HW-TSC at SemEval-2023 Task 7: Exploring the Natural Language Inference Capabilities of ChatGPT and Pre-trained Language Model for Clinical Trial

Xiaofeng Zhao, Min Zhang, MiaoMiao Ma, Chang Su, Yilun Liu, Minghan Wang, Xiaosong Qiao, Jiaxin Guo, Yinglu Li, Wenbing Ma, Shimin Tao, Hao Yang
Huawei Translation Services Center, Beijing, China
{zhaoxiaofeng14, zhangmin186, mamiaomiao, suchang8, liuyilun3 wangminghan, qiaoxiaosong, guojiaxin1, liyinglu, mawenbing taoshimin, yanghao30}@huawei.com

Abstract

In this paper, we describe a effective system for SemEval-2022 Task 7. This task aims to determine whether a given statement is supported by comparing one or two clinical trial reports, and to identify evidence that supports the statement. This is a task that requires high natural language inference capabilities. In Subtask 1, we compare our strategy based on prompt learning and ChatGPT with a baseline constructed using BERT in zero-shot setting, and validate the effectiveness of our strategy. In Subtask 2, we fine-tune DeBERTaV3 for classification without relying on the results from Subtask 1. We find that early stopping of the training can effectively prevent model overfitting, and this achieves a good performance in Subtask 2. In addition, we do not use any ensemble strategies. We have achieved the 10th place in Subtask 1 and the 2nd place in Subtask 2.

1 Introduction

In recent years, the introduction of the pre-trained masked language model (LM) BERT (Devlin et al., 2018; Vaswani et al., 2017) has been a significant milestone in the field of natural language processing (NLP). BERT achieves an absolute improvement of 7.7 points on the General Language Understanding Evaluation (GLUE) benchmark and produces state-of-the-art (SOTA) results in multiple tasks. Subsequently, numerous pre-trained language models (PLMs) based on the transformer architecture appear, such as XLNet (You et al., 2019), which addresses the issue of mask independence, BioBERT (Lee et al., 2020), a domain-specific model focused on biomedical text , and RoBERTa (Liu et al., 2019). More recently, in 2021, DeBERTa (He et al., 2020) has emerged as a powerful model that achieves scores surpassing human performance on the SuperGLUE leaderboard, thanks to its disentangled attention and enhanced decoding mechanisms. In addition to models that reinforce natural language understanding, there are also models that focus on natural language generation, such as GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020). For example, in 2022, ChatGPT and InstructGPT (Ouyang et al., 2022) have gained popularity for their remarkable performance using reinforcement learning and Proximal Policy Optimization. Their chain-of-thought (CoT) ability with super large models enables logical reasoning in zero-resource scenarios.

2 Task Description

The goal of this task (Jullien et al., 2023) is to evaluate the natural language inference capability of systems on clinical trial reports in the medical domain. In recent years, there has been a significant increase in the publication of clinical trial reports, making it impossible for healthcare practitioners to keep up-to-date with all existing reports to provide personalized care. In this context, natural language inference presents an opportunity to retrieve and interpret medical evidence to support personalized care. This paper presents a task based on a set of breast cancer clinical trial registry (CTR) data, with the aim to classify four paragraphs in each CTR - eligibility criteria, interventions, outcomes, and adverse events - based on CTR annotations, interpretations, and domain expert labels. And the data example as shown in Figure 1.

Subtask 1 is textual entailment, where a given statement is evaluated for the inferential relationship (Entailment or Contradiction) with 1-2 specified CTRs. This subtask can be viewed as a binary classification task, where a common approach is to concatenate the statement and all the evidence in the CTRs, assume a pre-defined classification relationship, and predict a label of "0" (Entailment) or "1" (Contradiction). It should be noted that in some cases where the statement may have inherent inconsistencies, the relationship can be predicted as contradiction without reference to the CTR.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1603–1608

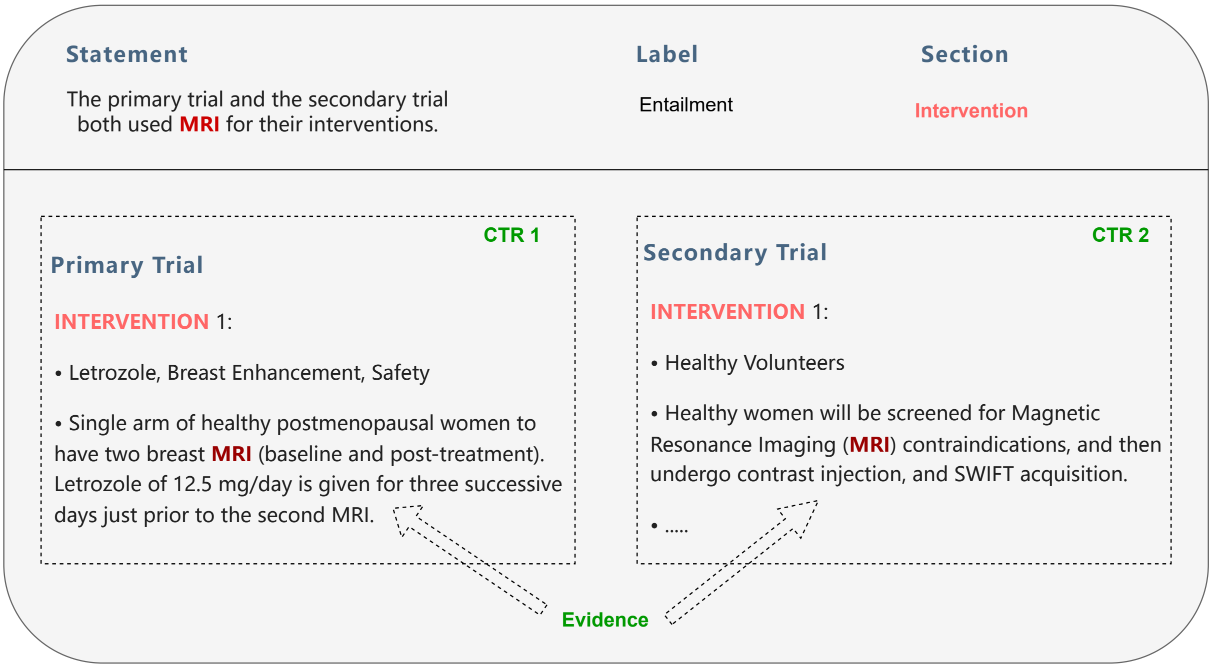

Figure 1: Data Example

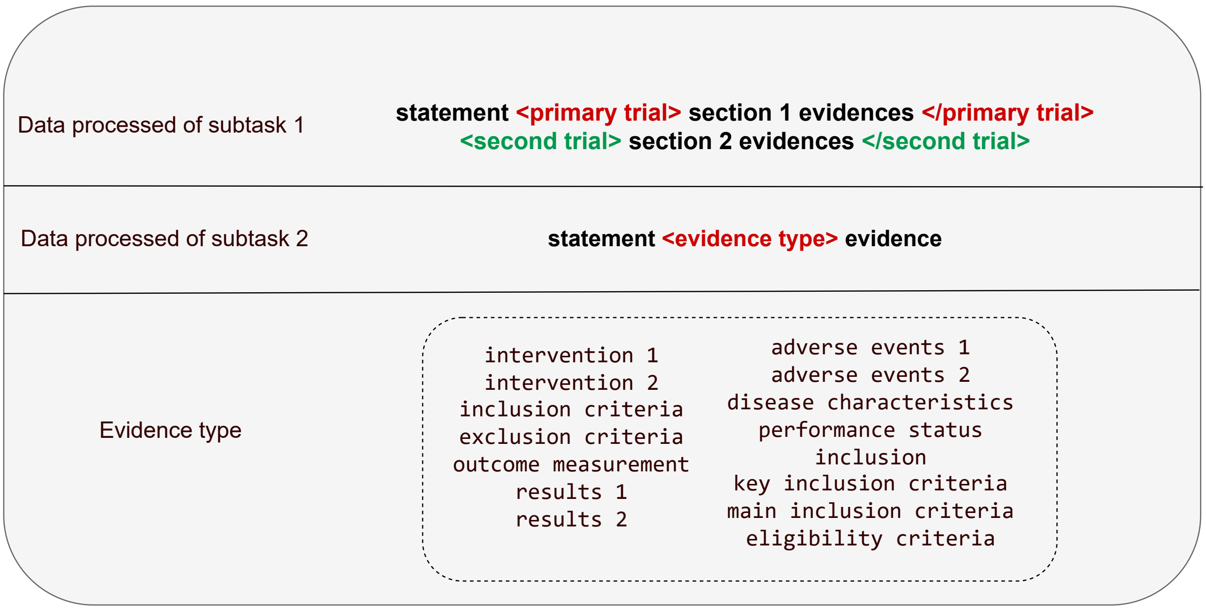

Figure 2: Data Example after processing

Subtask 2 is evidence retrieval, where a given statement and 1-2 CTRs are provided, and the output consists of evidence in the CTRs that is relevant to the statement and supports the label predicted by Subtask 1. Subtask 2 can be viewed as a classification task. For instance, one approach is to concatenate each evidence in the given CTRs with the statement, and predict a label of "0" (irrelevant to the statement) or "1" (relevant to the statement). The evidence relevant to each statement can be extracted to obtain the final results.

For Subtask 1, we believe strong logical reasoning ability is required due to the partially self-contradictory statements. Ultimately, we use ChatGPT with CoT capabilities and set prompts to achieve an F1 score of 0.679 on the zero-resource setting, which improves the official TF-IDF baseline and our PLM baseline by 3% to 17% (absolute value). For Subtask 2, we train a sentence classification model based on DeBERTaV3-large (He et al., 2021), and achieve an F1 score of 0.842 on the test dataset, ranking 2nd on the Subtask 2.

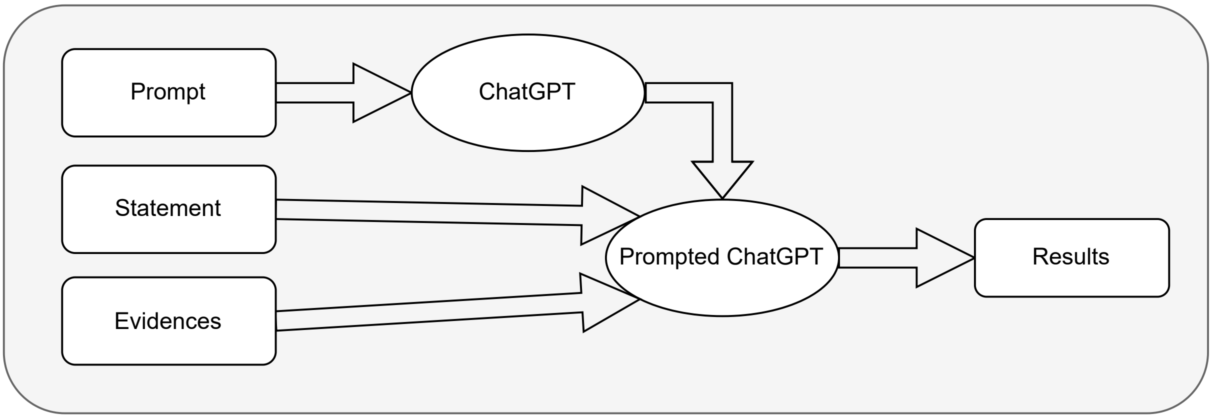

Figure 3: Flowchart for Subtask 1 ChatGPT

3 System

3.1 Data Process

The statements and evidence are generated by clinical domain experts, clinical trial organizers, and research oncologists from the UK's Manchester Cancer Research Institute and the Digital Experimental Cancer Medicine Team. In total, there are 2400 statements, evenly distributed across different sections and categories. The dataset includes a file containing statements, labels, section id (evidence type), CTR file index, and evidence index rows, as well as a number of indexed CTR files. We use the same partitioning method as the competition organizers, dividing the data into training, development, and test sets.

For Subtask 1, we concatenate the statement with the corresponding section in the given CTR, following the format "statement <primary trial> section 1 evidences </primary trial><second trial> section 2 evidences </second trial>", and concatenate multiple evidences within each section with space. The processed data is shown in Figure 2 and can be used for classification or generation tasks.

For Subtask 2, we concatenate each piece of evidence in the corresponding CTR with the statement one by one, and add the type information of the evidence, which is shown in the form of evidence type in Figure 2. The processed data takes the form of "statement <evidence type> evidence", which is also shown in Figure 2.

3.2 Prompt Learning for Subtask 1

For Subtask 1, the pre-processed data can be used to train a sentence classification model based on BERT. However, due to the limited data, the model is prone to overfitting. Moreover, LMs are not good at inference. Therefore, we attempt to use prompt learning based on ChatGPT. The flowchart is shown in Figure 3.

Specifically, we provide a prompt to stimulate the logical reasoning potential of ChatGPT. To determine whether the statement itself is contradictory, we add a prompt for checking the statement itself. Our prompt is as follows: [ChatGPT, you are an AI with reasoning and distinguishing abilities. Now I need you to help me with a classification task in the medical field. I'll give you a statement first, and then give you one or two sections from the clinical trial report, which may be one of the entry conditions, intervention, outcome, and adverse events. Each clinical trial report may contain one or two events. You need to check whether the section is consistent with the previous statement. If the section is consistent with the previous statement, you need to return "Entailment". If the section is inconsistent with the previous statement, you need to return "Contradiction". Note that the statement itself may conflict. It goes straight back to "Contradiction." Okay?]

Then, we batch-input the pre-processed sentences into ChatGPT and obtain the return results. We check the "Entailment" or "Contradiction" in the return results and obtain the final output. For the few cases without return parameters, we assume that the contradiction is not clear enough and uniformly process them as "Entailment". After the above processing, we obtain the test results under zero-shot scenarios.

3.3 PLM-based Classification for Subtask2

For Subtask 2, we build a sentence classification model based on DeBERTaV3-large, which is independent of Subtask 1. The input of this model does

Model	Precision	Recall	F1 Score
Baseline-tfidf	0.502	0.460	0.480
BERT-base	0.563	0.587	0.540
XLNet-base	0.606	0.660	0.632
BioBERT-base	**0.618**	0.680	0.647
ChatGPT	0.592	**0.796**	**0.679**

Table 1: Results of different models for Subtask 1 test

Model	Precision	Recall	F1 Score
BERT-base	0.754	0.769	0.761
BioBERT-base	0.738	0.786	0.761
DeBERTaV3-base	0.749	0.838	0.791
RoBERTa-large	0.755	0.806	0.780
DeBERTaV3-large	**0.816**	**0.871**	**0.842**

Table 2: Results of different models for Subtask 2 test

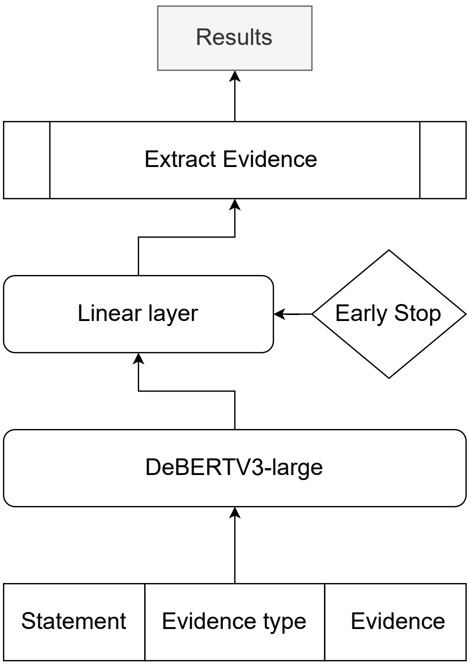

Figure 4: Model for Subtask2

not rely on the predictions from Subtask 1. The objective of this model is to determine whether the evidence is relevant to the statement. Specifically, we feed the pre-processed sentences into DeBERTaV3 and then a linear layer to predict their labels. The model is shown in Figure 4.

To prevent overfitting, we use early stopping and stop training after one epoch. We set the max_len to 233, which is the maximum sentence length in the pre-processed data. We also set batch_size to 8, random seed to 42, learning rate of the large model to 1e-5, and learning rate of the base model to 2e-5. For inference, we group all sentences with the same statement together and extract their evidence as the final output.

4 Results and Analysis

4.1 Subtask 1

Table 1 presents the results of prompt learning based on ChatGPT and other baselines on the Subtask 1 test set. The evaluation metrics used are macro F1 score, precision, and recall. As shown in the Table 1, the ChatGPT-based approach achieves an outstanding improvement compared with other baselines. Our method demonstrates a 19.9% (absolute value) increase in F1 score compared with the official TF-IDF baseline, and a 3.2% (absolute value) improvement in F1 score compared with the best-performing BioBERT-base among the sequence classification models based on PLMs. This demonstrates the effectiveness of using ChatGPT for clinical trial inference.

It should be noted that, although the pre-training data and model size of ChatGPT far exceed those of the other models, our ChatGPT-based method is developed under the zero-resource scenario. It does not require any additional annotated data and only needs appropriate prompts to outperform other common models. This is particularly attractive in scenarios with scarce data.

As shown in Figure 5, we find that some state-

User

ChatGPT, you are an AI with resoning and distinguishing abilities. Now I need you to help me ... Okay?

Statement: In the primary trial cohort 2 had more patients with Leukopenia than cohort 1, whereas in the secondary trial cohort 1 had more than cohort 2. Cohort 1 of the primary trial had the highest proportion of patients with leukopenia.

Please judge whether the statement has conflict ifself. If yes, return "Contradiction" and give me reasons. If no, continue to request section.

ChatGPT

There is a conflict in the statement. The statement says that cohort 1 of the primary trial had the highest proportion of patients with leukopenia, but also says that cohort 2 had more patients with leukopenia than cohort 1 in the primary trial. Therefore, the statement is contradictory.

Figure 5: Case for subtask1

Model	**Precision**	**Recall**	**F1 Score**
DeBERTaV3-large (1 epoch)	0.816	**0.871**	**0.842**
DeBERTaV3-large (10 epochs)	**0.818**	0.834	0.826 (-0.016)
DeBERTaV3-large (20 epochs)	0.789	0.838	0.813 (-0.029)

Table 3: Results of different epoch number for subtask 2 test

ments in the dataset are inherently contradictory. Therefore, we could conclude that the label for this data should be "Contradiction" without sections. We test these statements using ChatGPT (Jan 1 version) and find that when we strongly request ChatGPT to determine whether the statement is contradictory, ChatGPT could provide the correct answer and point out the errors. It should be noted that when using gpt-3.5-turbo, it may be necessary to repeatedly request the API to determine whether the statement is contradictory. This is because the problem is complex and gpt-3.5-turbo does not focus on statement judgement. In addition, we notice that ChatGPT is more likely to return the standard result "Contradiction" than gpt-3.5-turbo. Therefore, the prompt's requirements for the output format and data post-processing should be strengthened when using gpt-3.5-turbo.

4.2 Subtask 2

Table 2 shows the results of the DeBERTaV3-large-based approach and other baselines on the Subtask 2 test set, evaluated by macro f1 score, precision, and recall. As seen from the Table 2, the DeBERTaV3-large-based sequence classification model outperforms other baselines, achieving the SOTA performance. Compared with DeBERTaV3-base, DeBERTaV3-large shows an improvement of 5.1% (absolute value) in F1 score. And DeBERTaV3-large increases the F1 score by 6.2% (absolute value) compared with RoBERTa-large. These results demonstrate that the DeBERTaV3-large-based sequence classification model has stronger capabilities than other PLMs.

Moreover, the fact that this model ranks second in the final leaderboard of Subtask 2 demonstrates that exploring Subtask 2 independently of Subtask 1 is feasible, although there are some potential issues with this approach, for example, when the statement involves comparing or combining two pieces of evidence, it is difficult for the model to determine their relevance. We will explore these issues in future.

Table 3 presents the performance at different epochs related to early stopping. As shown in Table 3, DeBERTaV3-large achieves corresponding results under different epoch conditions in Subtask 2. Counterintuitively, the performance of model after 1 epoch is 1.6% (absolute value) higher than performance after 10 epochs, and the model's performance after 1 epoch is 2.9% (absolute value) higher than that after 20 epochs. The phenomenon indicates that DeBERTaV3-Large is prone to overfitting on Subtask 2. We believe this may be due to insufficient training data. In general, large models require more data for fine-tuning than base models since their deeper architectures and more hidden dimensions. Therefore, we propose to use an early stopping strategy to stop training after 1 epoch. The experimental results in Table 3 confirm the effectiveness of early stopping.

5 Conclusion

In this paper, we propose solutions for two subtasks of SemEval2023 Task 7. For Subtask 1, we explore the feasibility of using ChatGPT and prompt learning for logical reasoning. Although there are still many areas that can be improved, such as more refined prompts and inputting statements and evidence separately in multi-turn dialogues, ChatGPT's reasoning path is visible and analyzable compared with masked LMs. We believe that ChatGPT's logical reasoning ability in zero-resource scenarios has great appeal to NLP researchers.

For Subtask 2, we explore the feasibility of the DeBERTa-large classification model that does not depend on the predictions of Subtask 1. We find that the early stopping strategy can effectively prevent overfitting. In addition, the appropriate segmentation strategy can also enhance the model's performance. In future work, we will try other segmentation strategies, such as adding explanations and descriptions for different evidence types. Additionally, a prompt-based approach may be an interesting direction to explore.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Reducing BERT pre-training time from 3 days to 76 minutes. *CoRR*, abs/1904.00962.