HBUT at # SMM4H 2024 Task1: Extraction and Normalization of Adverse Drug Events with a Large Language Model

Yuanzhi Ke
Hubei University of Technology
keyuanzhi@hbut.edu.cn

Hanbo Jin
Hubei University of Technology
jinhanbo@hbut.edu.cn

Xinyun Wu
Hubei University of Technology
xinyun@hbut.edu.cn

Caiquan Xiong
Hubei University of Technology
xiongcq@hbut.edu.cn

Abstract

In this paper, we describe our proposed systems for the Social Media Mining for Health 2024 shared task 1. We built our system on the basis of GLM, a pre-trained large language model with few-shot learning capabilities, using a two-step prompting strategy to extract adverse drug events (ADEs) and an ensemble method for normalization. In the first step of extraction phase, we extract all the potential ADEs with in-context few-shot learning. In the second step for extraction, we let GLM to filter out false positive outputs in the first step by a tailored prompt. Then we normalize each ADE to its MedDRA preferred term ID (ptID) by an ensemble method using Reciprocal Rank Fusion (RRF). Our method achieved an excellent recall rate. It obtained 41.1%, 42.8%, and 40.6% recall rate for ADE normalization, ADE recognition, and normalization for unseen ADEs, respectively. Compared to the performance of the average and median among all the participants in terms of recall rate, our recall rate scores are generally 10%-20% higher than the other participants' systems.

1 Introduction

Extracting medical entities from social media is a challenging task. BERT (Devlin et al., 2019) is one of the most popular models used for named entity recognition (NER). Among its family, BioBERT (Lee et al., 2019) and clinicalBert (Huang et al., 2019) are especially reported useful for Medical Named Entity Recognition tasks. But without sufficient resources and data for fine-tuning, BERT tends to underperform. Some researches (Brown et al., 2020; Kojima et al., 2022) have demonstrated that tailored prompts can drive large language models (LLMs) to perform various downstream tasks well with only a small amount of data.

This paper describes our work in the Social Media mining for Health 2024 (SMM4H2024). SMM4H2024 task 1 was a mission for mining adverse drug events (ADE) from Twitter (X) and normalizing to their corresponding preferred terms within the Medical Dictionary for Regulatory Activities (MedDRA) terminology. Inspired by the works on Few-shot (Brown et al., 2020) and Zero-shot (Kojima et al., 2022) learning with LLMs, we introduce a system for ADE identification and normalization.

The identification phase of our system contains two steps: a potential ADE extraction step and a self-improvement step:

1. In the first step, we use a LLM with Few-shot in-context learning to find out all the potential ADEs.

2. In the second step, we use another prompt template that provides the potential ADEs and their original twitter sentences from the first step to let the LLM distinguish whether each ADE is caused by any drug mentioned in the twitter. In this way, false positive outputs in the first step are identified. Then a script drops such outputs to get the final text extraction result for normalization in the next phase.

In the normalization phase, we use an ensemble method, utilizing the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) method to integrate the relevance scores based on five pre-trained Embedding models and Levenshtein Distance (Yujian and Bo, 2007). The resulted fused score serves as the final measure of relevance. In this phase, each extracted ADE text in the previous phase is labeled with the preferred term ID (ptID) of the most related ADE in MedDRA measured in this way.

Our system scored above mean and median on multiple metrics among all the participants and had the advantage of not requiring fine-tuning on downstream tasks.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 48–54

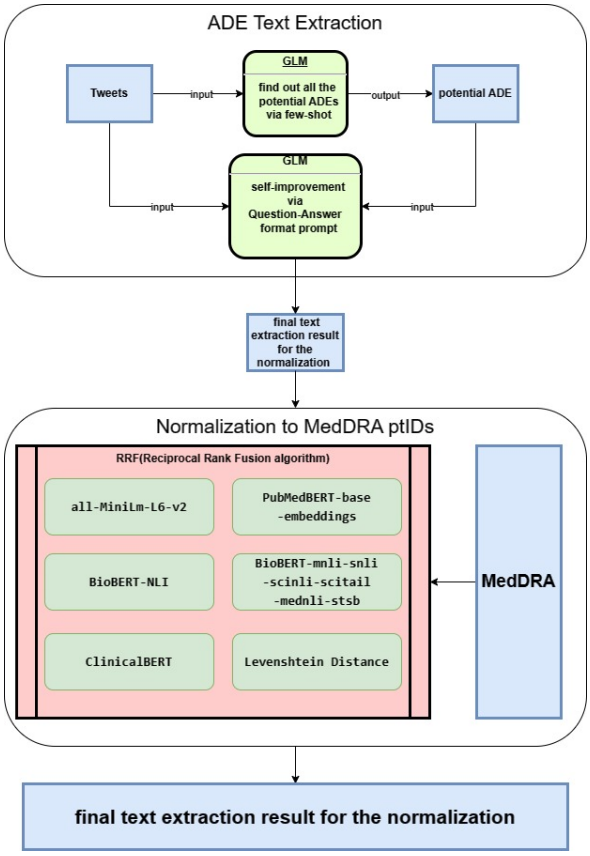

Figure 1: Framework of our system. In ADE Text Extraction, the green rectangle are GLM models with different prompt, and blue are the texts. In Normalization to MedDRA ptIDs, the red rectangle is RRF algorithm, including five language models and Levenshtein Distance.

2 Methodology

The entire framework is illustrated in Figure 1. We have two main steps in our system: a potential ADE Text Extraction phase and a normalization phase to MedDRA ptIDs. In this section, we will introduce the details of these phases.

2.1 ADE Text Extraction Phase

2.1.1 Potential ADE Text Extraction Step

In this section, considering the relatively small dataset size, we employ an approach that utilizes a pre-trained LLM to perform few-shot learning to extract potential ADE mentions. We used GLM-4 (Du et al., 2022) as our base model in this phase. In the first step, we input the tweets into GLM using few-shot prompts to extract potential ADEs.

Some tweets in the datasets do not contain any ADEs. We carefully designed our prompts to avoid outputting ADEs for such tweets. The prompt is shown in Appendix A.3.

2.1.2 Self-improvement Step

In our local experiments, we found that although our system captured many candidate ADE mentions in the first step, most of them were not caused by drugs regarding the context. To improve performance of our system, we designed a second step.

In the second step, inspired by conventional works that proposed a self-consistency check in zero-shot NER (Xie et al., 2023), we propose to utilize a question-and-answer style prompt to let the LLM model improve the outputs by itself.

We input the tweets containing the identified ADEs from the first step, along with their corresponding ADEs, into GLM in a question-and-answer format that ask the LLM to distinguish whether the ADEs identified in the fist step are caused by a drug rather than any other factors. The prompt can be found in Appendix A.3.

By employing this approach, our system filters out the tweets that do not have ADEs caused by drugs regarding the context of the input tweet. Then the remaining candidate ADE mentions are going to be mapped to the MedDRA ptIDs in the normalization phase.

2.2 Normalization Phase

We employ an ensemble approach involving multiple models and methods for normalization. The models and methods used in ensemble are summarized as follows:

- all-MiniLm-L6-v2, a general sentence-transformers (Reimers and Gurevych, 2019) model based on nreimers/MiniLM-L6-H384-uncased[1] and fine-tuned on a corpus including Reddit Comments, WikiAnswers, etc , with a total of 1B tokens [2].
- PubMedBERT-base-embeddings, a PubMedBERT-base model fine-tuned on the titles and abstracts of medical papers from the PubMed dataset [3].
- BioBERT-NLI, a BioBERT (Lee et al., 2019) model further fine-tuned on the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets [4].

[1] https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://huggingface.co/NeuML/pubmedbert-base-embeddings

[4] https://huggingface.co/gsarti/biobert-nli

- BioBERT-mnli-snli-scinli-scitail-mednli-stsb (Deka et al., 2022), a sentence-transformer model trained on SNLI, MNLI, SCINLI, SCITAIL, MEDNLI, and STSB datasets [5].
- ClinicalBERT (Wang et al., 2023), a BERT model trained on a 1.2B disease-related dataset [6].
- The Levenshtein Distance between the extracted results and the corresponding text of each ADE in MedDRA.

For Sentence-transformers and BERT models, we calculate the cosine similarity between the embeddings of the extracted results and the embeddings of each ADE as the relevance measure. For Levenshtein Distance, we directly use it as a relevant metric.

Due to the inconsistency in measures among these methods above, we employ Rank-based Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to fuse the relevance scores since it is a rank-based method.

Denote the embedding from embedding model i as $E_i(\cdot)$, the extracted text piece of a candidate ADE mention as t_{sys}, and any ADE term text in MedDRA as t_{dra}. The similarity based on embedding model i is

$$S_i(t_{sys}, t_{dra}) = \frac{(E_i(t_{sys})) \cdot (E_i(t_{dra}))}{|(E_i(t_{sys}))||(E_i(t_{dra}))|}. \quad (1)$$

We rank the ADEs in MedDRA separately based on the similarities and Levenshtein Distances obtained from the above models, and then fuse the rank results by RRF. Denote the ranking by method i as r_i, the rank of a MedDRA ADE preferred term t_{dra} as $r_i(t_{dra})$, the set of all candidate rankings as $\mathcal{R}$, and the set of all ADE texts in MedDRA as $\mathcal{D}$. The final score is

$$S(t_{dra} \in \mathcal{D}) = \sum_{r_i \in \mathcal{R}} \frac{1}{k + r_i(t_{dra})}. \quad (2)$$

k is a preset constant. We used $k = 5$.

The ptID corresponding to the ADE text with the highest final score is extracted and submitted as the ptID in the submission file.

[5] https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

[6] https://huggingface.co/medicalai/ClinicalBERT

3 Experiments

Besides our method described in Section 2, there are several prompting methods that used in similar tasks, including few-shot prompting (Brown et al., 2020), zero-shot prompting and least-to-most prompting (Zhou et al., 2022) as follows,

1. **Few-shot**: This method provides several examples along in the prompts. Our system uses this method in the first-step in the ADE extraction phase.
2. **Zero-shot**: This method input instructions to a LLM without examples.
3. **Least-to-most**: This prompting strategy proposes to divide a complex problem into subproblems in prompting. To apply it to this task in our experiments, we divided the NER task into three subtasks, including identifying whether the effect is negative, determining whether it is caused by the medication, and determining if it is an ADE caused by the medication based on the results of the previous two tasks.

We compared these methods in our local experiments.

Moreover, GLM-4 are reported as an improved version of GLM-3-Turbo. Thus, we also evaluated it for this task.

In details, we compared five methods that used different prompts and LLM models: few-shot prompt + GLM-3-Turbo WSI(without self-improvement), few-shot prompt + GLM-3-Turbo, few-shot prompt + GLM-4, zero-shot prompt + GLM-3-Turbo, and least-to-most prompt + GLM-3-Turbo. We evaluated these five methods in validation dataset in order to find the best method. The result is shown in Appendix A.1. The few-shot + GLM-4 achieved leading results across all metrics, so we ultimately decided to use the few-shot prompt + GLM-4 method on the test set.

4 Results and Discussions

4.1 Overall Results

The metrics used to evaluate results in the task are as follows,

- **F1-Norm**: The ADE Normalization F1 Score.
- **P-Norm**: The ADE Normalization Precision Score.

	Ours	Mean	Median
F1-Norm	20.5	28.3	29.3
P-Norm	13.7	29.2	33.9
R-Norm	**41.1**	33.4	32.6
F1-NER	21.6	32.7	37.6
P-NER	14.5	35.6	43.7
R-NER	**42.8**	34.0	37.4
F1-Norm-Unseen	10.6	20.9	14.1
P-Norm-Unseen	6.1	20.5	14.4
R-Norm-Unseen	**40.6**	28.7	36.5

Table 1: The final result of our system in the task in comparison with the mean and median scores among all the participants.

- **R-Norm**: ADE Normalization Recall Score.
- **F1-NER**: ADE Extraction F1 Score.
- **P-NER**: ADE Extraction Precision Score.
- **R-NER**: ADE Extraction Recall Score.
- **F1-Norm-Unseen**: ADE Normalization on Unseen MedDRA IDs F1 Score.
- **P-Norm-Unseen**: ADE Normalization on Unseen MedDRA IDs Precision Score.
- **R-Norm-Unseen**: ADE Normalization on Unseen MedDRA IDs Recall Score.

Table 1 presents the results of our method on the test dataset. F1-Norm scored 20.5, P-Norm scored 13.7, and R-Norm scored 41.1. F1-NER scored 21.6, P-NER scored 14.5, and R-NER scored 42.8. F1-norm-unseen scored 10.6, P-Norm-Unseen scored 6.1, and R-Norm-Unseen scored 40.6.

Our method works well without the need for large amounts of training data. At the same time, in terms of recall rate, we have achieved scores higher than both the median and the average scores among the participants.

5 Conclusion

Our method surpasses the median and average results of the SMM4H 2024 Task 1 in some metrics, with improvements of 10%-20% over the median in R-Norm, R-NER, and R-Norm-Unseen. This proves the feasibility of using large language models for social media text mining tasks with a well-designed prompting strategy, especially in the cases that high recall rate is required.

However, the decline in the precision of ADE mining is still an issue that needs to be addressed. We consider the reason for the low accuracy score in ADEs mining is that the large language model we used generalize based on their pre-trained data and prompt engineering, while the text of ADEs in social media comments can be very different from the standard ADE text. This gap leads to the failure of the large language model in identifying ADEs when it mines ADEs from social media.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335,

Dublin, Ireland. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. D19-1:3982–3992.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv e-prints*, arXiv:2205.10625.

A Appendix

A.1 Local Experiment Result on Evaluation Set

As shown in Table 2 and Table 3, our system achieved excellent results on the evaluation set (DEV set). Normalization F1: 0.465, Normalization Precision: 0.500, Normalization Recall: 0.435. We also achieved great performance in extraction: Extraction F1: 0.580, Extraction Precision: 0.627, Extraction Recall: 0.540. Other methods performed lower compared to the method we used. Therefore, we believe that Few-shot learning is more suitable for large language models in named entity recognition tasks compared to Zero-shot and least to most approaches. Model improvement is also crucial. Under the same conditions, replacing the model from GLM-3-Turbo to GLM-4 alone can yield a 25% performance improvement.

A.2 Dataset Details

The test dataset contains a total of 11439 tweets. According to the dataset, we removed all sentences which are too short (those with fewer than 5 words). Besides, we deleted all non-English and non-numeric characters.

A.3 Prompts for the Two-step ADE Extraction

Our prompts used for the two-step ADE extraction is shown in Table 4. Our system used prompts were written in Chinese because GLM works better with Chinese prompts. The translated version is provided in Table 5.

Method	F1-Norm	P-Norm	R-Norm	F1-NER	P-NER	R-NER
Few-shot + GLM-3-Turbo WSI	0.168	0.247	0.127	0.332	0.483	0.253
Few-shot + GLM-3-Turbo	0.300	0.320	0.282	0.580	0.627	0.540
Few-shot + GLM-4	0.471	0.514	0.435	0.580	0.627	0.540
Zero-shot + GLM-3-Turbo	0.134	0.101	0.200	0.267	0.202	0.391
Least-to-most + GLM-3-Turbo	0.172	0.197	0.153	0.222	0.258	0.153

Table 2: The overall results achieved by different combinations of prompting strategies and LLMs, evaluated using the DEV set in our local experiments.

Method	F1-Norm-Unseen	P-Norm-Unseen	R-Norm-Unseen
Few-shot + GLM-3-Turbo WSI	0.000	0.000	0.000
Few-shot + GLM-3-Turbo	0.095	0.054	0.400
Few-shot + GLM-4	0.214	0.130	0.600
Zero-shot + GLM-3-Turbo	0.000	0.000	0.000
Least-to-most + GLM-3-Turbo	0.000	0.000	0.000

Table 3: The results for unseen ADEs regarding the training set, achieved by different combinations of prompting strategies and LLMs, evaluated using the DEV set in our local experiments.

Prompts for Coarse-grained ADE Text Extraction
你是一个超级医药化学专家，请你找出我下列英文句子中的化学物品或药物的副作用，句子中不一定包含副作用。 若句子没有副作用则不需要输出。如果有副作用，请找到后仅提供最终对应副作用的结果，只输出第一个副作用，无需展示推理过程。 副作用可能有多个。我会分多次给你进行提示 有且只有一个副作用的句子请这样输出 例句：avelox has hurt your liver, avoid tylenol always,it further damages liver, eat grapefruit unless taking cardiac drugs 副作用：hurt your liver
Prompts for Self-improvement
我会给你一句话和多个词，你需要使用你的推理能力来判断，这些ADEs是否是由于药物导致的而不是原本就存在的问题。 例如：avelox has hurt your liver, avoid tylenol always,it further damages liver, eat grapefruit unless taking cardiac drugs 副作用：hurt your liver 推理：因为例句中avelox has hurt your liver 这句话所以可以推断出hurt your liver是由于avelox导致的。不需要输出原因，只需要输出是或者不是。 请你用你的逻辑能力回答：[tweet]中的[ADE]是否是药物导致，还是本身或者其他不良习惯导致，只需要输出是或者不是。

Table 4: Our prompts for the two steps for ADE Text Extraction, which is originally written in Chinese.

Prompts for Coarse-grained ADE Text Extraction in English
You are a super medical chemistry expert, please identify any chemical substances or drugs and their Adverse Drug Event (ADE) in the following English sentences. The sentences may or may not contain ADEs. If the sentence does not contain any ADE, you don't need output anything. If there are ADEs, please find them and provide only the result of corresponding side effects without showing the reasoning process. Side effects may be multiple. I will give you some examples. Example： avelox has hurt your liver, avoid tylenol always, it further damages liver, eat grapefruit unless taking cardiac drugs ADE： hurt your liver
Prompts for Self-improvement in English
I will give you a sentence and multiple words, and you will need to use your reasoning ability to determine whether these ADEs are caused by the medication rather than pre-existing conditions. Example： avelox has hurt your liver, avoid tylenol always, it further damages liver, eat grapefruit unless taking cardiac drugs ADE： hurt your liver reasoning： Because of the sentence "avelox has hurt your liver," it can be inferred that "hurt your liver" is caused by avelox. There is no need to output the reason, just yes or no. Please use your logical abilities to answer: In the [tweet], is the [ADE] caused by the medication, or is it due to the person's own issues or other bad habits? Just output "yes" or "no".

Table 5: Our prompts for the two steps for ADE Text, translated in English.