Open Access Library Journal
2021, Volume 8, e7602
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Summary of Research Methods on Pre-Training Models of Natural Language Processing

Yu Xiao[1], Zhezhi Jin[2]*

[1]Depertment of Mathematics, Yanbian University, Yanji, China
[2]Department of Economics and Management, Yanbian University, Yanji, China
Email: 1150457896@qq.com, *jinzhezhi@sina.com

How to cite this paper: Xiao, Y. and Jin, Z.Z. (2021) Summary of Research Methods on Pre-Training Models of Natural Language Processing. *Open Access Library Journal*, **8**: e7602.
https://doi.org/10.4236/oalib.1107602

Received: June 2, 2021
Accepted: June 28, 2021
Published: July 1, 2021

Abstract

In recent years, deep learning technology has been widely used and developed. In natural language processing tasks, pre-training models have been more widely used. Whether it is sentence extraction or sentiment analysis of text, the pre-training model plays a very important role. The use of a large-scale corpus for unsupervised pre-training of models has proven to be an excellent and effective way to provide models. This article summarizes the existing pre-training models and sorts out the improved models and processing methods of the relatively new pre-training models, and finally summarizes the challenges and prospects of the current pre-training models.

Subject Areas

Statistics

Keywords

Natural Language Processing, Pre-Training Model, Language Model, Self-Training Model

1. Introduction

Natural language processing is an interdisciplinary subject that combines linguistics, computer science, mathematics and other disciplines. Natural linguistics has many achievements in machine translation, speech recognition, intelligent voice assistants, and public opinion analysis. It involves many researches in related disciplines, especially in the field of artificial intelligence in recent years, there have been many breakthroughs.

Pre-training refers to the training of the data model before a relatively large-scale data processing, and after a wave of training, the previous training will fine-tune

*Corresponding author.

in the model of the downstream task. In most language models, pre-training models usually use self-supervised methods to train the models.

With the wide application of pre-training models in natural language processing, pre-training technology has also entered a new era. The pre-training models involved in this article are mainly traditional pre-training models. After an overview of traditional pre-training models, this article will introduce pre-training models. The training model gives the improved T5 model of the BERT model, and the model overview after combining the pre-training model and the self-training model.

2. Natural Language Processing

Natural language processing can express the language with discrete symbols and then form different sentences to express different semantics, to help the computer understand the language. Natural language processing is one of the most difficult problems in artificial intelligence.

The basic points of natural language processing are divided into corpus, Chinese word segmentation, part-of-speech tagging, syntactic analysis, stem extraction, morphological restoration, etc. Natural language usually undergoes feature extraction, feature selection, dimensionality reduction and other methods for feature processing, and then uses Markov model, hidden Markov model, conditional random field, Bayesian network and other methods are all used to classify and process languages [1].

The research on natural language processing is conducive to the development of personalized knowledge recommendation services, knowledge retrieval, intelligent question answering, speech recognition and other fields, making the relationship between artificial intelligence and language processing closer.

3. Traditional Pre-Training Model

3.1. ELMo Model [2]

The ELMo model (Embedding from Language Model) is based on the early pre-training model of the PTMs model that cannot solve the problem of complex context. After the improvement, a pre-training of the LSMT language model is added to the large-scale unsupervised corpus. The LSMT model is a two-way model. In pre-training, first use the language model to pre-process on the corpus, and then select the corresponding word network from the trained network China to embed each layer of words in the downstream task as new features, thereby effectively solving the problem The problem of ambiguity [3]. The structure of the ELMo model is shown below (**Figure 1**).

Compared with the previous model, the ELMo model adds a two-way language model LSTM (Long-Short Term Memory), so that the model can better connect the content between the context of the article in a complex context, effectively improving the performance of the model, but in Compared with subsequent models, the integrated fusion and some feature extraction capabilities also have obvious limitations.

3.2. GPT Model [4]

As mentioned above, the ELMo model has obvious limitations in many aspects, so based on the improvement of the EMLo model, Transformer proposed the GPT pre-training model of OpenAI. The common point of the GPT model and the EMLo model is that they both use two stages to train the model. The difference is that the first stage of the GPT model pre-trains the unsupervised language model on the corpus, so that the model parameters are converted to neural network The initial parameters are then fine-tuned for downstream tasks through a supervised model in the second stage. The GPT model structure is as follows (**Figure 2**).

The GTP model has improved some shortcomings in the later period, and proposed the GTP2 [5] model. Although the new model has better performance in the preprocessing of corpus, the model itself is still a one-way language model in essence, in terms of semantic information. There are still big limitations in the establishment.

3.3. BERT Model [2]

The BERT model is based on the stacked Transformer substructure proposed by the GPT model to establish a basic model (**Figure 3**). It can be said that it is an evolution of the GPT model to some extent. The unidirectional structure of the GPT model is improved to a bidirectional structure, which can be at a deeper level. Up to achieve the training of the data set, and then achieve the purpose of adjusting the model parameters.

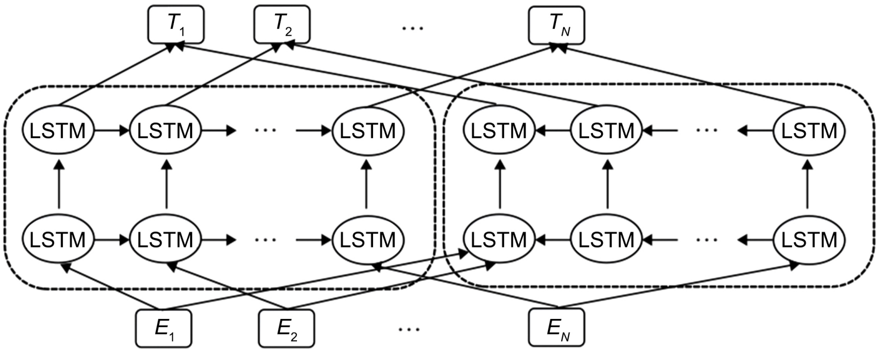

Figure 1. Structure diagram of ELMo model.

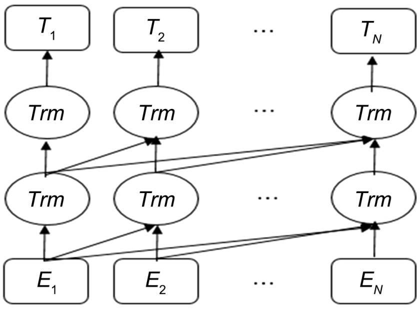

Figure 2. Structure diagram of GPT model.

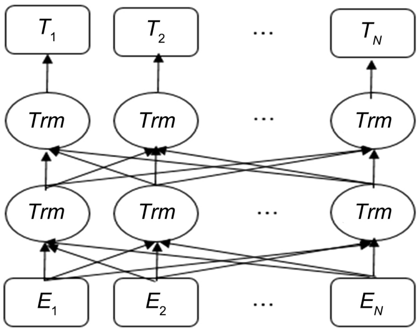

Figure 3. BERT model structure diagram.

The BERT model can formalize the following likelihood function to the maximum

$$\begin{aligned}\theta &= \arg\max \sum_{x\in Cor} P_\theta\left(x\middle|\overline{x}\right) = \arg\max \sum_{x\in Cor}\sum_{t} m(t)\cdot P_\theta\left(x_t\middle|\overline{x}\right)\\ &= \sum_{t=1}^{T} m_t \log\frac{\exp\left(H_\theta\left(\hat{x}\right)_t^T e\left(x_t\right)\right)}{\sum_{x'}\exp\left(H_\theta\left(\hat{x}\right)_t^T e\left(x'\right)\right)}\end{aligned} \tag{1}$$

among them

$$m(t) = \begin{cases}1, & x_t \text{ masked}\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

The emergence of the BERT model has greatly promoted the development of natural language processing, and at the same time promoted the transfer learning of natural language processing, but there are still shortcomings such as excessive time consumption and high hardware requirements.

3.4. New Pre-Trained Model

After the emergence of the BERT model, a large number of new pre-training models have emerged. Most of the new training models are generated after improvements based on the BERT model. They are all pre-training models in the deep learning era. The more representative models are ERNIE, SpanBERT, RoBERTa [6], ALBERT, etc. These models have good versatility in the application of large text corpora, and also provide better model initialization, which brings better generalization performance for subsequent machine learning, and avoids the problem of data overfitting in some experiments.

4. Related Improvements to the Model

4.1. T5 Model

Pre-training models appeared in the early stage [7], so that the development of pre-training models has reached a relatively complete level, so in order to be able to use a model to adapt to various NLP tasks, the T5 pre-training model was born in response to the times, the emergence of the T5 pre-training model provides a general framework model for almost all tasks involved in the entire pre-training model field. Although from the perspective of the model, the T5 pre-training model has not improved much in the innovation of the model, it is able to integrate many models involved in the previous single items and find the optimal combination scheme as quickly as possible so that more data can be processed.

For a simple example, when doing English-Chinese translation, you only need to add "translate English to Chinese" to the training data set. If you need to translate "university student", then add "translate English to Chinese: university student" input into the model, you can directly get the Chinese translation "大学生". For example, when we are preparing for emotional translation, we only need to add "sentiment" before the input to get the output result directly. For example, input "sentiment: This movie is terrible" to get "negative".

In CSL and LCSTS text generation tasks, the T5 model can become the best model of several known models (Table 1).

4.2. Combination of Self-Training Model and Pre-Training Model

4.2.1. Model Overview

The self-training model means to train one labeled data first, then label the remaining large-scale unlabeled data, and then use the result as pseudo-labeled data for target training. It can be seen that the self-training model is very similar to the pre-training model, except that the pre-training model is only trained and operated on one model, learning directly from unlabeled data, while the self-training model uses two models, first indirectly from the learning of the data. Therefore, the combination of the two models can often get better results.

First pre-train the model. After the pre-trained model is trained on the data, it is trained as the first model of the training model of the self-training model, and the result obtained is trained as the pseudo-labeled model of the second step of the self-training model. The training results are used in the inference test. The training results obtained in this way are greatly improved than those obtained by simply pre-training, and outstanding results can also be achieved in small-sample learning tasks.

Table 1. Comparison of CSL test results between T5 model and other models.

CSL summary generates experimental results					
	beam size	**rouge-l**	**rouge-1**	**rouge-2**	**bleu**
BERT	1	63.75	65.45	55.01	45.52
WoBERT	1	65.98	68.35	57.83	47.78
mT5	1	67.25	69.21	59.21	49.79
BERT	2	64.65	66.25	55.75	46.42
WoBERT	2	66.65	68.68	58.65	48.52
mT5	2	67.52	69.35	59.68	50.17
BERT	3	64.59	66.34	56.20	46.85
WoBERT	3	66.85	68.95	58.75	48.59
mT5	3	67.23	70.54	60.69	50.19

4.2.2. Model Checking

The method of combining self-training model and pre-training model for model training is mainly divided into four steps, The first step is to train a pre-trained model on the labeled data， as a teacher model f_T, the second step is to use f_T extract data in related fields from a massive general corpus, the third step uses f_T annotate the extracted data, the fourth step is to train the target student model with the labeled pseudo-labeled corpus f_S.

The first, third, and fourth steps in the training process are all deterministic steps. Therefore, the focus of the model is how to obtain the relevant corpus D' required by the model from the massive corpus D. Under normal circumstances, corpus D can directly divide the document into sentences, and then extract the data in sentence units. Therefore, you can use the method of sentence encoding, use the encoding vector to represent the sentence, and then use the sentence encoder in multiple data After training on the set, you can get the feature vector of each coded sentence. In this way, you only need to add a special task code as a query condition to judge whether the sentence meets our requirements by calculating the cosine value of the sentence code and the task code At the same time, it can also reduce the noise interference of downstream pending tasks and improve the confidence of f_T.

Table 2 shows that the result of the method of combining self-training model and pre-training model in the process of training sentence encoding is still considerable compared with a single pre-training model.

5. Model Follow-Up Development and Outlook

Pre-training models are models that have epoch-making significance, the Pre-training models based on various language models have been widely used [8] [9], but so far there is still a lot of room for the development of pre-training models [10] [11], and more training scenarios are needed. Besides larger corpus to improve the accuracy of the model, in addition to the training of the large corpus, the professional corpus is also trained, and the problems encountered in the fine-tuning of the model are improved. The main development direction of the future model is to solve the problem of deep neural networks being vulnerable to adversarial example attacks, so as to achieve the same defensive effect as the image processing problem. And the structure of the future model also needs

Table 2. Comparison of test results between self-training model and pre-training model.

MODEL	SEMANTIC TEXTUAL SINILARITY (STS)						STSB
	2016	2017	2018	2019	2020	AUG	
INFERSENT	61.1	51.4	68.1	70.9	70.7	64.4	70.6
USIF	68.3	66.1	78.4	79.0	72.8	70.9	79.5
WORD, TRIGRAM	67.8	62.7	77.4	80.3	78.1	73.3	79.9
SASE (OURS)	69.7	62.9	77.3	79.8	78.1	73.5	80.8

to be streamlined and improved, which is also very important in the construction of the evaluation system.

6. Conclusion

This article mainly sorts out and summarizes the pre-training involved in the language model, and sorts out the two new improved pre-training models to process data. Among them, the T5 model can perform comprehensive optimal combination model processing of more models on big data, and the method of combining the self-training model and the pre-training model can improve the effect of small-sample learning and knowledge distillation. Finally, this article summarizes the improvement space and expectations of the future pre-training model. In this regard, the author will also conduct a deeper research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lyu, L.C., Zhang, B., Wang, Y.P., Zhao, Y.J., Qian, L. and Li, T.T. (2021) Global Patent Analysis of Natural Language Processing. *Science Focus*, **16**, 84-95.

[2] Yu, T.R., Jin, R., Han, X.Z., Li, J.H. and Yu, T. (2020) Review of Pre-Treaning Models for Natural Language Processing. *Computer Engineering and Applications*, **56**, 12-22.

[3] Liu, Q., Kusner, M.J. and Blunsom, P. (2020) A Survey on Contextual Embeddings.

[4] Radford, A., Narasimhan, K., Salimans, T., *et al.* (2020) Improving Language Understanding by Generative Pretraning. https://www.cs.ubc.ca/~amuham01/LING530/papers/redford2018improving.pdf

[5] Radford, A., Wu, J., Child, R., *et al.* (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, **1**.

[6] Liu, Y., Ott, M., Goyal, N., *et al.* (2020) RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

[7] Martin, L., Muller, B., Suarez, P.J.O., *et al.* (2019) CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203-7219. arXiv:1911.034894 https://doi.org/10.18653/v1/2020.acl-main.645

[8] Alsentzer, E., Murphy, J.R., Boag, W., *et al.* (2019) Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72-78. arXiv:1904.03323 https://doi.org/10.18653/v1/W19-1909

[9] Huang, K., Altosaar, J. and Ranganath, R. (2019) Clinica-1BERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342

[10] Shoeybi, M., Pateary, M., Puri, R., *et al.* (2019) Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Paralleslism. arXiv:1909.08053

[11] Clark, K., Luong, M.T., Le, Q.V., *et al.* (2020) ELECTRA: Pretraining Text Encoders as Discriminators Rather than Generators. arXiv:2003.10555