HBUT at #SMM4H 2024 Task2: Cross-lingual Few-shot Medical Entity Extraction using a Large Language Model

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

Zhangju Yin
Hubei University of Technology
yinzhangju@hbut.edu.cn

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

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

Abstract

Named entity recognition (NER) of drug and disorder/body function mentions in web text is challenging in the face of multilingualism, limited data, and poor data quality. Traditional small-scale models struggle to cope with the task. Large language models with conventional prompts also yield poor results. In this paper, we introduce our system, which employs a large language model (LLM) with a novel two-step prompting strategy. Instead of directly extracting the target medical entities, our system firstly extract all entities and then prompt the LLM to extract drug and disorder entities given the all-entity list and original input text as the context. The experimental and test results indicate that this strategy successfully enhanced our system performance, especially for German language. Our code is available on Github [1].

1 Introduction

Discussions on drugs and their adverse drug reactions shared by users in social media, including the efficacy, side effects, and personal treatment journeys serve as valuable references for pharmacovigilance.

We participated in the 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks (Xu et al., 2024). The task targets both the extraction of drug and disorder/body function mentions (Subtask 2a) and the extraction of relations between those entities (joint Named Entity Detection and Relation Extraction, Subtask 2b). The paper is about the task 2a. Our task was to identify drug and disorder mentions in German, French, and Japanese datasets from X(Twitter) and a German patient forum.

In previous works on similar tasks, people commonly use BERT models (Kenton and Toutanova, 2019). For multilingual tasks, the m-BERT (multilingual BERT) model is often employed (Papadimitriou et al., 2021). However, in this particular open task, the provided dataset is relatively small, especially in the training set. Therefore, it is challenging to perform pre-training and fine-tuning on the traditional BERT model due to the limited amount of data available. Thus, we opted to use a large language model (LLM) for the task.

Because the training data is also not enough to fine-tune a LLM well, we opted for the prompt-based approach. Our system utilizes a two-step prompting strategy. A first step to get a list of all entities, and then a second to further extract drug and disorder entities from the list.

2 Methodology

Due to the small size of the provided dataset, we placed our focus primarily on using large language models (LLMs). We build our system based on a GLM (Du et al., 2022; Zeng et al., 2022) through its online API. It is trained on multilingual dataset and performs almost on a par with ChatGPT and worked well for multilingual tasks in our preliminary experiments.

We utilize few-shot prompt engineering technologies to adapt GLM to the drug and disorder mention extraction task as shown in Figure 1.

2.1 Prompting Method

The task is challenging for the numerous abbreviations and colloquial vocabulary in the dataset, which requires our system to be able to recognize such unofficial representations. In our preliminary experiments, we found that our system failed to work well with conventional prompts that guide the LLM to extract drugs and disorder mentions from the input text.

Thus, we address this issue by a novel strategy. Firstly, instead of instructing the LLM to extract drug/disorder mentions, our system prompts the model to extract all entities, just with additional

[1] https://github.com/YinZhangJu/SMM4H24code

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 58–62

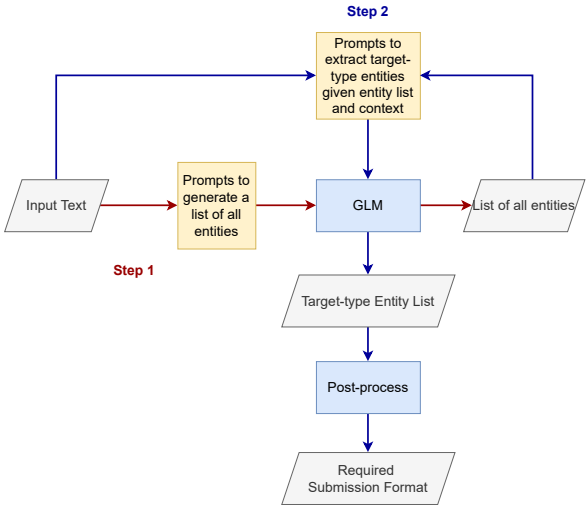

Figure 1: Overall architecture of our system. In the first step, the prompting guides the model to identify all entity lists from the dataset. In the second step, the prompting guides the model to combine the original data content and extract the target entities from the entity list.

attention on medicine-related entities. Then our system prompts the model to extract drugs and disorders given both the entity list and the original input as the context. In this way, our system is more robust for informal representations and ill-formed sentences. The detailed settings of our two-step prompting are described in Tables 1 and Tables 2.

2.2 Preprocessing for Sensitive Words

The GLM API raised errors for some samples in the test dataset, with error messages indicating that the input texts contained some sensitive words.

When our system detects such failures, it collects and writes the sample IDs, sample contents, and error messages in an error log. We manually check the error records, identify and remove the sensitive words in the corresponding samples, and then redo generation for them.

This method may produce incomplete and ill-formed sentences. However, because our two-step design is robust against ill-formed text, the adverse effects on the NER accuracy are relatively small. Replacing sensitive words with alternative words may result in better NER performance. However, we did not have enough time to work out a replacement table for the sensitive words in the shared task. We plan to complete it in future work.

2.3 The Post-Processing Step

We save the output results to a CSV file. Then, to match the submission format, a Python script is used to remove punctuation and symbols, and convert the results into the BRAT format.

3 Experiments and Results

3.1 Experiments on LLM Model Selection

Conventional works reported that Qwen-14B (Bai et al., 2023) performs well for factual tasks. Thus, we also conducted experiments with Qwen-14B in comparison to GLM.

We randomly sampled 100 samples from the

Our Prompt (Original in Chinese)
“role”: “user”, “content”: “作为一名精通日法德三语的语言学家，请你找出下列日语、法语或者德语语句中的实体，尤其注意医学领域的实体，比如药物名称以及副作用等，注意简称，只提供结果，不需要推理过程” [Examples of finding entity lists for German, French, and Japanese] “role”: “assistant”, “content”: “好的，我为您筛选出了语句中的各种实体” “role”: “user”, “content”: “请找出下列[list]中的实体名称”
Our Prompt (Translated into English)
“role”: “user”, “content”: “You are a linguist proficient in German, French, and Japanese. Please help me identify the entities in the following Japanese, French, or German sentences. Please pay attention to entities in the medical field, such as drugs and disorder mentions. Pleas also be aware of abbreviations. Provide only the results without the need for the reasoning process” [Examples of finding entity lists for German, French, and Japanese] “role”: “assistant”, “content”: “I have filtered out various entities from the sentences for you.” “role”: “user”, “content”: “Please identify the entity names in the following [list].”

Table 1: The prompt template used in the first step. In this step, the model is tasked with identifying the entity list from the dataset. “[list]" represents the examples randomly sampled from the dataset. We provide the English translation of our prompt together with the original prompt used in the system (which is written in Chinese).

Our Prompt (Original in Chinese)
“role”: “user”, “content”: “作为一名日法德三语药剂师，请你根据提供的日语、法语或者德语句子语境，按照下列格式找出实体列表中的药物以及副作用实体，注意简称，只提供结果，不需要推理过程” [Examples of target entities within the entity lists for German, French, and Japanese:] “role”: “assistant”, “content”: “好的，我为您筛选出了实体列表中的药品名称和副作用” “role”: “user”, “content”: “请根据[list]中的语境，仿照上面格式，找出下列[str1]中药物名称以及副作用”
Our Prompt (Translated into English)
“role”: “user”, “content”: “You are a pharmacist proficient in German, French, and Japanese. Please identify the drug and disorder mentions from the list, according to the context in Japanese, French, or German. Pleas also be aware of abbreviations. Provide only the results without the need for the reasoning process.” [Examples of target entities within the entity lists for German, French, and Japanese:] “role”: “assistant”, “content;”: “I have filtered out the drugs and disorder mentions from the entity list for you.” “role”: “user”, “content”: “Please, based on the context in [list], follow the format above to identify the drugs and disorder mentions in [str1].”

Table 2: The prompt template used in the second step. In this step, the model is fed with the original text with the generated entity list from the previous step to identify drugs and their corresponding disorder mentions. “[list]" represents the original text of the dataset, and “[str1]" denotes the generated entity list. We provide the English translation of our prompt together with the original prompt used in the system (which is written in Chinese).

outputs generated by Qwen-14B and GLM respectively, and manually checked the correctness. The accuracy scores achieved by the two models in this local experiment is as shown in Table 3. Qwen-14B failed to get satisfying results. Qwen-14B exhibited tokenization issues with German, French, and Japanese data, resulting in significant problems with fabricated entities and incorrect scope.

Model	Accuracy by Human
GLM-3-Turbo	0.4358
Qwen-14B	0.1538

Table 3: Human evaluation results of GLM-3-Turbo and Qwen-14B in our local experiments to choose the candidate base model in our system.

In case of GLM, there are several available versions. We tried GLM-3-turbo and GLM-4. However, we encountered some encoding issues with GLM-4. When answering questions in German and French, we observed that GLM-4 firstly translated the text into English and then provided an English response. This led to inaccuracies in the output of German and French words. In contrast, GLM-3-turbo exhibited minimal issues in this regard. Therefore, we used GLM-3-turbo in our system. We used the default generation parameters. The temperature is set to 0.95, top p is set to 0.7, and max tokens is set to 1024.

3.2 Experiments on Prompting Methods

We conducted experiments on three prompt strategies as follows,

- Single-word Prompting (Single Prompt): In this approach, we directly guided the model to perform named entity recognition on German, French, and Japanese data.
- Detached Step-by-step Prompting (DSBS Prompt): In this approach, we utilized a step-by-step prompting method. In the first step, we initially prompted the LLM model to generate a list of all entities in the input text, instead of just medical-related entities like that in conventional system. In the second step, we let the LLM model to extract drug and disorder entities from the entity list.
- Integrated Step-by-step Prompting (ISBS Prompt): This approach is also a step-by-step prompting method. The first step is as the same as DSBS prompt. But in the second step, we combined the entity list and the original input text in the prompts.

The results are shown in Table 4.

Single Prompt uses fewer tokens. However, it exhibited lower precision in capturing phrase scope, with larger discrepancies compared to the gold annotation.

DSBS Prompt resulted in more accurate identification of entity words and achieved higher precision. However, as it detached from the original text, it performed poorly in extracting drugs and disorder mentions from the entity list.

ISBS Prompt led to more accurate identification of entity words, higher precision, and effective extraction of drugs and disorder mentions from the entity list.

Prompting Method	F1
Single Prompt	0.3420
DSBS Prompt	0.3023
ISBS Prompt	**0.3533**

Table 4: Comparison of F1 scores achieved by Single Prompt, DSBS Prompt and ISBS Prompt, evaluated using the online scores in validation phase.

4 Final Results

Our final system employed GLM-3-turbo model with ISBS Prompt. The overall results are shown in Table 5. Our system achieved a better precision than the mean precision among all teams (including baseline).

Especially, our method demonstrated excellent performance for the German part. The results for the German part are shown in Table 6. The overall precision, recall and F1 scores are higher than the mean scores among all the participants. However, our system failed to address mentions of function.

Team	**Precision**	**Recall**	**F1**
Ours	0.6052	0.2654	0.3690
Mean	0.5942	0.3434	0.4291
Std	0.0156	0.1103	0.0850

Table 5: Overall results.

Entity Type	**Precision**	**Recall**	**F1**
DISORDER	0.5463	0.2744	0.3653
DRUG	0.7627	0.3629	0.4918
FUNCTION	0.0000	0.0000	0.0000
All	0.6228	0.2751	0.3817
Mean	0.4404	0.1945	0.2699

Table 6: Final result of our system for German data. The first three rows show the results of our system for different entity types. The last row shows the mean result among all participants for all entity types.

5 Conclusion

Due to the small size of our dataset, we found that the performance with conventional small language models (e.g. BERTs) is not able to give satisfying precision. Consequently, we build our system based on GLM-3-Turbo large language model. Besides, we found that conventional straight-forward prompting strategy encountered low precision for this task. To address this issue, we proposed a step-by-step prompting strategy. We firstly prompt to extract a list of all entities. Then we prompt to guide the model to choose drugs and disorder mentions from the list within the context of the corresponding original input text. This prompting approach significantly improved the effectiveness of our system.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Isabel Papadimitriou, Ethan A Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, et al. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

A Appendix

A.1 Dataset Details

The detailed information of the dataset is shown in Table 7.

	Training	**Validation**	**Test**
German	70	23	25
Japanese	392	168	118
French	4	0	96

Table 7: The Japanese data is sourced from X (Twitter), the German data is obtained from a German patient forum, and the French data is translated from the German data (distinct from the German data).