A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With "Same Semantics, Different Structure" After Fine Tuning

Chen Yin[1,*] and Zixuan Zhang[2]

[1] Department of Statistics and Data Science, Southern University of Science and Technology, Guangdong, 518000, China
[2] Department of artificial intelligence, Tianjin University of Technology, Tianjin, 300000, China
12112124@mail.sustech.edu.cn

Abstract. Traditional natural language processing models often find it difficult to distinguish between sentences with "similar structure and different semantics" and sentences with "different structure and similar semantics". Based on the all-MiniLM-L6-v2 and Bidirectional Encoder Representations from Transformers(BERT) model, this paper uses supervised learning and transfer learning methods to study the similarity of sentences with "similar structure, different semantics" and "different structure, similar semantics". New datasets in medical aspects with the same format as the hard datasets are artificially constructed and used as subdivided small-volume datasets to verify the model performance, thus simulating the needs of specific fields. On the basis of meta-learning and small number of shots learning, different models are fine-tuned, and good verification results are obtained and compared. For the fine-tuned models, the performance has been improved, among which the most significant improvements are: BERT model: accuracy: 0.51 to 0.65, all-MiniLM-L6-v2 model: precision:0.74 to 0.91 and so on. In this paper, the supervised learning method is used to provide effective ideas and directions for sentence similarity division of "semantically similar, structurally different" and "semantically different, structurally similar".This optimization can be proved to be effective and necessary.

Keywords: all-MiniLM-L6-v2, meta learning, few-shot learning, sentence similarity.

1 Introduction

With the rapid development of chat robots, machine reading and translation, the research demand for natural language processing is also increasing. Rational use of data and accurate and efficient are the research hotspots of information technology today [1]. However, in the actual application process, for chat robots or question-answering systems in various fields that use BERT, transformer and other models, it is necessary to learn a lot of data in advance, and there are problems that the research in some sub-

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_69

fields is not enough to get accurate learning results. Especially in the scene where man-machine conversation is needed, because of the complexity of structure and semantics of natural language and the influence of personal expression, there is a possibility of misunderstanding, which makes the final result inaccurate or quite different. Given such research difficulties, how to judge the similarity of sentences is the key to improving the application performance of natural language processing.

Semantic text similarity is an important branch of natural language processing, and related research is also developing. In 2020, Seol et al. [2] proposed an Natural Language Processing(NLP) algorithm based on an expert system for pediatric population electronic health records to identify children's asthma. Clustering whether there is asthma by predetermined diagnostic indicators, has a good effect on identifying children's asthma with unique characteristics. In 2021, Ormerod et al. [3] used the transformer model in the medical clinical field based on the characteristics that abbreviations are usually used in professional language and frequently. That article uses the average forecast similarity scores of several independently fine-tuned transformers. In this paper, the relationship between the final model losses is studied.Also in 2021, Suleman and Korkontzelos [4] studied Latent Semantic Analysis (LSA), which only used the semantic relationship between words to evaluate the similarity of texts without considering the influence of sentence structure, which made it impossible to distinguish sentences containing words with similar semantics but opposite meanings. This paper proposes an xLSA method, which can focus on the syntactic structure of sentences to a certain extent and alleviate the problem of sentence blindness in LSA.

In the task of text classification in natural language processing, the large-scale model which is widely used at present relies heavily on the analysis of sentence structure in identifying semantics, so it is not accurate enough to distinguish between "the same semantics but different structures" and "the same structure but different semantics". In this paper, the BERT model, BERT+Long Short-Term Memory(LSTM) model, and all-MiniLM-L6-v2 model are used to classify the hard data sets and the model is fine-tuned, and good results are obtained on the medical field data sets generated by using the large language model.

2 Data and methods

2.1 Data source

This paper started with the hard_similiarity dataset for this aspect. Weber et al. created a primitive dataset in which each sample has two types of sentence pairs: one pair of sentences that are semantically similar but structurally different, and the other pair of sentences that are semantically different but structurally similar. This dataset contains 230 event pairs. Ding et al. then extended this dataset to 1000 event pairs.[5]

At the beginning of the experiment, the hard data set is used to test the model.hard_similiarity is a dataset(denoted as"Original"), where each sample has two types of event pairs: one with events that should be close to each other but have very little lexical overlap and another with events that should be farther apart but have high overlap [5].

At the same time, IFlystar Fire language model was used to generate medical statements, and a new data set with the same format as the hard data set was manually constructed to train the model as a subdivision oriented, small-volume data set.

Then, using the idea of supervised learning.In this paper the hard_similarity data set was labeled and process it into the correct sentence pair + label form, so that it can be easily transmitted to the corresponding model for training or verification.

2.2 Modeling process

In this study, this paper takes this model as the main research object and compare it with the traditional model. The models involved are shown in Table 1:

Table 1. The model used and its operation.

Model	Application method
BERT	Evaluate
BERT+bidirectional LSTM	Evaluate
BERT+bidirectional LSTM(Fine tuning)	Train(Fine-tune using meta-learning and few shot learning methods)+Evaluate
all-MiniLM-L6-v2	Evaluate
all-MiniLM-L6-v2(Fine tuning)	Train(Fine-tune using meta-learning and few shot learning methods)+Evaluate

All-MiniLM-L6-v2 model: 'all-MiniLM-L6-v2' is a natural language processing model developed by Microsoft as part of the MiniLM family. Its specific structure includes:6 Transformer-encoder-layers, which is a reduction from larger models such as the 12 layers of BERT-base，Each layer has a smaller number of hidden units, usually fewer than the 768 units of BERT-base. Due to the reduction in the number of layers and hidden units, the total number of parameters for 'all-MiniLM-L6-v2' is also reduced, making the model lighter and faster to load and reason.This model is trained to support multiple languages, enabling it to be used in cross-language NLP tasks.

BERT&BERT+bidirectional LSTM model: Transformer-based pre-trained language models (PLMs) have started a new era in modern natural language processing (NLP). These models combine the power of transformers, transfer learning, and self-supervised learning (SSL) [6].

BERT model is a pre-trained language model. Google released it in 2018 and implements a novel approach to language representation through the Transformer architecture. BERT's model can be used to describe the global information of text in natural language processing, and it can comprehensively extract the features of context to obtain better semantic expression [7,8]. The main feature of BERT model is the use of Transformer encoder structure and the introduction of a bidirectional pre-training strategy. Traditional language models (such as those based on LSTM or RNN) only consider the words before each word, while BERT models can consider the context before and after a word so as to better capture the context information of the language. Bidirectional LSTM is added manually in order to take advantage of its ability to process sequence features and enhance the BERT model.Because of its unique long-term dependence, LSTM is often used as a collocation for text classification [9].

Download and use BERT's baseline model, and add a bidirectional LSTM to optimize the model. At the same time, the all-MiniLM-L6-v2 model is found as the main model of the study.

In this paper, some other machine-learning methods are introduced, including Meta-learning & few-shot learning and Evaluation method:

(1)Few shot sampling: In the training function, by calling the few_shot_sampler function, randomly select a small number of samples for training. This simulates the scenario of a Few Shot Learning, where only a few samples can be trained at a time, allowing the model to better generalize to new, unseen samples.

(2)Meta Learing: In the meta-learning part, the MAML framework is mainly used to enhance the generalization ability of the model through the steps of model parameter initialization, task sampling, inner loop training, loss calculation, outer loop training and repetitive training.

(3)Model training process: During training, each epoch is trained using a small number of samples, which is different from traditional training methods, which usually use the entire training data. In this way, the model can better adapt to a small amount of data and improve the generalization ability of the model in the case of small samples.

(4)Early Stop strategy: In the verification phase, use the early stop strategy to end the training early. When the accuracy on the verification set is no longer improved, the training is stopped to avoid overfitting the model. This strategy is also to better adapt to the situation of small amounts of data and avoid over-fitting in the case of small samples.

For BERT model, the traditional evaluation method is used, that is, the prediction accuracy of the model is calculated, and the performance of the model is evaluated.

For the all-MiniLM-L6-v2 model, if only the traditional evaluation method is used, the prediction accuracy of the model is calculated, and the performance of the model is evaluated, the results are basically the same with or without fine-tuning.This is because the model is a big data training model, and the fine tuning of small sample data is difficult to affect the accuracy of such an insensitive performance index. Therefore, this paper introduced ROC curve and the concepts of precision and recall rate, and evaluated the model performance.

2.3 Introduction of indicators

In this paper, several indexes are used to measure the model performance, including loss rate, accuracy, recall rate and ROC curve. The measurement of accuracy and recall rate is often used in the measurement of inspection classifier.

Accuracy (positive predictive value) is the ratio of true positive divided by the number of true positive and false positives. The range of values is [0,1], and the closer the accuracy is to 1, the better the performance of the model is. In the case of a large number of false positives, it will give a lower accuracy value. The recall rate, also known as sensitivity, is calculated by dividing the number of true positives by the ratio of true positives to false negatives.The range of values is [0,1], and the low recall rate may lead to a large number of false negatives, indicating that the model performance is not good [10]. The horizontal axis of ROC curve adopts false positive probability and the vertical axis adopts true positive probability. The closer the drawn curve is to the upper left of the coordinate system, the better the performance of the model is.

3 Results

3.1 The point of view of the model

This paper uses BERT model to call and verify directly. As a basic consideration of how data sets are represented in large-scale language models.

Then, a bidirectional LSTM layer is added to enhance the processing capability of the model for sequence structure. Direct verification again.

The modified BERT+ bidirectional LSTM model is adjusted, and the ideas of meta-learning and few shot learning are adopted to verify the model after fine tuning.

The effect on BERT model is very small, so this paper considers reducing the number of model layers and the number of hidden elements to make the model lighter and faster to load and reason. The all-MiniLM-L6-v2 model was considered. First, direct verification is performed.

Then, based on the idea of meta-learning and less shot learning, the model is fine-tuned, the evaluation standard of the model is optimized, and more sensitive evaluation indicators are introduced.

3.2 Current model performance effect

Based on theTable 2, this paper finds that the hard_similarity performance of BERT model is poor for data sets with "same semantics, different structure" and "different semantics, similar structure", with an accuracy of 0.5, which is very bad for a binary classification problem.

After adding bidirectional LSTM to the BERT model, the results are slightly better, but the accuracy is still low, maintaining around 0.5083, and the results are still poor.

After fine-tuning the BERT+bidirectional LSTM model based on meta-learning and few shot learning methods, this paper shows that the effect is very outstanding, and the training accuracy of the model is close to 0.8, and the verification accuracy is also increased to 0.65. However, after many tunings and cross-validations, the performance of the model was difficult to improve, so this paper considers replacing the model.

For the all-MiniLM-L6-v2 model, this paper shows that it is very powerful relative to the task in this study, so that the precision of the model before and after fine-tuning seems unchanged, and it is very high. (0.8208). In the fine-tuned training section, the accuracy of the model reached an amazing 0.9375.

Table 2. Current model performance effect.

	Train (hard_extend)	Evaluate(hard)	Evaluate(Self-designed datasets)
BERT	×	Loss: 0.69 Accuracy: 0.5	Loss: 0.69 Accuracy: 0.48
BERT + bidirectional LSTM	×	Loss: 0.69 Accuracy: 0.51	Loss: 0.69 Accuracy: 0.5
BERT + bidirectional LSTM(Fine tuning)	Loss: 0.51 Accuracy: 0.79	Loss: 0.66 Accuracy: 0.65	1Loss: 0.79 Accuracy: 0.52
all-MiniLM-L6-v2	×	Accuracy: 0.82 ROC AUC: 0.81 Precision: [0.5 0.74 1.] Recall: [1. 0.95 0.]	Accuracy: 0.44 ROC AUC: 0.44 Precision: [0.5 0.44 1.] Recall: [1. 0.47 0.]
all-MiniLM-L6-v2(Fine tuning)	Loss：0.66 Accuracy: 0.94	Loss：0.67 Accuracy: 0.82 ROC AUC: 0.89 Precision: [0.55 0.91 1.] Recall: [1. 0.91 0.]	Loss：0.74 Accuracy: 0.44 ROC AUC: 0.76 Precision: [0.55 0.87 1.] Recall: [1. 0.64 0.]

After the introduction of the AUC, precision, and recall indicators, this paper shows the effect of fine-tuning. Recall rates under the same threshold were all above 0.9, but fine-tuned precision and AUC both increased. This shows that the fine-tuning is effective and moving in a better direction.

This paper uses a self-defined dataset from the medical field for validation, and according to the evaluation results, it appears that the BERT model performs better and is more accurate than the all-MiniLM-L6-v2 model. In addition, this paper also finds that the ideas of meta-learning and less shot learning are effective, and the verification indicators of the model after such training have been improved.

4 Conclusion

In this paper, BERT, BERT+LSTM, and all-MiniLM-L6-v2 models are adopted, and the ideas of transfer learning and supervised learning are adopted to fine-adjust and train the model based on meta-learning and few shot learning for text classification tasks of "different structure-similar semantics" and "different semantically similar structures".

Finally, a good result is obtained in the verification. For BERT + bidirectional LSTM model, the accuary on test data is increased from 0.51 to 0.65. All-MiniLM-L6-v2 model's test precision and recall rate also increased, The most significant was precision, which grew from 0.74 to 0.91 under a certain threshold.

In this paper, supervised learning and transfer learning are adopted, and meta-learning and few shot learning are used to fine-tune some existing large models of natural language processing, trying to verify and explore the similarity of sentences with "different semantics, similar structure" and "similar semantics, different structure" in specific fields.

It can be seen from the index values in Table 2 that the accuracy of 0.8 can be achieved by using the all-MiniLM-L6-v2 model, and the model accuracy can reach above 0.9 after fine-tuning.This study shows that fine-tuning based on meta-learning and few shot learning is effective for existing large models of natural language processing, and can significantly improve the judgment ability of sentence similarity for "semantically similar, structurally different" and "semantically different, structurally similar" in specific domains.

The data loading and processing part of this paper is relatively simple, and only supervised learning and transfer learning methods are used. In the future, this paper tries to add data preprocessing and enhancement steps when processing more complex data formats from supervised learning, so as to improve the robustness of the model. In addition, because the research in this field is limited and the training and testing samples are small, this paper adopts grid search, random search or Bayesian optimization to adjust the learning rate, batch size and other parameters to further improve the performance of the model.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.
References

References

1. Huang Z.M.: Question Answering System in Financial Field based on Machine Reading Comprehension. Guangdong University of Technology(2021)
2. Seol, Hee Yun, et al: Expert artificial intelligence-based natural language processing characterises childhood asthma. BMJ open respiratory research 7(1) (2020)
3. Ormerod, Mark, Jesús M.D.R., and Barry D.: Predicting semantic similarity between clinical sentence pairs using transformer models: Evaluation and representational analysis. JMIR Medical Informatics 9(5) (2021)
4. Suleman, Raja M., and Ioannis K.: Extending latent semantic analysis to manage its syntactic blindness. Expert Systems with Applications 165, (2021)
5. Gao J., Wang W., Yu C.L., Zhao H., Wilfred N., and Xu R.F..: Improving Event Representation via Simultaneous Weakly Supervised Contrastive Learning and Clustering. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 3036–3049. Computational Linguistics, Dublin, Ireland (2022)
6. Kalyan, Katikapalli S., Ajit R., and Sivanesan S..: AMMU: a survey of transformer-based biomedical pretrained language models. Journal of biomedical informatics 126 (2022)

7. Wang S.J., Ma Z.T., Wu Z., et al.: An automatic recognition method for traffic named entities considering context. Engineering of Surveying and Mapping 33(02), 65-70 (2024)
8. Hu X.T., Li D., Song H., et al.: Memory interactive network for aspect-based sentiment analysis. Computer Applications and Software 41(02), 188-194+237 (2024)
9. Khataei M., Hamed, et al.: A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification. Mathematics 10(3) (2022)
10. Xu S.Z., Zhang C.L., and Hong D..: BERT-based NLP techniques for classification and severity modeling in basic warranty data study. Insurance: Mathematics and Economics 107, 57-67 (2022)