基于大模型数据增强的药物相互作用关系抽取

朱 海[1] 邵山武[1]
ZHU Hai SHAO Shanwu

摘 要 药物治疗是治愈疾病的重要手段之一，药物种类繁多且关系复杂，在诊断用药时药物间的相互作用关系可能会影响患者的健康，因此研究自动化的药物相互作用关系抽取方法是必要的。目前，相关研究中存在高质量标注数据较少且语义表达不够丰富的问题。针对该问题，文章提出一种简易的数据增强方法，更好地捕捉数据的不变性并增加样本量。具体而言，利用大型语言模型（Llama2）作为生物医学文本数据生成器，通过合理设计提示工程来增加数据的多样性，并且使用一致性检验确保生成数据的正确标注。实验结果表明，使用该方法生成的数据在多个关系抽取模型中均取得了不错的性能。

关键词 大语言模型；Llama2；数据增强；提示工程；关系抽取

doi：10.3969/j.issn.1672-9528.2024.12.004

0 引言

在患者治疗中，药物相互作用（DDI）对治疗方案和疗效有重要影响，可能对患者健康或公共卫生构成威胁。DDI指患者同时服用两种或多种药物时产生的反应，如药物之间的对抗或协同作用。大量药物上市前需仔细研究[1]，如果药物相互作用未被发现，可能导致不良药物反应（ADR）。例如，同时服用头孢曲松和兰索拉唑的患者可能会出现危及生命的心律失常[2]。此外，随着人口老龄化严重，在老年群体发生ADR的概率大大增加[3]。基于此，挖掘DDI信息对于药物研究者和临床医生至关重要。

目前，已有多种DDI数据库可供研究人员、医生和患者使用，如DrugBank、KEGG等。然而，一方面这些数据库的更新速度往往滞后，许多DDI信息未能得到充分提取和利用[4]。另一方面，随着生物医学出版物数量的迅速增长，使得从文章中手动提取有价值的DDI信息变得越来越困难。因此，开发自动化的DDI提取方法至关重要。

为了从海量文本中提取关键的DDI信息，此前许多基于机器学习的方法被提出。而近年来随着深度学习的快速发展，基于模型的方法已成为主流。主要方法包括基于卷积神经网络（CNN）[5]和循环神经网络（RNN）[6]等。

尽管深度学习模型在DDI关系提取中取得了显著进展，但其性能始终受限于DDI数据集人工标注的质量和数量。虽然数据增强可以缓解该问题，但是如何生成高质量的增强文本仍是一个关键。

目前，大语言模型在生成新的标注数据上已被证明有效，因此本文提出了一种基于大语言模型的文本数据增强方法来解决上述问题。主要方法为：（1）使用Llama2作为数据生成器来确保所生成数据的多样性；（2）设计一种大模型专用的提示模板以及使用一致性检验来提升生成数据的质量。实验结果表明，该方法在关系预测上取得了不错的性能。

1 相关工作

1.1 关系抽取

此前，研究人员开发了许多优秀的DDI抽取模型。这些模型主要分为两类：基于机器学习的方法和基于深度学习的方法。

传统机器学习方法中，基于特征的方法最为常见，通常包括词汇特征、上下文特征和句法特征等。KIM等人[7]提出了一种基于SVM的药物相互作用预测模型，并定义了5种特征类型：单词、单词对、句子依赖图、句子解析树和名词短语，从而能够通过丰富的特征来获取文本中的隐藏信息。Huang等人[8]采用基于特征的分类器并结合SVM和LSTM来提取DDI。然而，基于机器学习的方法需要冗余的特征工程，并且其在DDI提取中的性能表现不足。

为应对传统机器学习的挑战，利用神经网络的深度学习技术成为DDI抽取研究的一个重要方向。Liu等人[9]首次使用卷积神经网络模型进行DDI抽取，该模型将词汇转化为词向量，并结合位置信息作为特征输入，从而打破了传统方法对特征工程的依赖。Zhang等人[10]提出了一种基于分层递归神经网络的方法，将SDP和句子序列整合用于DDI提取任务，并且额外引入注意力机制来更好识别文本中的关键词。Deng

1. 广州南方学院商学院 广东广州 510000

等人[11]则提出了一个 DDIMDL 深度学习框架，将多种药物特征融合在一个模型中以预测 DDI。

2018 年，Google 推出了预训练语言模型 BERT[12]，并且在所有 11 项自然语言处理任务中都取得了最佳表现。不久后，专门用于生物医学领域的语言模型问世，如 PubmedBERT[13]和 SciBERT[14]，这些模型在生物医学语料库上进行训练，并在 DDI 抽取任务上取得了优秀的效果。然而，随着对预训练语言模型的深入研究，研究学者发现预训练的目标形式与下游任务之间存在较大差距，导致普通微调未能充分利用模型中的先验知识。因此 Wang 等人[15]针对 DDI 抽取任务的特点，设计了一种新的基于提示微调和数据增强的 DDI 抽取模型，通过上下文词嵌入替换来解决数据集中类别不平衡问题。

1.2 数据增强

数据质量的重要性在于确保模型学习信息的清晰性。然而，获取高质量数据的过程充满挑战，其成本高昂、耗时，且会导致不准确性。为应对这些挑战，研究人员致力于开发数据增强（DA）技术以减轻这些问题。数据增强通过应用多种增强策略，不仅丰富数据集的多样性，还能扩展数据集的范围，从而增强模型的鲁棒性并提升其泛化能力。

数据增强主要包括基于规则和基于模型的方法。EDA（easy data augmentation）技术是基于规则的方法，如同义词替换、随机插入、随机交换和随机删除等。除此之外，Abdollahi 等人[16]还提出了两种新的基于规则的方法：一种是本体引导方法，另一种是结合了本体和词典的方法，通过替换本体来增强医疗文本数据。然而，这些方法也面临一些局限，如现有数据分布与实际数据分布不一致、生产的数据信息丢失或失真以及标签不一致等。

随着深度学习技术的发展，专家学者开始探索基于模型的方法，Wu 等人[17]提出了 C-BERT 模型，其通过 BERT 随机替换单词来增强标注文本。近年来，大语言模型已经成为一个研究热点，如 GPT-3[18]、Llama2[19]。这些模型在复杂任务中展示了非凡的能力，并推动了大语言模型的流行。因此有专家学者将其用于数据增强研究中以探究模型的性能。Dai 等人[20]提出了一种基于 ChatGPT 的文本数据增强方法，其将训练样本中的每个句子改写为多个概念相似但语义不同的样本，确保生成数据的正确性和多样性。Cai 等人[21]同样使用大语言模型作为数据生成器来创建高质量的科学文本数据，旨在解决数据不平衡的挑战。Zhang 等人[22]提出了 3 种简单有效的策略，即词汇级、句法级和篇章级的数据增强策略。这些策略使大语言模型能够在保持文本依存结构的同时，确保准确性并增加不同层级的多样性。

2 模型方法

2.1 整体框架

本文所使用方法的框架主要包括两个部分：（1）Text Augmentation：使用大语言模型扩充原始数据集，所使用的大语言模型为 Llama2；（2）Text Classification：使用生物医学领域的预训练语言模型抽取数据集中的关系。

整体框架如图 1 所示。首先，使用 Llama2 进行数据增强生成增强数据集。之后，再使用增强后的数据集和原始数据一起微调 BERT 并用于 DDI 预测。这一过程的核心在于使用增强数据来提升原始数据的多样性，确保 BERT 模型可以学习到不同的语义表达。

2.2 基于 Llama2 的数据增强

为了生成优质的文本数据，本文使用 Llama2 作为数据生成器，通过构建提示的方法来生成文本数据。Llama2 采用和 GPT 系列类似的 Transformer 架构，该架构通过自注意力机制（Self-Attention）实现对输入序列的全局依赖建模，使得模型能够捕捉到序列中各个位置之间的关系。

2.2.1 注意力机制

自注意力机制是 Transformer 的核心组件，其计算过程可以表示为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (1)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别代表查询（Query）、键（Key）和值（Value）矩阵，d_k 为键的维度。在 Llama2 中，自注意力机制被应用于每一层的多头注意力（multi-head attention）模块中，使得模型能够从不同的子空间中提取特征。

多头注意力机制则是通过多个注意力头并行处理输入序列，从而捕捉到不同子空间中的特征。其计算过程为：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\boldsymbol{W}^O \quad (2)$$

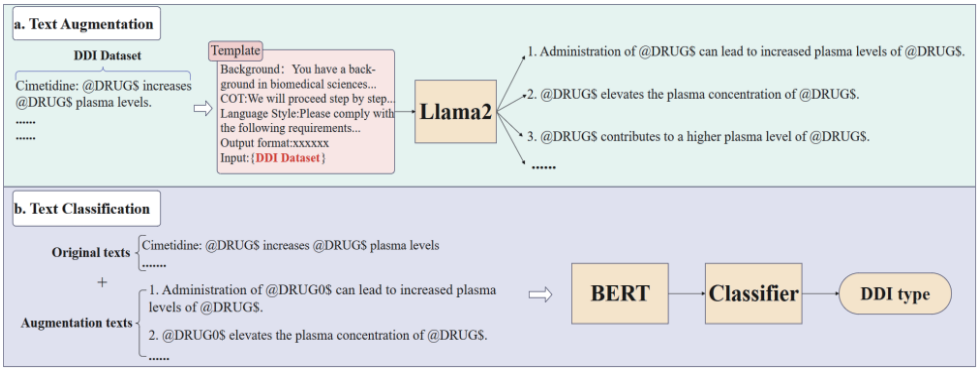

图 1 模型整体框架

其中，每个注意力头的计算方式为：

$$\text{head}_i = \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{3}$$

式中：$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$ 为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 各自的线性变换矩阵；$\boldsymbol{W}^O$ 为输出线性变换矩阵。

2.2.2 位置编码和激活函数

Llama2 采用旋转位置编码（rotary position embedding，RoPE）来编码 token 的位置信息。RoPE 通过将词向量旋转特定角度来隐式地编码位置信息 , 其数学表达式为：

$$\text{RoPE}\left(x_m, \theta_m\right) = \begin{bmatrix} \cos\left(m\theta\right) x_m - \sin\left(m\theta\right) y_m \\ \sin\left(m\theta\right) x_m + \cos\left(m\theta\right) y_m \end{bmatrix} \tag{4}$$

式中：x_m 为第 m 个位置的词向量；θ_m 为旋转角度。

在前馈神经网络中，Llama2 使用了 SwiGLU 激活函数，这是对传统 GELU 函数的改进。SwiGLU 的数学表达式为：

$$\text{SwiGLU}\left(x\right) = x \cdot \sigma(\beta x) \cdot (\alpha x) \tag{5}$$

式中：σ 表示 sigmoid 函数；α 和 β 为可学习参数。

2.2.3 模型训练及微调

Llama 2 的预训练采用大规模的无监督学习。模型在约 2 万亿个 token 上进行训练，数据来源包括网页、书籍、文献等多种语料。训练目标为最小化下一个 token 的负对数似然：

$$L = -\sum_{t=1}^{T} \log P(x_t | x_{<t}) \tag{6}$$

式中：x_t 表示第 t 个 token；$x_{<t}$ 表示 t 之前的所有 token。

为提升模型在特定任务上的表现，Llama2 进行了指令微调和基于人类反馈的强化学习（RLHF）。这一过程使用了人类标注的高质量指令回复对，通过监督学习方式优化模型参数。其中 RLHF 方法是通过拒绝采样和近端策略优化对模型进行迭代优化。在 RLHF 阶段，累积迭代奖励建模数据与模型改进同时进行，从而确保奖励模型的分布始终保持正确。

3 文本分类

为了突出数据增强的有效性，本文仅使用在生物医学领域预训练好的 BERT 模型作为关系分类器。其中 BERT 的顶层输出特征 h 可以表示为：

$$z = [z_c, z_1, z_2, ..., z_n] \tag{7}$$

式中：z_c 是类别特定的 CLS 标记的表示。在文本分类中，通常将 z_c 输入到一个任务特定的分类器中以进行最终预测。然而，在正样本数据量少的场景下，通过 BERT 微调难以获得令人满意的性能。因为整体样本较少会导致模型过拟合和缺乏泛化能力。

为了解决该问题，本文使用原始数据集和通过 Llama2 生成的文本数据集一起来微调 BERT。并且所使用的目标函数为交叉熵损失，因此，将 z_c 输入到一个全连接层中，用于最终预测的分类器：

$$\hat{y} = W_c^T z_c + b_c \tag{8}$$

式中：W_c 和 b_c 是可训练参数，并且目标函数为：

$$L_{CE} = -\sum_{d \in D'} \sum_{c=1}^{C} y_{dc} \ln \hat{y}_{dc} \tag{9}$$

式中：C 是输出维度，指的是整体数据集的标签分布；y_d 是真实标签。

3.1 提示工程设计

医疗文本是一种由特定领域相关的关键词所组成的文本类型。为有效生成此类数据并增强生成数据的多样性，采用提示工程（prompt engineering）方法与大语言模型进行交互。如表 1 所示，这种“提示”由 4 个主要部分组成，即背景、思维链（COT）、格式和语言风格。

表 1 提示工程组成要素及解释

背景	提供大模型需了解的领域相关的背景知识
思维链	定义大模型生成文本的详细思考过程
语言风格	约束大模型生成的语调和语言风格
输出格式	限制大模型所输出的生成文本的格式

在本文中，基于上面四要素设计了一种大模型通用的提示模板，具体内容如图 2 所示。最终使用该模板对数据集中的每一条正样本生成了 3 条以上的文本数据。

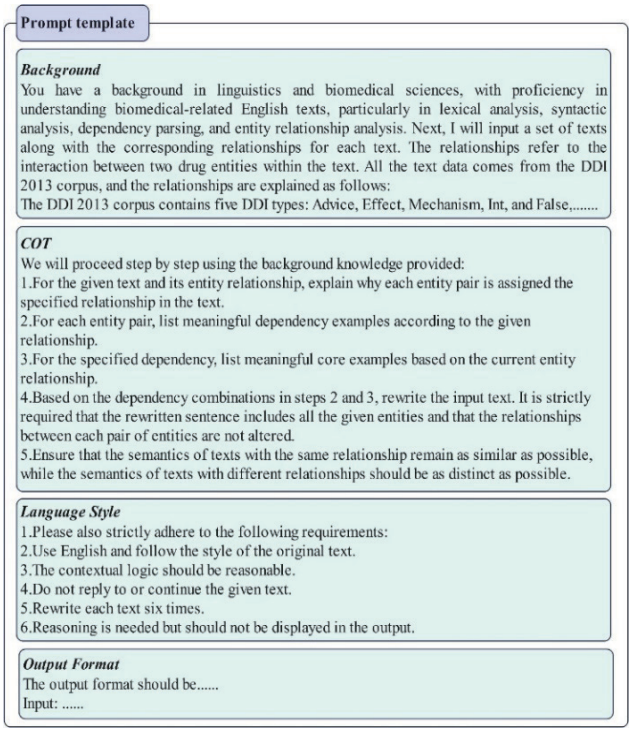

图 2 提示模板

3.2 文本评估

本文参考 Alberti 等人[23]所提出的回合一致性方法（roundtrip consistency approach，RCA）来提升文本质量。具体而言，该方法利用现有的分类模型来预测由 Llama2 生成的文本数据的药物相互作用关系，从而确保预测的结果与原始文本的关系一致。使用 3 个不同参数的生物医学预训练语言模型来预测文本中药物对实体的关系。如果其中两个模型的预测结果与原始数据的关系一致，则保留该条文本数据。最终通过该方法为每条原始正例保留 3 条增强的文本数据。

4 实验

4.1 数据集和评价指标

为了验证模型的关系抽取效果，本章使用目前药物间相互作用关系抽取任务中常用的评估数据集，它来自于 DDIExtraction 2013 任务。数据集包含 5 种关系类型：Mechanism、Effect、Int、Advice 和 False。其中 False 表示药物之间不具有相互作用关系。

本文采用精确率 Precision、召回率 Recall 以及 F_1 值作为实体关系抽取的评价指标。这些指标在数据集上通过微平均的方式计算，具体公式为：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

$$F_1 = \frac{2PR}{P + R} \tag{12}$$

式中：TP 指正确识别的关系数；FP 表示预测为某一类型但识别错误的关系数，即假正例；FN 表示未能预测出关系数。

4.2 实验结果

为了更好地评估所提出方法的有效性，本文使用 PubmedBERT 和 SciBERT 作为基础的预训练模型。首先，在 Llama2 上生成增强文本，并且将这些文本和原始数据集进行合并。然后，使用合并后的数据集来微调预训练模型。为了评估数据增强方法的效果，设置了两个不同的实验配置。第一个是使用原始数据集进行微调；第二个是使用添加了新的增强文本数据集进行微调。在所有实验中，使用的批次大小为 32，最大序列长度为 256，学习率为 4e-5。

表 2 展示了本文所使用的各个模型方法的实验结果，其中 DA 代表使用了数据增强的方法。可以看出本文提出的数据增强方法在 PubmedBERT 和 SciBERT 上均取得了不错的性能提升。其中在 PubmedBERT 上，该方法相比于原始模型的 F_1 值提升了 1.63%，SciBERT 则提升了 1.19%。这些结果表明，本文所提出的方法在提升模型性能方面具有显著效果。

表 2 实验结果

方法	Precision	Recall	F_1
SciBERT	80.19	77.68	78.92
SciBERT+DA	**80.58**	**79.65**	**80.11**
PubmedBERT	81.03	81.69	81.36
PubmedBERT+DA	**83.21**	**82.76**	**82.98**

同时还测试了使用不同数据的增强文本对于模型性能的影响，如图 3 所示，其中 None 表示只使用原始文本，可以看出使用 6000 条增强样本时具有最好的性能。

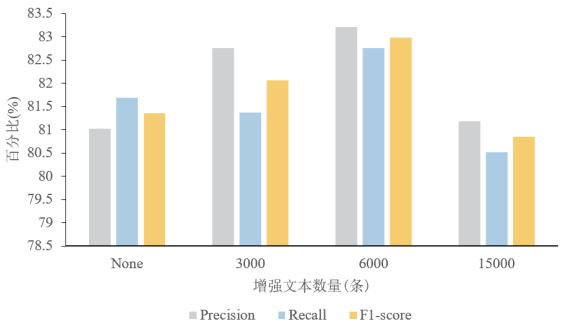

图 3 不同数量增强文本的效果对比

当样本数量为 3000 时，虽然总体性能略有下降，但仍然比不使用文本数据增强的模型性能更优。然而，当样本数量为 15 000 时，模型的性能有所下降，这有可能是更多的数据带来了噪音。总体来说，使用大语言模型进行数据增强在解决数据量较少的问题中展示出很大的潜力。但在选择增强样本数量时需要谨慎考虑，以保证模型性能最优。未来研究可以进一步探索优化合成样本数量，并改进生成过程，以提升模型性能。

5 结论

在本文中，提出了一种新的数据增强方法用于药物相互作用关系抽取。与其他方法不同，数据增强方法更加简洁有效，使用大模型在语义层面扩展有限的数据，以增强数据的一致性和鲁棒性，并且通过一致性检测来提升文本质量。实验结果表明，使用大型语言模型进行数据增强可以有效提升关系抽取模型的性能。

尽管这项工作在缓解生物医学领域文本数据不足问题上展示了不错的结果，但仍有值得进一步探讨的地方。由于 Llama2 缺乏专业领域知识，其可能会产生错误的增强结果，因此可以探索融合领域专业知识的提示工程技术，以进一步提高生成数据的质量和多样性。如加入药物信息等其他元数据，以更好地模拟真实数据。总体而言，这项工作为缓解生物医学领域中的数据稀缺问题提供了简洁有效的参考方法。

参考文献：

[1] 张亚飞，于琦，王于心，等 . 基于药物论坛中潜在不良反应与适应症的知识发现体系构建 [J]. 中华医学图书情报杂志 , 2020, 29(7): 38-43.

[2]LORBERBAUM T, SAMPSON K J, CHANG J B, et al. Coupling data mining and laboratory experiments to discover drug interactions causing QT prolongation[J]. Journal of the american college of cardiology, 2016, 68(16): 1756-1764.

[3] 吴明智 , 崔雷 . 生物医学实体关系抽取的研究 [J]. 中华医学图书情报杂志 , 2010,19(5):5-10.

[4] 刘悦 , 龚辉 , 单青 , 等 . 老年患者长期用药的潜在药物相互作用及其影响因素分析 [J]. 解放军医学院学报 , 2023, 44(6): 587-593.

[5]LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[6]WILLIAMS R J, ZIPSER D. A learning algorithm for continually running fully recurrent neural networks [J]. Neural computation, 1989, 1(2): 270-280.

[7]KIM S, LIU H B, YEGANOVA L, et al. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach[J]. Journal of biomedical informatics, 2015, 55: 23-30.

[8]HUANG D G, JIANG Z C, ZOU L, et al. Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks[J]. Information sciences, 2017, 415: 100-109.

[9]LIU S Y, TANG B Z, CHEN Q C, et al. Drug-drug interaction extraction via convolutional neural networks[J]. Computa-tional and mathematical methods in medicine, 2016, 2016(1): 6918381.

[10]ZHANG Y J, ZHENG W, LIN H F, et al. Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths[J]. Bioinformatics, 2018,34(5): 828-835.

[11]DENG Y F, XU X R, QIU Y, et al. A multimodal deep learning framework for predicting drug–drug interaction events [J]. Bioinformatics, 2020; 36(15): 4316-4322.

[12]DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Conference on the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg: ACL, 2019: 4171-4186.

[13]GU Y, TINN R, CHENG H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare(HEALTH), 2022, 3(1): 1-23.

[14]BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg:ACL, 2019:3613-3618.

[15]WANG K L, FU X F, LIU Y P, et al. PTDA: improving drug-drug interaction extraction from biomedical literature based on prompt tuning and data augmentation[J]. IAENG international journal of computer science, 2024, 51(5):463-476.

[16]ABDOLLAHI M, GAO X Y, MEI Y, et al. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques[J]. Artificial intelligence in medicine, 2021, 120: 102167.

[17]WU X, LV S W, ZANG L J, et al. Conditional bert contextual augmentation[C]//Computational Science,Part IV. Cham: Springer, 2019:84-95.

[18]BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[19]TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundationand fine-tuned chat models[DB/OL]. (2023-07-19) [2024-07-10].https://arxiv. org/abs/2307.09288.

[20]DAI H X, LIU Z L, LIAO W X, et al. Auggpt: leveraging chatgpt for text data augmentation[DB/OL]. (2023-03-20) [2024-07-12]. https://arxiv.org/abs/2302.13007.

[21]CAI X X, XIAO M, NING Z Y, et al. Resolving the imbalance issue in hierarchical disciplinary topic inference via LLM-based data augmentation[C]//2023 IEEE International Conference on Data Mining Workshops. Piscataway :IEEE, 2023:1424-1429.

[22]ZHANG M S, JIANG G Y, LIU S, et al. LLM–assisted data augmentation for chinese dialogue–level dependency parsing[J]. Computational linguistics, 2024, 50(3):867-891.

[23]ALBERTI C, ANDOR D, PITLER E, et al. Synthetic QA corpora generation with roundtrip consistency[C]//57th Annual Meeting of the Association for Computational Linguistics,vol. 10. Stroudsburg, PA: ACL, 2019:6168–6173.

【作者简介】

朱海（1998—），男，河南信阳人，硕士，助教，研究方向：文本挖掘、大数据分析。

邵山武（1997—），男，湖南长沙人，硕士，助教，研究方向：图像加密、计算机视觉。

（收稿日期：2024-08-19）