Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/18/20230358

Chinese news topic prediction using bidirectional encoder representation from transformers

Yifan Bi

Department of Computer Science and Technology, Nanjing University, Nanjing, 210023, China

191220001@smail.nju.edu.cn

Abstract. Nowadays, there are many researches on natural language processing (NLP). Through the research of NLP method, many problems in machine learning field have been solved. However, since the study of Chinese NLP has not developed rapidly until recent years, there is still much to be studied on Chinese NLP. As an excellent pre-training model, whether Bidirectional Encoder Representation from Transformers (BERT) performs well on specific Chinese NLP remains to be studied. Therefore, this paper uses BERT for Chinese NLP, and trains BERT model by collecting news title data to achieve Chinese text classification. Finally, the prediction results are studied by statistical methods. The research shows that BERT method performs well on Chinese NLP and can predict different types of news headlines well. Although it performs differently on different kinds of titles, its performance is satisfactory on the whole, and the prediction results are relatively balanced in different categories. Therefore, BERT can be used as a very practical and efficient NLP method. At the same time, it can also be predicted that it will play a great role in Chinese NLP.

Keywords: deep learning, natural language processing, BERT, Chinese news topic prediction

1. Introduction

The growth of the Internet has risen along with the advancement of science and technology. At present, Internet technology has entered the era of big data, and the surge of data has caused people's demand for deep learning. Under this premise, the related technologies have developed rapidly. The pre-training approach in the NLP area has assumed a very significant role in recent years. Through the pre-training technique, researchers can reduce time and energy expenditure by easily using the pre-training model. This makes information processing easier, which can help people in various aspects, such as text classification, machine translation and other fields.

Large-scale unlabelled text corpora are used by the pre-training technique in NLP to optimize the structure of deep network and generate a set of parameters. It is common to refer to this structure as a "pre-training model". Because pre-training models are often based on large-scale data, most of them have good performance, and some even reach the SOTA (sate of the art) results. Several areas have used the pre-training model, such as Image-Net [1] in the field of computer vision, ELMo [2] which proposed context sensitive text representation, and subsequent GPT [3] and BERT [4]. BERT, a deep bidirectional representation pre-training model based on Transformer [5], can more thoroughly

DOI: 10.54254/2753-8818/18/20230358

retrieve text's semantic content and performs well in the area of NLP. As a result, it has been used in many different contexts, and several studies have been done on it.

The BERT model has been employed in several languages. Delobelle et al. used the BERT model in Dutch [6], and carried out robust optimization to train the Bob BERT Dutch model. Through experimental comparison, it was found that compared with the existing BERT-based Dutch language model, Bob BERT method achieved better results. However, the study of Chinese NLP has not developed rapidly until recent years. Whether the pre-training model such as BERT can achieve satisfactory results in specific aspects remains to be studied. Therefore, in this study, BERT pre-training model is used for text classification. Results are acquired in the verification set after training in the training set. This paper mainly analyses the classification effect under different categories, and obtains the performance of the model. This study tries to get BERT's performance through the classification of news headlines, so as to test BERT's performance in Chinese NLP.

2. Method

This part introduces the experimental materials used in the experiment and their characteristics, the construction of the model, the setting of relevant parameters and the indicators of the evaluation results.

2.1. Experimental materials

The dataset is sourced from THUCNews (http://thuctc.thunlp.org/), which contains 208921 news headlines. These titles are divided into 20-30 words in advance to improve the efficiency of learning. There are 10 categories of data, with 10000 entries for each category in the training set. This can ensure that each category achieves a quantitative balance, so that the accuracy of the model will not be affected. Categories include sports, military, entertainment, politics, education, disaster, society, science and technology, finance and crime. The data is input into the model in words, which is convenient for learning. Training, verification, and test sets are created from the data collection. Most of them are training sets to guarantee the training effect. Meanwhile, verification set and test set are also indispensable. The verification set is used to establish the network structure's parameters or the complexity of the control model, while the training set is applied for estimating the model. Meanwhile, the test set is applied for evaluating the performance of the optimal model that was eventually chosen. In Table 1, the precise data amount is displayed.

Table 1. Data composition

Training set	Verification set	Test set
100000	50000	58921

2.2. Model building

The Google team first introduced Transformer, a classical NLP technique, in June 2017 [5]. Transformer beats RNN and CNN in machine translation applications. It employs an attention mechanism, encodes decoder, and produces positive outcomes. The fact that it can be parallelized effectively is its major benefit. The realm of visual images saw the earliest use of the attention mechanism. To classify images, researchers utilized an attention mechanism on an RNN model. [7] The attention mechanism started to be used in the field of NLP when Bahdanau et al. [8] later extended it to problems involving machine translation. Different from conventional CNN and RNN is Transformer. Attention mechanisms make up the network's whole structure. The effect improvement brought by this change is also subversive. The emergence of attention mechanism has completely solved the problem of how to use neural networks for "poor memory", and created a new deep learning paradigm, called "pre-training + fine tuning". Through this mode, it can train a more general model, and then fine tune according to different downstream tasks to improve performance. This is considered to be the most likely path to GAI at present. Transformer was first invented in the NLP

DOI: 10.54254/2753-8818/18/20230358

field. It was first applied to machine translation and obtained the SOTA of that year. Later, the most popular natural language processing model: BERT was born.

In October 2018, the Google AI research institute introduced the BERT pre-training approach. [4] Bidirectional encoder representation from transformers is the full name of BERT. BERT is a classic example of a bidirectional coding paradigm since it connects via transformer encoder blocks. Figure 1 depicts its architecture.

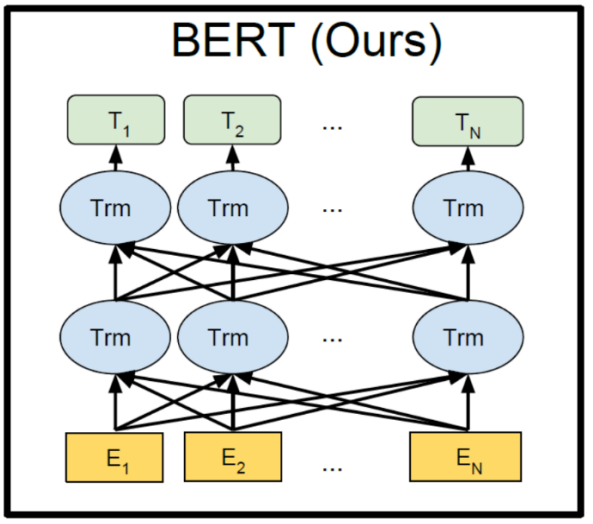

Figure 1. BERT architecture diagram [4].

BERT is a pre trained language model that can be used by all developers in the future, which is one of its highlights. BERT directly refers to the encoder module in the transformer architecture and discards the decoder module, which means it has the bidirectional coding ability and strong feature extraction ability.

The advantage of BERT is that it is built on transformer, so it has strong language representation ability and feature extraction ability, achieving SOTA in 11 NLP benchmark tasks. BERT also leverages the task of foreseeing the following sentence in Skip-thoughts [9] in order to understand the semantic relationship at the sentence level in order to better manage the relationship between numerous phrases. The combination approach suggested by GPT is used by BERT to combine two phrases into a sequence. The aim of modelling and forecasting the next sentence is to determine if the following sentence is that of the preceding sentence. Its disadvantage is that only 15% of the data in each branch participate in the prediction during its training [4], so the model converges slowly.

The program runs on python 3.10.4 with a running memory of 16GB. The model proposed in this paper is built by TensorFlow. TensorFlow is a deep learning framework developed by Google AI team, which is frequently employed in the implementation of various machine learning algorithms. The model uses dropout as regularization to prevent overfitting problems. The following are the experimental parameters: the epochs is 5, the hidden size equals to 768, the learning rate is 5×10^{-5}, dropout rate equals to 0.15. For BERT model, Adam [10] optimization algorithm is applied in the experiment to adjust the learning rate. Recording the batch number of loss drop in the verification set can determine whether there is no effect improvement for a long time. If the loss of the verification set does not drop in more than 1000 batch, the training will be ended early, which will speed up the training.

2.3. Evaluating indicator

The performance of BERT in Chinese semantic classification can be obtained by analysing the precision, recall and the f1-score of different categories. The performance can also be obtained by analysing the overall accuracy, macro average and weighted average. Precision is the percentage of

DOI: 10.54254/2753-8818/18/20230358

correctly predicted items in all positively predicted things as determined by the model. Recall is the percentage of accurate things that the model correctly predicted among all real positive items.

$$V_{precision} = \frac{TP}{TP+FP} \tag{1}$$

$$V_{recall} = \frac{TP}{TP+FN} \tag{2}$$

The weighted average of recall and precision is known as the F1-score.

$$V_{F1} = \frac{2 \times V_{precision} \times V_{recall}}{V_{precision} + V_{recall}} \tag{3}$$

Accuracy is the accuracy of the judgment of the model in the test. The average of all categories is the macro average. The weighted average is the weighted average of all samples across all categories.

3. Result

This chapter obtains the application effect of BERT in Chinese NLP by analysing the prediction results of models under different categories.

3.1. Forecast accuracy of different categories

The accuracy of this model using BERT on Chinese NLP is 0.9764. In view of the possible data imbalance, it is not appropriate to simply use accuracy to evaluate the effect. In order to indicate the accuracy of the model's prediction outputs, the experiment employed several sorts of precision and recall in addition to accuracy. Table 2 shows the forecast results under the ten categories of sports, military, entertainment, politics, education, disaster, society, science, finance, and crime. In addition, Table 2 also shows macro averaging and weighted averaging to reduce the negative impact of imbalance. It can be concluded from the data in the table that the BERT model performs well in classifying Chinese news headlines. This demonstrates that it performs well on specific Chinese NLP tasks.

Table 2. Experimental result of different categories.

	Precision	Recall	F1-score
sports	0.9942	0.9935	0.9938
military	0.9924	0.9155	0.9524
entertainment	0.9789	0.9845	0.9817
politics	0.9379	0.9697	0.9535
education	0.9719	0.9622	0.9670
disaster	1.0000	0.9920	0.9960
society	0.9402	0.9486	0.9444
science	0.9866	0.9752	0.9809
finance	0.9791	0.9670	0.9730
crime	1.0000	0.9968	0.9984
macro average	0.9781	0.9705	0.9741
weighted average	0.9766	0.9764	0.9764

The data in Table 2 show that the accuracy of prediction results varies with different categories. The performances of the two categories, disaster and crime, are particularly outstanding. This may be due to different data sizes. Since different categories of titles may have different characteristics, such

DOI: 10.54254/2753-8818/18/20230358

accuracy changes are acceptable. In general, the accuracy of different categories is consistent with the overall accuracy. Macro average and weighted average values are similar, indicating that the data is balanced for different categories.

3.2. Confusion matrix

The confusion matrix of model prediction results is shown in Table 3, where ENT means entertainment. The row of each data represents its real classification, while the list represents the classification predicted by the model.

Table 3. Confusion matrix.

	sports	military	ENT	politics	education	disaster	society	science	finance	crime
sports	13164	0	44	20	6	0	14	2	0	0
military	0	260	0	24	0	0	0	0	0	0
ENT	42	0	9149	32	10	0	20	35	5	0
politics	9	1	21	6010	17	0	62	60	18	0
education	6	0	21	45	4050	0	64	18	5	0
disaster	0	0	0	3	0	370	0	0	0	0
society	4	0	39	116	37	0	4907	60	10	0
science	14	1	64	124	37	0	124	15788	37	0
finance	2	0	8	34	9	0	28	39	3518	0
crime	0	0	0	0	1	0	0	0	0	313

The diagonal line from the top left to the bottom right of the matrix indicates the number of correctly predicted titles. The remaining part is the number of wrong prediction results. The data of the matrix are mainly concentrated on the diagonal, which indicates that the model receives a great accuracy. The number of model errors varies with the category and quantity. On small-scale data, the error of the model is not significant. Small data size may lead to large experimental error.

3.3. Practical examples

Table 4 shows some translated Chinese news headlines, their prediction by the model and their actual classification. These attempts to categorize news headlines are all correct.

Table 4. Some examples of classification of Chinese news headlines (translated).

news title	forecast classification	actual classification
The new technologies, new products and new business formats displayed by the exhibitors at the 2020 Smart Expo are too numerous to see.	science and technology	science and technology
According to the official measurement of China Seismological Network, a magnitude 3.9 earthquake occurred at 8:00 on August 25 in Longmatan District, Luzhou City, Sichuan Province.	disaster	disaster

In the actual application scenario, the model performs satisfactorily on the classification of real news titles.

DOI: 10.54254/2753-8818/18/20230358

4. Discussion

The outcome of the experiment is in line with expectations, that is, BERT model can indeed play its good performance on Chinese NLP. The experimental findings demonstrate that the BERT model may produce ideal accuracy, recall, and F1-score values. In addition, when viewing the accuracy rate of each category in the text, it is also obvious that BERT's accuracy rate results are high, and the classification effect of each category is relatively balanced. BERT has achieved some success in language processing other than English, so its performance in Chinese NLP is understandable. For example, researchers have proposed a monolingual BERT (Pars BERT) model [11] for Persian for text categorization, sentiment analysis, etc. This strategy outperforms the multilingual BERT model and other ways, according to experiments on a large number of datasets.

However, the performance of BERT's prediction results in some categories is not as good as that in other categories. Especially on such humanistic topics as politics and society. Some studies [12,13] have shown that up to now, since the pre-training model might not have absorbed the true information on semantics, it struggles to solve challenges requiring logic and common sense. This may be one of the reasons why it cannot get good performance on humanistic topics. A certain level of cognitive ability or the capacity to apply common sense information in reality is necessary for the majority of natural language processing activities in the real world. In order for the pre-training model to be more capable of handling tasks requiring common sense and reasoning, the subsequent improvement must enhance the pre-training model's capability in this regard.

It takes a lot of resources to get good results because the pre-training model represented by BERT is actually a simple and crude way to handle natural language by using big data, big models, and large amounts of computation, as opposed to being able to fully comprehend and use natural language in a flexible manner like human beings. Due to a shortage of computational capacity, this approach of comparing experimental findings with enormous data and massive models prevents regular researchers from fully training a pre-training model which is large in scare. A barrier of computer power this high could prevent further investigation of alternative approaches. Limited by computational power, the data size of this study is small, which may lead to a certain degree of accuracy error. At the same time, the inability to start training from scratch also leads to only fine tuning the model. It is vital to investigate ways to develop a more affordable, quicker, and smaller pre-training model.

Most professionals are only able to conduct research using pre-training models that have been publicly distributed and trained at the moment because the classical pre-training models have the issues of needing too much data, too large models, and too much demand for computing power. As a result, they lack the resources to investigate new models. This high standard prevents many academics from suggesting novel pre-training technique approaches, which is detrimental to the field's overall growth. Additionally, the industry is unable to use the pre-training technique broadly, for instance, it cannot be deployed on portable devices, due to excessive resource usage and low efficiency. Therefore, one potential research path may be to investigate ways to lower the pre-training model's training costs and investigate smaller, quicker pre-training models with as little accuracy loss as feasible. It will be a development direction in the future if researchers can further explore the model volume, performance and efficiency.

5. Conclusion

Through research, this paper finds that BERT performs well in the classification task of Chinese news headlines. It has good performance for different categories of titles, and its accuracy is relatively balanced. This can be attributed to BERT's unique pre-training model and its model structure. It refers to the Encoder module in the Transformer architecture and discards the Decoder module, so it has the bidirectional coding ability and strong feature extraction ability. Therefore, BERT can be a good model choice for different Chinese NLP tasks. Taking full advantage of BERT can increase the accuracy of NLP tasks and improve the training effect. However, due to the computational power and data size, the distribution of test data is not ideal, and the experimental results fluctuate on small-scale data. A smaller, quicker pre-training model should be investigated in order to maintain as much

DOI: 10.54254/2753-8818/18/20230358

accuracy as feasible while attempting to lower the training costs of the pre-training model. At the same time, due to the weakness of BERT model in the field of humanities, the prediction performance of some titles in this field is not as good as that of others. In this regard, the BERT model needs to be improved. To make the pre-training model's representation of semantics more accurate and help it tackle challenges involving common sense and reasoning, researchers might combine the knowledge map with it. This paper mainly explores the possibility of the application of BERT model in Chinese NLP, provides a simple attempt for a larger scale application in the future, and puts forward the existing problems and the direction for future improvement. This will help practitioners to avoid their weaknesses and make full use of their advantages when applying the BERT model in the future. Although the BERT model has made significant strides in the area of NLP, it still has a number of shortcomings, including the inability to handle difficulties involving logic and common sense, high resource usage, etc. Pre-training technique may be used in the future to address the aforementioned challenges, with an emphasis on how to enhance the model's performance in relation to human issues and how to use less computational resources.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

[2] Peters M E, Neumann M, Iyyer M, et al. (2018). Deep contextualized word representations. arXiv:1802.05365.

[3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[6] Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286.

[7] Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. Advances in neural information processing systems, 27.

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

[9] Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in neural information processing systems, 13.

[10] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[11] Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). Parsbert: Transformer-based model for persian language understanding. Neural Processing Letters, 53(6), 3831-3847.

[12] Niven, T., & Kao, H. Y. (2019). Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355.

[13] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007.