Numerical Reasoning with mT5-Small: Addressing Cloze Test Challenges

Jianing Lei
jianingl@chalmers.se

Hongliang Zhong
zhongh@chalmers.se

Abstract

This report focused on Subtask 2 of SemEval 2024 Task 7: Comprehension of Numerals in Text, using the NQuAD dataset (Chen et al., 2021), which is in traditional Chinese and consists of over 70,000 questions. To address the task, we fine-tuned the mT5-Small model with a structured instruction template, achieving an accuracy of 87.00% [1], outperforming several BERT-based methods while closely approaching the top-performing T5-based model (Chen et al., 2024a). Despite computational limitations, our approach demonstrated the effectiveness of instruction fine-tuning in enabling the model to reason about numerical information. Furthermore, we introduced specific evaluation rules to handle edge cases in the dataset, ensuring the validity and reliability of the model's outputs. These findings highlight the potential of multilingual models in handling complex numerical reasoning tasks efficiently, providing a foundation for future research.

1 Introduction

The ability to accurately comprehend and generate numerical information embedded within text is crucial across numerous domains such as medicine, engineering, and finance, while numerical data plays an important role in decision-making and effective planning. However, existing language models face some challenges in representing and processing textual numerals effectively. This limitation can result in inaccuracies on tasks requiring numerical understanding (Chen et al., 2023).

[1] Our code is available here.

Recognizing the importance of numeracy in language tasks, Task 7 of SemEval 2024 (Chen et al., 2024b) emphasizes enhancing models' capabilities to comprehend and generate numerically-aware text. This task encompasses diverse objectives, including quantitative analysis (Chen et al., 2023), comprehension of numerals within text (Chen et al., 2021), and the generation of numeral-sensitive headlines (Huang et al., 2023). In this project, we focus on Subtask 2 of Task 7: Reading Comprehension of Numerals in Text. Our approach utilizes the mT5-small model, which was pre-trained on a multilingual corpus. To improve performance, we designed a task-specific instruction template for inputs and applied instruction fine-tuning.

2 Problem Statement

Task 2 focuses on evaluating a model's ability to perform numerical reasoning by selecting the correct numerical value from a set of four options. This task adopts a cloze test format, requiring the model to understand a news article to accurately answer a corresponding question. Table 1 illustrates an example drawn from the original paper. Based on this format, we design a suitable instruction template for the model to improve the performance in this task.

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)

Table 1: An example question in NQuAD.

3 Methodology

3.1 Dataset

The dataset utilized for this task is NQuAD[2] , a publicly available dataset in Traditional Chinese, specifically designed for tasks involving numerical reasoning. It comprises over 70,000 questions, divided into a training set with 57,750 samples and a test set with 14,438 samples. Each data entry in the dataset includes a news article, a question stem, multiple-choice answer options, the correct answer, and a target number. This comprehensive dataset provides a foundation for exploring various approaches to solving the problem, such as instruction fine-tuning and formulating the task as a classification problem.

3.2 Instruction Tuning

To enhance the model's ability to process and reason about numerical reasoning tasks, we structured the data using a well-defined instruction template. Each input begins with the directive: "*Choose the correct answer to the following questions.*" This is followed by the context, the question, and the answer options, presented in a clear and consistent format. Notably, due to memory constraints, we used only the sentences containing numerical information instead of the full article.

Table 2 illustrates an example of the input and output from the training set, presented here as a translated version of the original data for understanding. By providing the model with this structured format, we enable it to better comprehend the relationships between the context and the question, ultimately improving its performance on the task.

3.3 Model Selection

The mT5 (Multilingual Text-to-Text Transfer Transformer) model is a multilingual variant of the T5 model, designed specifically to handle tasks in multiple languages. It was pre-trained on a massive dataset derived from Common Crawl, covering 101 languages, including Chinese (Xue et al., 2021). One of its key features is treating every natural language processing task as a text-to-text problem, making it suitable for tasks like classification, translation, and cloze tests. This makes mT5 particularly well-suited for the challenges of our task.

[2]The NQuAD dataset is available here.

(Input) Choose the correct answer to the following questions. **Context:** ['The cooperation between the two parties has lasted for over **20** years.', 'Aircraft landing gear and subsystem-related component manufacturer Sheng Tian (4541) will officially list on the emerging stock market on June **25**.', 'Gross margin of **27%** and after-tax net profit of 47 million yuan.', 'Gross margin of **29%** and after-tax net profit of 82 million yuan.'] **Question:** Sheng Tian will list on the emerging stock market on June ___, planning to apply for a main board listing by the end of the year. **Options:** 20, 25, 27, 29 **Answer:**
(Output) **Targets:** 25

Table 2: An example of instruction tuning (translated version).

Due to computational constraints, we selected the mT5-Small model, the smallest variant in the mT5 family. This variant has significantly fewer parameters than its larger counterparts, resulting in reduced computational requirements, faster training times, and lower costs. Despite its compact size, the mT5-Small model achieves competitive performance, offering an effective balance between efficiency and accuracy.

3.4 Hyper-parameter Selection

For fine-tuning the mT5-Small model, we used a learning rate of 3e-4, with a training batch size of 16 and an evaluation batch size of 8. To address computational constraints, gradient accumulation was set to 4 steps, effectively simulating a larger batch size of 64. The model was trained for up to 10 epochs, with evaluations conducted every 50 steps and logging every 100 steps to monitor progress and performance. To prevent overfitting, a weight decay of 0.01 was applied, and the best model was automatically selected based on validation loss. Additionally, we applied early stopping with a patience of 2, which allowed the training process to halt early if no significant improvement was observed. As a result, training concluded at step 1250 during Epoch 1, ensuring efficient and effective resource utilization while maintaining model performance.

4 Experiments and Results

4.1 Evaluation Metrics

To assess the performance of the model, accuracy was used as the evaluation metric. Accuracy measures the proportion of correct predictions out of the total number of samples and provides a straightforward and effective means of evaluating the model's ability to select the correct answer.

During the evaluation process, certain rules were introduced to enhance the accuracy of the evaluation process by addressing potential limitations in the dataset and mitigating issues with generative outputs. Upon reviewing the dataset, we found some entries in which the target value was not included among the provided answer options. Therefore, these entries were excluded from the evaluation process, as the model's predictions are restricted to the four given options. Additionally, in cases where the model generated outputs that were invalid (e.g., None) or not among the provided answer options, a regeneration mechanism was applied. The model was allowed to regenerate its output up to a maximum of three attempts to ensure valid predictions. This setting was implemented to prevent infinite loops while maintaining the validity of the outputs.

4.2 Results

Our model achieved an accuracy of 87.00%, demonstrating competitive performance in the numerical reasoning task. As shown in Table 3 (Chen et al., 2024a), this result surpassed two other competing methods, JN666 (79.40%) and CYUT (77.09%), both of which relied on BERT-based approaches with various enhancements. However, our performance shows a narrow gap with YNU-HPCC, who achieved a slightly higher accuracy of 89.71% with Randeng-T5-77M, despite both approaches utilizing T5-based models. This good performance of our model highlights the effectiveness of the instruction fine-tuning strategy, which enables the model to accurately comprehend and reason about numerical information embedded within text.

Team	Method	Accuracy (%)
YNU-HPCC	Randeng-T5-77M	89.71
JN666	BERT + Pre-Finetuning with Comparing Number Task	79.40
CYUT	BERT + Number Augmentation + Features	77.09
Ours	**mT5-Small + Instruction Fine-Tuning**	**87.00**

Table 3: Performance comparison with other teams

5 Conclusion

In this project, we successfully completed the task, achieving high accuracy to demonstrate the application of mT5-Small for numerical reasoning in a multilingual context. Throughout the process, we implemented instruction fine-tuning to develop the system, identified certain defects in the original dataset, and introduced specific rules to evaluate the model's results. These measures ensured that the generated outputs were both valid and accurate, contributing to the reliability and effectiveness of our approach.

6 Limitations

This project has some limitations that should be addressed. Due to constraints in computational resources, we simplified the input by only using sentences containing the numerical values found in the answer options. While this approach reduces resource requirements, it may overlook other important contextual information. If sufficient computational resources were available, extracting all sentences containing numerical values from the article as input could provide a more comprehensive context. This improvement would help the model generate more accurate predictions, particularly in cases where the correct answer is not included in the provided options.

7 Ethics Statement

The dataset used in this study, NQuAD, is publicly available and does not contain sensitive or personally identifiable information. However, as this work involves numerical reasoning in a multilingual context, there is a potential risk of model misuse in contexts where numerical precision is critical. We advocate for the responsible use of this technology, ensuring it is applied within appropriate domains and ethical guidelines.

References

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. 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*, pages 2925–2929, New York, NY, USA. Association for Computing Machinery.

Chung-chi Chen, Jian-tao Huang, Hen-hsen Huang, Hiroya Takamura, and Hsin-hsi Chen. 2024a.

SemEval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1482–1491, Mexico City, Mexico. Association for Computational Linguistics.

Chung-Chi Chen, Jian-Tao Huang, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2024b. Semeval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023. Improving numeracy by input reframing and quantitative pre-finetuning task. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 69–77, Dubrovnik, Croatia. Association for Computational Linguistics.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Numhg: A dataset for number-focused headline generation. *arXiv preprint*, arXiv:2309.01455.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 483–498. Association for Computational Linguistics.