Check for updates

A Generative Drug-Drug Interaction Triplets Extraction Framework Based on Large Language Models

Hu, Haotian Nanjing University, China | hhtdlam@126.com
Yang, Alex Jie Nanjing University, China | alexjieyang@outlook.com
Deng, Sanhong Nanjing University, China | sanhong@nju.edu.cn
Wang, Dongbo Nanjing Agricultural University, China | db.wang@njau.edu.cn
Song, Min Yonsei University, Republic of Korea | min.song@yonsei.ac.kr
Shen, Si Nanjing University of Science and Technology, China | shensi@njust.edu.cn

ABSTRACT
Drug-Drug Interaction (DDI) may affect the activity and efficacy of drugs, potentially leading to diminished therapeutic effect or even serious side effects. Therefore, automatic recognition of drug entities and relations involved in DDI is of great significance for pharmaceutical and medical care. In this paper, we propose a generative DDI triplets extraction framework based on Large Language Models (LLMs). We comprehensively apply various training methods, such as In-context learning, Instruction-tuning, and Task-tuning, to investigate the biomedical information extraction capabilities of GPT-3, OPT, and LLaMA. We also introduce Low-Rank Adaptation (LoRA) technology to significantly reduce trainable parameters. The proposed method achieves satisfactory results in DDI triplet extraction, and demonstrates strong generalization ability on similar corpus.

KEYWORDS
Drug-drug interaction; Large language model; Biomedical triplet extraction; Entity recognition; Relation extraction

INTRODUCTION
When patients take two or more drugs at the same time, there is a potential for Drug-Drug Interaction (DDI). This interaction may pose life-threatening risks to patients in severe cases (Zhang et al., 2020). The rapid growth of biomedical literature makes it increasingly challenging to manually update the DDI databases. Therefore, automatic DDI extraction is important for both pharmacological research (Asada et al., 2023) and clinical treatment.

Currently, most DDI triplets joint recognition research is based on BERT-like models, which usually shares a single set of parameters (Li et al., 2023). However, the decoding of entities and relations still relies on independent linear layers, and tends to use self-defined rules (Luo et al., 2020; Sun et al., 2022). The GPT-like Large Language Models (LLMs) can simultaneously complete sequence labeling and text classification tasks (Luo et al., 2022), which is precisely what is needed for DDI triplets joint extraction. Therefore, we proposed a generative DDI triplets extraction architecture (Figure 1) based on the current state-of-the-art LLMs, which transforms traditional entity recognition and relation extraction tasks into a unified seq2seq DDI triplets generation task.

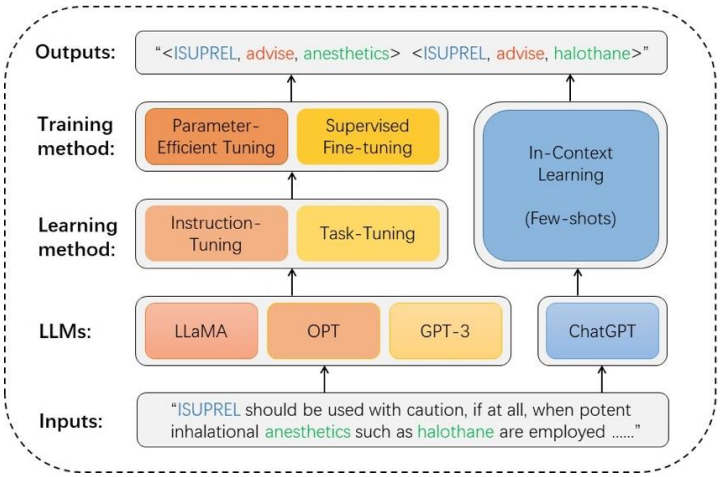


Figure 1. Generative DDI triplets extraction framework

METHODS
For LLMs, the GPT-3 (Brown et al., 2020) is an autoregressive model and performs well in text generation tasks. ChatGPT (GPT-3.5 and GPT-4) is a general-purpose dialogue model built on GPT-3. OPT (Zhang et al., 2022) and LLaMA (Touvron et al., 2023) are open source models adopt similar architecture and pre-training method as GPT.

86th Annual Meeting of the Association for Information Science & Technology | Oct. 27 – 31, 2023 | London, United Kingdom. *Author(s) retain copyright, but ASIS&T receives an exclusive publication license.*

For learning methods, the In-context Learning is applied to ChatGPT group, we compared the DDI triplets extraction performance of them in the 5-shots and 20-shots scenarios. We conducted Task-tuning (fine-tuning) for GPT-3, OPT, and LLaMA models with different parameter sizes. The Instruction tuning (Wei et al., 2021) is applied to OPT and LLaMA, improving the model's understanding of user instructions and provide accurate feedback.

For Training methods, in addition to Supervised Fine-Tuning (SFT), which trains all parameters of LLMs, we also adopted the Parameter-Efficient Tuning method: Low-Rank Adaptation (LoRA). It freezes model parameters during training, only trains low-rank matrix, thus significantly reducing trainable parameters (Hu et al., 2021). Since the weights of GPT-3 are not publicly available, we adopted the LoRA method on OPT and LLaMA.

RESULTS AND DISCUSSION

Data and Model Setup

The Drug-Drug Interaction (DDI) dataset (Herrero-Zazo et al., 2013) is a widely-used benchmark containing 1,017 documents, 18,502 entities and 5,028 DDI relations selected from DrugBank and MedLine. We trained the models on DrugBank training set and validated performance on the test set. We further validated the best model on MedLine's test set to verify the model's generalization ability on similar data.

The hyperparameters for each model are as follows: the LLaMA trained for 10 epochs, with a learning rate of 3e-4, and a LoRA rank of 8. The OPT trained for 15 epochs, with a learning rate of 2e-4, and a LoRA rank of 16. The GPT-3 trained for 2 epochs with a learning rate of 0.1.

DDI Triplets Extraction Performance

Due to space constraints, we only reported the F1-score for two ChatGPT models in the In-context learning scenario. GPT-3.5 achieved F1-scores of 58.33% and 68.09% in the case of 5-shots and 20-shots, respectively, and the corresponding F1-scores for GPT-4 were 64.00% and 65.22%. We used ChatGPT's F1-score of 68.09% as the baseline for subsequent experiments.

Table 1 lists the performance of LLMs with different parameter sizes and different training methods on DDI triplets extraction. The F1-score of all models exceeded the score of in-context learning, especially GPT-3 (175B) achieved an absolute improvement of 14.11%. In addition, most of the models have billions of parameters, while ChatGPT has a parameter size of more than 100 billion, which shows that using biomedical domain text for model fine-tuning can significantly enhance the information extraction ability of small models.

Models	LLaMA 7B	LLaMA 13B	LLaMA 7B	OPT 6.7B	OPT 6.7B	GPT-3 6.7B	GPT-3 13B	GPT-3 175B
Instruction	√	√	×	√	×	×	×	×
Precision	78.98%	80.75%	76.54%	75.25%	79.88%	79.39%	82.40%	84.78%
Recall	66.67%	71.10%	67.92%	69.85%	73.15%	73.61%	77.25%	79.86%
F1-score	72.30%	75.62%	71.97%	72.45%	76.37%	76.39%	79.74%	**82.20%**

Table 1. DDI Triples extraction performance

For the GPT-3 and LLaMA group, performance was positively correlated with parameter size. The extraction performance of LLaMA was improved after introducing instruction-tuning technology. However, the fine-tuning approach yielded better results for the OPT model, indicating that instruction-tuning is not applicable to all LLMs. We also found that the performance of OPT (6.7B) and GPT-3 (6.7B) with the same parameters was comparable, and the former reduced the amount of training parameters by introducing LoRA during the fine-tuning process. This provided insight into training low-resource biomedical information extraction models. Admittedly, training only on the DDI corpus without external features may not provide enough information for capturing entities and relations.

In order to explore the automatic labeling and generalization capabilities of the model, we used the best-performing GPT-3 (175B) to automatically extract DDI triplets from a similar source, the MedLine test dataset. After calculation, the Precision, Recall, and F1-score reached 82.00%, 86.32%, and 84.1%, respectively, showing strong knowledge transfer and generalization capabilities.

CONCLUSION AND FUTURE WORK

We proposed a generative DDI triplets extraction framework based on LLMs for the joint extraction of entities and relations in the biomedical field. We introduced various technologies such as instructing-tuning, task-tuning, LoRA, etc., and validate the superior performance of LLaMA, OPT, and GPT-3. Next, we will expand the scope of the task and introduce expert knowledge to further improve performance. Also, the human evaluation will be considered.

23739231, 2023, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/pra2.918 by Nanjing University, Wiley Online Library on [06/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

REFERENCES

Asada, M., Miwa, M., & Sasaki, Y. (2023). Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. *Bioinformatics*, 39(1), btac754.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5), 914-920.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Li, D., Wu, P., Dong, Y., Gu, J., Qian, L., & Zhou, G. (2023). Joint Learning-based Causal Relation Extraction from Biomedical Literature. *Journal of Biomedical Informatics*, 104318.

Luo, L., Yang, Z., Cao, M., Wang, L., Zhang, Y., & Lin, H. (2020). A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, 103, 103384.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), bbac409.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2022). Mrc4bioer: joint extraction of biomedical entities and relations in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 125, 103956.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A.M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, T., Leng, J., & Liu, Y. (2020). Deep learning for drug–drug interaction extraction from the literature: a review. *Briefings in bioinformatics*, 21(5), 1609-1627.

23739231, 2023, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/pra2.918 by Nanjing University, Wiley Online Library on [06/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License