JN666 at SemEval-2024 Task 7: NumEval: Numeral-Aware Language Understanding and Generation

Xinyi Liu, Xintong Liu and **Hengyang Lu**
School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
1130174701@qq.com
liuxintong@stu.jiangnan.edu.cn
luhengyang@jiangnan.edu.cn

Abstract

This paper is submitted for SemEval-2027 task 7: Enhancing the Model's Understanding and Generation of Numerical Values. The dataset for this task is NQuAD [1], which requires us to select the most suitable option number from four numerical options to fill in the blank in a news article based on the context. Based on the BertForMultipleChoice model, we proposed two new models, MC BERT and SSC BERT, and improved the model's numerical understanding ability by pre-training the model on numerical comparison tasks. Ultimately, our best-performing model achieved an accuracy rate of 79.40%, which is 9.45% higher than the accuracy rate of NEMo [1].

1 Introduction

In the field of Natural Language Processing (NLP), the understanding and analysis of textual data have always been the main focus. However, the numerical information in these textual data is often overlooked. Numerals play a significant role in our daily life and work, providing rich information such as dates, times, quantities, proportions, and money. Although numerals may not occupy a large proportion in the text, their existence is crucial for understanding the meaning of the text.

To better evaluate the numerical understanding ability of models, [1] proposed the Numeral-related Question Answering Dataset (NQuAD) [1], which is specifically designed to evaluate and enhance the model's ability in Reading Comprehension of the Numerals in Text. In our work, our goal is to improve the performance of the BERT model on the NQuAD [1]. For this purpose, we propose a new method that can effectively handle numerical information. Our experimental results show that our method has achieved significant performance improvement on the NQuAD [1]. This proves the effectiveness of our method in handling numerical information and also shows the potential of our method in enhancing the numerical understanding ability of the model.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the dataset and tasks. Section 4 describes our method in detail. Section 5 reports our experimental results and analysis. Finally, we conclude our work.

2 Related Work

In recent years, pre-training tasks have become increasingly prevalent in the field of Natural Language Processing (NLP). The design goal of pre-training tasks is to enhance the model's understanding of natural language through learning from a large amount of unlabeled data, thereby improving the model's performance on specific tasks in the subsequent fine-tuning stage. In this trend, [2] shows that by pre-training on numerical comparison tasks, the model's understanding of numerals can be significantly improved. The model, after pre-training, also shows a significant performance improvement on other numeral-related tasks.

Multiple choice [3] format represents a category within machine reading comprehension tasks. In this context, [4] proposed a Multi-stage Multi-task learning framework (MMM) aimed at enhancing the performance of multiple choice tasks, [5]introduced the Dual Co-Matching Network (DCMN+), a model that emulates human problem-solving strategies, and [6] presents a two-step strategy for enhancing the performance of Large Language Models (LLMs) on multiple choice tasks.

Recent research has increasingly recognized the significance of numerical data within text. The NQuAD [1] dataset has been instrumental in examining the relationship between numerical values in news headlines and the corresponding figures within the articles. Similarly, the FNXL [7] dataset has brought attention to the numerical data con-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 497–502

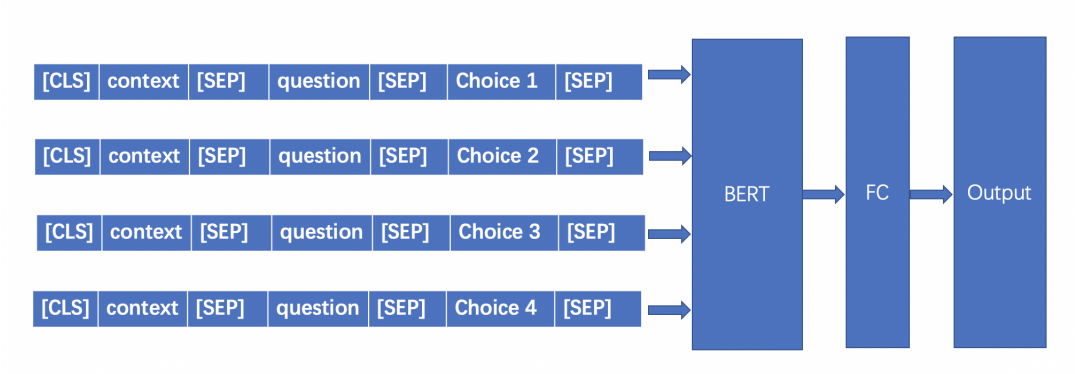

Figure 1: **Architecture of BertForMultipleChoice model**

tained within the periodic financial reports submitted by publicly listed companies.

The BertForMultipleChoice model is a variant of the BERT [8], specifically designed for multiple-choice tasks. Figure 1 is the architecture of BertForMultipleChoice, it adds a multiple-choice classification head on top of the BERT model, mainly for handling multiple-choice tasks. The input to the BertForMultipleChoice model includes a question and multiple potential answers (options), and the goal of the model is to select the most reasonable answer. This model has shown excellent performance in tasks that require choosing one answer from multiple options, such as reading comprehension and sentiment analysis.

3 Dataset and Tasks

Figure 2 shows an example in NQuAD [1] dataset, including a news article, a question stem and four answer options. Our task is predicting the correct the option.

NQuAD [1] collects news articles from the data vendor, MoneyDJ1, and get the news articles within the period from June 22, 2013 to June 20, 2018. A total of 75,448 Chinese news articles are collected. A total of 43,787 news articles are selected, and 46.97% of the headlines contain more than one numeral. The average number of numerals in the headline and in the content are 1.65 and 29.48, respectively. Each numeral in each headline is used to form a question, thus NQuAD [1] dataset finally obtain 71,998 questions and separate 80% of the instances as the training set and the rest of the instances form a test set.

News Article:
Major banks take the lead in self-discipline. The five major banks' newly imposed mortgage interest rates climbed to **1.986**% in May. ... Also approaching **2**% integer alert ... Up to **2.5**% ... Also increased by **0.04** percentage points from the previous month ... Prevent the housing market bubble from fully starting.

Question Stem: Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly___%.

Answer Options: (A) 0.04 (B) 1.986 (C) 2 (D) 2.5

Answer: (C)

Figure 2: **An example question in NQuAD.**

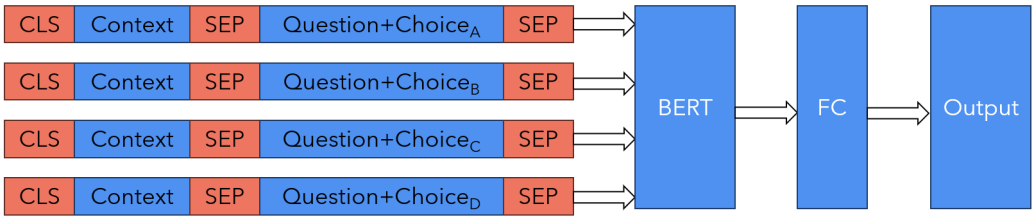

Figure 3: **Architecture of the proposed model, MC BERT**

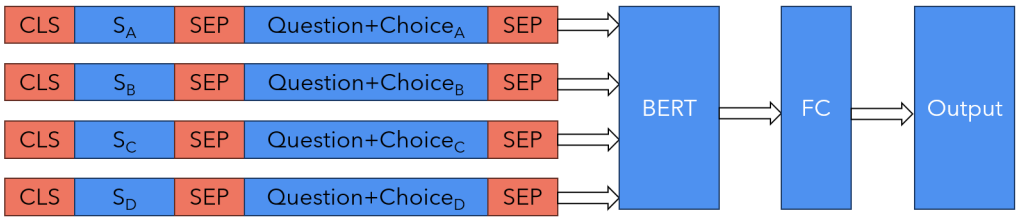

Figure 4: **Architecture of the proposed model, SSC BERT**

MC BERT Input:
[[CLS]Also increased by **0.04** percentage points from the previous month;he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.;Also approaching **2**% integer alert;Up to **2.5**[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **0.04**%.[SEP],
[CLS]Also increased by **0.04** percentage points from the previous month;he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.;Also approaching **2**% integer alert;Up to **2.5**[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **1.986**%.[SEP],
[CLS]Also increased by **0.04** percentage points from the previous month;he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.;Also approaching **2**% integer alert;Up to **2.5**[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **2**%.[SEP],
[CLS]Also increased by **0.04** percentage points from the previous month;he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.;Also approaching **2**% integer alert;Up to **2.5**[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **2.5**%.[SEP]]

SSC BERT Input:
[[CLS]Also increased by **0.04** percentage points from the previous month[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **0.04**%.[SEP],
[CLS]he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **1.986**%.[SEP],
[CLS]Also approaching **2**% integer alert[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **2**%.[SEP],
[CLS]Up to **2.5**[SEP]Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **2.5**%.[SEP]]

Figure 5: **Examples of MC BERT Input and SSC BERT Input.**

4 Methods

4.1 Multiple Choice

Multiple Choice [3] format represents a category within machine reading comprehension tasks. Our present endeavor aligns with this format. The BertForMultipleChoice model, an adaptation of the BERT [8] framework, is expressly engineered to process tasks involving this format. So we choose the BertForMultipleChoice model as our baseline model.

Resent research [5] suggests that existing multi-choice MRC models learn the passage representation with all the sentences in one-shot, which is inefficient and counter-intuitive. Their research indicates that the model should be extremely beneficial if it focuses on a few key evidence sentences. At the same time, the "sentences_containing_the_numeral_in_answer_options" in the NQuAD [1] dataset are four sentences that contain each of the options. So, our strategy is using this sentences as the context of inputs. In the BertForMultipleChoice model, we connect this sentences as the context, and the Question Stem and Answer Options are used as the question and Choices. After that, we find that humans often employ the method of substituting potential solutions into the given problem in the context of problem-solving. Be inspired by this, we fill in the blanks of the question with the options. As shown in Figure 3, we named this new model MC BERT. Furthermore, when humans solve problems, they will finally compare the question and options with key sentences one by one, and then choose the most matching option. So we further propose an improvement strategy, that is, changing the context to the sentence corresponding to each option to increase differentiation. As shown in Figure 4, we named this improved model SSC BERT. Here are examples of input:

- Context:Also increased by **0.04** percentage points from the previous month,he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.,Also approaching **2**% integer alert,Up to **2.5**%
- S_A: Also increased by **0.04** percentage points from the previous month
- S_B: he five major banks' newly-imposed mortgage interest rates climbed to **1.986**% in May.
- S_C: Also approaching **2**% integer alert
- S_D: Up to **2.5**%
- Question: Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly ___%.
- $QC_A/QC_B/QC_C/QC_D$: Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly **0.04/1.986/2/2.5**%.

Figure 5 are final input examples of MC BERT and SSC BERT.

4.2 Pre-training

Chapter 2 mentioned that pre-training models on simple numerical comparison tasks can enhance the model's numerical understanding. Therefore, we pre-train 'bert-base-chinese' on numerical comparison tasks according to the method mentioned in the [2]. The pre-training model is then trained according to the three frameworks mentioned above. We finally obtain six models: BertForMultipleChoice, MC BERT, SSC BERT, PreBertForMultipleChoice, Pre-MC BERT and Pre-SSC BERT.

4.3 Implementation Details

Trained on the ComNum dataset [2]. Given that NQuAD [1] is a Chinese dataset, we first replaced the English in the ComNum dataset [2] with the corresponding Chinese. The pre-trained model we used is bert-base-chinese, the maximum sequence length is 32, batch size is 32. The loss function employed is the cross-entropy loss, with the AdamW optimizer. The learning rate is set at 5e-6, and epsilon is 1e-8.

Trained on the NQuAD dataset [1]. We employed the BertForMultipleChoice from the transformers library to train the model and divided the training set into a training set and a validation set at a ratio of 4:1. The maximum sequence length is 128, batch size is 32. The loss function employed is the cross-entropy loss, with the AdamW optimizer. The learning rate is set at 5e-6 (If the model used was trained on the ComNum dataset [2], the learning rate is set at 5e-5), and epsilon is 1e-8. After applying softmax function to the model's output, we selected the index with the highest probability as the output.

Model	Accuracy
BERT Embedding Similarity	57.30%
Vanilla BERT	66.41%
BERT-BiGRU	67.15%
BERT-CNN	63.92%
NEMo	69.95%
BertForMultipleChoice	74.48%
MC BERT(ours)	76.83%
SSC BERT(ours)	78.21%
Pre-BertForMultipleChoice	74.91%
Pre-MC BERT(ours)	77.16%
Pre-SSC BERT(ours)	**79.40**%

Table 1: Experimental results.

5 Result

As shown in Table 1, the accuracy of our six models all exceed NEMo [1], which indicates that the BertForMultipleChoice model framework can effectively handle and solve the specific requirements and challenges of this task. Among the three models obtained by directly training with bert-base-chinese, the accuracy of MC BERT is higher than that of BertForMultipleChoice, indicating that directly filling in the blanks of the question with options is effective. Furthermore, the accuracy of SSC BERT surpasses that of MC BERT, suggesting that comparing the question and options with key sentences individually can also enhance the model's performance. Meanwhile, The accuracy of the three models trained using pre-trained models are higher than that of the models obtained by directly training with bert-base-chinese, which further confirms the conclusion that in some simple numerical related tasks, pre-trained models can enhance the model's numerical understanding ability [2].

6 Conclusion

In this work, we select the BertForMultiple model as the baseline model and propose two new models based on it: MC BERT and SSC BERT. The accuracy of these models all exceeded the results of NEMo [1], which confirmed the applicability of the BertForMultiple model framework for our task, the results also indicate that the method of inserting options into the blanks of the question and individually comparing them with key sentences is effective. We also refer to the method in reference [2], pre-trained the bert-base-chinese on numerical comparison task, and use the pre-trained model for subsequent training. The experimental results show that this method can improve the model's numerical understanding ability, thereby making the model perform better on numerical related tasks. Finally, we found that the Pre-SSC BERT model had the highest accuracy, and its accuracy was 9.45% higher than the NEMo model [1], which further proved the effectiveness of our method.In summary, our research proposes a new model training and optimization strategy, which has been proven to be effective and superior in experiments.

However, in analyzing the cases where our model made incorrect judgments, we identify some shortcomings. We notice that humans can easily understand the equivalence between different descriptions of the same thing, but the model cannot. For example, our model cannot understand that '%' and 'Cheng'('into')are equivalent, 'EPS' and 'Earning Per Share' are equivalent, and it cannot understand that 'January' is 'Q1'. To address this issue, we tried to add some specific examples to the dataset of numerical comparison tasks, such as '10% is equal to 1 Cheng' and 'EPS 100 is equal to Earning Per Share 100', hoping that the model could understand the equivalence between two different descriptions of things through this method. However, the final result did not meet our expectations, indicating that our model still needs improvement in handling these types of problems. We will continue to explore this issue in future research, with the aim of improving the model's understanding ability and accuracy.

7 ACKNOWLEDGMENTS

This research was funded by the China Postdoctoral Science Foundation (Grant No. 2022M711360), in part by the Laboratory for Advanced Computing and Intelligence Engineering.

References

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "NQuAD: 70,000+ Questions for Machine Comprehension of the Numerals in Text". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM '21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 2925–2929. ISBN: 9781450384469. DOI: 10.1145/

3459637.3482155. URL: https://doi.org/10.1145/3459637.3482155.

[2] Chung-Chi Chen et al. "Improving Numeracy by Input Reframing and Quantitative Pre-Finetuning Task". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 69–77. DOI: 10.18653/v1/2023.findings-eacl.4. URL: https://aclanthology.org/2023.findings-eacl.4.

[3] Shanshan Liu et al. "Neural Machine Reading Comprehension: Methods And Trends". In: *APPLIED SCIENCES-BASEL* 9.18 (2019).

[4] Di Jin et al. "Mmm: Multi-Stage Multi-Task Learning For Multi-Choice Reading Comprehension". In: *National Conference on Artificial Intelligence* 34.05 (2020), pp. 8010–8017.

[5] Shuiliang Zhang et al. "DCMN+: Dual Co-Matching Network for Multi-Choice Reading Comprehension." In: *THIRTY-FOURTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, THE THIRTY-SECOND INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE AND THE TENTH AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE* 34.05 (2020), pp. 9563–9570.

[6] Chenkai Ma and Xinya Du. "POE: Process of Elimination for Multiple Choice Reasoning". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4487–4496. DOI: 10.18653/v1/2023.emnlp-main.273. URL: https://aclanthology.org/2023.emnlp-main.273.

[7] Soumya Sharma et al. "Financial Numeric Extreme Labelling: A dataset and benchmarking". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3550–3561. DOI: 10.18653/v1/2023.findings-acl.219. URL: https://aclanthology.org/2023.findings-acl.219.

[8] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.