_2, x_3...x_n represents the input sequence, the encoding phase converts the sequence into a fixed-length vector. And y_1 , y_2, y_3 ... y_n represents the output sequence, that is, after the decoding stage, the previously generated fixed vector is converted into an output sequence.

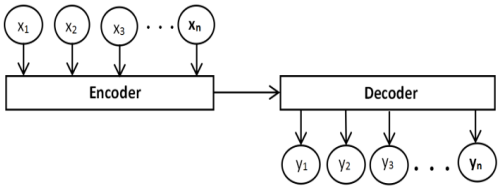

Figure 4. Transformer

The BERT pre-training model structure is the Encoder layer of the Transformer, which has the nature of a Transformer. As shown in the figure below (plotted in Figure Several), the input sequence with length n is represented by three different Embeddings vectors, and the last three are summed to complete the coding input link of the BERT pre-training model.

The characteristics of these three embeddings are:

a.Token Embedding: dividing words into a limited set of common word units, converting individual words into fixed-dimensional vectors. In BERT, each word is converted into a 768-dimensional vector representation.

b.Segment Embedding: Used to distinguish between two sentences, such as whether B is the following of A.

c.Position Embedding: Encode the position information of words into feature vectors, Position embedding can effectively introduce the position relationship of words into the model and improve the model's ability to understand sentences.

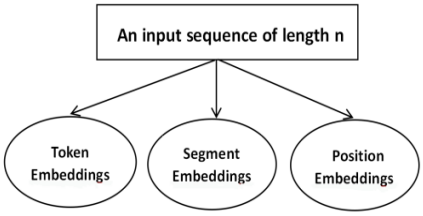

Figure 5. The BERT pre-training model structure

Furthermore, the BERT pre-training method can be divided into two modules, one is to use Masked LM to learn the representation of words in context; The second is to use Next Sentence Prediction to learn sentence-level representation.

In summary, the general flow of the BERT pre-training model is as follows:

The above are the two major models often used in natural language processing (NLP): sequence labeling model and BERT pre-training model, with the help of these two models, a number of word processing tasks can be completed, such as part-of-speech tagging, named entity recognition, and so on. This paper mainly introduces the operation mechanism of these two models, explains the word processing methods, and analyzes the advantages and disadvantages of different processing methods.

This paper introduces two text processing methods based on natural language processing technology, namely sequence labeling model and BERT pre-training model. Sequence labeling model is a method to predict labels for each position in a given text sequence, which can be used for tasks such as word

segmentation, part-of-speech tagging, named entity recognition, etc. BERT pre-training model is a method to learn deep bidirectional representations from large-scale unlabeled text data, which can be used for tasks such as natural language understanding, question answering, text summarization, etc. The paper analyzes the principles, advantages and disadvantages of these two models, and compares them with other models. The paper concludes that sequence labeling model and BERT pre-training model are both effective and powerful methods for text processing, but they also have some limitations and challenges that need to be addressed in the future.

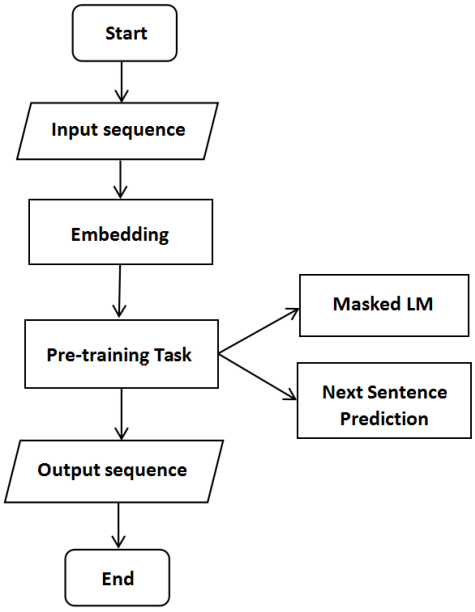

Figure 6. The general flow of the BERT pre-training model

Acknowledgments

This thesis is funded by the following projects:

1. Research on the Application of Competency Model Theory in Legal Education for Digital Society (Guangdong Provincial Education Science Planning Leading Group Office Project No. 2021GXJK231)

2. Research on the integration of intellectual property law teaching with ideological elements of the curriculum (Department of Academic Affairs, Shenzhen University, Project No. 000030110214) Instructor Zhang Xianwei

References

[1] REN Zhihui, XU Haoyu, FENG Songlin, et al. Sequence labeling Chinese word segmentation method based on LSTM networks[J]. Application Research of Computers, 2017, 34(5): 1321-1324.

[2] LIN Jiarui, CHENG Zhigang, HAN Yu, et al. Classification method of disaster tweet based on BERT pre-training model [J]. Journal of Graphics,2022,43(03):530-536.

[3] Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 591-598).

[4] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[5] WEI Xinyang, TANG Xianghong. Research on Extractive Judgment Document Abstract Generation Method based on BERT[J]. SOFTWARE ENGINEERING,2022:1-4.

[6] LIN Jia-rui, CHENG Zhi-gang, HAN Yu, YIN Yun-peng. Disaster tweets classification method based on pretrained BERT model[J]. JOURNAL OF GRAPHICS,2022,43(3):530-536.