Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230297

Language sense classification model based on neural network

Letao Gu[1,4,†], Yuxiang Wang [2,†], Yihan Wu[3,†]

[1]School of artificial Intelligence, Guilin University of Electronic Technology, Guilin, Guangxi,541004, China
[2]College of Information Technology, Shanghai Jian Qiao University, Shanghai,201306, China
[3]Information Engineering Institute,Henan University Minsheng College, Kaifeng, Henan,475001, China

[4]2001630407@mails.guet.edu.cn
[†]These authors contributed equally.

Abstract. The common international language, English, is playing an increasingly important role in various fields with the rapid development of artificial intelligence in recent years. Artificial intelligence can improve students' English abilities as an additional teaching tool. Therefore, this study seeks the English language sense between different types of sentences based on Long Short-Term Memory (LSTM) and BERT model analysis sentences and generates a model to distinguish the types. This paper adapts the LSTM model and BERT model: first, this paper crawls the sentences from British Broadcasting Corporation (BBC) documentaries, podcasts, and YouTube and then constructs a data filter to remove the sentences with low quality and short. This paper analyzes the data set through the BERT module and LSTM model. this paper then compares the differences between different sentences in a large-scale corpus to generate a language model without long-term dependence. A model is expected to be generated after corpus analysis, and the model can be used to analyze new input statements and give their types. This study can help English learners improve their sense of the English language and the types of sentences they need to say in the face of different situations.

Keywords: natural language process, LSTM, BERT, classification, language sense.

1. Introduction

With the continuous maturity of artificial intelligence technology in recent years, users can fully utilize internet resources to collect as much excellent knowledge and content as possible, gather better learning methods, and form a comprehensive knowledge and skills-sharing platform [1, 2]. In addition, artificial intelligence, as an auxiliary teaching tool, can improve students' learning efficiency and enhance their learning ability. As an international language, English is increasingly important in various fields. English education in China has always been a focus of attention. Although in big cities and developed areas, English education has been highly valued and supported with high-quality resources, in some underdeveloped areas, English education still needs to catch up, and many students encounter many difficulties in English learning. These difficulties may lead to their inability to improve their English language sense, thus affecting their future career development. Therefore, improving the English

DOI: 10.54254/2755-2721/27/20230297

language sense in underdeveloped areas of China's education has become a critical issue. This can help students in these areas better adapt to future social development and improve their competitiveness in international communication and competition.

With the continuous development of artificial intelligence technology, natural language processing technology has become a good way to improve English learning. In recent years, chatbots have been widely used in various fields. Chatbots can interact with users in language through natural language generation and understanding technology and provide personalized learning and entertainment services [3-5]. Therefore, this study analyzes sentences based on long-term and short-term memory (LSTM) and BERT models, seeks the English language sense between different types of sentences, and generates a model to distinguish these types. Help students in underdeveloped areas of China's education improve their English language sense and achieve personalized English learning.

The significance of this study is to explore the application of natural language processing technology based on BERT and LSTM models in improving English language sense, which is conducive to the research of natural language processing methods. Exploring the application of natural language processing technology based on BERT and LSTM models in improving English language sense has certain significance for the scientific research of natural language processing.

This paper analyzes sentences based on long-term and short-term memory (LSTM) and BERT models, seeks the English language sense between different types of sentences, and generates a model to distinguish these types of sentences [6-8]. This study will use the BERT model for natural language understanding and the LSTM model for natural language judgment. It will provide personalized English learning services by analyzing students' learning history and language habits and help students improve their English language sense.

The methods used in this study include: 1) Establishing a dataset for logical processing and data storage management. 2) Utilizing the BERT model for natural language understanding to extract students' learning history and language habits. 3) Using the LSTM model for natural language generation to produce personalized English learning content and interactive responses.

2. Data collection and pre-processing

In the data collection, this paper selects three kinds of data sets crawled from YouTube, BBC, and Podcast, processes the crawled corpus, and eliminates sentences with less than 20 words in the data. YouTube corpus comes from downloading English and Chinese subtitles of various anchor videos on YouTube, with a total of 5,000 pieces of data. BBC Corpora is derived from sentence crawling in many BBC documentaries, with 5000 data pieces. The podcast corpus comes from downloading the original subtitles of the video in the Podcast, with a total of 5000 data. A total of 15,000 pieces of data, 12,000 pieces of data for model training, and the rest for testing. Table 1 indicates the average sentence duration of the dataset.

Table 1. The average sentence length of the dataset.

Dataset	Average Length
BBC Corpora	24.89
YouTube Corpora	27.05
Podcast Corpora	25.69

The longest sentence in the YouTube data set is shown in Figure 1, which has 125 words.

"Funding for this program is provided by Additional Funding provided by this is course about Justice, and we begin with a story. Suppose you're the driver of a trolley car, and your trolley car is hurdling down the track at sixty miles an hour, and at the end of the track, you notice five workers working on the track you tried to stop but you can't your brakes don't work you feel desperate because you know that if you crash into these five workers they will all die let's assume you know that for sure and so you feel helpless until you notice that there is off to the right a side track at the end of that track there's one worker working on track you're steering wheel works so you can turn the trolley car if you want to onto this side track killing the one but sparing the five."

Figure 1. YouTube longest sentence.

DOI: 10.54254/2755-2721/27/20230297

The longest sentence in the BBC data set is shown in Figure 2, which has 105 words.

> *" I was a 10-year-old and one day I happened to be looking in my local public library and I found a book on math and it told a bit about the history of this problem, that someone had resolved this problem 300 years ago, but no-one had ever seen the proof, no one have ever seen a prove, and people ever since have looked for the proof and here was a problem that I, a 10-year-old, could understand, but none of the great mathematicians in the past had been able to resolve, and from that moment of course I just tried to solve it myself ."*

Figure 2. BBC longest sentence.

The longest sentence in the Podcast data set is shown in Figure 3, which has 65 words.

> *"Archive Recommended How to Listen About Overview Staff Announcements Fellowships Jobs Music Make Radio On The Road FAQ Submissions Store Contact Us Our Other Shows Store Contact Serial S-Town © 1995 - 2023 This American Life Privacy Policy | Terms of Use If you are able, we strongly encourage you to listen to the audio, which includes emotion and emphasis that's not on the page."*

Figure 3. Podcast longest sentence.

3. Methods

BERT is a pre-trained deep learning model that can learn context representation from many texts and achieves excellent performance in multiple natural language processing tasks. LSTM is a kind of cyclic neural network, which can capture the time dependence of text, and has achieved good results in the task of text classification. Therefore, this paper studies the reasons for using BERT and BERT+LSTM models to classify language sense. The data sets were tested separately on the two models, and accuracy and loss were used to evaluate the performance of the models. The experimental results show that BERT and BERT+LSTM models perform well in English language sense classification.

3.1. BERT

BERT is a pre-trained language model, which is implemented based on Transformer [9]. BERT contains many Transformer modules, which can be understood as a neural network module with complex network structures inside the module. This module realizes fast parallel through the self-attention mechanism. It improves the most criticized shortcoming of slow training of RNNs. It can be increased to a profound depth to fully explore the DNN model's characteristics and improve its accuracy. Figure 4 shows the BERT model structure.

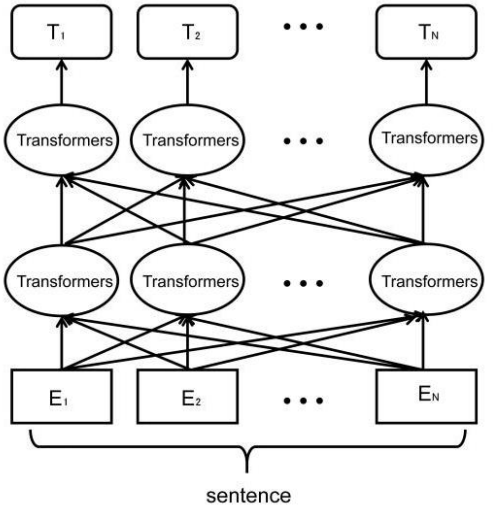

Figure 4. BERT model structure.

DOI: 10.54254/2755-2721/27/20230297

Because BERT is a pre-training model, it must be suitable for integration into various natural language projects, so the input sequence of this model must contain one sentence (emotional classification of text, sequence marking tasks) or more than two sentences (text summarization, natural language reasoning, question answering tasks). So, how can the model determine the scope to which sentence A and sentence B belong? BERT adopts two methods to solve this problem: 1) To separate different sentences called TOKENS, insert a marker ([SEP]) after each sentence. 2) Point out each TOKEN and add learnable regions to embed to indicate whether it belongs to sentence A or B.

BERT is an end-to-end model that does not require the user to adjust the network structure, and the user adds an output layer at the end specific to downstream tasks. Based on Transformer, fast parallelism can be realized. It can also be increased to an intense depth to fully explore the characteristics of the DNN model and improve model accuracy.

BERT has two sizes, the base version has 110M parameters, and the large version has 340M. In other words, whether base or large, the number of BERT parameters is hundreds of millions, which is quite large. The reason why Bert is chosen as a pre-training model. It is because he does not need to use a large amount of corpus for training, which saves time, is efficient, and has strong generalization ability.

3.2. Long short term memory

To address the issues of gradient fading and gradient exploding during lengthy sequence training, a special sort of RNN called long short-term memory was developed [10]. LSTMS, in contrast to conventional RNNS, regulates the transmission status by triggering the status, remembering the information that needs to be remembered over a long period of time, and forgetting the information that doesn't matter that can perform better in longer sequences. The LSTM was specifically created to prevent issues with long-term reliance.

The structure of all RNNs is a series of repeating neural network modules. This repeated module has a simple structure in a typical RNN. The repeating modules have a different structure than LSTM, which has a similar structure. The structure of the repetition modules is different from the LSTM, which has a similar structure. The four neural network layers interact in a very specific way rather than one. Figure 5 shows the LSTM model structure.

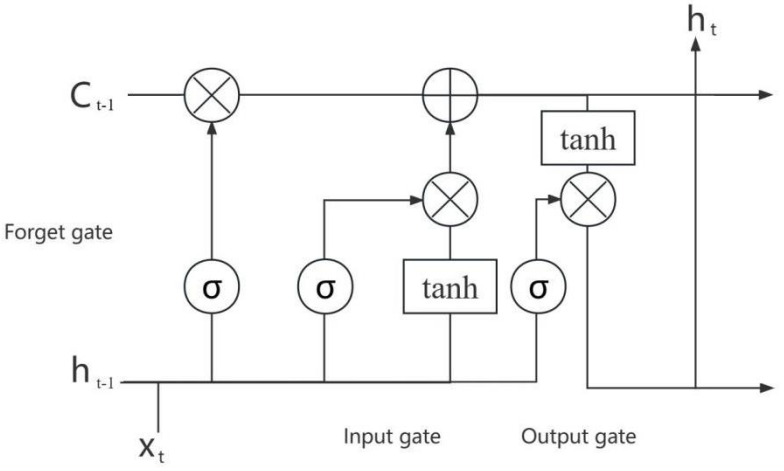

Figure 5. LSTM model structure.

LSTM is controlled by three gates, called forget gate, input gate, and output gate: 1) Forgetting phases. At this step, we decide what data to remove from the cell state and selectively forget the input from the preceding node. Put, "forget what's not important and remember what's important." The forget gate, a sigmoid component, manages this operation. 2) Choose the memory phase. The model selects the additional data to be added to the cell state. There are two parts to this phase. First, an action called

DOI: 10.54254/2755-2721/27/20230297

an input gate determines what information to update. After fresh prospective cell information is gathered through a Tanh layer, the cell information can be updated. After fresh prospective cell information is gathered through a Tanh layer, the cell information can be updated. It is required to determine which state features of the output cell after updating the cell state. To get the judgment conditions, the input must be sent via a sigmoid layer referred to as the output gate. Finally, what is essential is recorded absolutely, and what is not necessary is remembered less. 3) Output phase. What will be regarded as the current state's output is decided during this phase.

4. Experimental results and analysis

The 5000 data sets organized by the three modules are split, in which 4000 data of each module are used for training, and 1000 data are used for model testing. This paper conducts a three-classification test to determine the difference in language sense between the three English platforms.

Firstly, the model is initialized, and then the pre-training model bert-base-uncased is used to complete the training task. The number of Transformer layers in the pre-training model is 12, and the number of hidden Transformer neurons in each layer is 768. Specifically, the pre-training method is to extract each line in the document into a sentence and conduct word segmentation for each sentence to get a word vector. Then, these words are weighted to calculate the sentence vectors for these sentences and generate the vectors for the entire document. When BERT processes our document, it will generate a 768-dimensional vector, which is the input to the model.

In the process of network training, the initial learning rate is 5*10-4 and the maximum number of iterations is 10. We use Adam optimizer to optimize the parameters of the model. Loss function We use the cross entropy loss function, and we use the dynamic attenuation of the learning rate to solve the three-classification problem. When using the cross entropy loss function, a softmax layer is automatically added internally, through which subtle differences can be measured. The optimization function uses the gradient descent method to solve the optimal solution.

In terms of model selection, we use BERT and BERT+LSTM to process the dataset. BERT+LSTM connects the LSTM network at the output layer of BERT. After hierarchical feature extraction through LSTM, we can obtain a more detailed semantic representation, to accurately classify sentence intentions. This paper compares the average accuracy of the two models. Both models can perform well in the English language sense. In terms of model selection, we use BERT and BERT+LSTM to process the data set. BERT+LSTM connects the LSTM network at the output layer of BERT. After hierarchical feature extraction through LSTM, we can obtain a more detailed semantic representation, so as to accurately classify sentence intentions. This paper compares the average accuracy of the two models. Both models can classify English language sense well, but the BERT model is slightly better than the BERT+LSTM model.

Table 2. Experimental results.

Network model	Set of verification ACC (%)	Verification set loss
BERT	0.867	0.681
BERT+LSTM	0.866	0.682

As shown in Table 2, BERT and BERT +LSTM reached more than 0.86.

5. Discussion

The higher learning rate in the later period may lead to a slight increase in the training set loss LSTM is not linear accuracy, which is different from CNN, which considers the output to be explicit. In contrast, LSTM requires indefinite iterations before it begins to improve accuracy. Also, the majority of attention modules have parameters, therefore including more attention modules will make the model more complex. Adding parameters is helpful for model learning and will enhance performance if the model is underfitting before attention is added. The performance may stay the same or worsen if the model is already overfitting before attention is given due to this overfitting issue.

DOI: 10.54254/2755-2721/27/20230297

The complexity of the model is increased due to the addition of the attention module, or the current data is not suitable to be described by the attention mechanism. The model structure, loss function, label style, and many iterations should be changed, and the model situation before adding attention should be checked before the test, or the learning rate should be reduced, and the data set should be re-generated to increase the amount of data, and the training should be re-conducted. However, in practice, statements are subject to complex functions and may belong to written and spoken language in different contexts. In addition, our future work will consider combining contextual information to improve classification and model accuracy.

6. Conclusion

Based on the prepared dataset, this paper conducts experiments using BERT and LSTM models to identify the English language sense between different types of sentences. After training the model, this paper tested them with new input sentences, and the results showed that the accuracy of sentence classification was improved compared to traditional methods. In particular, using only BERT is better and more accurate than using BERT and LSTM together. In addition, this paper found that the accuracy of sentence classification is influenced by factors such as sentence length, syntactic structure, and word frequency. For example, longer sentences are more complex and, therefore, more difficult to classify accurately.

In summary, this paper uses the BERT and LSTM models to analyze the English language sense between different types of sentences. The results show that these models effectively improve the accuracy of sentence classification and can help English learners better understand the types of sentences required in different situations. This research is significant for English education in China, especially in underdeveloped areas where students have limited access to high-quality English resources. By providing personalized English learning services based on the BERT and LSTM models, this paper can help these students improve their English language sense and facilitate future scholars to continue using this research for innovation. However, this study still has some limitations. For example, this paper only used limited data to train and test our model. In the future, more data can be collected and used to improve the accuracy of sentence classification further. In addition, the model's performance can be further optimized by adjusting hyperparameters and exploring different model architectures.

References

[1] Decuypere, M., Alirezabeigi, S., Grimaldi, E., Hartong, S., Kiesewetter, S., Landri, P., ... & Broeck, P. V. (2023). Laws of Edu-Automation? Three Different Approaches to Deal with Processes of Automation and Artificial Intelligence in the Field of Education. Postdigital Science and Education, 5(1), 44-55.

[2] Peters, M. A. (2018). Deep learning, education, and the final stage of automation. Educational Philosophy and Theory, 50(6-7), 549-553.

[3] Deng, X., & Yu, Z. (2023). A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education. Sustainability, 15(4), 2940.

[4] Solanki, R. K., Rajawat, A. S., Gadekar, A. R., & Patil, M. E. (2023). Building a Conversational Chatbot Using Machine Learning: Towards a More Intelligent Healthcare Application. In Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines (pp. 285-309). IGI Global.

[5] Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35.

[6] Choi, H., Kim, J., Joe, S., & Gwon, Y. (2021, January). Evaluation of BERT and alBERT sentence embedding performance on downstream NLP tasks. In 2020 25th International conference on pattern recognition (ICPR) (pp. 5482-5487). IEEE.

[7] Masala, M., Ruseti, S., & Dascalu, M. (2020, December). RoBERT – A Romanian BERT Model. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6626-6637).

DOI: 10.54254/2755-2721/27/20230297

[8] Yao, L., & Guan, Y. (2018, December). An improved LSTM structure for natural language processing. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI) (pp. 565-569). IEEE.
[9] Rahali, A., & Akhloufi, M. A. (2023). End-to-End Transformer-Based Models in Textual-Based NLP. AI, 4(1), 54-110.
[10] Graves, A., & Graves, A. (2012). Long short-term memory. Supervised sequence labeling with recurrent neural networks, 37-45.