



计算机科学

COMPUTER SCIENCE

高低资源下思维链增强的药物不良反应关系抽取

李浩, 杨雨濛, 赵博扬, 郑朴琪, 林鸿飞

引用本文

李浩, 杨雨濛, 赵博扬, 郑朴琪, 林鸿飞. 高低资源下思维链增强的药物不良反应关系抽取[J]. 计算机科学, 2025, 52(12): 224-230.

LI Hao, YANG Yumeng, ZHAO Boyang, ZHENG Puqi, LIN Hongfei. Adverse Drug Reaction Relationship Extraction Based on Chain of Thought Enhancement Under High and Low Resources [J]. Computer Science, 2025, 52(12): 224-230.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

一种面向SIMD的控制流投机向量化方法

Speculative Control Flow Vectorization Method for SIMD

计算机科学, 2025, 52(11A): 241100012-7. <https://doi.org/10.11896/jsjcx.241100012>

大语言模型驱动的多智能体协同代码生成技术

Multi-agent Collaborative Code Generation Technology Driven by Large Language Models

计算机科学, 2025, 52(11A): 241200033-9. <https://doi.org/10.11896/jsjcx.241200033>

基于语义变化的缺陷生成与缺陷预测模型测试

Semantic Variations Based Defect Generation and Prediction Model Testing

计算机科学, 2025, 52(11A): 241200059-7. <https://doi.org/10.11896/jsjcx.241200059>

基于信息融合的结构化剪枝算法研究

Research on Structured Pruning Algorithm Based on Information Fusion

计算机科学, 2025, 52(11A): 241000041-6. <https://doi.org/10.11896/jsjcx.241000041>

结合图检索与上下文排序的检索增强生成技术研究

Research on Retrieval-augmented Generation Technology Combining Graph Retrieval and Contextual Ranking

计算机科学, 2025, 52(11A): 250100011-4. <https://doi.org/10.11896/jsjcx.250100011>

高低资源下思维链增强的药物不良反应关系抽取

李浩 杨雨濛 赵博扬 郑朴琪 林鸿飞

大连理工大学计算机科学与技术学院 辽宁 大连 116081

(dutlihao@163.com)

摘要 药物不良反应(Adverse Drug Reactions, ADRs)是指正常剂量的药物用于预防、诊断、治疗疾病或调节生理机能时出现有害的、与用药目的无关的反应,这些反应是预期之外的,会严重影响患者的身体状态与健康情况。为了从社交媒体数据与医药文献数据中发现药物不良反应关系三元组,从而为患者和医疗系统提供警示作用,并为药物研发机构提供药物迭代的参考,结合生成式大语言模型分别针对有标注数据低资源和高资源的情况提出两个框架:思维链提示下 GPT 的药物不良反应关系抽取 ADR-GPT 框架,以及思维链增强下微调的药物不良反应关系抽取 CADRE-LLM 框架。分别在 Twitter 和 Pubmed 数据集中评估了两个框架的性能结果,其中 CADRE-LLM 相比于之前最先进的模型, F1 值分别提高 1.51 个百分点和 1.74 个百分点;低资源下的 ADR-GPT 框架也取得了优秀的效果,在 Pubmed 数据集中超越了 Qwen2.5 模型的全监督微调。最后,通过消融实验证明了两个框架中各模块的有效性。

关键词: 关系抽取; 药物不良反应; 大语言模型; 思维链; 低资源

中图分类号 TP311.13; TP309

Adverse Drug Reaction Relationship Extraction Based on Chain of Thought Enhancement Under High and Low Resources

LI Hao, YANG Yumeng, ZHAO Boyang, ZHENG Puqi and LIN Hongfei

School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116081, China

Abstract ADRs refer to harmful and unintended responses that occur when drugs are administered at normal doses for prevention, diagnosis, treatment of diseases, or regulation of physiological functions. These reactions are unexpected and can significantly impact patients' health and physical condition. To identify ADR-related relational triples from social media and biomedical literature data—providing early warnings for patients and healthcare systems, as well as references for pharmaceutical research and drug development—this paper proposes two frameworks based on generative large language models for both low-resource and high-resource annotated data scenarios: the ADR-GPT framework, which extracts ADR relations using Chain-of-Thought prompting, and the CADRE-LLM framework, which leverages Chain-of-Thought-enhanced fine-tuning. It evaluates the performance of these two frameworks on the Twitter and PubMed datasets. CADRE-LLM achieves F1 score improvements of 1.51 percentage points and 1.74 percentage points over previous state-of-the-art models on the respective datasets. The low-resource ADR-GPT framework also demonstrates strong performance, outperforming fully supervised fine-tuned Qwen2.5 on the PubMed dataset. Ablation studies further validate the effectiveness of each module within the two proposed frameworks.

Keywords Relation extraction, Adverse drug reactions, Large language models, Chain of Thought(COT), Low resource

1 引言

药物不良反应(ADR)是指正常剂量的药物用于预防、诊断、治疗疾病或调节生理机能时出现有害的、与用药目的无关的反应^[1]。这些反应是预期之外的,会严重影响患者的生理状态与健康状况,甚至会有致命的后果。及时准确地发现潜在的不良反应不仅可以给患者和医疗系统提供警示作用,而且使得药物研发机构更清晰地了解到药物缺点,以方便药物

的进一步迭代升级。随着社交媒体的快速发展和医药文献的积累,很多用户在社交媒体中发布自己的用药感受,而医药文献中本身就含有很多药物不良反应的相关描述,故社交媒体和医药文献文本成为了药物不良反应发现的重要来源。目前,很多药物不良反应的工作集中在检测与命名体识别范围内^[2-3],此类工作仅可以简单识别药物不良反应信息,但不能将药物信息和症状很好地关联起来。本文将药物不良反应的发现工作视为关系抽取问题,将药物、症状及其关系从文本

到稿日期:2025-06-20 返修日期:2025-09-23

基金项目:国家自然科学基金(62302076,62276043)

This work was supported by the National Natural Science Foundation of China(62302076,62276043).

通信作者:林鸿飞(hflin@dlut.edu.cn)

中提取出来。图1对该任务进行了描述,为增加任务的实用性,本文设置3种关系类型,包括消极的不良反应关系(Outcome Negative)、积极意义上对症状起到治疗作用的关系(Outcome Positive),以及因为症状而使用药物但并不清楚是否起到作用的关系(Reason to Use)。

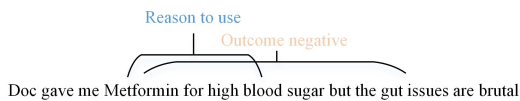


图1 药物不良反应关系抽取任务

Fig.1 Adverse drug reaction relationship extraction task

随着生成式模型的发展,生成式模型的大规模参数和学习到的海量知识使得模型能够深入理解文本语义,并且由于生成式的天然优势,不需要做复杂的解码过程即可端到端地得到关系三元组。得益于这一技术优势,大语言在关系抽取任务中的使用被广泛研究,并且有效提升了关系抽取的准确性和泛化性。但由于大语言生成式模型的随机性,目前模型通常存在幻觉现象,仅仅使用提示语句来引导大模型做药物不良反应实体识别并不能取得很好的效果,因此本文重点探索如何激发生成式大模型在药物不良反应关系抽取中的能力并缓解大模型幻觉现象的产生。在药物不良反应关系抽取任务中,幻觉现象一般表现为:不按提示语中的固定格式输出,导致无法按照相应指标评测三元组的生成质量;生成错误的内容,导致三元组中出现错误的实体或关系;甚至可能生成文本中不存在的实体或者与预先设定不符的关系。

为解决以上幻觉问题,本文结合思维链的思想进行药物不良反应关系抽取:首先,让大模型模拟人类逐步思考的过程并展示思考过程,深入理解文本的语义;随后,生成药物不良反应三元组答案,从而在一定程度上减轻幻觉现象,提高识别精确性,达到“三思而后答”的效果。基于上述思想,本文提出了两种框架来解决社交媒体和医药文献中药物不良反应的关系抽取问题:

1)思维链增强的上下文学习药物不良反应关系抽取框架;

2)思维链增强微调的药物不良反应关系抽取框架。

这两种框架分别用于解决在社交媒体文本和医药文献中低资源和高资源情况下,药物不良反应关系三元组抽取的问题。

2 相关工作

药物不良反应关系抽取,是信息抽取中的一项重要任务。关系抽取的研究方法主要经历了以下3个阶段:基于传统流水线的方法、基于判别式端到端的联合抽取、基于生成式模型的端到端三元组抽取。

目前,大多数药物不良反应关系抽取的工作都还集中在使用流水线的方法,即先进行命名体识别,然后将所得到的命名体进行关系分类。例如,Zhang等^[4]提出了一种基于多跳的自注意力机制网络来提高药物和不良反应间关系的分类性能,允许模型在每次迭代时关注文本的不同部分,获取药物与

不良反应的深层关系。

尽管流水线方法具有灵活性,但它忽略了这两个任务之间的偏差,并且存在错误传播的问题,导致最后三元组抽取的性能大大降低。近年来,关系抽取的研究主要针对联合抽取框架,端到端地提取出头实体、尾实体和关系。基于判别式端到端的联合抽取模型可以分为以下两类。一是基于序列标注的方法,利用多个相关的序列标注模块来标注头实体、尾实体甚至关系。例如,Zheng等^[5]提出的 Novel Tagging Scheme,首先将关系三元组抽取任务转换为一个标注问题。BiRTE提出了一种基于双向抽取框架的方法,该方法基于从两个互补方向提取的实体对来抽取三元组^[6]。二是基于表格填充的方法,通过分类词对的关系来确定头实体和尾实体。表格填充方法的典型代表是 TPLinker^[7],其引入了一种握手标注方案,将每种关系类型下的实体对的边界词对齐。另外,Shang等^[8]提出的 OneRel 将联合抽取问题转换为细粒度的三元组分类问题。

随着生成式模型能力的发展,越来越多的方法使用生成方式进行关系抽取。Ye等^[9]提出了一种基于生成式 Transformer 的对比学习三元组抽取框架,用于解决序列生成时难以捕获长期依赖关系并生成幻觉三元组的问题。Instruct-UIE通过对生成式大语言模型基座进行关系抽取等下游任务的指令微调,得到了一个统一的信息抽取框架,并在多个数据集上得到了与 BERT^[10]相当的性能^[11]。在医学垂直领域,大一模型通过指令微调 Qwen^[12]基座得到了一个具有信息抽取能力的医疗大模型^[13]。随着生成式模型训练参数的增加,模型能力越来越强,如何更好地提高其在关系三元组抽取的性能越来越受到关注,如 Wadhwa等^[14]提出的使用思维链(Chain of Thought, COT)的方式极大地增强了生成式模型进行关系抽取的性能。思维链是一种用于提升大语言模型推理能力的技术,其核心思想是将复杂问题分解为逐步推理的中间步骤,使模型能够更准确地理解问题并找到解决方案^[15]。思维链通过显式地展示推理过程,有效增加了大模型在药物不良反应关系抽取中的可解释性并缓解了大模型的幻觉问题;除此之外,大规模参数模型使用思维链能促进模型的理解能力,激发模型进行药物不良反应关系抽取的潜能。

3 思维链增强的上下文学习药物不良反应关系抽取

本文将低资源的药物不良反应关系抽取视为一个基于上下文的条件文本生成任务。为让大语言模型在少量样本中更加充分地进行上下文学习并具备更强的泛化能力,受 Auto-COT^[16]工作的启发,首先使用 Sentence-BERT^[17]方法对训练集中的社交媒体文本与医药文献文本计算向量表示;然后使用 K-means 聚类算法将训练集数据划分至 N 个簇中,并且对每个簇中的文本根据中心点距离进行升序排序,选取每个簇中离中心点距离最近的文本进行人工 COT 解释注释。这样, GPT 在每个要抽取的目标文本进行目标关系三元组抽取的过程中,都能在上下文中找到相近的训练集案例进行上下文学习,从而更加精准地思考三元组抽取的相关问题,从而增加

三元组生成的准确性和鲁棒性,以提高框架的抽取性能。GPT 低资源关系抽取模型的整体框架如图 2 所示。

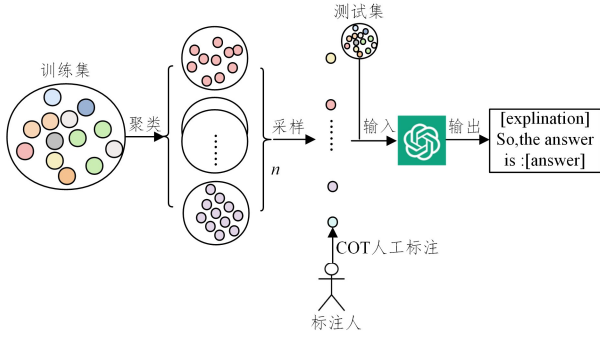


图 2 GPT 低资源关系抽取

Fig. 2 Low-resource relation extraction in GPT

模型的输入内容主要包括任务描述、 N 个不同簇的案例以及目标文本,输出包括目标文本提取三元组的解释和最终按照格式输出的答案。所有上下文字符串以及输出文本示例如图 3 所示。本文将此药物不良反应低资源抽取框架命名为 ADR-GPT (Adverse Drug Reaction Generative Pre-trained Transformer)。

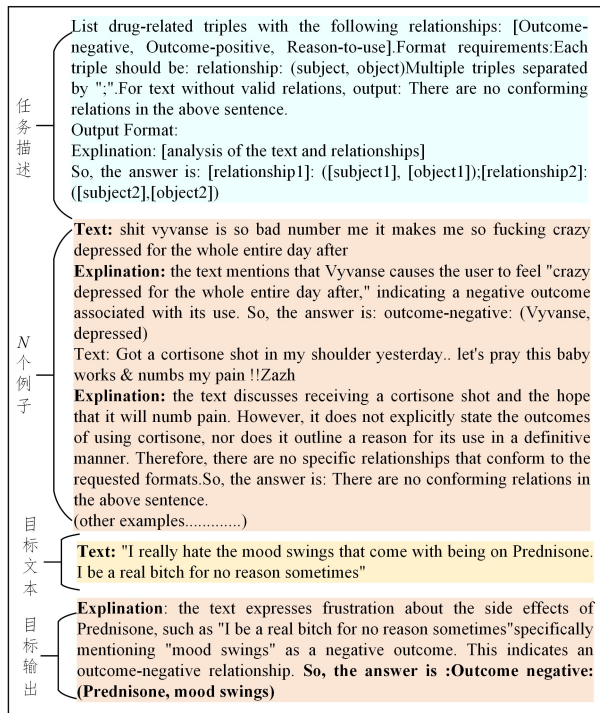


图 3 低资源思维链下的提示模板

Fig. 3 Prompt template under low-resource COT

4 思维链增强下微调的药物不良反应关系抽取

在有监督的情况下,本文通过引入药物不良反应关系三元组抽取的思维链解释来丰富用于微调模型的监督信息。思维链解释是在训练集实例上自动从 GPT-4o 中获取的,构建流程如下:首先,将文本和关系三元组输入 GPT-4o,从中获取文本中关系三元组的解释信息,同时人工设计多条高质量的提示语;随后,将提示语与目标文本拼接结合作为问题,COT 解释和关系三元组拼接结合作为

答案,从而得到指令问答对数据;最后,使用高效微调算法 QLoRA 对 Llama3-8B 模型进行微调,得到思维链增强的药物不良反应抽取大语言模型——CADRE-LLM (COT Enhanced Adverse Drug Reaction Relation Extraction Large Language Model)。

4.1 思维链数据生成

本文使用指令,从 GPT 中获取思维链解释。该指令与上文引导 ADR-GPT 输出三元组的指令类似,同样包括一个任务的提示语、 n 个案例和目标文本与关系三元组,输出为文本中为什么含有该关系三元组的解释。目标是让 GPT 根据目标文本和目标关系生成相关的解释数据,从而获得含有思维链增强的药物不良反应关系抽取数据集。

4.2 指令对数据构建

为了增强指令对的泛化性,设计了 5 条高质量的前置提示语,将提示语和待发现文本进行拼接作为问答对的问题,将获取到的思维链数据和三元组答案进行拼接作为答案,中间使用连接词:“So, the answer is:”进行平滑过渡,以增强回答的连贯性。最后所获得的药物不良反应关系三元组问答对数据集示例如图 4 所示。

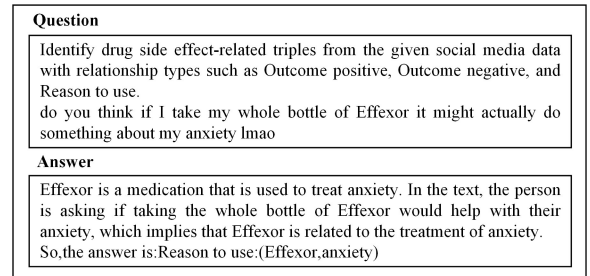


图 4 QA 问答对

Fig. 4 Question-answer (QA) pairs

4.3 基于 QLoRA 的指令微调

使用构建好的问答对对 Llama3-8B 的基座 Base 模型进行指令微调,以获得能够先思考后回答关系三元组的模型 CADRE-LLM。为了节省训练的成本与训练时间,使用高效的方式 QLoRA 进行模型的指令微调训练。

QLoRA (Quantized Low-Rank Adaptation)^[18] 是一种用于大语言模型的高效微调方法。其核心思想是通过低秩分解和量化技术,显著减少微调过程中所需的计算资源和存储开销,同时保持模型的性能,旨在突破传统微调方法在计算资源和存储需求上的瓶颈。

模型的原始权重矩阵为 W ,低秩分解是将权重矩阵的更新量分解为两个较小的矩阵 A 和 B 的乘积,然后得到更新后的权重矩阵 W' ,具体如式(1)所示。

$$W' = W + A \cdot B \quad (1)$$

这种方式只需要微调 A 和 B 两个较小的矩阵,而不是整个权重矩阵,从而大大减少了微调所需的参数量和计算量。另外,量化技术是将原模型 W' 的浮点数值表示的权重和激活值转换为低精度的整数表示,因此可以有效减少模型的存储需求并加速计算过程。整个 CADRE-LLM 的训练过程如图 5 所示。

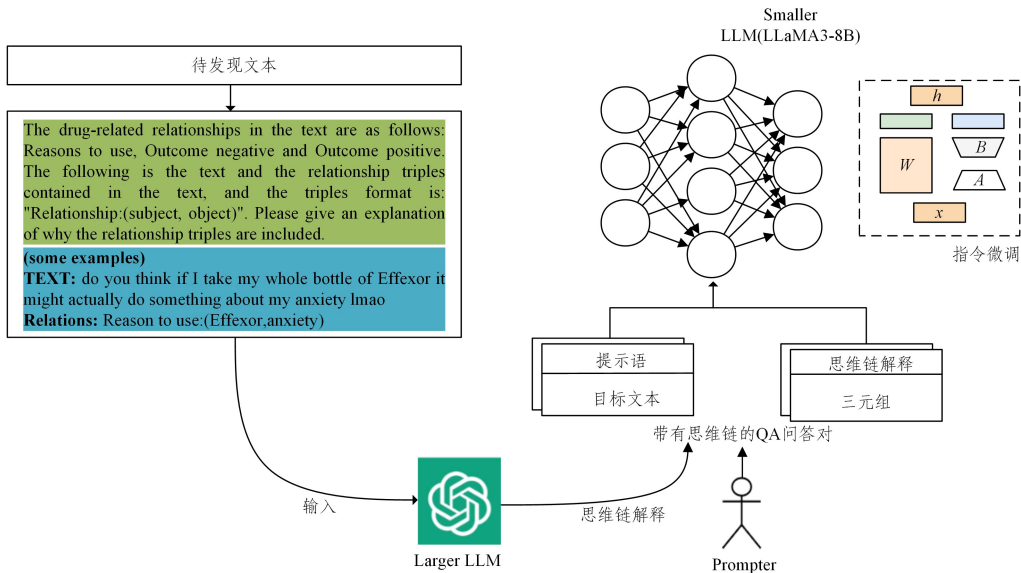


图5 思维链下的大语言模型微调

Fig. 5 COT enhanced fine-tuning of large language models

5 实验及分析

5.1 数据集

为公平评估两个框架的性能,在社交媒体数据集 Twitter 和医药文献数据集 Pubmed 上进行框架的泛化性验证。数据集情况如表 1 所列。

1) Twitter^[19]:来自社交媒体平台 Twitter,每个文档都附有疾病、症状、药物及其关系的注释,3 种类型的关系为 Outcome-negative, Outcome-positive 和 Reason-to-use。原数据集中只给出了相应的推特 ID 供用户爬取,共收集到 625 条相关的 Twitter 语料作为社交媒体数据集。

2) Pubmed^[19]:来自于医学领域的 Pubmed 文献,文档同样有疾病、症状、药物以及关系注释,同样包括 Outcome-negative, Outcome-positive 和 Reason-to-use 这 3 种关系,共有 1000 条文本数据。

表 1 药物不良反应关系抽取数据集介绍

Table 1 Introduction to ADR relationship extraction dataset

数据集	Outcome negative	Outcome positive	Reason to use	总计
Twitter	311	7	294	612
Pubmed	963	18	264	1245

5.2 实验参数

针对有监督微调的方法统一训练 4 个 epoch,QLoRA 的精度量化为 4 bit,Dropout 使用 0.05,学习率使用 2×10^{-4} ,优化器使用 Adamw,生成的温度设为 0.3,使用 NVIDIA RTX 3090 进行训练。对于上下文提示生成的 GPT,温度同样设为 0.3。

5.3 评价指标

实验采用精确率(Precision, P)、召回率(Recall, R)和 F1 值作为评估模型的指标,其中 F1 值可以通过平衡精确率和召回率来量化模型的整体性能。

$$P = \frac{TP_{\text{num_triple}}}{TP_{\text{num_triple}} + FP_{\text{num_triple}}} \quad (2)$$

$$R = \frac{TP_{\text{num_triple}}}{TP_{\text{num_triple}} + FN_{\text{num_triple}}} \quad (3)$$

$$F1 = \frac{2P * R}{P + R} \quad (4)$$

其中, $TP_{\text{num_triple}}$ 表示预测正确的关系三元组, $FP_{\text{num_triple}}$ 表示假阳性的关系三元组, $FN_{\text{num_triple}}$ 表示假阴性的关系三元组。

5.4 对比实验分析

为评估低资源下和高资源下的两个药物不良反应抽取框架的有效性,本节选取目前较为先进的关系抽取框架进行对比。基线模型主要分为基于编码器的判别式模型和基于生成式的大语言模型。

1) CasRel^[20]:采用流水线方式的级联二分类框架,先识别主体实体,再识别与主体相关的客体实体及关系。

2) TPLinker^[7]:通过构造全局信息矩阵直接链接实体和关系,能够同时处理实体和关系的抽取,避免了传统流水线模型中可能出现的错误传播问题,端到端地抽取关系和实体三元组。

3) OneRel^[8]:通过增加实体相互之间的交互,同时减少冗余信息,进行一步解码方式,减少误差传播。

4) GPT-4o-zero-shot^[21]:基于 OpenAI 开发的 GPT-4o,在零样本学习任务中直接抽取药物不良反应关系三元组答案。

5) GPT-4o-few-shot^[21]:对 OpenAI 的 GPT4o 进行少量高质量样本的提示,让模型在上下文中学习药物不良反应关系。

6) Qwen2-SFT^[12]:阿里发布的中英文开源大语言模型,经过监督微调,针对药物不良反应三元组抽取任务进行任务优化。

7) Llama3-SFT^[22]:经过指令数据对的监督微调,针对药物不良反应三元组抽取进行任务优化。

实验结果如表 2 和表 3 所列。可以看到,本文提出的高资源下的药物不良反应检测模型 CADRE-LLM 相比于之前最先进的判别式 OneRel 模型,分别在 Twitter 和 Pubmed 数

据集中将 F1 值提高了 1.51 个百分点和 1.74 个百分点;对于单纯使用指令微调的 Llama3-SFT 模型, CADRE-LLM 引入了增强的思维链,在 Twitter 和 Pubmed 数据集中将 F1 值分别增加了 3.37 个百分点和 3.57 个百分点。这和先前的理论一致,即使用 GPT 增强的思维链数据增加了 Llama3 对文本中药物不良反应关系三元组的理解能力,可以更精确地抽取文本中的药物不良反应关系三元组。

表 2 Twitter 数据集中的性能报告

Table 2 Performance report on the Twitter dataset (%)

模型	P	R	F1
CasRel	44.19	42.33	43.24
TPLinker	47.33	46.25	46.78
OneRel	48.11	46.53	47.31
GPT-4o-ZeroShot	28.73	34.59	31.39
GPT-4o-FewShot	32.01	39.67	35.43
Qwen2-SFT	42.11	41.08	41.59
Llama3-SFT	41.89	49.66	45.45
ADR-GPT	37.01	39.22	38.08
CADRE-LLM	45.76	52.31	48.82

表 3 Pubmed 数据集中的性能报告

Table 3 Performance report on the Pubmed dataset (%)

模型	P	R	F1
CasRel	48.21	49.33	48.76
TPLinker	52.77	48.24	50.40
OneRel	51.26	49.88	50.56
GPT-4o-ZeroShot	34.01	39.67	36.62
GPT-4o-FewShot	39.15	44.82	41.79
Qwen2-SFT	42.77	48.21	45.33
Llama3-SFT	48.69	48.77	48.73
ADR-GPT	42.86	49.21	45.82
CADRE-LLM	48.22	57.13	52.30

本文提出的低资源下的药物不良反应检测框架 ADR-GPT 在 Twitter 数据集中的 F1 值低于 Qwen2-SFT 模型 3.51 个百分点,而在 Pubmed 数据集中的 F1 值高于 Qwen2-SFT 模型 0.49 个百分点。尽管在低资源下基于聚类思维链上下文学习的 GPT 框架相较于先进的有监督微调数据仍然有一定差距,但随着生成式大语言模型的发展,模型自身能力越来越强,由于无需耗费大量的训练资源即可得到有效抽取结果,因此在未来低资源下的抽取框架是药物不良反应抽取的趋势,且值得进一步探究。

另外,可以看到在药物不良反应的关系抽取任务中,生成式模型几乎普遍呈现高召回率而低精确率的特性。例如, CADRE-LLM 框架在 Twitter 和 Pubmed 数据集上的 R 值相较于 P 值分别高出 6.55 个百分点与 8.91 个百分点,其原因在于生成式模型的预训练是基于概率生成机制的本质,通过自回归式语言建模生成式模型倾向于最大化序列生成的可能性,从而广泛覆盖输入文本中潜在的关系表述,这使其能够检测到传统判别式模型可能忽略的长距离依赖与弱信号关系,使得 TP_{num_triple} 上升,显著提升召回率 R 值。然而,这种生成过程缺乏严格的边界约束,模型基于统计先验而非文本证实,可能生成大量语义合理但缺乏直接证据的三元组,导致假阳性数量 FP_{num_triple} 上升。另外,预训练阶段学习的通用语言模

式与特定领域关系逻辑之间的偏差进一步加剧了生成结果的噪声,使得精确率相对降低, P 值下降。

5.5 ADR-GPT 框架的消融实验分析

为了进一步验证本文提出的低资源下药物不良反应抽取框架 ADR-GPT 中各模块的有效性,本节进行了消融实验分析。消融实验包括:从框架中消融掉聚类算法;使用从训练集中随机选取文本进行思维链的提示学习代替;消融掉 COT 解释,保留聚类算法,让 GPT 进行无解释的上下文少案例提示学习;消融掉相关案例提示,直接让 GPT 自己进行思维链推理,并生成三元组答案。

在社交媒体数据集 Twitter 中的消融实验结果如表 4 所列。可以看到,在消融掉聚类模块后,框架性能 F1 值下降了 1.30 个百分点,证明了通过细粒度聚类后从不同类别中选取有代表性的案例使得模型的提示框架更加鲁棒,让每个文本都能在提示案例中找到与自己相近的案例上下文学习,从而提升框架的性能。消融掉 COT 模块后,框架性能 F1 值下降了 1.77 个百分点,表明了在大语言模型中让模型先思考再进行抽取可以在一定程度上缓解大模型幻觉现象,让模型更准确地预测到文本中的关系。直接消融案例提示后,模型框架 F1 值下降了 2.55 个百分点,由于测试的文本数据中药物不良反应标注标准不同,直接消融掉提示案例后可能导致 GPT 对文本中关系的理解与数据集中的标注维度有偏差,故下降更为明显。

表 4 ADR-GPT 消融实验

Table 4 Ablations study of ADR-GPT (%)

消融模块	P	R	F1
ADR-GPT	37.01	39.22	38.08
-聚类	35.41	38.27	36.78
-COT	35.22	37.48	36.31
-案例提示	34.17	36.98	35.53

5.6 CADRE-LLM 框架的消融实验分析

为证明本文提出的高资源下药物不良反应抽取框架 CADRE-LLM 中各模块的有效性,本节同样进行了消融实验分析,包括:从框架中消融掉 GPT 生成的思维链 COT 部分;消融掉多条 Prompt 指令,只使用单一的 Prompt 指令进行提示。

在社交媒体数据集 Twitter 的消融实验结果如表 5 所列。可以看出,消融掉思维链 COT 解释后,模型 F1 值下降了 3.37 个百分点,证明了在指令微调的范式下思维链数据对药物不良反应关系抽取起到提示思考的作用,使得模型对文本中的关系识别更加准确。消融掉多条 Prompt,使用单一的提示语作为问答对的指令模型的 F1 值下降了 0.69 个百分点,这说明多条高质量的 Prompt 可以更好地激发大语言模型的学习能力,提高其泛化性能,防止收敛到单一域而使得模型性能下降。

表 5 CADRE-LLM 消融实验

Table 5 Ablation study of CADRE-LLM (%)

消融模块	P	R	F1
CADRE-LLM	45.76	52.31	48.82
-COT	41.89	49.66	45.45
-Multiple Prompts	45.72	50.82	48.14

5.7 不同聚类数量对 GPT 低资源抽取的影响

本节进一步对 ADR-GPT 框架的聚类数量超参数进行分析,分别选取 5,10,15 作为聚类类别,对社交媒体中的药物不良反应关系抽取数据集 Twitter 的训练集进行不同聚类划分,以评价 ADR-GPT 的性能结果。实验结果如图 6 所示。

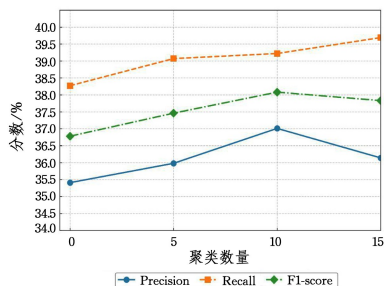


图 6 聚类数量对 ADR-GPT 性能的影响

Fig. 6 Influence of the number of clusters on the performance of ADR-GPT

可以看出,选取不同的聚类对 ADR-GPT 框架的影响不同:当聚类数量过少时,性能影响较大,所要预测的文本可能没有找到合适的案例进行参考,导致关系三元组抽取有误差;当聚类数量过多时,模型性能会出现收敛甚至下降的趋势,这是因为当聚类过多时,类别过于细化而模糊不清,对 GPT 原有的知识造成干扰,社交媒体数据噪声较大,过多的示例会导致示例引入不一致,从而给模型造成困扰。在本次社交媒体的药物不良反应关系抽取中,当聚类数量为 10 时,F1 值达到最优。

结束语 本文面向社交媒体和医药文献提出了两种基于思维链的生成式大模型药物不良反应关系抽取框架,分别解决低资源和高资源场景下的药物不良反应关系三元组抽取问题。

低资源框架 ADR-GPT 结合聚类算法与上下文学习,通过自动筛选代表性样本并引入人工标注的思维链解释,提升了 GPT 模型在少样本条件下的抽取性能。其性能虽略低于全监督微调模型,但显著优于零样本和少样本基线,验证了低资源场景下聚类增强与思维链提示对药物不良反应关系抽取的有效性。高资源框架 CADRE-LLM 基于 Llama3-8B 模型,利用 GPT-4o 生成的思维链数据增强监督微调,结合多指令提示优化模型泛化能力,超越了传统判别式模型和普通指令微调模型,证明思维链数据能显著提升模型对药物不良反应关系的理解。另外,本文对低资源框架中的聚类、思维链和案例提示进行消融分析,对高资源框架中的思维链解释和多指令提示进行消融分析,证明了两个框架的有效性。

本文框架虽然在药物不良反应关系抽取任务上取得了不错的性能,但仍然有很大的提升空间。未来将进一步结合生成式框架,融入外部已有的先验知识来缓解大模型的幻觉现象,提高药物不良反应关系抽取的性能,警示患者药物的不良反应症状,助力药物研发机构改进药物质量。

参考文献

[1] NATIONAL CENTER FOR ADR MONITORING. Annual re-

port of national adverse drug reaction monitoring(2023) [J]. Chinese Journal of Viral Diseases,2023,13(4):245-251.

- [2] LI H, QIU Y Z, LIN H F. Adverse drug reaction detection with multi-feature enhancement for social media [C]// Proceedings of the 23rd Chinese National Conference on Computational Linguistics. 2024:515-525.
- [3] ZHANG T, LIN H, REN Y, et al. Adversarial transfer network with bilinear attention for the detection of adverse drug reactions from social media [J]. Applied Soft Computing, 2021, 106: 107358.
- [4] ZHANG T, LIN H, REN Y, et al. Adverse drug reaction detection via a multihop self-attention mechanism [J]. BMC Bioinformatics, 2019, 20(1): 1-11.
- [5] ZHENG S, WANG F, BAO H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1227-1236.
- [6] REN F, ZHANG L, ZHAO X, et al. A simple but effective bidirectional framework for relational triple extraction [C]// Proceedings of the 15th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2022:824-832.
- [7] WANG Y, YU B, ZHANG Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking [C]// Proceedings of the 28th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2020:1572-1582.
- [8] SHANG Y, HUANG H, MAO X. Onerel: Joint entity and relation extraction with one module in one step [C]// Proceedings of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022:11285-11293.
- [9] YE H, ZHANG N, DENG S, et al. Contrastive triple extraction with generative transformer [C]// Proceedings of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021:14257-14265.
- [10] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conf of the North American Chapter of the ACL: Human Language Technologies. Stroudsburg, PA: ACL, 2019:4171-4186.
- [11] WANG X, ZHOU W, ZU C, et al. InstructUIE: Multi-task instruction tuning for unified information extraction [J]. arXiv: 2304.08085, 2023.
- [12] BAI J, BAI S, CHU Y, et al. Qwen technical report [J]. arXiv: 2309.16609, 2023.
- [13] LUO L, NING J, ZHAO Y, et al. Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks [J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1865-1874.
- [14] WADHWA S, AMIR S, WALLACE B. Revisiting relation extraction in the era of large language models [C]// Proceedings of the 61st Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2023:15566-15577.
- [15] KOJIMA T, GU S, REID M, et al. Large language models are

zero-shot reasoners [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213.

- [16] SHUM K, DIAO S, ZHANG T. Automatic prompt augmentation and selection with chain-of-thought from labeled data [C]// *Findings of EMNLP 2023*. Stroudsburg, PA: ACL, 2023: 12113-12139.
- [17] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks [C]// *Proceedings of EMNLP-IJCNLP 2019*. Stroudsburg, PA: ACL, 2019: 3982-3992.
- [18] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: Efficient finetuning of quantized LLMs [J]. *Advances in Neural Information Processing Systems*, 2023, 36: 10088-10115.
- [19] ALVARO N, MIYAO Y, COLLIER N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations [J]. *JMIR Public Health and Surveillance*, 2017, 3(2): e6396.
- [20] WEI Z, SU J, WANG Y, et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction [C]// *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics. 2020: 1476-1488.

- [21] ACHI AM J, ADLER S, AGARWAL S, et al. GPT-4 technical report [J]. arXiv:2303.08774, 2023.
- [22] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The LLaMA 3 herd of models [J]. arXiv:2407.21783, 2024.



LI Hao, born in 2001, postgraduate. His main research interests include natural language processing, large language models and information extraction.



LIN Hongfei, born in 1962, Ph.D. professor, Ph.D supervisor. His main research interests include natural language processing, affective computing and computational humor, information retrieval and recommendation systems.

(责任编辑:柯颖)