PGA-SciRE：基于大语言模型的数据增强框架进行科学领域的关系抽取

周洋[1]，单世民[1*]，魏宏夔[2]，赵哲焕[1*]，冯文铄[1]
[1]大连理工大学 软件学院，大连市 116620
[2]北京电子工程总体研究所 复杂产品智能制造系统技术国家重点实验室，北京市 100854
{zy22217034, fengws}@mail.dlut.edu.cn
weihongkui89@hotmail.com
{ssm, z.zhao}@dlut.edu.cn

摘要

关系提取旨在识别文本中提到的实体对之间的关系。大语言模型的进步对自然语言处理任务产生了巨大的影响。在这项工作中，我们针对科学领域的关系抽取任务，提出一个名为PGA的数据增强框架，用于提升模型在科学领域的关系抽取的性能。框架引入了两种数据增强的方式，利用大语言模型通过转述原训练集样本，得到句意相同但具备不同表述和形式的伪样本。以及指导大语言模型根据原训练集样本的关系和实体标签，生成暗含对应标签信息的句子。这两种伪样本分别与原数据集共同参与关系抽取模型的训练。实验中PGA框架提高了三个主流模型的科学领域内关系抽取的F1分数。同时，使用大语言模型获得样本也能有效减少人工标注数据的成本。

关键词： 数据增强；关系抽取

PGA-SciRE: Harnessing LLM on Data Augmentation for Enhancing Scientific Relation Extraction

Yang Zhou[1], Shimin Shan[1*], Hongkui Wei[2], Zhehuan Zhao[1*], Wenshuo Feng[1]
[1]Software school, Dalian University of Technology, Dalian 116620
[2]State Key Laboratory of Intelligent Manufacturing System Technology,
Beijing Institute of Electronic System Engineering, Beijing 100854
{zy22217034,fengws}@mail.dlut.edu.cn
weihongkui89@hotmail.com
{ssm, z.zhao}@dlut.edu.cn

Abstract

Relation Extraction aims at recognizing the relation between pairs of entities mentioned in a text. Advances in LLMs have had a tremendous impact on NLP tasks. In this work, we propose a textual data augmentation framework called PGA for improving the performance of models for relation extraction task in the scientific domain. The framework introduces two ways of data augmentation, utilizing a LLM to obtain pseudo-samples with the same sentence meaning but with different representations and forms by paraphrasing the original training set samples. As well as instructing LLM to generate sentences that implicitly contain information about the corresponding labels based on the relation and entity of the original training set samples. These two kinds of pseudo-samples participate in the training of the relation extraction model together with the original dataset, respectively. The PGA framework in the experiment improves the F1 scores of the three mainstream models for relation extraction within the scientific domain. Also, using a large language model to obtain samples can effectively reduce the cost of manually labeling data.

*共同通讯作者 Co-corresponding Author.

Keywords: Data Augmentation , Relation Extraction

1 引言

大语言模型(Large Language Model, LLM)的跨越式进展极大地提高了各项自然语言处理(Natural Language Processing, NLP)任务的性能和效率，而新的上下文学习方法的提出也为利用大语言模型解决各种NLP任务创造了新的机遇(Brown et al. (2020))。GPT系列是目前最流行的LLM，以其准确理解用户意图和产生类似人类反馈的能力而闻名。大量工作研究了LLM的性能、推理和解释能力，提出了系统的分析(Li et al. (2023a))。Wei et al. (2023)通过研究LLM在各种信息抽取(Information Extraction, IE)任务上的性能，对LLM的能力进行全面研究和详细分析。虽然LLM的性能无法与当前的SOTA模型相提并论，但证明了LLM拥有与众不同的解决NLP各种任务的能力。Han et al. (2023)提出新的对LLM的IE评价标准，即一个软匹配的评估策略，以更准确地反映LLM的性能。这些工作都证明了LLM在NLP任务中的巨大潜能。

关系抽取(Relation Extraction, RE)是从非结构化文本中获取结构化知识，是识别文本中实体之间的语义关系的任务。通常基于跨度的关系抽取方法是学习实体的跨度，再对枚举出的跨度进行多分类(Eberts and Ulges (2019))，亦或是用精心设计的特定的标记对将带有标签的实体包裹起来，以便预训练模型更好地学习实体之间的关系(Ye et al. (2022); Zhong and Chen (2021))。

一直以来，为NLP各种任务获取大规模且高质量的带有标注信息的数据是一件非常有挑战性、代价极高的事情(Beltagy et al. (2019))。特别是在科学知识图谱领域中，因为相对通用领域的的数据集，科学文献数据集往往会面临着独特的挑战。比如科学数据集可能会有更多的噪音；包含大量专业词汇、以及特殊概念的特有缩写，使得模型难以学习和泛化；由于整理它们需要花费大量时间和人力成本，因此这些数据集往往缺乏大量标注。

在过去的工作中，基于LLM来解决关系抽取任务的方法一般分为两个方向。第一种是利用上下文学习(In-Context Learning, ICL)和思维链(Chain of Thought, CoT)的方法，通过将关系抽取视为条件文本生成任务(Wan et al. (2023); Ma et al. (2023a))或问答任务(Zhang et al. (2023); Li et al. (2023b))。输入测试样例、示例、任务描述，由LLM直接给出答案。这些方法主要应用在Low-shot设置下，同时由于LLM输出的语言的灵活性，导致评估其输出的结果也成为了一种挑战(Wadhwa et al. (2023))。同时，由于科学领域的专业概念每年被持续地定义和提出，而训练LLM的语料库中并不一定包含这些缩写和概念，所以这些方法并不一定擅长工作于科学领域的关系抽取。

另外一种方法则是从模型的训练数据入手。除开模型开发，关系抽取始终涉及一定程度的标注数据。不仅是few-shot的设置下模型，全样本训练设置下的模型的性能也很大程度上被训练数据的质量和数量所影响，但是标注科学领域的数据所需的成本却又较高，这是一个在实践中非常常见的问题。LLM为解决这一问题带来新的方法，为生成标注数据提供了可能(Møller et al. (2024))。过去基于LLM的数据增强方法主要聚焦于解决文本分类(Dai et al. (2023); Piedboeuf and Langlais (2023); Tang et al. (2023))，与之不同的是，关系抽取通常包含更加丰富的预定义标签信息(Xu et al. (2023c))，以及更广的分类空间，并且科学领域数据集上的关系抽取则是更具有挑战性。此外，这些工作都致力于few-shot设置下的RE，数据增强只需要产生少量的样本就能让模型的性能得到提升。与之相反的是，通常的全样本模型在实践中需要的数据的规模和质量要求远远超过few-shot设置。

为此，我们从数据方面入手，将LLM难以理解的科学领域的专有词汇作为LLM的输入信息，弥补其在这一领域的知识，使其合成额外的域内数据样本。这项工作提出了名为PGA(Paraphrasing and Generating Augmentation)用于加强科学领域的关系抽取的框架，框架利用大语言模型GPT-3.5通过转述和生成两种方法来合成伪数据以加强关系抽取模型的性能，其贡献如下。

1. 这项工作的研究表明，LLM可以通过我们设计的简单而有效的提示，能够在无需人工标注的情况下生成较优的、带有标注的科学领域的关系抽取伪样本。

2. 本文提出的PGA框架通过转述和生成的方式获得的伪样本分别能够让几个主流关系抽取模型的F1分数得到不同程度的提高。

2 相关工作

关系抽取 关系提取旨在识别文本中提到的实体对之间的关系。主流的方法可分为流水线法和联合法。(1)流水线方法依次处理两个子任务，命名实体识别(Named Entity Recognition, NER)和关系提取，但会存在错误传播的问题(Ye et al. (2022); Zhong and Chen (2021);Miwa and Bansal (2016))。(2)联合方法可以通过同时处理两个子任务来缓解此问题。Yan et al. (2021); Wang et al. (2021); Wang and Lu (2020)将两个任务都作为表项的标记来处理。Sun et al. (2019); Fu et al. (2019); Nguyen et al. (2021)则利用图卷积网络(Graph Convolution Networks, GCN)从实例依赖图入手，设计模型进行联合关系抽取。Yan et al. (2023)利用超图神经网络(Hypergraph Graph Neural Networks, HGNN)来进行高阶建模，并以Ye et al. (2022)的流水线模型为基础，设计联合的关系抽取模型。Kong and Xia (2023)提出了联合抽取的协同关注网络，共同关注模块捕捉两个子任务之间的双向互动，利用实体信息进行关系预测。Chen et al. (2022); Zhang et al. (2022)使用位置感知注意力设计联合提取模型。Zaratiana et al. (2024)基于文本跨度的生成线性化图，在跨度和关系类型的动态词汇表上采用了带有指向机制的转换器编码器-解码器架构。此外，Ren et al. (2023)使用协方差优化方法将特征的协方差最小化，增强了各流水线和联合模型的特征的表征能力。

在科学领域关系抽取方面，Luan et al. (2018)提出了一个具有共享跨度表示的框架，减少了任务之间的级联错误。Jiang et al. (2020)利用句法规则作为远距离监督的一种形式来连接相关的科学术语对，训练分类器来进一步识别每对术语对的关系类型。Santosh et al. (2021)用词性标记(POS tags)的嵌入对transformer生成标记的上下文嵌入进行丰富，改进了科学关系抽取。

LLM改善NLP任务 最近，LLM在信息抽取任务的各方面指标上都展现出了可观的性能，这也为NLP任务的进步注入了新的活力(Li et al. (2023a); Zhu et al. (2024))。过去的工作除了研究ChatGPT的信息抽取性能之外，还对LLM独有的评价标准进行了研究，以求更准确地反映其性能(Han et al. (2023); Wadhwa et al. (2023))。此外，Wei et al. (2023)为关系抽取任务提出一个两阶段的ChatGPT框架，性能与过去的微调模型而言较有竞争力。Yuan et al. (2023)研究了ChatGPT在zero-shot时序关系提取方面的能力，发现其性能与有监督的方法有很大的差距。Gao et al. (2023b)同样证明了ChatGPT用于事件提取还有一定挑战。在医学领域方面，Agrawal et al. (2022)研究的临床提取任务中，GPT-3的表现超过了现有的zero-shot和few-shot的基线。Jimenez Gutierrez et al. (2022)研究了相较于微调较小的PLM，GPT-3 ICL方法在两个生物医学信息提取任务上性能明显不足。

在关系抽取方面，Ma et al. (2023a)提出了一种使用LLM进行zero-shot关系提取方法，利用特定任务和知识诱导LLM生成证据，并将其纳入CoT提示以进行关系提取。Wan et al. (2023)通过在演示检索中加入任务感知表征，以及使用金标签诱导推理逻辑来丰富演示，从而进行few-shot关系抽取。Zhang et al. (2023)使用LLM将关系抽取与问题解答(QA)结合起来，以解决zero-shot关系抽取。Li et al. (2023b)利用LLM递归的输入提示，将zero-shot关系抽取输入转换为有效的QA格式。Tao et al. (2024)利用LLM和ICL通过基于示例的详细推理来显著增强few-shot关系提取，图形推理方法将关系提取分解为连续的子任务，提高了处理复杂关系数据的精度和适应性。Li et al. (2024)将LLM与检索语料协同作用，从而实现相关检索和可靠的上下文推理。

另外，模型协同(Model Collaboration)也能提高特定任务的性能。综合LLM和基于微调的传统模型共同解决关系抽取。Ma et al. (2023b)提出了一种自适应“过滤-重排”(Filter-then-Rerank)范式，LLM对小语言模型(Small Language Model, SLM)识别出的一小部分困难样本进行重排，以提升few-shot关系抽取性能。Xu et al. (2023b)提出了超级上下文学习(SuperICL)，允许LLM与局部微调的SML一起工作，提高性能，同时解决ICL的不稳定问题。Tang et al. (2024)采用“训练-指导-预测”(Training-Guide-Predict)策略，预训练语言模型(Pre-trained Language Model, PLM)充当导师，将任务知识传输给LLM，并指导LLM进行关系抽取。

数据增强 RE任务下的数据增强是在不改变原始数据标签的设置下，增加数据集的规模和多样性，帮助模型更好地学习实体以及关系标签的表达方式。合成新的优质数据能改善和加强RE模型的训练，提高模型的泛化能力。常见的文本数据增强方法包括同义词替换(Hsu et al. (2021); Wei and Zou (2019))、随机插入、随机交换、随机删除(Belinkov and Bisk (2018))和回译(Sennrich et al. (2016); Hayashi et al. (2018))等。这些方法对原始文本进行变换，得到新的

第二十三届中国计算语言学大会论文集，第352页-第369页，太原，中国，2024年7月25日至28日。
卷1：主会论文

句子。

转述式数据增强(Paraphrasing-based Data Augmentation)方法是对原始文本进行重写或重新表述，以生成具有相同意义但表达形式不同的新文本，转述技术被广泛应用于各NLP任务(Kumar et al. (2019); Okur et al. (2022))，包括关系抽取(Yu et al. (2020))。由于LLM的兴起，转述技术也因此得到蓬勃发展，Piedboeuf and Langlais (2023)研究了ChatGPT在文本分类任务上的转述数据的性能。

生成式数据增强(Generating-based Data Augmentation)方法要求生成的数据需要尽可能接近原数据集的数据分布，并且保持与原始数据相似的信息。Meng et al. (2022); Gao et al. (2023a)聚焦于利用预训练语言模型(BERT)或生成式预训练语言模型(GPT-2)进行生成数据，但受数据多样化和需不能完全脱离原数据集分布的双重限制，它们的模型设计和微调往往具有挑战性(Pouran Ben Veyseh et al. (2023))。然而，基于GPT3.5系列的方法不再需要微调，更接近于一种端到端的方式进行生成数据。特别是在低资源领域(Xu et al. (2023a))，GPT3.5经过海量语料库训练，拥有广阔的数据储备，这些先验知识极大促进了生成数据的准确性，以提高文本分类任务(Dai et al. (2023); Piedboeuf and Langlais (2023); Tang et al. (2023))和关系抽取任务的性能(Yoo et al. (2021); Xu et al. (2023c))。

3 方法

3.1 任务定义

关系抽取(Relation Extraction, RE) 关系抽取任务的目标是根据给定的句子，抽取句子中（如果存在）提到的若干成对实体之间的关系。通常关系抽取数据集由句子语料库、实体对和关系类型构成。一个关系抽取样本示例中包含一个由多个单词w组成的句子$S = \{w_l\}_{l=1}^{L}$，以及S所提及的成对的头实体e_s和尾实体e_o，其中$e_s, e_o \in \mathbb{E}$和关系$r \in \mathbb{R}$，其中$\mathbb{E}$和$\mathbb{R}$分别代表一组预定义的实体类型和关系类型。具体来说，关系抽取任务是指给定一个句子S，模型需要从$\mathbb{R}$中，预测出S中包含的e_s 和e_o之间的对应关系（如果存在），得到关系抽取三元组(e_s, r, e_o)。即模型需要学习每个训练样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$，来预测其对应的关系标签$r_i$：

$$p_{LM}\left(r|S, e_s, e_o\right) = \prod_{i=1}^{N} p\left(r_i|S, e_{s_i}, e_{o_i}\right) \tag{1}$$

模型预测可能的关系集合的概率分布，并根据每个训练样本$\mathcal{D}$来预测概率最高的关系：

$$r^* = \arg\max_{r \in \mathbb{R}} \prod_{i=1}^{N} p\left(r_i|S, e_{s_i}, e_{o_i}\right) \tag{2}$$

跨度级的实体和关系抽取(Entity and Relation Extraction, ERE) 本工作旨在解决基于跨度的实体和关系抽取任务，即上述RE任务中给定的所有实体信息由模型中的NER模块预测获得。更具体地说，基于跨度指的是形式上，给定的句子S由多个分词(token)构成：$S = \{t_i\}_{i=1}^{n}$。跨度定义为连续的token序列，跨度有开始和结束的索引代表跨度的边界，定义$Span = \{x_1, x_2, x_3, ..., x_m\}$为所有可能的跨度，实体定义为标有实体类型的跨度，关系定义为标有关系类型的实体跨度对。对NER任务来说，模型需要预测每个跨度的实体类型，即映射每个跨度到实体类型中：$y_e(x_i) \in \mathbb{E}$代表跨度是实体，并具有相应实体类型，而$y_e(x_i) \notin \mathbb{E}$代表跨度不是实体，即模型NER部分输出为$Y_e = \{(x_i, e) : x_i \in Span, e \in \mathbb{E}\}$。对RE任务来说，模型需要对每一对实体跨度$x_s, x_o \in Span$进行预测其关系类型，即映射每对实体跨度到关系类型中：$y_r(x_s, x_o) \in \mathbb{R}$代表代表这对实体跨度存在关系，并具有相应关系类型，$y_r(x_s, x_o) \notin \mathbb{R}$代表它们没有关系，即模型RE部分输出为$Y_r = \{(x_s, x_o, r) : s_s, s_o \in Span, r \in \mathbb{R}\}$。

3.2 方法概述

如图1所示，这项工作提出了一个科学领域关系抽取的数据增强框架PGA(Paraphrasing-based and Generating-based Augmentation)，通过生成和转述的方式来指导LLM生成伪样本来提升关系抽取模型的性能。更具体地来说，首先针对LLM易理解的样本数据格式，分别设计用于转述和生成方式的数据增强地提示语，提示由基础的3部分构成：Task Formulation、Demonstration Instruct、Sample Input。需要注意的是，生成式数据增强方法还额外包

含Label Interpretation。通过这些精心设计的提示，指导LLM生成大量带有标签信息的伪样本。接着，经过数据后处理步骤，过滤错误样本，以控制样本质量，并将伪样本变换成原有的数据格式，利用这些额外的伪样本与原有的训练数据集或验证集共同对各骨干RE模型进行微调，让模型提升性能。

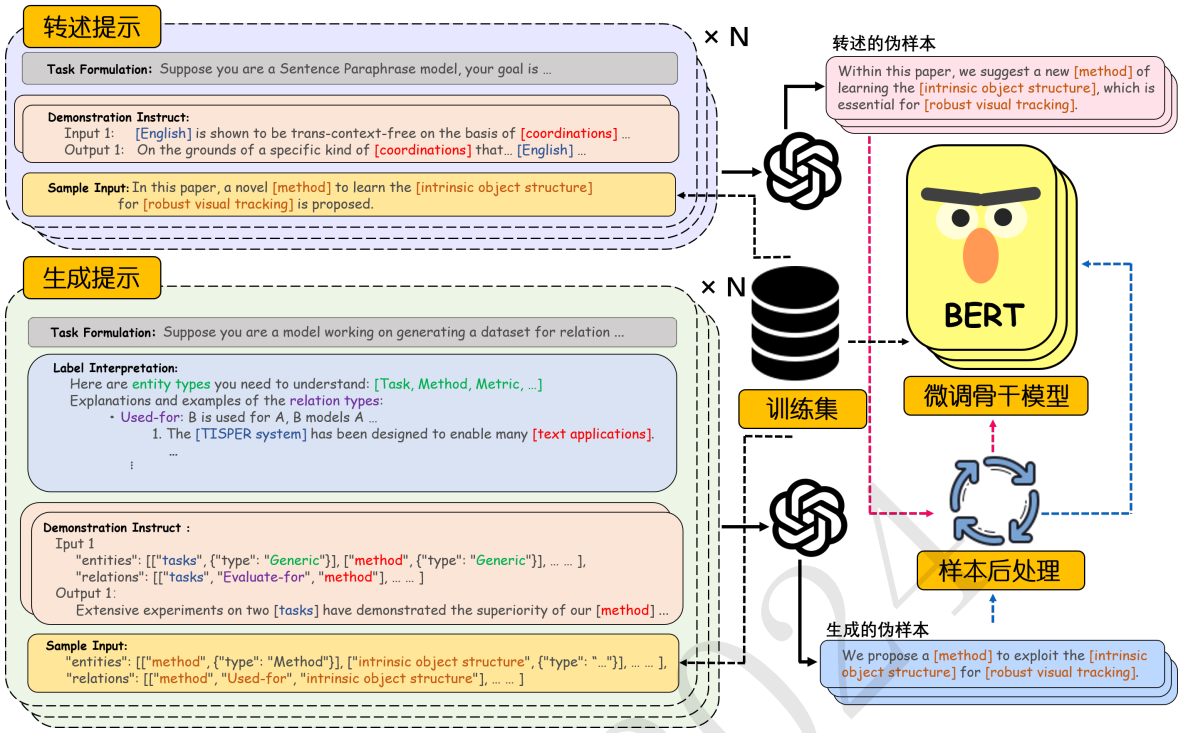

Figure 1: PGA的总体框架。图1左部分是框架的转述式和生成式数据增强方式的提示构成。分别将含有原训练样本的两种提示迭代输入至GPT-3.5中，合成两种伪样本，这些伪样本经过数据后处理，过滤之后转换成各模型需要的格式，连同原有训练样本一起微调各骨干关系抽取模型。

3.3 转述(Paraphrase)的数据增强方式

对于每个原训练集的样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$，其中，文本$S$的句意中隐含着关系语义，将$S$和实体跨度的边界信息$\mathcal{E}' = \{e'_i\}_{i=1}^{N}$（不包含实体类型标签）输入至LLM。通过句子转述，在保持原有实体不变的条件下，指导LLM产生一个隐含同样的关系语义但同时与原句子表述、形式不同的新文本S_P，最后将其与原有的实体信息连接，得到一个转述产生的伪样本$\mathcal{D}_\mathcal{P} = \{(S_P, e_{s_i}, e_{o_i})\}_{i=1}^{N}$。获得的伪样本$\mathcal{D}_\mathcal{P}$能让RE模型从中学习到比原有样本更多的信息包括同一种关系类型的更多表述和呈现方式，增大正确预测对应关系类型的概率。通过将文本的转述过程视为对字符串y的文本序列生成任务，其过程可以写成如下：

$$p_{LLM}\left(S_P|S, e'_s, e'_o\right) = \prod_{l=1}^{L} p\left(y_l|S, e'_s, e'_o, y_{<l}\right) \tag{3}$$

即使用LLM来生成长度为L的字符串y（转述的句子）的概率。

首先，为每个给定的训练集样本构建转述的数据增强方式的提示，并将其输入至GPT-3.5。如图1左上部分所示，每个转述提示具体由以下3部分组成，完整的转述提示详见附录。

Task Formulation：首先命令LLM进行指定的样本转述模型的角色扮演，明确其任务下的身份背景。再提供转述关系抽取样本的任务描述、详细的注意事项，防止LLM转述出不符合规定的样本、增加或者删减实体信息，使LLM进一步明确任务的要求。

Demonstration Instruct：为LLM更好学习指定的数据格式和理解转述伪样本的任务，这部分提示为LLM提供了任务中原样本的输入示例和伪样本的输出示例。实践中发现，LLM由

于预训练阶段所学习的先验知识大多是易于人类阅读的形式的文本语料库，其思维方式也更接近于人类的思维方式，所以输入的数据样本中如果出现过多的结构化信息，则会影响LLM对句子的理解。因为基于模型属于跨度级别的RE方法，导致样本全部是JSON格式的数据，嵌套的各类括号、引号，冗长复杂且无用的信息过多，会使LLM将部分样本信息丢失。因此，本文在转述方法的提示中的输入和输出样本格式中，并非使用了数据集原本的易于程序所理解的JSON格式，而是对数据集进行简单改造。将token还原为完整的句子，用方括号将句子中对应实体包裹起来，将实体融合于句中，并不携带实体类型标签($\mathbb{E}$)，仅仅利用实体跨度的边界信息($\mathcal{E}' = \{e'_i\}_{i=1}^{N}$)。这样既能缩短提示中重复的信息，也能便于LLM理解和学习。同时受提示长度限制，我们随机地从原训练集中选择两条样本作为演示，每个示例的输入和输出格式遵循前面提到的数据格式。

Sample Input：这部分提示包含一条来源于原训练数据集的样本的的句子和实体跨度边界信息，其作为转述式数据增强任务的输入。每条数据样本需要参照实例中的数据格式，迭代插入转述提示中，后重复输入至LLM中。

3.4 生成(Generate)的数据增强方式

将每个原训练集样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$中的句子$S$对应的实体信息$\mathcal{E} = \{e_i\}_{i=1}^{N}$，加上最后模型需预测的结果，即关系标签$\mathcal{R} = \{r_i\}_{i=1}^{N}$，一起输入至LLM。通过LLM强大的文本序列生成能力，得到蕴含全部实体以及关系语义的句子S_G。如上文所述，关系抽取的目的是通过模型学习每个训练样本$\mathcal{D}$，预测样本句子中隐含的的关系类型$\mathcal{R}$。而生成的数据增强方式，可以视为这个任务的部分逆序过程——即LLM根据提供的答案标签来生成可学习的样本。对于每个样本，通过其所有已知的标签信息$\mathcal{M} = \{e_{s_i}, e_{o_i}, r_i\}_{i=1}^{N}$，以生成新的句子$S_G$，最后将它与原有的实体信息连接，得到一个生成方式的伪样本$\mathcal{D}_G = \{(S_G, e_{s_i}, e_{o_i})\}_{i=1}^{N}$。这些样本中的生成句子呈现出与原句子不同的描述和表达，暗含完整的标签信息，使模型学习到更多额外的域内标注样本。同样，通过将生成方式的数据增强方法视为对字符串y的文本序列生成任务，其过程可以写成如下：

$$p_{LLM}\left(S_G | e_s, e_o, r\right) = \prod_{l=1}^{L} p\left(y_l | e_s, e_o, r, y_{<l}\right) \tag{4}$$

如图1左下部分所示，框架为每个给定的训练集样本构建生成的数据增强方式的提示，提示具体由以下4部分组成，完整的生成提示详见附录。

Task Formulation：这部分提示指示LLM扮演RE样本生成模型，规定生成样本的任务定义和任务目标、强调输出的注意事项，防止LLM生成错误样本、改变原实体信息，并且阐述输入文本的格式以及方括号含义，使LLM进一步明确输入输出要求。规定输出文本的体裁和风格，使其更加贴合原数据集句子。由于生成方式输入的实体信息有可能为空，本文规定了没有标签信息的句子经过生成方式获得的伪样本是*"No result can be generated with the given information."*，其包含的实体信息和关系信息为空。

Demonstration Instruct：一定数量的输入和输出示例能够帮助LLM理解生成样本的任务，这部分提示为LLM提供了两条随机的样本输入和它们对应的输出。在生成方式的提示中，输出同样采用了方括号包裹句子中的实体的格式。但不同于转述方式的是，生成方式的输入使用格式仍是采用JSON格式，这是因为生成方式的输入不是句子和实体信息，而是完整实体标签信息和关系标签信息。因为输入不包含句子，所以无法使用前面所提出的方括号包裹实体的格式，同时由于JSON格式作为输入的弊端，因此生成方式合成的伪样本错误率会更高。

Label Interpretation：数据增强需要确保生成的样本仍然保持实体和关系的标注准确性，保持生成样本的质量和与原始数据尽可能地接近。但同时也要避免生成过于简单和重复的文本，以确保模型能够从中学习到新的信息和模式。由于生成式数据增强任务的特殊性，提示中必须要为LLM提供所有实体标签集合、关系类型标签信息及诠释、数条关系标签的案例以约束LLM，避免其生成的文本的含义与对应的标签信息不符，影响伪样本质量。其中，对应的关系标签描述来源于Luan et al. (2018)。

Sample Input：同样，将原训练数据集里的每条样本对应的所有实体和关系信息迭代插入此部分提示中,作为生成方式数据增强任务的输入。

3.5 数据后处理

得到LLM输出的句子后，为了保持格式统一，并且减少不必要的API的token消耗，其输出格式同样使用易于阅读的方括号格式。在数据后处理步骤，将得到的句子与其对应的原数据集实体和关系标签信息合并为完整的伪样本，同时与其对应的原样本信息进行比对，识别并且过滤掉错误的伪样本，确保RE模型输入的数据是准确的。最后，将经过筛选的伪样本还原成各RE模型能读取的格式，确保数据的准确性和一致性，保证模型在训练中能够取得更好的效果。

4 实验

4.1 数据集

本文在一个科学领域关系抽取数据集评估PGA框架。SciERC(Luan et al. (2018))数据集收集了500篇人工智能论文摘要，为其标注了关系和实体信息。我们合成了大量转述和生成的伪样本，以便模型在测试集上评估性能。其合成的两个数据集统计数据如表1所示。

Dataset	#Train	#Dev	#Test	#Ents(Types)	#Rels(Types)	#Defect Rate
SciERC	1,861	275	551	8,089(6)	4,716(7)	-
$SciERC_P$	1,861	-	-	5,598(6)	3,219(7)	21.60%
$SciERC_G$	1,589	-	-	4,341(6)	2,402(7)	14.61%

Table 1: 原数据集、转述($SciERC_P$)和生成($SciERC_G$)数据集的统计数据

4.2 评估指标

对于命名实体识别任务，实体跨度的边界和实体类型需要模型同时正确预测。对于关系抽取任务，本文分别报告了两个评价指标：(1)**边界评估(Rel)**要求模型正确预测每个主体实体和客体实体对的跨度边界和它们之间的关系类型；(2)**严格评估(Rel+)**在Rel的设置下，进一步要求模型对实体的类型也要预测正确(Taillé et al. (2020))。

4.3 基线和骨干模型

为了对PGA框架进行评估，本文分别选择以下先进的关系抽取模型作为PGA框架的骨干模型：(1)**SpERT**(Eberts and Ulges (2019))是主流联合实体关系抽取模型，列举文本中所有可能的片段跨度，编码生成跨度的向量表示，并进行跨度的实体标签分类预测。对提取的所有实体跨度两两配对，通过一个多分类模块判断实体对之间的关系类型。(2)**PL-Marker**(Ye et al. (2022))属于流水线模型，为命名实体识别模块提出了面向邻域打包策略，综合考虑了邻域跨度，以更精确地建模实体边界信息。为关系抽取设计了主客实体打包策略，使用悬浮标记和固定标记，让不同客体实体可以共享同一主体实体。(3)**PURE**(Zhong and Chen (2021))是端到端的流水线模型，命名实体识别模块和关系抽取模块独立训练，它们共享相同的预训练编码器。并为关系抽取任务提出悬浮标记，独立处理每对主客实体。还可以选择将其统一放在样本末尾加快训练速度(PURE-Approx)(Zhong and Chen (2021))。此外，本文额外加入了6个主流的关系抽取模型作为基线进行比较。

4.4 实施细节

LLM方面，由于截止于进行实验的日期时，OpenAI最先进的基于GPT-3.5长文本处理模型是*text-davinci-003*，同时实践中表明*gpt-3.5-turbo*更适用于聊天对话领域而并非长文本生成领域，所以本文选择*text-davinci-003*作为PGA框架的基干LLM。

在转述式数据增强中，LLM温度设置为0.5，而生成方式里，由于需要更好的样本的多样性，温度被设置为1。本文将原训练集的所有样本进行了转述和生成，在实验中出现了LLM输出错误的样本的情况。因为由转述方式得到伪样本句子拥有对应原样本句子作为清晰的答案对照，所以在数据后处理步骤中，错误的转述样本被过滤筛选出，然后将对应转述错误的样本利用LLM重新合成，直到获得正确的样本为止。而由生成方式合成的伪样本句子对应的原样本输入是标签信息，不便对照，所以错误样本直接被丢弃。两种方式合成伪样本的瑕疵率详

见表1。在将伪样本句子还原成相应模型读取的token格式时，对于那些模型输入需要的但不参与NER和RE任务的无关紧要的样本属性信息（如样本中句子的来源，即ID），我们使用程序统一生成伪属性将其替代。

骨干模型方面，对于SpERT(Eberts and Ulges (2019))，按照原工作，使用*scibert-scivocab-cased*(Beltagy et al. (2019))作为预训练编码器，并遵照设置，使用SciERC的训练集和验证集训练模型；对于PL-Marker(Ye et al. (2022))和PURE(Zhong and Chen (2021))，按照原工作使用的*scibert-scivocab-uncased*(Beltagy et al. (2019))作为预训练编码器，使用SciERC的训练集用以训练模型。并为主要实验和联合伪样本实验中框架下的骨干模型优化模型的超参数，其它实验按照对应原工作的模型参数进行实验，不再优化超参数。

4.5 主要结果

将PGA框架主要实验结果与各基线方法和骨干模型性能进行了比较，结果如表2所示。在PGA的两种数据增强方法中，PGA_P代表转述方式的数据增强，PGA_G代表生成方式的数据增强。从表中可以得知，数据增强对关系抽取来说是一种简单而有效的方法，它可以借用LLM的先验知识来合成伪样本以训练基于微调的模型提升性能：PGA框架基本上取得了Ent，Rel，Rel+三个指标下最高的F1分数，也优于之前的所有基线。在于骨干模型对比方面，基于SpERT(Eberts and Ulges (2019))的PGA_P相比原来，在Ent和Rel+的F1分数上取得了最佳，最高提升1.78%，而PGA_G相比原来在Rel+上取得1.53%F1分数提升，但在Rel指标下有一定下滑；基于PL-Marker(Ye et al. (2022))的PGA_P相比原来，在每个指标上都有一定提升，Rel+的F1分数最高提升2.05%，并且PGA_G同样所有指标都有提升，在Ent和Rel的指标上取得所有模型最佳，Rel和Rel+F1分数都提升了1.26%；基于PURE(Zhong and Chen (2021))的PGA_P相比原来，同样所有指标都有提升，Rel提升了1.81%，Rel+提升了1.67%，另外PURE的PGA_G所有指标也都有提升，Rel+最高提升了1.9%。

同时，我们注意到各个模型转述和生成效果不尽相同，极个别指标的甚至性能会降低。也可以发现，PGA框架在关系抽取方面有主要的提升，尤其是在严格的指标下。

Model	Encoder	Ent			Rel			Rel+		
		Precision	Recall	**F1**	Precision	Recall	**F1**	Precision	Recall	**F1**
SciIE(Luan et al. (2018))	BiLSTM	67.20	61.50	64.20	47.60	33.50	39.30	-	-	-
DyGIE(Luan et al. (2019)	ELMo	-	-	65.20	-	-	41.60	-	-	-
DyGIE++(Wadden et al. (2019))	$BERT_{Base}$	-	-	67.50	-	-	48.40	-	-	-
UniRE(Wang et al. (2021))		65.8	71.10	68.40	-	-	40.20†	37.30	36.60	36.90
TriMF(Shen et al. (2021))		70.18	70.17	70.17	52.63	52.32	52.44	-	-	-
PFN(Yan et al. (2021))		-	-	66.80	-	-	-	-	-	38.40
SpERT*(Eberts and Ulges (2019))		68.54	70.09	69.31	52.66	50.82	**51.72**	39.15	37.78	38.45
SpERT+PGA_P		69.43	69.67	**69.55**	50.72	50.51	50.62	40.31	40.14	**40.23**
SpERT+PGA_G	SciBERT	68.72	69.38	69.05	52.27	47.33	49.68	42.06	38.09	39.98
PL-Marker(Ye et al. (2022))		-	-	69.90	-	-	53.20	-	-	41.60
PL-Marker+PGA_P		71.15	68.78	69.95	-	-	53.40	-	-	**43.65**
PL-Marker+PGA_G		70.11	70.86	**70.48**	-	-	**54.46**	-	-	42.86
PURE*(Zhong and Chen (2021))		69.33	68.55	68.93	49.69	49.69	49.69	37.37	37.37	37.37
PURE+PGA_P		70.46	68.07	**69.24**	55.09	48.36	**51.50**	41.75	36.65	39.04
PURE+PGA_G		70.50	67.95	69.21	55.58	44.97	49.72	43.91	35.52	**39.27**

Table 2: PGA框架加上各个骨干模型在SciERC测试集上总体的实验结果。性能的评估指标包括查准率、召回率和F1分数。**粗体** 数据表示各骨干模型下的最高性能，带有下划线的数据表示次优的性能。请注意：带有∗符号的模型表示为尽可能获取更多的评估指标，我们根据原作者公布的参数和代码重新运行的实验结果。带有†符号表示引用(Ren et al. (2023))对原模型的复现结果。

5 分析

5.1 联合两种伪样本

联合两种方式产生的伪样本，与原训练集共同用于微调各骨干模型。将框架联合伪样本PGA_C实验结果与各骨干模型性能进行比较，结果如表3所示。在SpERT(Eberts and Ulges

(2019))上，PGA_C表现不佳，多数性能指标出现不同程度下降；在PURE(Zhong and Chen (2021))上甚至所有指标的性能有都下降；仅在PL-Marker(Ye et al. (2022))上Rel+有0.99%的提升。我们推测是否因为框架合成的伪样本的标注质量存在问题，加之由表1可知，联合起来的伪样本规模远超过原数据集的训练样本规模，其引入大量的实体却未标注、对科研数据集的专有词汇理解不足而产生大量的噪声以致于整个参与训练的数据集的质量低于原训练集的质量，这就导致了对噪声敏感的模型的性能下降。

Model	Ent			Rel			Rel+		
	Precision	Recall	**F1**	Precision	Recall	**F1**	Precision	Recall	**F1**
SpERT*(Eberts and Ulges (2019))	68.54	70.09	**69.31**	52.66	50.82	**51.72**	39.15	37.78	38.45
SpERT+PGA_C	69.40	67.72	68.55	53.00	44.46	48.35	42.72	35.83	**38.97**
PL-Marker(Ye et al. (2022))	-	-	69.90	-	-	**53.20**	-	-	41.60
PL-Marker+PGA_C	70.61	69.44	**70.02**	-	-	52.83	-	-	**42.59**
PURE*(Zhong and Chen (2021))	69.33	68.55	**68.93**	49.69	49.69	**49.69**	37.37	37.37	**37.37**
PURE+PGA_C	69.61	67.41	68.50	49.62	47.43	48.50	37.16	35.52	36.33

Table 3: 在SciERC测试集上联合两种伪样本与原数据集用以微调骨干模型的性能，**粗体** 数据表示各骨干模型下的最高性能。PGA_C代表各骨干模型额外地联合了两种数据增强方式合成的伪样本进行训练。

5.2 伪样本的质量

根据上面的猜测设计实验，将原数据集从模型的训练里剔除，仅用单独的伪样本以微调各骨干模型。将单独仅含伪样本实验结果与各骨干模型性能进行比较，结果如表4所示。总体上，分别使用单独转述和生成的伪样本进行微调的各模型的性能都出现下降。这些差距表明伪样本的质量与原样本的质量相比还有一定提升的空间。即使在最好情况下，SpERT(Eberts and Ulges (2019))在转述的伪样本单独训练下，Ent的F1分数也下降了3.06%，Rel的F1分数下降了5.88%，Rel+则下降了4.12%。然而，也许是因为PURE(Zhong and Chen (2021))对噪声更加敏感或此实验没有对其超参数进行进一步优化的缘故，在生成的伪样本单独训练中，其F1分数是所有模型中下降得最多的，Ent的F1分数降幅达到了25.56%，Rel的F1分数甚至下降了29.34%，Rel+指标则下降了22.95%。PL-Marker(Ye et al. (2022))的下降幅度则处于两者之间。

总体来说，框架的两种数据增强方式中，转述伪样本质量高于生成伪样本质量。可以发现，转述伪样本微调下的骨干模型的每一个F1分数都超过了对应的生成的伪样本微调下的骨干模型的F1分数，并且每一个指标都相差10%以上。我们考虑造成以上现象的原因主要在于：在实体方面，实体方面的信息主要依靠输入样本提供，而LLM只会在合成样本的时候引入一定量的噪声实体。然而，在关系方面，转述方法产生关系信息几乎只依靠LLM对句子整体含义理解后的关系的推理，因为没有明确的关系信息的输入，LLM需要推导出文本中隐含的关系，这对其识别和推理信息的能力有一定挑战。在产生转述样本的时候，LLM需要生成暗含实体间语义关系的句子，这要求其具备高质量的文本生成能力，以便准确地表达关系信息。此外，生成数据扩增方法产生关系信息的条件更为苛刻，仅输入两种标签信息，没有句子供其参考。这对LLM的条件约束就远远不如转述方式，所以可能在生成句子里LLM就不会引入更隐式的关系语义供模型学习，且其样本还保有一定的噪声，导致其质量不如前者。

5.3 伪样本参与微调的数量对性能的影响

分别分割两种方式产生的伪样本，把它们各自以不同数量和原训练集样本一起参与训练，确定是否伪样本数量也会对性能有一定影响，结果如图2所示。可以观察到，对于命名实体识别任务来说，随着转述伪样本的数量的提升，性能先是短暂提升后再次因为过多伪样本加入而下降。然而生成伪样本则经过初始的跌落而后缓慢上升。

另一方面，对于关系抽取任务来说，随着转述伪样本的数量的提升，Rel和Rel+指标的性能分数都在一个特定的数量下达到顶峰，而后随着更多伪样本数量加入，性能逐渐下降，此时其伪样本规模与原训练集规模接近。注意到，之所以命名实体识别和关系抽取的性能拐点不同，可能是因为SpERT(Eberts and Ulges (2019))是一个联合实体关系抽取模型，能一定程度减

Model	Ent			Rel			Rel+		
	Precision	Recall	**F1**	Precision	Recall	**F1**	Precision	Recall	**F1**
SpERT*(Eberts and Ulges (2019))	68.54	70.09	69.31	52.66	50.82	51.72	39.15	37.78	38.45
SpERT+PGA_P^{Sole}	67.24	65.28	66.25	48.67	43.33	45.84	36.45	32.44	34.33
SpERT+PGA_G^{Sole}	50.34	56.74	53.35	36.73	21.46	27.09	26.01	15.20	19.18
PL-Marker(Ye et al. (2022))	-	-	69.90	-	-	53.20	-	-	41.60
PL-Marker+PGA_P^{Sole}	67.00	59.41	62.98	-	-	44.68	-	-	34.63
PL-Marker+PGA_G^{Sole}	52.67	41.61	46.49	-	-	26.60	-	-	19.42
PURE*(Zhong and Chen (2021))	69.33	68.55	68.93	49.69	49.69	49.69	37.37	37.37	37.37
PURE+PGA_P^{Sole}	59.25	55.13	57.12	40.37	31.42	35.33	29.95	23.31	26.21
PURE+PGA_G^{Sole}	52.21	37.09	43.37	34.62	14.68	20.62	24.21	10.27	14.42

Table 4: 在SciERC测试集上仅有单独伪样本参与的骨干模型微调性能。PGA_P^{Sole}代表仅有单独转述的伪样本参与训练，PGA_G^{Sole}代表仅有单独生成的伪样本参与训练。请注意，SpERT(Eberts and Ulges (2019))原本设置中，原训练集和验证集共同参与了训练。

少误差传递的问题，并且关系抽取总体的性能趋势并未与命名实体识别的性能变化相脱离。这一点在生成伪样本的关系抽取性能上也有所体现。总体上，Rel+性能在更多的伪样本参与下性能再次下降，由此可推测出，命名实体识别和关系抽取的性能并不能随更多伪样本的加入得到持续的提升。

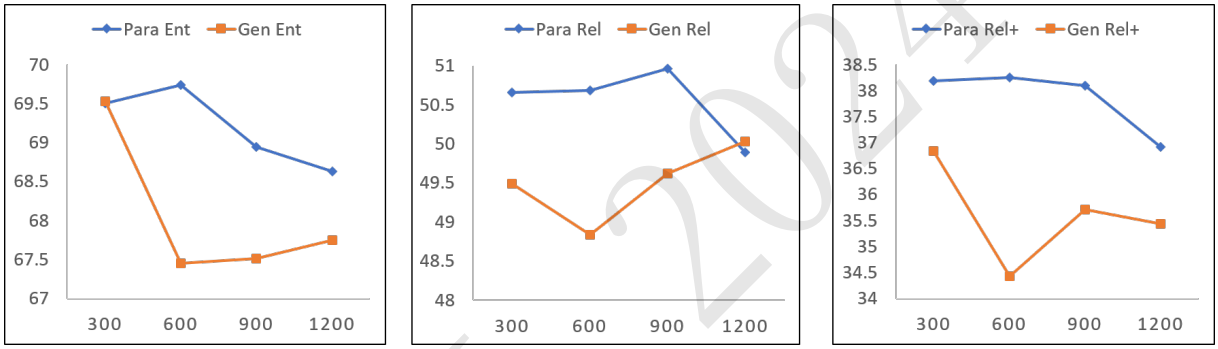

Figure 2: 分别由不同数量伪样本连同原训练集参与训练下，SpERT(Eberts and Ulges (2019))模型的性能变化，横坐标代表参与伪样本数量，纵坐标代表F1分数。

5.4 伪样本对原样本的忠实度

为了衡量上面的实验中伪样本具体是否真正接近于真实的原训练集样本，我们使用*sentence-transformers*中的*all-MiniLM-L6-v2*模型，分别对400条转述和生成产生的伪样本，连同它们各自所对应的原训练集样本中的句子进行文本嵌入，再使用t-SNE算法对向量空间中高维的句子嵌入进行降维和可视化处理，得到样本嵌入的空间分布，以观察伪样本对原样本的忠实度。由图3可知，转述的伪样本与原样本保持了相对较高的重叠程度。对于每一个原样本，包括图中边缘零星的样本都有一个非常接近的转述伪样本。然而，生成的伪样本与原样本大致上重叠和贴近，但是从贴近程度来看，其弱于前者。另外从其分布图中右上角可以发现，已经出现了伪样本分布远离对应原样本分布的情况，而转述样本分布未出现此现象。所以，总体上生成样本对原样本的忠实度不及转述样本对原样本的忠实度。

6 结论

在这项工作中，我们提出了一个名为PGA的基于大语言模型的数据增强框架，该框架能通过转述和生成两种数据增强方式合成与原数据集对应伪样本，在更具挑战性的科学领域的数据集上，我们对主流的关系抽取模型进行了数据增强，提升了模型的关系抽取的性能。在对产生的伪样本进行具体的分析实验后，本文认为，PGA转述的伪样本在保持对原样本的忠实度较高

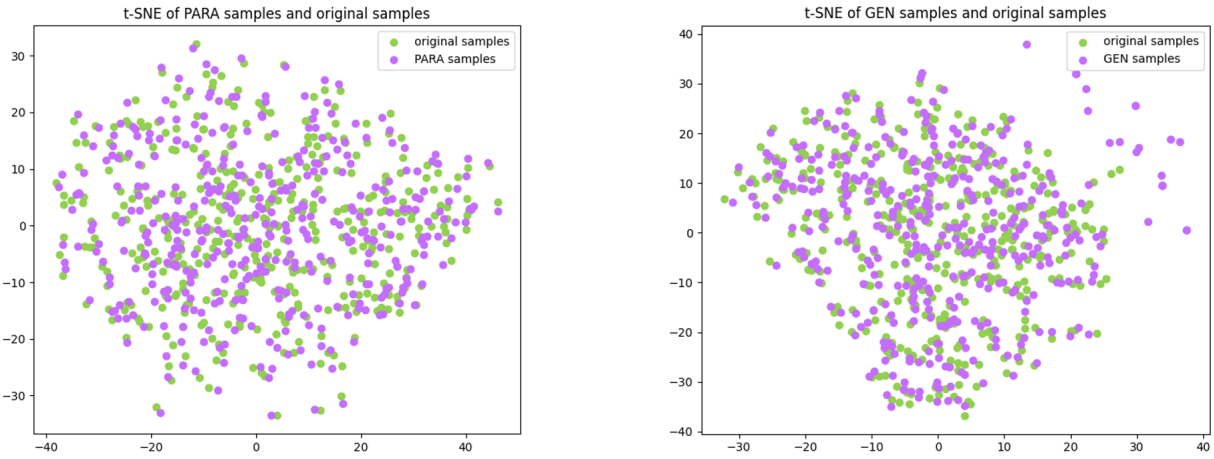

Figure 3: 伪样本与原数据集样本句子的嵌入在向量空间中的分布。

的前提下，能有效引入同种关系的更多语义信息来供模型学习。即使生成的伪样本能提升模型一部分性能，但是其质量还有一定提升的空间。在未来我们需要进一步研究是否可以通过更先进的上下文学习方法或有效的伪数据过滤策略来提升利用大语言模型进行数据增强的精确性和忠实度。

致谢

感谢复杂产品智能制造系统技术国家重点实验室开放基金课题《基于多源数据挖掘的复杂产品性能推理与评估方法研究》对本工作提供的资助。

参考文献

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Tiantian Chen, Lianke Zhou, Nianbin Wang, and Xirui Chen. 2022. Joint entity and relation extraction with position-aware attention and relation embedding. *Applied Soft Computing*, 119:108604.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023a. Self-guided noise-free data generation for efficient zero-shot learning.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023b. Exploring the feasibility of chatgpt for event extraction.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.

Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. 2018. Back-translation-style data augmentation for end-to-end asr. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 426–433.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Ming Jiang, Jennifer D'Souza, Sören Auer, and J Stephen Downie. 2020. Targeting precision: A hybrid scientific relation extraction pipeline for improved scholarly knowledge organization. *Proceedings of the Association for Information Science and Technology*, 57(1):e303.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Wenjun Kong and Yamei Xia. 2023. Care: Co-attention network for joint entity and relation extraction. *arXiv preprint arXiv:2308.12531*.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore, December. Association for Computational Linguistics.

Guozheng Li, Peng Wang, Wenjun Ke, Yikai Guo, Ke Ji, Ziyu Shang, Jiajun Liu, and Zijie Xu. 2024. Recall, retrieve and reason: Towards better in-context relation extraction.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore, December. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38, Online, June. Association for Computational Linguistics.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore, December. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2023. Generating labeled data for relation extraction: A meta learning approach with joint GPT-2 training. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11466–11478, Toronto, Canada, July. Association for Computational Linguistics.

Lin Ren, Yongbin Liu, Yixin Cao, and Chunping Ouyang. 2023. CoVariance-based causal debiasing for entity and relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2627–2640, Singapore, December. Association for Computational Linguistics.

Tokala Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Sanyal, and Parthapratim Das. 2021. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. 09.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A trigger-sense memory flow framework for joint entity and relation extraction.

Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. 2019. Joint type inference on entities and relations via graph convolutional networks. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, Florence, Italy, July. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online, November. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining?

Xuemei Tang, Jun Wang, and Qi Su. 2024. Small language model is a good guide for large language model in chinese entity relation extraction.

Yicheng Tao, Yiqun Wang, and Longju Bai. 2024. Graphical reasoning: Llm-based semi-open relation extraction.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore, December. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November. Association for Computational Linguistics.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A unified label space for entity relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online, August. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023a. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada, July. Association for Computational Linguistics.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023b. Small models are valuable plug-ins for large language models.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023c. How to unleash the power of large language models for few-shot relation extraction?

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zhaohui Yan, Songlin Yang, Wei Liu, and Kewei Tu. 2023. Joint entity and relation extraction with span pruning and hypergraph neural networks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7512–7526, Singapore, December. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Junjie Yu, Tong Zhu, Wenliang Chen, Wei Zhang, and Min Zhang. 2020. Improving relation extraction with relational paraphrase sentences. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1687–1698, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada, July. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. An autoregressive text-to-graph framework for joint entity and relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19477–19487.

Chenglong Zhang, Shuyong Gao, Haofen Wang, and Wenqiang Zhang. 2022. Position-aware joint entity and relation extraction with attention mechanism. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4496–4502. International Joint Conferences on Artificial Intelligence Organization, 7. Main Track.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada, July. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

附录

Paraphrase Prompt	
Task Formulation	Suppose you are a sentence paraphrase model. Your goal is to follow a given sentence and the entities in the sentence, describe the sentence in different words than the original sentence while keeping the entities in the sentence and the meaning of the sentence intact, then output it. **Note:** In the sentence to be processed, the part enclosed in parentheses is the entity; do not add, delete, or change the entity, or output any redundant information. Please fully understand the meaning of the sentence before paraphrasing.
Demonstration Instruct	Please refer to the following examples for the format of the output, and please study the following examples to accomplish the above tasks more accurately. **Iput1:** '[English] is shown to be trans-context-free on the basis of [coordinations] of the respectively type that involve [strictly syntactic cross-serial agreement].' **Output1:** On the grounds of a specific kind of [coordinations] that require [strictly syntactic cross-serial agreement], [English] is demonstrated to be trans-context-free. **Iput2:** 'The [agreement] in question involves number in [nouns] and [reflexive pronouns] and is syntactic rather than semantic in nature because [grammatical number] in [English], like [grammatical gender] in [languages] such as [French] , is partly arbitrary.' **Output2:** The number in [nouns] and [reflexive pronouns] that must match in this [agreement] is based on syntax rather than meaning, since [English] [grammatical number], like [grammatical gender] in [languages] like [French], is partly random.
Sample Input	**Sentence:** In this paper we show how two standard [outputs] from [information extraction (IE) systems]-[named entity annotations] and [scenario templates]-can be used to enhance access to [text collections] via a standard [text browser].

Table 5: PGA转述的数据增强方法的详细提示

Generate Prompt	
Task Formulation	Suppose you are a model working on generating a dataset for relation extraction. One data sample of a relation extraction dataset is a sentence, and its corresponding entity information and relation information. Entities are embedded in the sentence, entities definitely have entity types, and certain pairs of entities will also contain corresponding relation types based on semantics. **Your goal:** to generate sentences that contain information about entities and relations, given that information.
Label Interpretation	Here is the information you need to understand: 6 entity types: **[Task, Method, Metric, Material, Generic, OtherScientificTerm]** Below are explanations and examples of the relation types, with pairs of entities where a relation exists enclosed in parentheses. • **Used-for:** B is used for A, B models A, A is trained on B, B exploits A, A is based on B.E.g. 1. The [TISPER system] has been designed to enable many [text applications]. 2. Our [method] models [user proficiency]. 3. Our [algorithms] exploits [local soothness]. • **Feature-of:** B belongs to A, B is a feature of A, B is under A domain. E.g. 1. [prior knowledge] of the [model] 2. [genre-specific regularities] of [discourse structure] 3. [English text] in [science domain] • **Hyponym-of:** B is a hyponym of A, B is a type of A.E.g. 1. [TUIT] is a [software library] 2. [NLP applications] such as [machine translation] and [language generation] • **Part-of:** B is a part of A... E.g. 1. The [system] includes two models: [speech recognition] and [natural language understanding] 2. We incorporate [NLU module] to the [system]. • **Evaluate-for:** B is evaluated for A, A use B to get estimation, E.g. 1. [Intra-sentential quality] is evaluated with [rule-based heuristics] 2. We describe a new [system] that enhances Criterion's capability , by evaluating multiple aspects of [coherence in essays]. • **Compare:** Opposite of conjunction, compare two models/methods, or listing two opposing entities. E.g. 1. Unlike the [quantitative prior], the [qualitative prior] is often ignored... 2. We compare our [system] with previous [sequential tagging systems]... • **Conjunction:** Function as similar role or use/incorporate with. E.g. 1. obtained from [human expert] or [knowledge base] 2. NLP applications such as [machine translation] and [language generation]

续表

Generate Prompt	
Demonstration Instruct	**Caution:** Do not leave out any entities, if the 'entities' is empty, output 'No result can be generated with the given information.', and please try to avoid generating entities in the output sentence that are not provided by the input, such as 'Problem', 'Task', 'Method' and so on. Do not leave out any relation or add relation between pairs of entities that don't have given relation. The genre of each sample sentence generated is the abstract of an AI paper. Please study the following examples and refer to them for the output format (entities enclosed in square brackets in the output sentences) **Input1:{'entities':** [['agreement', {'type': 'Generic'}], ['nouns',{'type': 'OtherScientificTerm'}], ['reflexive pronouns', {'type': 'OtherScientificTerm'}], ['grammatical number', {'type': 'OtherScientificTerm'}], ['English', {'type': 'Material'}], ['grammatical gender', {'type': 'OtherScientificTerm'}], ['languages', {'type': 'Material'}], ['French', {'type': 'Material'}]], **'relations':** [['nouns', 'Conjunction', 'reflexive pronouns'], ['grammatical gender', 'Feature-of', 'languages'], ['French', 'Hyponym-of', 'languages']]} **Output1**: 'The [agreement] in question involves number in [nouns] and [reflexive pronouns] and is syntactic rather than semantic in nature because [grammatical number] in [English], like [grammatical gender] in [languages] such as [French] , is partly arbitrary.' **Input2:{'entities':** [['tasks', {'type': 'Generic'}], ['method', {'type': 'Generic'}], ['state-of-the-art methods', {'type': 'Generic'}]], **'relations':** [['tasks', 'Evaluate-for', 'method'], ['method', 'Compare', 'state-of-the-art methods']]} **Output2**: 'Extensive experiments on two [tasks] have demonstrated the superiority of our [method] over the [state-of-the-art methods].'
Sample Input	Now please generate the output based on my input. **Input: {'entities':** [['method', {'type': 'Method'}], ['intrinsic object structure', {'type': 'OtherScientificTerm'}], ['robust visual tracking', {'type': 'Task'}]], **'relations':** [['method', 'Used-for', 'intrinsic object structure'], ['intrinsic object structure', 'Used-for', 'robust visual tracking']]}

Table 6: PGA生成的数据增强方法的详细提示

第二十三届中国计算语言学大会论文集，第352页-第369页，太原，中国，2024年7月25日至28日。
卷1：主会论文