Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/132/2024.20851

Comparative Analysis of Improved Versions of BERT Models on Chinese NLP Tasks

Yifan Hu[1,*,†], Sibo Tao[2,†], Ziyan Miao[3,†], Cary Cao[4,†] , John Su[5,†]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510641, China,
[2] School of Mathematical Science, Fudan University, Shanghai, 200433, China,
[3] Lord Byng Secondary School, Vancouver, V6R 2C9, Canada,
[4] Britannia Secondary, Vancouver, V5L 3T4, Canada,
[5] Harrow International School Hong Kong, Hong Kong, 999077, China,

[1] whohyf@163.com, [2] bob916380963@163.com, [3] andrewzymiao@gmail.com,
[4] sailakedi444@gmail.com, [5] Johnyisu08@gmail.com
*corresponding author
[†]co-first authors

Abstract. BERT is a pre-trained language representation model that has received a lot of attention for its impressive results in Natural Language Processing (NLP) tasks, such as sentiment analysis and text classification. This inspired many BERT-based models to be created that improved on the original in different ways. Examples of these models include RoBERTa, which improves on BERT's pretraining; K-BERT, which uses Knowledge Graphs to improve BERT's domain-specific knowledge; and many others. In this paper, we will test these BERT-based models with various datasets and present a comparative analysis of the models.

Keywords: Artificial Intelligence, Natural Language Processing, BERT

1. Introduction

As the framework of Transformer demonstrated outstanding capabilities in NLP (Natural Language Processing) tasks and gradually became the dominant framework in this field, Transformer based models with excellent performance have continuously emerged. Among them, BERT [1] (Bidirectional Encoder Representation from Transformers) is a significant breakthrough. Thanks to the development of hardware devices, it has been able to train on a large amount of training data through MLM+NSP pre training tasks, and has demonstrated unprecedented performance in various natural language understanding tasks such as sentiment analysis and named entity recognition.

Since the creation of BERT, many BERT-based models have been created to improve on the original in different ways. For example, researchers found that BERT was severely undertrained, which led to the creation of RoBERTa [2], which improves on BERT's pretraining. However, the models are often created to target specific areas, and therefore may be far better at certain tasks than others. Additionally, for the most part, these models have only been tested against BERT, not against each other. As such, a

DOI: 10.54254/2755-2721/132/2024.20851

comparative analysis of BERT-based models is necessary to see which models perform best at certain tasks as knowledge of this may be useful in further research.

In this paper, we will compare and analyse five BERT-based models: RoBERTa, K-BERT [3], StructBERT [4], MacBERT[5], and DeBERTa. They will be tested on two sentiment analysis datasets and two named entity recognition datasets, containing data from Chinese sources, such as book reviews on Chinese websites. We will then discuss the results and show which models are the most effective and on which tasks.

The paper will be structured as follows. Section 2 will introduce the models and explain them in more detail. Section 3 talks about the datasets and evaluation metrics that we will be using Section 4 shows how all the models did in each task. Section 5 discusses the results concludes the paper.

2. Literature Review

Natural Language Processing (NLP) is a field of study that revolves around combining deep learning models and computational linguistics to process human language. The introduction of BERT's improved models have greatly accelerated the efficiency of completing NLP tasks. These models include RoBERTa, K-BERT, StructBERT, MacBERT and DeBERTa, where each stands out for their unique modifications upon the original BERT. This section dives into each specific model and explores their innovations.

Through the addition of larger datasets, dynamic masking and next sentence prediction, RoBERTa improves upon BERT's pre-training process. Several NLP benchmarks displayed better performance because of these modifications. Mainly aimed to build upon BERT's text classification ability, Roberta's dynamic masking is a key component that allows the model to come across more diverse contexts, leading to overall richer language representations.

K-BERT improves BERT's contextual comprehension by integrating structured knowledge from knowledge graphs into BERT. K-BERT is able to perform well in tasks that require in-depth knowledge of specific professions, such processing legal documents and biomedical text analysis, by offering domain-specific information. Domain-specific tasks are where K-BERT excels in, as it requires contextual knowledge which K-BERT's semantic information is enhanced through the incorporation of triples from knowledge graphs.

BERT's framework is introduced to structural information by the means of StructBERT, which focuses on preserving sentence order and word relationships during pre-training. This method helps StructBERT understand and generate coherent text. StructBERT' s unique sentence order prediction task enhances its ability to understand and grasp logical flow of information. This change allows StructBERT to significantly outperform BERT in tasks that require a deep understanding of text structure and consistency.

MacBERT uses whole world masking along with a word correction task in order to modify the masked language model's pre-training target. Since word segmentation is crucial for Chinese NLP applications, these adjustments allow MacBERT to excel in such areas. Named improvements also guarantee the modified BERT to learn more meaningful world representations, while increasing the model's resilience.

DeBERTa separates positional and content information to overcome constraints in BERT's attention mechanism. Due to this breakthrough, DeBERTa can now more effectively capture relationships and dependencies within the text, which improves contextual understanding. Natural language inference and text categorizations are DeBERTa' s specialties due to its improved position encoding methodology, leading to higher capabilities of handling long-range dependencies. DeBERTa outperforms BERT and other versions in terms of accuracy and contextual comprehension.

Each different NLP task and domain produce a different result in performance for these BERT-based models. An example being RoBERTa' s improved accuracy in general NLP tasks due to its dynamic masking and larger training dataset. Chinese NLP tasks will be used to create a comparative and thorough analysis on all of these models. Through evaluation on specific datasets, we seek to determine their effectiveness and identify the best-performing models for various NLP applications.

DOI: 10.54254/2755-2721/132/2024.20851

3. Methodology

3.1. Datasets

We chose two tasks - sentiment analysis and named entity recognition - with two datasets each. These datasets were obtained from the Chinese K-BERT GitHub code. Specifically:

3.2. Sentiment Analysis:

Book Review: This dataset includes some comments on different books.

ChnSentiCorp: This dataset includes some user reviews in the area of e-commerce.

3.3. Named Entity Recognition (NER):

Financial NER: This dataset includes financial texts with labeled entities such as names of organizations, person names and locations

MSRA NER: This dataset includes various types of named entities such as names of people, locations, and organizations in Chinese text.

Additionally, the Chinese knowledge graph CnDbpedia is used for knowledge injection in K-BERT.

3.4. Baseline Model

BERT is used as the baseline model for all experiments. The baseline performance is established using the standard BERT model without any additional enhancements or modifications.

3.5. Training and Fine-tuning

As the figure 1 shows, we directly fine-tuned RoBERTa, DeBERTa, MacBERT and StructBERT on the datasets using the pre-trained models and tokenizers available from the Transformer library provided by Huggingface.

In particular, K-BERT is fine-tuned with dataset by following procedures:

(1) Knowledge Injection: External knowledge from the provided knowledge graph is integrated into the training data.

(2) Soft-position Embedding: Embed training data with position information, injecting structural information to model.

(3) Seeing Layer and Visible Matrix: A visible matrix is created to control the visibility of tokens to each other, ensuring that injected knowledge does not introduce noise and alter the original sentence meaning.

(4) Mask-Transformer: Modified transformers with mask-self-attention mechanisms are used to limit the self-attention region based on the visible matrix, preserving the integrity of the original sentence's semantics during fine-tuning.

DOI: 10.54254/2755-2721/132/2024.20851

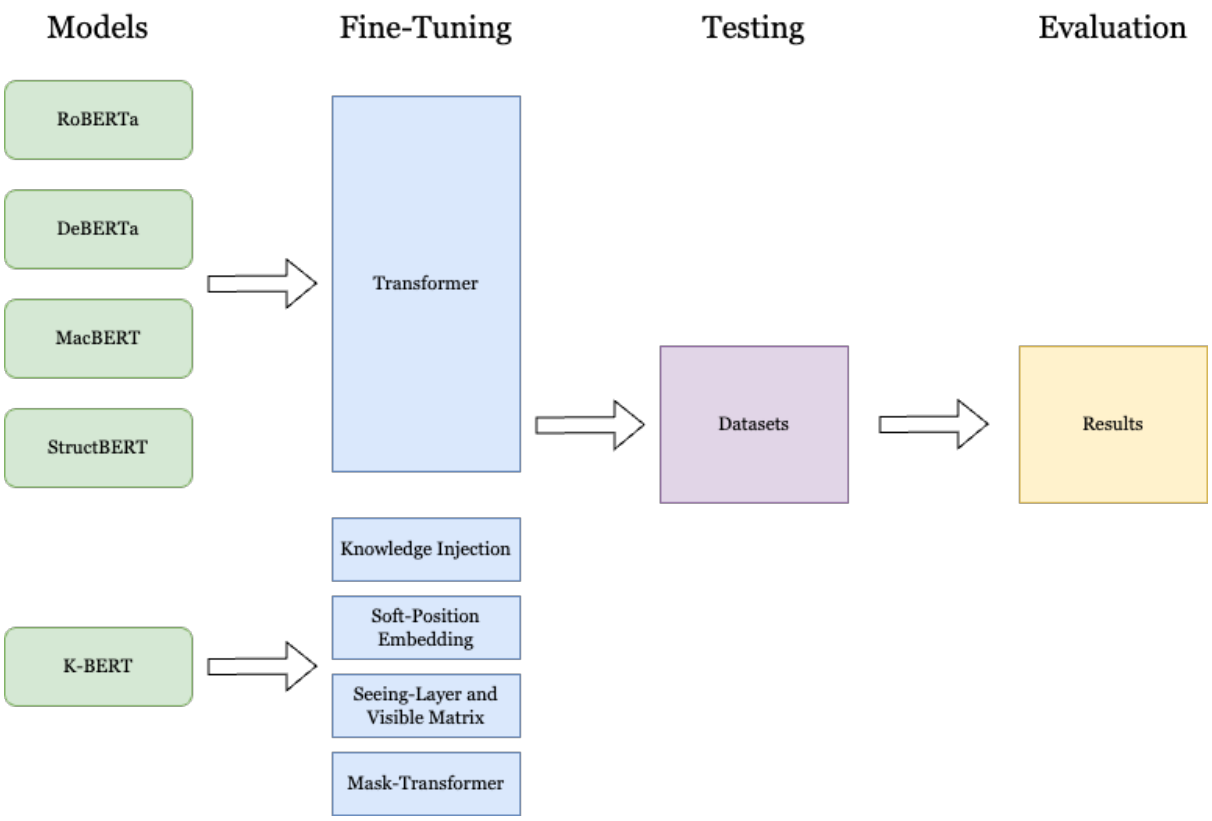

Figure 1. the finetune process of RoBERTa, DeBERTa, MacBERT, StructBERT and K-BERT

For all models and tasks, we evaluated the fine-tuned models on validation datasets and recorded the best version every epoch. In the final evaluation process, we test these best models with the same test dataset.

3.6. Evaluation Metrics

For sentiment analysis datasets, we evaluate these models using accuracy and F1-score. For NER tasks, models are evaluated using precision, recall, and F1-score.

4. Results

The comparison of the BERT-based models in Chinese NLP tasks has revealed that the performances are capable of varying significantly given the datasets and tasks. We evaluated these models in terms of their F1 scores on two sentiment analysis datasets: Book Review, and ChnSentiCorp; as well as their performance for the NER task with Financial NER and MSRA data. The results are shown in the table below to give an insight into the advantages and disadvantages of each model.

Table 1. Comparison of different model's performance under several dataset

(F1 score)	StructBERT	MacBERT	K-BERT	BERT	DeBERTa
Book Review	0.8932	0.8776	0.8708	0.8108	0.8036
Chnsenticorp	0.9583	0.9356	0.9417	0.864	0.9341
financial NER	0.801	0.9715	0.876	0.8527	0.8501
MSRA NER	0.8413	0.878	0.955	0.9415	0.9573

Highlighted bar indicates the model with best performance.

DOI: 10.54254/2755-2721/132/2024.20851

4.1. Sentiment Analysis

In the sentiment analysis task on the Book Review dataset, StructBERT achieved the top among the rest of the models as presented in the table with a score of 0.8932, followed by RoBERTa with a score of 0.8817. K-BERT and MacBERT achieved slightly lower results with scores of 0.8708 and 0.8776, respectively. DeBERTa performed the worst with scores of 0.8087 and 0.8036.

Similarly, in the ChnSentiCorp dataset, StructBERT achieved the highest score of 0.9583. RoBERTa and K-BERT followed by only a slightly lower rating of 0.9516 and 0.9417, respectively. MacBERT and DeBERTa were next with a score of 0.9356 and 0.9341. The baseline BERT model got a 0.864, which was significantly lower than all of the other models.

4.2. Named Entity Recognition (NER)

In the Financial NER task, MacBERT became an outlier and achieved the highest score of 0.9715, significantly higher than all the other models. K-BERT and DeBERTa got scores of 0.876 and 0.8501. RoBERTa and StructBERT fell behind with scores of 0.8045 and 0.801. The baseline BERT model scored 0.8527, which outperformed RoBERTa and StructBERT.

As for the MSRA NER dataset, DeBERTa emerged at the top with a score of 0.9573, along with K-BERT whom also showed convenient results with a score of 0.955, while the baseline BERT model yielded a score of 0.9415. On the contrary, MacBERT, RoBERTa and StructBERT delivered very weak performance with scores of 0.878, 0.874 and 0.8413, respectively.

5. Conclusion

The comparative analysis of the results of the NLP tasks reveals that both StructBERT and RoBERTa are somewhat equally consistent in their performances across all tasks, and are also among the most efficient models, specially for sentiment analysis which StructBERT consistently outperforms in. The performance of DeBERTa was generally acceptable and stable, while it excelled significantly in the MSRA NER task. On the other hand, the excellent performance of MacBERT in the Financial NER task revealed theories about its applications in specialized fields. The utilization of external knowledge is one of the great advantages of K-BERT in the NER tasks, particularly, in MSRA NER dataset. The baseline BERT model, though is already strong, but is constantly surpassed by its improved variants in every task.

These results help to decide, for example, which model is suitable for a certain Chinese NLP task and how to choose a particular model based on the characteristics of the dataset and the task. For the further research in the next step, we should focus the reasons behind these disparities in the performances and the directions to improve these models for better and MORE Chinese language understanding.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, 2019

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, “RoBERTa: a Robustly Optimised BERT Training Approach,” arXiv preprint arXiv:1907.11692, 2019

[3] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, “K-BERT: Enabling Language Representation with Knowledge Graph,” arXiv preprint arXiv:1909.07606, 2019

[4] W. Wang, B. Bi, M. Yang, C. Wu, Z. Bao, J. Xia, L. Peng, L. Si, “StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding,” arXiv preprint arXiv:1908.04577, 2019

[5] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, G. Hu, “Revisiting Pre-trained Models for Chinese Natural Language Processing,” arXiv preprint arXiv:1908.04577, 2020

[6] P. He, X. Liu, J. Gao, W. Chen “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” arXiv preprint arXiv:2006.03654, 2021