Text Classification for Marine Natural Products Literature

Xuebing Dai¹, Wei Liu¹*, Chunxiao Gao¹ and Zhiqiang Wei¹, ²

¹Department of Computer Science and Technology, Ocean University of China, Qingdao, China
²National Laboratory for Marine Science and Technology (Qingdao), Qingdao, China

*Corresponding author e-mail: 21170231326@stu.ouc.edu.cn

Abstract. We propose a text classification model MNPC-BERT for marine natural product literature based on BERT pre-training model. The marking task of Masked LM performed at the input stage of the pre-training phase is moved to the encoder of the transformer, which solves the problem of inconsistency between the pre-training phase and the fine-tuning phase. We have constructed a data set from a number of literature data sources containing different types of marine natural products literature, as well as a variety of different types of non-marine natural products literature. Compared with the classical algorithm, the algorithm proposed in this paper can better complete the classification based on existing data set.

1. Introduction

The ocean accounts for about 71% of the earth's surface area. The special environment of high salt and high pressure inside it also gives birth to many marine natural products[1]. It can be said that the unique structure and biological activity of marine natural products play an important role in the development of new drugs. The marine natural product literature[2] contains a large amount of marine natural product information, and we can learn more about marine natural product information through these marine natural product literature. Due to the limited number of existing marine natural product literature and the data source isn’t single, manually dividing the search for documents is time consuming and labor intensive. The traditional manual classification methods obviously cannot meet our requirements for the classification of marine natural product literature texts.

The challenges faced by marine natural product literature classification are as follows: 1. Because the literature data we obtained is cumbersome, it is not a standard natural language processing data set, how the model obtains data input and extracts the characteristics of the data; 2. Marine natural product literature and the land source product literature data has a great intersection in the content, how to obtain the characteristics of the data text and complete the effective representation of the data text word vector. Aiming at the above problems, this paper uses the current representative Text-CNN[3] convolution model and BERT[4] pre-training model to conduct comparative experiments, and proposes a marine natural products literature classification method based on BERT pre-training model MNPC-BERT.

First, this paper constructs a data set for the vertical field of marine natural product literature. We achieve a balance of positive and negative samples through a combination of multiple data sources.

Second, on the basis of comparing the classical text classification methods, we optimize and achieve a classification model MNPC-BERT for marine natural product literature, which improves the accuracy of literature classification.
Finally, we get a text classification model with high classification accuracy, which can better complete the classification of marine natural product literature.

2. Related Work

Text classification is one of the more common tasks in the field of natural language processing. From the perspective of training methods, the training methods of text classification task include supervised learning[5], semi-supervised learning[6] and unsupervised learning[7]. David D. Lewis et al. completed heterogeneous uncertainty sampling work for supervised learning[8]. R Caruana et al. completed the experience comparison of supervised learning algorithms[5]; O Chapelle et al. introduced the theoretical knowledge of semi-supervised learning[9] and DP Kingma et al. Generating semi-supervised learning of models[10]; M Weber et al. used unsupervised learning to propose a method of learning object class models from unmarked and undivided chaotic scenes to achieve visual object recognition[11]. A Radford et al. Convolution generates an unsupervised representation learning study against the network[12]. Supervised learning generally requires manual labeling of data. This learning method has high requirements for the quality of data annotation. Unsupervised learning is characterized by self-learning data of the model. It is generally used for pre-training of models and can learn the general underlying characteristics of data; Semi-supervised learning has great advantages in dealing with small data sets, and can improve the classification results based on small data sets.

From the perspective of the evolution of text classification technology, it is roughly divided into three stages. The first stage is the feature engineering and classifier period. The representative model is benchmark[13], which completes the training of supervised learning based on the LSHTC data set; the second stage is after the emergence of word2Vec[14]. Feature engineering incorporates word segmentation technology. The representative model is Text-CNN, which realizes end-to-end learning based on IMDB dataset. The third phase is the emergence of pre-training model, which allows the model to perform unsupervised learning on a large number of unlabeled data sets, and finally fine-tuned to specific natural language processing tasks to achieve a better training effect. The representative model of this period is BERT.

3. Method

3.1. Data set construction

Because the data sources of marine natural product literature data are scattered, in order to obtain the literature data under different data sources, this paper proposes a reptile system oriented to the vertical field of marine natural product literature. The experimental data set includes two types of data: positive and negative samples. The positive sample data is composed entirely of marine natural product literature, that is the text data we want to obtain after classification; The negative sample data is partly composed of land-source product literature, and the other part is composed of literature in other fields. The composition of the negative sample data is mainly considering that there will be some intersections between the marine natural product literature and the terrestrial product literature. The integration of the land source product literature and the literature in other fields can better train the generalization ability of the model. This has a certain improvement in the experimental results.

In order to improve the quality of the data set and adapt to the experimental model, the data set is processed as follows:

1. Each piece of data consists of the title, author and abstract of each document;
2. Convert all the contents of the data text to lowercase, removing numbers, stop words and special symbols;
3. Construct a dataset for the Text-CNN model and the MNPC-BERT model separately.

Finally, the experimental data set contains 3000 positive sample data texts and 3000 negative sample data texts.
3.2. The training process of Text-CNN model

CNN is originally used for image classification in computer vision, target detection, and achieves good results. In 2014, Kim.Y et al. proposed the Text-CNN model and used CNN for text classification features extraction. The traditional One-Hot word bag model[15] has many disadvantages in processing the text. Since One-Hot sets the feature position at the time of encoding and sets all the other positions to 0, the processed vector has high-dimensional and sparse features, which is not suitable as a feature of CNN input. In this paper, Word2Vec is used to embed the word vector in the text, which is used as the input of Text-CNN. At the same time, the three-size filter is used to convolve the text word vector. Each type of filter has 100 filters, and then convolution. As a result, the maximum pooling is performed, and then the pooled result is fully connected by the SoftMax layer. Finally, the classified result is obtained. The Text-CNN model has a simple structure and a fast training fitting time. However, due to the loss of a large amount of data feature information during the pooling process, the final classification result is not very satisfactory.

The experimental parameters of the Text-CNN model are set as follows:

| Description                  | Values     |
|------------------------------|------------|
| Word vector model            | Word2Vec   |
| Filter size                  | (4,5,6)    |
| Number of filters per size   | 100        |
| Activation function          | ReLU       |
| Pooling strategy             | 1-Max pooling |
| Dropout rate                 | 0.5        |
| L2 regularization            | 3          |

3.3. The training process of MNPC-BERT model

The MNPC-BERT model proposed in this paper is based on the BERT pre-training model proposed by Google in 2018. The BERT model completes the unsupervised learning training process based on massive data sets in the pre-training phase, which can better learn the semantic features of
the text. The bottom layer of the BERT is composed of a bidirectional encoder based on Transformer [16].

![Figure 2. The MNPC-BERT network structure.](image)

Unlike other recent language representation models, BERT aims to pre-train deep two-way representations by jointly adjusting the context of all layers, which allows the BERT model to better capture the semantic features of the text. One training task of the BERT model is: randomly select 15% of the training text data with [Mask] tag, where 80% probability of the marked part is marked with [Mask], 10% probability keeps the text data itself unchanged, and 10% probability is replaced by other words in the data. The model introduces the [Mask] tag on the input, but there is no such [Mask] tag in the Fine-tuning phase of the downstream task, which leads to inconsistencies between the pre-training phase and the Fine-tuning phase. In order to solve this problem, we move the [Mask] tag on the input side into the encoder of the Transformer layer, and do not display the [Mask] tag on the input side. Finally, based on the improved MNPC-BERT model, we have trained the marine natural product text classification task and achieve good results.

The experimental parameters of the MNPC-BERT model are set as follows:

| Description       | Values |
|-------------------|--------|
| Train batch size  | 32     |
| Learning rate     | 2e-5   |
| Num epochs        | 3      |

![Table 2. Experimental parameters of MNPC-BERT.](image)
4. Experiments and Results

4.1. Experimental environment and data set

This article is based on the Ubuntu system equipped with GeForce GTX 1080 graphics card. The deep learning framework of the application is TensorFlow. The programming language is python. The data set contains 3000 positive sample data and 3000 negative sample data.

4.2. Evaluation indicators

Based on the existing positive and negative sample data set, we complete the training of Text-CNN and MNPC-BERT models, and finally express the results of the training as follows:

| Model      | Precision | Recall | F-Measure |
|------------|-----------|--------|-----------|
| Text-CNN   | 0.72      | 0.69   | 0.70      |
| MNPC-BERT  | 0.87      | 0.86   | 0.86      |

Table 3. Model training results.

4.3. Experimental results and analysis

After the model training is completed, we select the test data set prepared in advance to test the effect of our model. The test results are as follows:

| Class | Samples                                                                 |
|-------|-------------------------------------------------------------------------|
| 1     | As a part of our search for novel histamine H3 receptor agonists, we designed |
| 1     | The objective of this work was to evaluate the biological properties of a new |
| 0     | As a continuation of our earlier study we conformationally restricted the |
| 1     | Extensive structural modification of immepyr led to the discovery of trans |
| 1     | The synthesis and structure-activity relationship SAR studies of a series of |
| 1     | Previously we have reported the discovery of active enantiomer of a |
| 0     | Use of automated synthesis led to the discovery of several 6-membered nitrogen |

Table 4. Test results of Text-CNN model.

| Positive | Negative | Class     | Samples                                                                 |
|----------|----------|-----------|-------------------------------------------------------------------------|
| 0.134015 | 0.865985 | 0         | The inducible transcription factor nuclear factor kappa                   |
| 0.991756 | 0.008244 | 1         | Cyanobacteria from marine and freshwater habitats                        |
| 0.981061 | 0.018939 | 1         | Mytilin is a 34-residue antibacterial peptide from the                    |
| 0.991277 | 0.008723 | 1         | Trimusculus peruvianus. Their structures were                           |
| 0.143514 | 0.856486 | 0         | a single enantiomer is described. Novel steps include a                 |
| 0.192258 | 0.807742 | 0         | Organelles are sequestered in the mollusc's digestive epi                 |

Table 5. Test results of MNPC-BERT model.

In the test result, 0 represents negative sample data, and 1 represents positive sample data. The test result of Text-CNN directly gives the test data category judged by the model; The MNPC-BERT model gives the probability of the predicted result which is predicted by a float number, and the actual category of the data is also listed. From the test results, we can see that the improved MNPC-BERT model can better complete the classification of test data set.

5. Conclusion

In this paper, a high quality marine natural product literature data set is constructed independently, and an improved MNPC-BERT model is proposed based on the BERT pre-training model. In order to improve the classification ability of the model, we have enhanced the negative sample data, and the
fusion of the land source product literature and other unrelated literature as negative sample data can improve the generalization ability of the model processing. At the same time, we change the location of the Masked LM task execution and achieve good results. Of course, our experiments also have certain deficiencies. From the experimental results data, we can see there is still much room for improvement in our accuracy and recall rate.

The BERT model is a new beginning in the field of natural language processing and a trend in the future. In the future, we will also increase the experimental data set based on this model, and constantly optimize and adjust the experimental model to strive for better experimental results.

6. Acknowledgments

This paper is supported by Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao)(No.2018SDKJ0402).

References

[1] J. W. Blunt, B. R. Copp, M. H. G. Munro, P. T. Northcote, and M. R. Prinsep, “Marine natural products,” Nat. Prod. Rep., vol. 28, no. 2, pp. 196–268, 2011.
[2] T. F. Molinski, D. S. Dalisay, S. L. Lievens, and J. P. Saludes, “Drug development from marine natural products,” Nat. Rev. Drug Discov., vol. 8, no. 1, pp. 69–85, Jan. 2009.
[3] X. Sun, X. Ma, Z. Ni, and L. Bian, “A New LSTM Network Model Combining Text-CNN,” in Neural Information Processing, Cham, 2018, pp. 416–424.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” ArXiv181004805 Cs, May 2019.
[5] R. Caruana and A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning Algorithms,” in Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 2006, pp. 161–168.
[6] X. (Jerry) Zhu, “Semi-Supervised Learning Literature Survey,” University of Wisconsin-Madison Department of Computer Sciences, Technical Report, 2005.
[7] Q. V. Le, “Building high-level features using large scale unsupervised learning,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8595–8598.
[8] D. D. Lewis and J. Catlett, “Heterogeneous Uncertainty Sampling for Supervised Learning,” in Machine Learning Proceedings 1994, W. W. Cohen and H. Hirsh, Eds. San Francisco (CA): Morgan Kaufmann, 1994, pp. 148–156.
[9] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions,” p. 8.
[10] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semi-supervised Learning with Deep Generative Models,” in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3581–3589.
[11] M. Weber, M. Welling, and P. Perona, “Unsupervised Learning of Models for Recognition,” in Computer Vision - ECCV 2000, Berlin, Heidelberg, 2000, pp. 18–32.
[12] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” ArXiv151106434 Cs, Jan. 2016.
[13] D. E. Taylor and J. S. Turner, “ClassBench: A Packet Classification Benchmark,” IEEEACM Trans. Netw., vol. 15, no. 3, pp. 499–511, Jun. 2007.
[14] Y. Goldberg and O. Levy, “word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method,” ArXiv14023722 Cs Stat, Feb. 2014.
[15] R. Johnson and T. Zhang, “Effective Use of Word Order for Text Categorization with Convolutional Neural Networks,” ArXiv141212058 Cs Stat, Mar. 2015.
[16] M. Jaderberg, K. Simonyan, A. Zisserman, and koray kavukcuoglu, “Spatial Transformer Networks,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025.