如何释放大语言模型在少样本关系抽取中的潜力?

Xin Xu、Yuqi Zhu、Xiaohan Wang、Ningyu Zhang*浙江大学 & AZFT知识引擎联合实验室 {xxucs, wangxh07, zhangningyu}@zju.edu.cn

arXiv:2305.01555v4 [cs.CL] 9 Jun 2023

摘要

语言模型规模的扩展推动了众多自然语言处理任务的变革，但目前对于利用大语言模型进行少样本关系抽取的研究仍缺乏全面探索。本文通过大量实验，研究了利用GPT-3.5 进行少样本关系抽取的两种主要方法：上下文学习和数据生成。为提升少样本性能，我们进一步提出了与任务相关的指令和模式约束数据生成方法。我们观察到，上下文学习能够取得与先前提示学习方法相当的性能，而利用大语言模型进行数据生成可以进一步提升已有方法，在四个广泛研究的关系抽取数据集上获得新的少样本关系抽取最佳结果。我们希望本工作能够激发未来对大语言模型在少样本关系抽取方面能力的研究[1]。

1 引言

少样本关系抽取（RE）受到许多自然语言处理（NLP）研究者的关注，因为它能够在仅提供少量支持示例的情况下抽取文本信息（Han 等，2018年；Yang 等，2021年；Han 等，2021a；Brody 等，2021年；Ma 等，2023年）。以往的大多数工作都侧重于使用相对较小的语言模型进行微调（Soares 等，2019年；Ye 等，2022年）或提示微调（Chen 等，2022；Han 等，2021b），例如RoBERTa（Liu 等，2019年）。近年来，随着模型规模和语料库规模的扩大，大语言模型（LLMs）如 ChatGPT（OpenAI，2022年）和 GPT-4（OpenAI，2023a）仅通过展示少量实例便展现出强大的能力，即上下文学习（Dong 等，2023年）。尽管大语言模型在许多自然语言处理任务中取得了显著成果，但它们的 po-少样本关系抽取中的潜力尚未得到充分探索。

本文以 GPT-3.5（OpenAI，2023b）为例，探究如何通过上下文学习和数据生成，最大限度地利用大语言模型完成少样本关系抽取任务。与文本分类不同，关系抽取任务包含丰富的预定义模式（例如实体和关系类型约束），且分类空间相对较大、较为复杂，并存在噪声数据。我们还设计了两种简单而有效的策略，以更好地释放大语言模型的潜力：**与任务相关的指令**和**受模式约束的数据生成**。我们在四个知名关系抽取数据集上进行了全面实验。实证结果表明，大语言模型有望在少样本关系抽取中发挥优势，并提升以往提示学习方法的表现。

2 背景

2.1 少样本关系抽取

关系抽取任务旨在从纯文本上下文中抽取头实体和尾实体之间的关系。具体而言，关系抽取任务的一个实例由上下文 $\boldsymbol{x} = \{x_1, x_2, ..., h, ..., t, ..., x_{|\boldsymbol{x}|}\}$、头实体和尾实体提及 $\boldsymbol{h}$ 和 $\boldsymbol{t}$、实体类型 $\boldsymbol{t}_h$ 和 $\boldsymbol{t}_t$，以及 $\boldsymbol{h}$ 和 $\boldsymbol{t}$ 之间的关系 $\boldsymbol{y} \in \mathcal{Y}$ 组成，其中 $\mathcal{Y}$ 是候选关系集合。关系抽取系统将根据 $\boldsymbol{x}$、$\boldsymbol{h}$、$\boldsymbol{t}$、$\boldsymbol{t}_h$ 和 $\boldsymbol{t}_t$ 预测 $\boldsymbol{y}$。对于少样本关系抽取，微调预训练语言模型（预训练语言模型）是一种直接的解决方案（Han 等，2019；Yamada 等，2020；Joshi 等，2020；Lyu 和 Chen，2021；Zhou 和 Chen，2022）。为了减小预训练目标与下游应用之间的差距，提示调优最近已被应用于关系抽取，尤其是在低资源场景下（Chen 等，2022；Han 等，2021b、2022）。大多数此类方法

* 通讯作者。[1]代码可在以下地址获取 https://github.com/zjunlp/DeepKE/tree/main/example/llm。



图1：释放大语言模型在少样本关系抽取中潜力的策略。头实体类型 和 尾实体类型 是模式。头实体 和 尾实体 是实体提及。 关系 指口头表达的关系标签词。

利用相对较小的语言模型（RoBERTa（Liu 等，2019年）、GPT2（Radford 等人，2019年）），在少样本关系抽取性能方面取得了实证成功。迄今为止，大语言模型已展现出强大的能力，即通过提示少量实例而无需进行微调（Ding 等人，2022年）；然而，LLM 在少样本关系抽取方面的潜力仍鲜为人知。

2.2 大语言模型

大语言模型使用规模极大的语料库进行训练，通常包含大量参≥10数（B），已在众多下游自然语言处理任务中取得出色的性能（Taylor等人，2022年；Zhang等人，2022年；Zeng等人，2022年；Chowdhery等人，2022年；Ouyang等人，2022年）。与相对较小的语言模型（小型语言模型）相比，大语言模型通常不开源，也无法进行微调，这给下游任务适配带来了挑战。因此，研究者提出了上下文学习（Brown等人，2020年），通过提示和少量示例来实现少样本学习。以往的研究（Yoo等人，2021年；Wang等人，2021年）探讨了使用大语言模型进行文本分类和生成。在这项工作中，我们首次尝试研究基于大语言模型的少样本关系抽取，并由此带来了新的挑战与见解。

3 使用大语言模型进行少样本关系抽取

本节介绍两种利用大语言模型进行关系抽取的策略：1）上下文学习（§3.1）；2）使用大语言模型进行数据生成（§3.2），如图1所示。

3.1 使用大语言模型进行上下文学习

第一种策略通过在提示中为大语言模型提供示例来应用上下文学习（ICL），以引导大语言模型理解关系抽取任务。为此，我们手动构建并设计了具体且有说服力的关系抽取提示及其示例，以指导大语言模型理解关系抽取任务及其执行方式。考虑到关系抽取任务的各个方面和特征，包括任务定义、候选关系（标签）词、实体类型（模式）等，我们设计了不同表述方式和复杂度的提示，以研究提示如何帮助大语言模型释放少样本关系抽取的潜力。首先，**文本提示**仅包含关系抽取的基本要素，包括关系类别、上下文以及相应的头实体和尾实体。受 InstructGPT（Ouyang等人，2022年）和 ChatGPT（OpenAI，2022年）出色表现的启发，我们设计了**与任务相关的指令**来描述关系抽取任务，并将其加入提示中，称为**指令提示**。同时，根据以往的少样本关系抽取研究（Zhou和Chen，2022年），实体类型（模式）很有帮助；因此，我们也探究了在提示中使用模式的效果。

3.2 使用大语言模型进行数据生成

为弥补标注数据的稀缺性，我们引入了另一种策略：通过大语言模型生成数据。具体而言，我们使用包含数据形式描述的特定提示，引导大语言模型自主生成更多领域内标注数据，随后利用这些数据以及现有的少样本标注训练数据，对一个相对较小的语言模型进行微调。我们设计提示以说明一个关系抽取训练实例的基本组成部分（$\boldsymbol{x}$、$\boldsymbol{h}$、$\boldsymbol{t}$、$\boldsymbol{t}_h$、$\boldsymbol{t}_t$ 和 $\boldsymbol{y}$），并展示少量训练实例作为示例，以教会大语言模型理解

	方法	TACRED K=8	TACRED K=16	TACREV K=8	TACREV K=16	RE-TACRED K=8	RE-TACRED K=16	SciERC K=8	SciERC K=16
Baselines	SpanBERT（Joshi 等人，2020年）	8.4	17.5	5.2	5.7	14.2	29.3	29.0	38.7
	LUKE（Yamada 等人，2020年）	9.5	21.5	9.8	22.0	14.1	37.5	33.2	48.9
	GDPNet（Xue 等人，2021年）	11.8	22.5	8.3	20.8	18.8	48.0	33.5	42.3
	TANL（Paolini 等人，2021年）	18.1	27.6	18.6	28.8	26.7	50.4	32.4	38.7
	TYP Marker (Zhou和Chen, 2022年)	26.5	29.9	26.7	29.5	44.8	54.1	50.4	59.0
	KnowPrompt（Chen 等，2022年）	29.4	32.1	29.8	34.1	56.1	61.4	50.2	57.1
GPT3	上下文学习†	31.9		32.4		49.9		46.6	
	上下文学习†（含指令）	31.0		31.9		51.8		48.8	
	数据生成（TYP Marker）	35.8	36.6	36.7	36.5	58.4	60.6	63.2	64.3
	数据生成（KnowPrompt）	37.9	37.4	42.6	41.0	62.7	66.2	58.6	67.8

表 1：少样本性能的微平均F1（%）。† 表示单样本示例下的性能。

提示	TACRED	TACREV	RE-TACRED	SciERC
TEXT	31.9	32.4	49.9	46.6
文本 +模式	36.9	37.7	54.3	45.9
指令	31.0	31.9	51.8	48.8
指令 + 模式	38.3	36.7	58.5	50.2

表 2：不同提示（文本提示和指令提示）的性能微平均F1（%）。

带标签关系抽取数据的特征。需要注意的是，模式（例如关系类型和实体类型）是关系抽取数据中的重要结构信息。因此，我们提出**模式约束数据生成**，通过在提示中添加实体类型作为模式引导（见图1），以提升性能。随后，利用该提示引导大语言模型创建增强的关系抽取数据，并将其转换为预期格式，以供后续使用。

4 实验设置

4.1 方法与数据集

GPT-3.5 通过 OpenAI API 调用，[2] 作为我们实验中的大语言模型。我们在四个关系抽取数据集上开展实验，包括 TACRED（Zhang等人，2017年）、TACREV（Alt 等，2020年）、RE-TACRED（Stoica 等，2021年）和 SciERC（Luan 等，2018年）。与大语言模型相比，我们采用相对较小的模型实现了六种基线方法（详情见附录 A）。

4.2 少样本设置

每种关系抽取 K 个实例（K-样本）用于训练和验证。对于所有基线方法，我们使用随机抽取的 8 样本和 16 样本数据集用于训练和验证。对于上下文学习，由于 GPT-3.5 的请求令牌数上限为 4097，且 TACRED 数据集系列包含 40 多种关系，**只能使用单样本示例**，单样本效果见表1。出于同样的原因，为每种关系单独生成更多标注数据时，提示中仅添加该关系的三个示例。

在四个完整测试集上进行上下文学习。每次都会从打乱顺序的训练集中随机抽取不同的示例，以避免示例排列造成的影响（Lu 等，2021年）。在数据生成方面，我们将 GPT-3.5 生成的数据与原始少样本训练数据结合，用于微调两种基线方法：TYP Marker（Zhou和Chen，2022年）和 KnowPrompt（Chen 等，2022年）。使用不同数量的生成数据会产生不同结果。因此，我们逐步增加生成的 k样本（$k \in \{8, 16, 32, 48\}$）数据，并分别将其加入原始的 8 样本和 16 样本训练数据中，报告 k 表 1中的最佳性能。更多详情见附录 A.3。

5 结果与讨论

5.1 关系抽取的主要发现

大语言模型上的上下文学习无需调整相对较小的预训练语言模型，即可在关系抽取任务中取得相当的性能。从表 1中我们注意到，仅使用单样本示例的上下文学习，也能取得与基于全参数微调的提示学习基线方法相当的性能。通过上下文学习使用大语言模型无需进行任何参数更新，这具有以下潜在价值：使

[2] https://platform.openai.com/docs/models/gpt-3-5

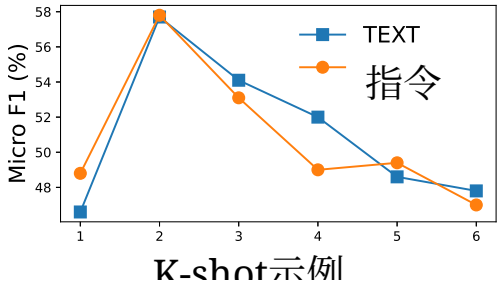

图2：SciERC中 k 上下文示例的微平均F1（%）。

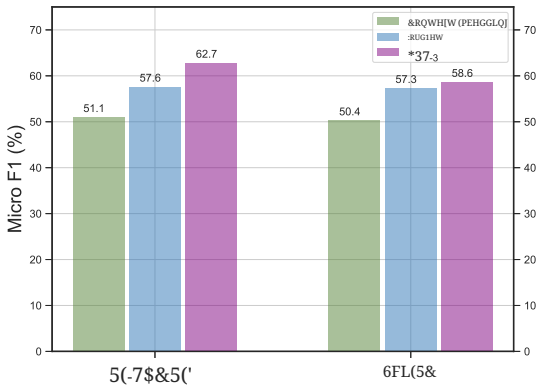

图3：使用大语言模型进行数据生成及采用不同数据增强方法的性能。*Roberta* 和*SciBERT* 分别用于 RE-TACRED 和 SciERC 中的基于上下文嵌入的数据增强方法。

使模型能够适应不同场景，而不像监督学习那样需要进行参数优化。

使用大语言模型进行数据生成，可提升以往解决方案，取得新的最先进的少样本关系抽取结果。 我们发现，仅使用表格中 GPT-3.5 生成的数据，就能显著提升以往的基线方法：**10.7%**，用于 SciERC 的 16-shot 设置；**6.6%** ，用于 RE-TACRED 的 16-shot 设置。见表1。值得注意的是，数据生成是一种简单而有效的方法，能够激发大语言模型的能力，助力以往的方法。我们还证明，使用大语言模型进行模式约束生成，能够使所有以往采用小型语言模型的方法受益。

5.2 大语言模型上下文学习中的提示

指令和模式在使用大语言模型进行关系抽取的上下文学习中发挥着至关重要的作用。从表2中，我们注意到，在大多数情况下，使用指令提示的模型比使用文本提示的模型表现更好，这表明与任务相关的信息确实有助于释放大语言模型在关系抽取方面的更多能力。TACRED 和 TACREV 中出现了异常结果，这是因为这两个数据集中的错误标注示例会违背输入大语言模型的正确指令，从而使大语言模型产生困惑，导致其表现甚至不如不使用指令的上下文学习。此外，添加模式信息能够显著提升性能，这体现了预定义结构信息对于关系抽取的重要性。

反直觉的是，增加示例数量未必能提升大语言模型进行关系抽取的性能。我们发现，性能不但不会提升，甚至还会下降；而且随着上下文示例数量增加，图 指令提示 与 文本提示 之间的差距会逐渐缩小，如图 2所示。我们认为，这可能有两个原因：1）很难选择代表性的代表性示例；2）对于大语言模型而言，理解输出（关系）空间更大的结构预测任务并非易事。附录中提供了更多关于 GPT-3.5 的案例研究B.1。

5.3 大语言模型生成数据的效用

将大语言模型生成的数据与原始训练数据相结合，能够比传统数据增强方法获得更好的关系抽取性能。我们在图3 中比较了通过大语言模型进行数据生成与以往广泛使用的数据生成方法（例如用 WordNet 的同义词替换训练集中的词语，以及使用上下文词嵌入）（详见附录A.3 ）。使用大语言模型进行数据生成的效果优于所有其他方法，这表明引导大语言模型生成数据，是弥补标注数据不足的有效方法。

不断增加来自大语言模型的生成数据，只能在一定程度上提升关系抽取性能，而无法持续改善。从图 4中，我们观察到，随着生成数据增多，结果先上升后下降，但始终高于未使用生成数据时的结果。根据附录B.2中对生成数据的分析，我们认为低质量的生成数据会在训练过程中引入大量噪声，而语言模型可能具有抗噪能力（Song etal., 2020年）。

6 讨论与结论

本文首次探讨如何利用大语言模型进行少样本关系抽取。我们发现，任务相关信息（包括指令或

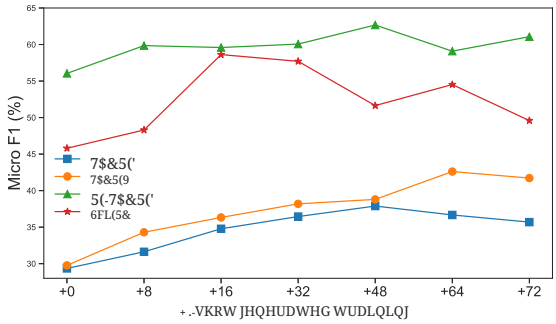

图 4：KnowPrompt 使用生成的训练数据和原始 8 样本数据时的微平均 F1（%）。

模式）有助于激发大语言模型的能力，并提升少样本关系抽取的效果。在现阶段，利用大语言模型生成数据，或许是增强基础模型（相对较小的预训练语言模型）在实际应用中能力的一种简单而有效的方案。我们希望这项工作能为自然语言处理领域带来使用大语言模型的益处。请注意，大语言模型只能根据上下文以及作为示例的少量训练样本进行预测。我们认为，它有潜力为适应不同场景（例如低样本量和任意领域）的关系抽取设计复杂的人类可读提示。

致谢

我们谨向匿名审稿人的宝贵意见表示感谢。本研究得到国家自然科学基金（编号：62206246）、浙江省自然科学基金（编号：LGG22F030011）、宁波市自然科学基金（2021J190）、甬江人才引进计划（2021A-156-G）以及 CAAI-华为 MindSpore 开放基金的资助。

局限性

尽管我们已尽最大努力，本文仍可能存在一些局限性。

大语言模型： 由于预算有限，我们无法承担使用各种大语言模型的费用，因此仅评估了 GPT-3.5（*Text-davinci-003*）。我们将尝试使用更多大语言模型研究关系抽取，例如 OPT（Zhang等人，2022年）、GLM-130B（Zeng等人，2022年）或代码语言模型（Bi等人，2023年），例如 Codex。

利用大语言模型的其他方法： 还有其他一些利用大语言模型的技术，例如黑盒优化（Sun 等人，2022年）和基于特征的学习（Lang 等人，2022年）；然而，我们发现，由于标签空间庞大且模式结构复杂，这些方法大多无法直接应用于关系抽取。我们将来会探索如何利用大语言模型和其他方法进行关系抽取。

数据集： 我们仅评估了四个关系抽取数据集，并将尝试研究大语言模型在更多不同领域和语言的数据集上的关系抽取性能。

参考文献

Christoph Alt、Aleksandra Gabryszak 和 Leonhard Hennig。2020年。重新审视 TACRED：对 TACRED 关系抽取任务的全面评估。收录于 *ACL* 论文集。

Iz Beltagy、Kyle Lo 和 Arman Cohan。2019年。SciBERT：一种用于科学文本的预训练语言模型。收录于 *2019*年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第 3615–3620 页，中国香港。计算语言学协会。

Zhen Bi、Jing Chen、Yinuo Jiang、Feiyu Xiong、WeiGuo、Huajun Chen 和 Ningyu Zhang。2023年。CodeKGC：用于生成式知识图谱构建的代码语言模型。*CoRR*，abs/2304.09048。

Sam Brody、Sichao Wu 和 Adrian Benton。2021年。迈向现实场景中的少样本关系抽取。载于 *2021*年自然语言处理实证方法会议论文集，第 5338–5345 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D 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 Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Xiang Chen、Ningyu Zhang、Xin Xie、Shumin Deng、Yunzhi Yao、Chuanqi Tan、Fei Huang、Luo Si 和 Huajun Chen。2022年。Knowprompt：结合协同优化的知识感知提示微调，用于关系抽取。载于 *WWW '22*：国际计算机学会 *Web*

*2022*年会议，线上会议，法国里昂，*2022*年*4*月*25*日至*29*日，第2778–2788页。国际计算机学会。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vin- odkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael I sard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、DaphneIppolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agra wal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Mor- eira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。2022年 PaLM：利用路径实现语言建模扩展。*arXiv* 预印本，abs/2204.02311。

Bosheng Ding、Chengwei Qin、Linlin Liu、Lidong Bing、Shafiq R. Joty 和 Boyang Li。2022年。GPT-3 是一个好的数据标注员吗？*arXiv* 预印本，abs/2212.10450。

Qingxiu Dong、Lei Li、Damai Dai、Ce Zheng、Zhiyong Wu、Baobao Chang、Xu Sun、Jingjing Xu、LeiLi 和 Zhifang Sui。2023年。上下文学习综述。*arXiv* 预印本，abs/2301.00234。

Jiale Han、Bo Cheng 和 Wei Lu。2021a。探索少样本关系抽取的任务难度。载于*2021*年自然语言处理经验方法会议论文集，*EMNLP 2021*，线上会议 / 多米尼加共和国蓬塔卡纳，*2021*年*11*月*7*日至*11*日，第2605–2616页。计算语言学协会。

Jiale Han、Shuai Zhao、Bo Cheng、Shengkun Ma 和 Wei Lu。2022年。用于关系分类的生成式提示调优。载于计算语言学协会论文集：*EMNLP 2022*，第3170–3185页，阿联酋阿布扎比。计算语言学协会。

Xu Han、Tianyu Gao、Yuan Yao、Deming Ye、Zhiyuan Liu 和 Maosong Sun。2019年。Opennre：用于神经关系抽取的开放式可扩展工具包。载于*2019*年自然语言处理经验方法会议和第九届自然语言处理国际联合会议论文集，*EMNLP-IJCNLP 2019*，香港，中国，*2019*年*11*月*3*日至*7*日——系统演示，第169–174页。计算语言学协会。

Xu Han、Weilin Zhao、Ning Ding、Zhiyuan Liu 和 MaosongSun。2021b。PTR：提示调优 with rules f或文本 c lassificati开启. ar Xiv prepr i nt，abs/2105.11259。Xu Han、Hao Zhu、Pengfei Yu、Ziyun Wang、Yuan Yao、Zhiyuan Liu 和 Maosong Sun。2020年。FewRel：具有最先进评测的大规模有监督小样本关系分类数据集。载于*2018*年自然语言处理实证方法会议论文集，第4803–4809页，比利时布鲁塞尔。计算语言学协会。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S. Weld、Luke Zettlemoyer 和 Omer Levy。2020年。Spanbert：通过表示和预测片段改进预训练。计算语言学协会会刊，8：64–77。Hunter Lang、Monica N. Agrawal、Yoon Kim 和 David A. Sontag。2022年。协同训练改进大语言模型的基于提示的学习。载于国际机器学习会议，*ICML 2022*，*2022*年*7*月*17*日至*23*日，美国马里兰州巴尔的摩，第162卷机器学习研究论文集，第11985–12003页。PMLR。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。Roberta：一种稳健优化的 BERT 预训练方法。*arXiv:1907.11692*，abs/1907.11692。Yao Lu、Max Bartolo、Alastair Moore、Sebastian Riedel 和 Pontus Stenetorp。2021年。奇妙排序的提示及其寻找之处：克服小样本提示顺序敏感性。*arXiv* 预印本，abs/2104.08786。Yi Luan、Luheng He、Mari Ostendorf 和 Hannaneh Hajishirzi。2018年。用于科学知识图谱构建的实体、关系和共指多任务识别。载于自然语言处理实证方法会议论文集（*EMNLP*）。Shengfei Lyu 和 Huanhuan Chen。2021年。具有实体类型限制的关系分类。载于计算语言学协会论文集：*ACL/IJCNLP 2021*，线上活动，*2021*年*8*月*1*日至*6*日，第ACL/IJCNLP 2021卷*ACL* 论文集，第390–395页。计算语言学协会。Yubo Ma、Yixin Cao、YongChing Hong 和 Aixin Sun。2023年。大语言模型不是优秀的小样本信息抽取器，却是困难样本的优秀重排序器！*CoRR*，abs/2303.08559。OpenAI。2022年。ChatGPT：针对对话优化语言模型。https://openai.com/blog/chatgpt/。OpenAI。2023a年。GPT-4 技术报告。*arXiv* 预印本，abs/2303.08774。OpenAI。2023b年。Text-davinci-003。

https://platform.openai.com/docs/models/text-davinci-003。

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F. Christiano、Jan Leike 和 Ryan Lowe。2022年。训练语言模型遵循人类反馈的指令。 *arXiv* 预印本，abs/2203.02155。

Giovanni Paolini、Ben Athiwaratkun、Jason Krone、Jie Ma、Alessandro Achille、Rishita Anubhai、Cícero Nogueira dos Santos、Bing Xiang 和 Ste- fano Soatto。2021年。结构化预测作为增强型自然语言之间的转换。载于第*9*届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督多任务学习器。*OpenAI* 博客，1(8):9。

Livio Baldini Soares、Nicholas FitzGerald、Jeffrey Ling 和 Tom Kwiatkowski。2019年。填补空白：用于关系学习的分布相似性。载于第*57*届计算语言学协会年会论文集，*ACL 2019*，佛罗伦萨，意大利，*2019*年*7*月*28*日至*8*月*2*日，第*1*卷：长篇论文，第2895–2905页。计算语言学协会。

Hwanjun Song、Minseok Kim、Dongmin Park 和 Jae- Gil Lee。2020年。使用深度神经网络从噪声标签中学习：综述。*arXiv* 预印本，abs/2007.08199。

George Stoica、Emmanouil Antonios Platanios 和 Barnabás Póczos。2021年。Re-tacred：解决 TACRED 数据集的不足。载于 第三十五届*AAAI*人工智能大会、*AAAI 2021*、第三十三届人工智能创新应用大会、*IAAI 2021*、第十一届教育领域人工智能进展研讨会、*EAAI 2021*、线上会议，*2021*年*2*月*2*日至*9*日，第13843–13850页。AAAI出版社。

Tianxiang Sun、Yunfan Shao、Hong Qian、Xuanjing Huang 和 Xipeng Qiu。2022年。面向语言模型即服务的黑盒调优。载于 国际机器学习大会，*ICML 2022*，*2022*年*7*月*17*日至*23*日，美国马里兰州巴尔的摩，第162卷《机器学习研究论文集》，第20841–20855页。PMLR。

Ross Taylor、Marcin Kardas、Guillem Cucurull、Thomas Scialom、Anthony Hartshorn、Elvis Saravia、Andrew Poulton、Viktor Kerkez 和 Robert Stojnic。2022年。Galactica：面向科学的大语言模型。*arXiv* 预印本，abs/2211.09085。

Shuohang Wang、Yang Liu、Yichong Xu、Chenguang Zhu 和 Michael Zeng。2021年。想降低标注成本？GPT-3 可以提供帮助。载于 计算语言学协会论文集：*EMNLP 2021*年，线上会议 / 多米尼加共和国蓬塔卡纳，*2021*年*11*月*16*日至*20*日，第4195–4205页。计算语言学协会。

Fuzhao Xue、Aixin Sun、Hao Zhang 和 Eng Siong Chng。2021年。GDPNet：用于关系抽取的潜在多视图图优化方法。发表于第三十五届*AAAI*人工智能大会（*AAAI 2021*）、第三十三届人工智能创新应用会议（*IAAI 2021*）、第十一届人工智能教育进展研讨会（*EAAI 2021*），线上会议，*2021*年*2*月*2*日至*9*日，第14194–14202页。AAAI出版社。

Ikuya Yamada、Akari Asai、Hiroyuki Shindo、Hideaki Takeda 和 Yuji Matsumoto。2020年。LUKE：采用实体感知自注意力的深度上下文化实体表示。发表于*2020*年自然语言处理实证方法会议论文集（*EMNLP 2020*），线上，*2020*年*11*月*16*日至*20*日，第6442–6454页。计算语言学协会。

Shan Yang、Yongfei Zhang、Guanglin Niu、Qinghua Zhao 和 Shiliang Pu。2021年。实体概念增强的少样本关系抽取。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集，*ACL/IJCNLP 2021*（第*2*卷：短论文），线上会议，*2021*年*8*月*1*日至*6*日，第987–991页。计算语言学协会。

Hongbin Ye、Ningyu Zhang、Hui Chen 和 Huajun Chen。2022年。生成式知识图谱构建：综述。载于*2022*年自然语言处理实证方法会议论文集，第1–17页，阿拉伯联合酋长国阿布扎比。计算语言学协会。

Kang Min Yoo、Dongju Park、Jaewook Kang、Sang-Woo Lee 和 Woo-Myoung Park。2021年。Gpt3mix：利用大规模语言模型进行文本增强。载于计算语言学协会论文集：*EMNLP 2021*成果，线上会议／多米尼加共和国蓬塔卡纳，*2021*年*11*月*16*日至*20*日，第2225–2239页。计算语言学协会。

Aohan Zeng、Xiao Liu、Zhengxiao Du、Zihan Wang、Hanyu Lai、Ming Ding、Zhuoyi Yang、Yifan Xu、Wendi Zheng、Xiao Xia、Weng Lam Tam、Zixuan Ma、Yufei Xue、JidongZhai、Wenguang Chen、Peng Zhang、Yuxiao Dong 和 Jie Tang。2022年 GLM- 130B：开放式双语预训练模型。 *CoRR*，abs/2210.02414。

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shus- ter, Daniel Simig, Punit Singh Koura, Anjali Srid- har, Tianlu Wang 和 Luke Zettlemoyer。2022年。OPT：开放式预训练 Transformer 语言模型。*CoRR*，abs/2205.01068。

Yuhao Zhang、Victor Zhong、Danqi Chen、Gabor Angeli 和 Christopher D. Manning。2017年。位置感知注意力和监督数据可改进槽位填充。载于*2017*年自然语言处理实证方法会议论文集（*EMNLP 2017*），第35–45页。

Wenxuan Zhou 和 Muhao Chen。2022年。句子级关系抽取的改进基线。载于《计算语言学协会亚太分会第二届会议暨第十二届自然语言处理国际联合会议论文集》，*AACL/IJCNLP 2022*——第 *2* 卷：短论文，*2022* 年 *11* 月 *20* 日至 *23* 日，仅在线举行，第 161–168 页。计算语言学协会。

A 实验细节

A.1 数据集

TACRED[3] 是广泛使用的关系抽取数据集，包含 42 个关系标签，其中包括无_关系，表示未发现关系。TACREV[4] 包含与 TACRED 相同的训练集，以及重新标注的开发集和测试集。

RE-TACRED[5] 是 TACRED 的重新标注版本，包含 40 种关系。SciERC[6]包含七种关系类别，构建于科学领域。所有数据集均取自其官方网站，未经修改，包括内容以及训练集、测试集和开发集划分。

A.2 基线方法

我们使用规模相对较小的模型，将大语言模型与近期的基线方法进行比较。1）常规微调方法：**SpanBERT**（Joshi 等人，2020年），一种基于跨度的预训练语言模型；**LUKE**（Yamada 等人，2020年），一种基于双向变换器的词语和实体的上下文化预训练表示；**GDPNet**，一种高斯动态时间规整池化网络，能够选择对关系预测重要的词语；**TYPMarker**（Zhou和Chen，2022年），使用实体类型标记进行微调。2）生成式方法：**TANL**（Paolini 等人，2021年），将结构化预测语言任务构建为增强自然语言之间的翻译任务。3）提示微调方法：**KnowPrompt**，通过协同优化进行知识感知的连续提示微调。

A.3 实现细节

生成数据与现有训练数据随后在 KnowPrompt 上进行评估。数据增强通过 Word-Net 的同义词和上下文词嵌入实现的方法由 *nlpaug*[7] 完成。OpenAI API 中的参数 温度参数 设置为 0，以确保上下文学习中的精确率；设置为 1，以生成多样化的关系抽取数据。所有实验均使用一块配备 24GB 显存的 NVIDIA GeForce RTX 3090 图形处理器运行。除 SciERC 数据集外，我们按照基线方法的原始设置重新运行了其官方代码。由于 SciERC 属于垂直领域，出于公平性考虑，在 TYP Marker 和 KnowPrompt 中使用了 *SciBERT*(Beltagy 等人，2019年)。对于另外三个数据集，TYP Marker 和 KnowPrompt 均采用 *RoBERTa-large* 。*TYP Marker*和 *KnowPrompt*。

B 案例分析

B.1 上下文学习中的错误案例

从表4中可以看出，一些关系抽取实例对大语言模型来说颇具挑战，大语言模型也存在若干局限：1）大语言模型不擅长清晰地区分头实体和尾实体的顺序。2）相同的头实体和尾实体提及会使大语言模型困惑。4）如果上下文中的头实体和尾实体相距较远，大语言模型就难以正确判断关系。5）语义相似的关系标签词和实体提及会令大语言模型感到困惑，因为它们的嵌入相似。6）由于关系抽取的标签空间很大，大语言模型无法处理过长的实例。7）大语言模型往往无法抽取那些难以确定或标注错误的关系；这些关系对人类来说也很有挑战性。更多高质量的示例或许有助于缓解这些问题。我们还认为，有必要开发基于大语言模型的逐步（聊天式）方法，在一个阶段内抽取有限数量的关系。

B.2 大语言模型生成的数据

表 5 中列出了一些来自 GPT-3.5 的生成数据案例。通过对每个数据集中 100 个生成样本进行人工检查，我们发现约 78% 的生成数据标注正确且质量较高（TACRED 为 85%，TACREV 为 82.5%，RE-TACRED 为 72%，SciERC 为 75%）。与此同时，我们分别将生成数据和原始金标准训练数据加入八样本数据集，并微调*Know-Prompt*，以评估生成数据的质量，如表3 所示。我们观察到，GPT-3.5 生成的标注数据大多正确。对于 TACRED 和 TACREV，生成数据带来的提升大于金标准标注数据。由于

[3] https://nlp.stanford.edu/projects/tacred/[4] https://github.com/DFKI-NLP/tacrev[5] https://github.com/gstoica27/Re-TACRED[6] http://nlp.cs.washington.edu/sciIE/

[7] https://github.com/makcedward/nlpaug

8-shot 数据集	TACRED		TACREV		RE-TACRED		SciERC	
	生成	gold	生成	gold	生成	gold	生成	gold
添加 0-shot	29.35	29.35	29.77	29.77	56.05	56.05	45.80	45.80
添加 8-shot	31.63	30.73	34.30	33.16	59.85	60.92	48.30	57.08
添加 16-shot	34.78	31.88	36.33	33.49	59.59	61.30	**58.62**	65.15
添加 32-shot	36.45	33.35	38.19	33.98	60.06	64.65	57.70	72.11
添加 48-shot	**37.89**	33.97	38.80	35.06	**62.67**	65.56	51.64	74.29
增加 64-shot	36.67	34.36	**42.61**	35.57	61.07	67.28	54.52	75.36
添加 72-shot	35.69	**34.58**	41.72	**35.96**	59.09	**67.43**	49.59	**75.87**

表 3：在 8-shot 数据集上添加由 GPT-3.5 生成的标注数据或金标准标注数据后，*KnowPrompt* 的微平均 F1（%）。

TACRED 和 TACREV 中存在许多错误标注数据（Zhang等人，2017年；Alt 等人，2020年），我们认为，更好的性能得益于 GPT-3.5 的帮助。不过，我们也发现，GPT-3.5 生成的一些数据质量不如金标准数据。对于 RE-TACRED 和 SciERC，使用更多金标准数据的效果优于使用生成数据。人工检查发现，一些生成样本过短，且由若干语义无关的句子拼接而成。同时，SciERC 上显著的性能差异表明，GPT-3.5 不擅长科学等垂直领域。

数据集	Case	金标准关系	上下文学习
TACRED	上下文：奇怪的是，Cain 短暂的三年..... 他在美国全国步枪协会的任期显然是他唯一的一段时期。 数十年的职业生涯，在此期间，他被指控曾是 一名性侵犯者。 头实体类型：组织。头实体：美国全国步枪协会。 尾实体类型：人物。尾实体类型：Cain	组织:top_成员/员工	人物:employee_of
	上下文：“我向学生学习，也挑战他们”， Heloise（58岁）说，她接手了家族提示业务 当她的母亲 ，也称为 Heloise ，于1977年去世。 头实体类型：人物。头实体：Heloise。 尾实体类型：人物。尾实体：Heloise。	人物：父母	人物：别名_名称
TACREV	上下文：Anna Mae Pictou Aquash 是一名米克马克族印第安人 来自加拿大，于1975年惨遭杀害。 头实体类型：人物。头实体：Anna Mae Pictou Aquash. 尾实体类型：国家。尾实体：加拿大。	per:出生国家_的_	per:countries_of_residence
	上下文：Messina Denaro 一直试图强加他的 在西西里首府巴勒莫掌权，并成为 the 新任西西里黑手党首领，因被捕而势力削弱 Provenzano于2006年4月。 头实体类型：人物。头实体：他。 尾实体类型：城市。尾实体：巴勒莫。	无关系_	per:cities_of_residence
RE-TACRED	上下文：据说弗拉基米尔·拉德任斯基于周六晚些时候去世了—— 在芬兰南部比赛期间的星期六，而他的 芬兰选手蒂莫·考科宁被紧急送往医院。 头实体类型：人物。头实体：弗拉基米尔·拉德任-skiy。 尾实体类型：人物。尾实体：他。	人物：身份	人物：死亡日期_of_
	中美洲议会议长（中美洲议会） Jacinto Suarez 周一表示， 中美洲国家不支持巴拿马的 退出中美洲议会的退出请求。 头实体类型：组织。头实体：中美洲 议会。 尾实体类型：人物。尾实体：Jacinto Suarez。	org:top_members/employees	per:title
SciERC	上下文：我们在两个语料库（电话对话语音 和新闻广播语音）上分别使用 人工转录文本和语音识别输出进行评估。 头实体类型：OtherScientificTerm。头实体：transcrip- 关系。 尾实体类型：其他科学术语。尾实体：输出。	合取	比较
	上下文：我们在九个标准...上验证了这种新方法 行人重识别数据集，包括两个大规模的 Market-1501 和 CUHK03 数据集，并表明我们 在所有方面超越当前最先进的方法 其中的。 头实体类型：材料。头实体：CUHK03 数据集。 尾实体类型：材料。尾实体：数据集。	下位词关系	部分关系

表 4：GPT-3.5 预测错误的案例。金标准关系类别列于第三列，上下文学习预测的结果列于第四列。

数据集	Case	纠正数据
TACRED	上下文：美国癌症协会总部位于亚特兰大，并且是 由15名受过培训的女性志愿者于1913年创立。 头实体类型：组织。头实体：美国癌症协会。 **尾实体类型：组织**。尾实体：15名受过培训的女性志愿者。 关系：org:founded_by。	尾实体类型：人物
	上下文：Mary Brown，Brown Corp 的首席执行官兼知名女商人， 是行业会议和活动上的常客演讲嘉宾。 头实体类型：人物。头实体：Mary Brown。 **尾实体类型：人物**。尾实体：首席执行官。 关系：per:title。	尾实体类型：职称
TACREV	上下文：古斯塔夫·马勒于1860年7月7日出生在波希米亚的卡利施特。 头实体类型：人物。头实体：古斯塔夫·马勒。 **尾实体类型：人物**。尾实体：1860。 **关系：per:country_of_birth**。	尾实体类型：日期 关系：per:date_of_birth
	上下文：MTN Nigeria（南非 MTN Group 的子公司）已 开始在尼日利亚证券交易所上市其股份。 头实体类型：组织。头实体：MTN Group。 尾实体类型：组织。尾实体：MTN Nigeria。 关系：组织：子公司。	-
RE-TACRED	上下文：教皇约翰·保罗二世是一位广受欢迎的天主教领袖，他 在其教皇任期的大部分时间里驻于梵蒂冈城。 头实体类型：人物。头实体：教皇约翰·保罗二世。 **尾实体类型：人物**。尾实体：梵蒂冈城。 **关系：per:countries_of_residence**。	尾实体类型：城市 关系： per:cities_的_居住地
	上下文：中国制药商赛诺菲-安万特解散了其在中国的 子公司广州医药，此前该公司卷入了贿赂丑闻。 头实体类型：组织。头实体：赛诺菲-安万特。 尾实体类型：组织。尾实体：广州医药。 关系：org:dissolved。	-
SciERC	上下文：两种方法的比较表明，神经 方法的结果远优于基于规则的系统。 头实体类型：方法。头实体：神经方法。 尾实体类型：方法。尾实体：基于规则的系统。 关系：比较。	-
	上下文：色谱法与质谱法的结合已经 使科学家能够实现前所未有的蛋白质组分析水平。 头实体类型：方法。头实体：质谱法。 尾实体类型：方法。尾实体：色谱法。 **关系：特征-属于**。	关系：合取

表 5：大语言模型生成的数据。第二列中的错误以粗体标出，并在第三列中予以更正。