NumEval: Numeral-Aware Language Understanding and Generation

Reading Comprehension of the Numerals in Text (Chinese)

1. Abstract:

In the realm of advanced natural language processing (NLP) and artificial intelligence (AI), understanding numerals within text, especially in languages like Chinese, remains a formidable challenge. This research addresses this challenge by employing the cloze test paradigm to identify correct numerical values from options based on news articles. Leveraging the NQuAD dataset tailored for numeral comprehension in Chinese text, the study proposes innovative solutions through progressive models, including Vanilla Bert-QT-semantic, BERT Bi-Gru Modified-CM Model, BERT-fully-encoded BiGRU Model, and BERT-Q-Semantic-GRU. The discussion explores potential improvements such as dataset refinement and transformer-based architectures, contributing to NLP advancement and real-world applications reliant on numerical data comprehension.

2. Introduction:

In the era of advanced natural language processing (NLP) and artificial intelligence (AI), the task of reading comprehension of numerals within text presents a significant challenge, particularly in Chinese language processing. This research project focuses on addressing this challenge through the exploration of the cloze test paradigm, where models are tasked with identifying the correct numerical value from a set of options based on a given news article. The dataset chosen for this endeavour is NQuAD (Numerals Question Answering Dataset), a valuable resource tailored for evaluating machine comprehension of numerals in Chinese text.

The motivation behind this project stems from the growing importance of numerical comprehension in textual data across various domains, including finance, healthcare, and education. In today's data-driven world, accurate interpretation of numerical information is crucial for making informed decisions and deriving meaningful insights. However, existing NLP models often struggle to effectively comprehend numerical data, especially in languages with complex linguistic structures like Chinese.

By focusing on reading comprehension of numerals in Chinese text, this research project seeks to address this gap in NLP research and development. Through the utilization of the cloze test methodology, the authors aim to devise innovative solutions that enhance machine understanding of numerical content within textual contexts. This not only contributes to the advancement of NLP technology but also holds significant implications for real-world applications where numerical data plays a pivotal role.

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 1: An example question in NQuAD.

3. Related Work

Chen, Chung-Chi, Hen-Hsen Huang, and Hsin-Hsi Chen introduced the NQuAD dataset, designed specifically for evaluating machine comprehension of numerals in text, providing a valuable resource for training and assessing models in tasks related to reading comprehension of numerals. Yiming Ju, Yuanzhe Zhang, Zhixing Tian, and Kang Liu proposed a method to enhance multiple-choice machine reading comprehension (MRC) models by penalizing illogical interpretations, offering insights into model reasoning through a case study. Their approach addresses challenges posed by noisy ground-truth interpretations and yields consistent improvements across strong and trivial baselines, suggesting new avenues for NLP

research and enhancing our understanding of model interpretability in complex tasks. MRC involves training models to select the correct answer from a set of options given a passage of text and a question, with various datasets such as RACE, MULTIRC, and DREAM serving as benchmarks for evaluating model performance (Lai et al., 2017; Khashabi et al., 2018; Sun et al., 2019). Mohit Iyyer et al. explored reading comprehension using multiple-choice questions (MCQs) as a benchmark task, discussing different approaches for tackling MCQ-based reading comprehension tasks and comparing their effectiveness on standard datasets.

4. Dataset

The dataset used in this research paper is NQuAD (Numerals Question Answering Dataset), which comprises Chinese news articles collected from the data vendor MoneyDJ. These articles span from June 22, 2013, to June 20, 2018, resulting in a total of 75,448 articles. The dataset primarily focuses on the comprehension of numerals within textual contexts and aims to address the task of selecting the correct numerical value from a set of options based on a given news article.

The dataset features news articles written by professional journalists, ensuring satisfactory quality and reliability of both headlines and content. It encompasses a wide range of topics and events, providing diverse textual contexts for evaluating machine comprehension of numerals.

Here are the various columns of the dataset:

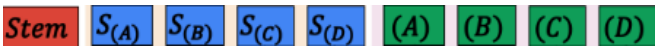
- News Article: This column contains the textual content of the news article. It provides information about the events, announcements, or developments discussed in the article.
- Question Stem: This column presents the stem of the cloze-test question related to the news article.
- Answer Options: This column provides the options for the answer to the question presented in the question stem.
- Correct Answer: This column indicates the correct answer to the question.
- Target Numeral: This column specifies the numerical value that needs to be identified or completed in the question.'.
- Sentences Containing the Numeral: This column contains the 4 separate sentences from the news article that mention the target numeral. It helps provide context for understanding the significance of the numeral within the article.

5. Methodology

In our project, we're building a model to fill in the blanks in a cloze test. We're following a method suggested in the NQuAD paper to train this model. Here's how it works:

1. The task has three inputs: question stem, article text, and options.
2. In the question stem, a masked sentence is given from the article text, and the task is to predict the masked number using the four numerical options given.
3. In further sections, "Stem" represents the question stem, S(x) represents the set of sentences containing that content numeral x, and (A), (B), (C), and (D) represent the options.
4. Options are handled in different ways in different models.

Thus, for each of the models, our data loader for model inputs looked like the the one shown below:



We then implemented the following models progressively for our task:-

5a. Vanilla Bert- QT-semantic

The Vanilla Bert-Question-Text-Semantic model works by using basic understanding and a simple method to answer questions about what's been read. In this simple method, you first read and understand the text, then read the questions, and finally use what you've understood to find the answers among the choices given.

For finding the answers, we used a pre-trained model called Bert trained on a chinese text corpora, which is really good at understanding language. Then, we added a layer called a linear layer to help pick out the answers

from what Bert understood. You can see how the model is set up in Figure 2 of our paper. For the option input a 4x1 tensor has been given with the float representation of that particular option.

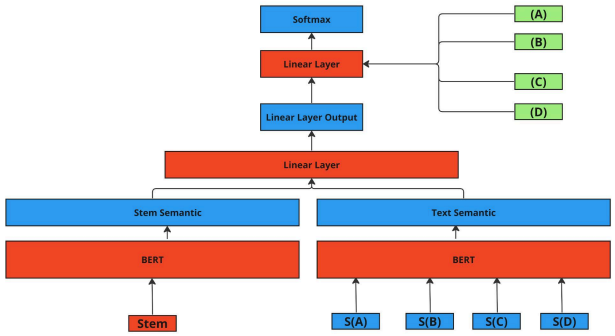


Fig. 2. Vanilla Bert QT-semantic

5b BERT Bi-Gru Modified-CM Model

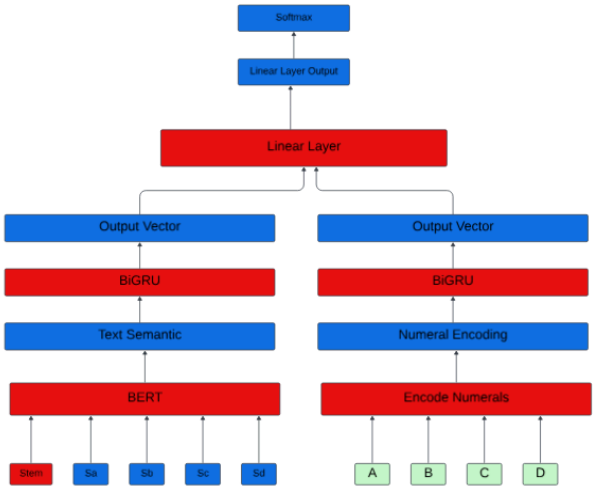


Fig. 3.BERT Bi-GRU Modified-CM Model

In this model we are first creating BERT embedding for all the sentences containing the numeral in the ans--options and question stem which are then further put into a BiGRU model. For answer options we noticed that max length of each answer option in the the dataset is 10, so for each character of our answer option we made one hot vector of size 12, where 0 to 9 are reserved for digits 0-9 , 10 place is for decimal point '.' and 11 place is for negative sign '-'. Therefore each answer option embedded into dimension of 10× 12. We are using right padding in our answer option embedding for options which have length less than 10. After getting output embedding, we put it into a BiGRU model.
Then we concatenate all vectors and add it to a linear layer which gives predictions.

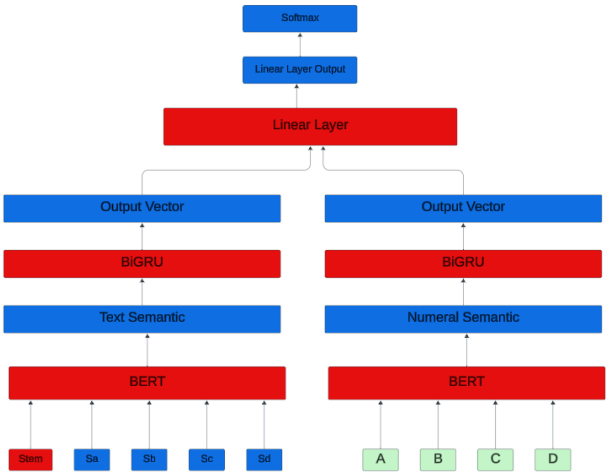


Fig. 4. BERT-fully-encoded BiGRU Model

5c BERT-fully-encoded BiGRU Model

This model is similar to our previous model, except that this time the numerical representation is also done by BERT. The intuition behind this was that since both Numerals and Sentence Words would get represented in the same "BERT" domain, the correlation between both of them would be established in a better way. Since the NQuAD authors didn't test this, we could test whether the correlation of options with their corresponding sentences would help.

Again, these BERT embeddings would be sent to GRU layer, and then similar to the previous model, we concatenate all outputs to put into Linear Layer.

5d BERT-Q-Semantin-GRU

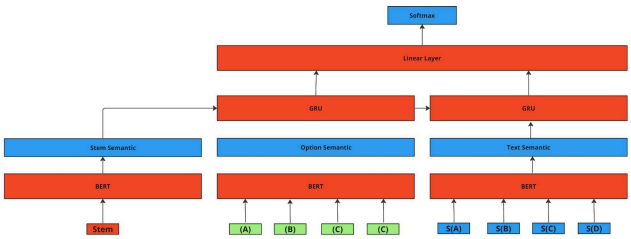


Fig. 5. BERT-Q-Semantin-GRU

This model is like how students approach reading comprehension questions in exams. First, they read and understand the question and choices, then they read the text.

Similarly, our model uses a BERT encoder to understand the meaning of the text. The meaning of the question is used as a starting point for a type of model called GRU, which helps process the choices. Then, the output from processing the choices is used to process the article text.

Finally, we use the last important piece of information from both the choices and the article text to figure out the answer using a special calculation called linear extraction.

6. Results/Findings:

We have used accuracy as our evaluation metric to test our models for the task, which will tell how accurate the options were matched.

The following were our performances for each of the models(based of accuracy of option prediction):-

Table 1

Model	Accuracy
Vanilla Bert- QT-semantic	48.54%
BERT Bi-Gru Modified-CM Model	**62.41%**
BERT-fully-encoded BiGRU Model	59.22%
BERT-Q-Semantin-GRU	57.1%

After training each model on a full training dataset, the testing was done on the model using the test dataset. On this, the model **BERT Bi-Gru Modified-CM Model** performed best, with the overall accuracy of **62.4%.**

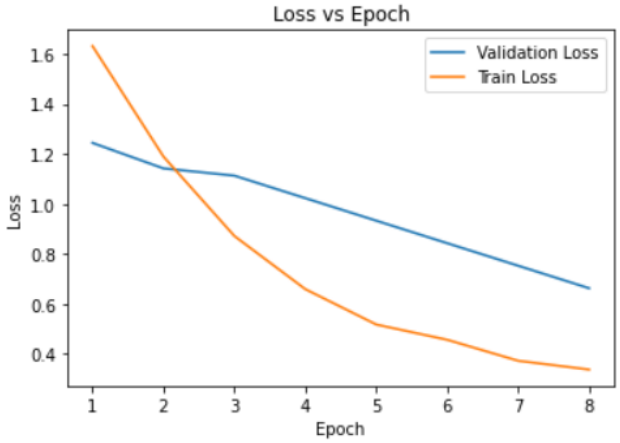


Fig. 5. Plot for BERT Bi-GRU Modified-CM

7. Discussion, observation and analysis:-

From our analysis, one thing was indeed indicative:- BiGRU layer onto BERT embeddings proved effective onto the model, since Vanilla Bert performed lesser than other three models having GRU layer(s) in them.

Other inferences that can be drawn is that the Modified-CM encoding for numerals was much more effective for this task, than the pre-trained embeddings, Since Modified-CM worked better than BERT-fully connected. This hints that the inter-relation between the option-numerals can indeed hinder the prediction, and thus our Modified-CM one-hot encoding of numerals proved better since it conceals its relation with other numbers.

Our models have been able to follow the results of the NQuAD paper, where they also showed how one-hot Numeral encodings of options are better at representation than pre-trained embeddings. The way their model worked better for CM-embeddings is

similar to how our modified-CM worked better than our other models.

8. Conclusion

In conclusion we focused on addressing the challenge of solving numeral related questions. The experimental results show that the BERT Bi-GRU Modified-CM performs better than other models. Our findings contribute to the advancement of NLP technology by offering insights into the complexities of numeral question answering and providing guidance for future research endeavors in this domain.

9. Future Work and Possible Extensions:

There are many ways through which we can improve and extend this task for better use cases, keeping in mind the primary goal of the task, which was "Answering Numeral-related questions from dataset". One such improvement, as pointed out by authors of NQuAD themselves, was to change the dataset in the perspective of its question-asking. Currently, the NQuAD only has masked headlines of the articles of the questions, which should be changed to actual question-style format. For instance, in the Figure-1, the question-stem in the usual use-case could be "What are the new mortgage-interest of the five-banks". This way, the model can get trained for a more general use-case. (Chen, Huang, & Chen, 2021)
Another thing that can be done was the change of style in the training architecture. As described above, we are performing overall training only on those filtered-out sentences having option numerals. However, on analysis, we also found that on the samples wrongly predicted, the semantics could've been drawn from the left-over sentences, which would add the remaining practical semantics to the model, although, as stated by the authors, this may also confuse the model on a high level due to too many sentences.

10. References:

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "Comprehension of the Numerals in Text." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2017.

[2] Yiming Ju, Yuanzhe Zhang, Zhixing Tian, and Kang Liu. "Enhancing Multiple-choice Machine Reading Comprehension by Punishing Illogical Interpretations." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[3] Mohit Iyyer et al. "Reading Comprehension with Multiple Choice Questions." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018.