自动化学报
ACTA AUTOMATICA SINICA

基于大语言模型的中文实体链接实证研究

徐正斐 辛欣

An Empirical Study of Chinese Entity Linking Based on Large Language Model

XU Zheng-Fei, XIN Xin

在线阅读 View online: https://doi.org/10.16383/j.aas.c240069

基于大语言模型的中文实体链接实证研究

徐正斐 [1, 2] 辛 欣 [1, 2]

摘 要 近年来, 大语言模型 (Large language model, LLM) 在自然语言处理中取得重大进展. 在模型足够大时, 大语言模型涌现出传统的预训练语言模型 (Pre-trained language model, PLM) 不具备的推理能力. 为了探究如何将大语言模型的涌现能力应用于中文实体链接任务, 适配了以下四种方法: 知识增强、适配器微调、提示学习和语境学习 (In-context learning, ICL). 在 Hansel 和 CLEEK 数据集上的实证研究表明, 基于 Qwen-7B/ChatGLM3-6B 的监督学习方法超过基于小模型的方法, 在 Hansel-FS 数据集上提升 3.9% ~ 11.8%, 在 Hansel-ZS 数据集上提升 0.7% ~ 4.1%, 在 CLEEK 数据集上提升 0.6% ~ 3.7%. 而当模型参数量达到720 亿时, Qwen-72B 的无监督方法实现与监督微调 Qwen-7B 相近的结果 (−2.4% ~ +1.4%). 此外, 大语言模型 Qwen 在长尾实体场景下有明显的优势 (11.8%), 且随着参数量的增加, 优势会更加明显 (13.2%). 对错误案例进行分析 (以下简称错误分析) 发现, 实体粒度和实体类别相关错误占比较高, 分别为 36% 和 25%. 这表明在实体链接任务中, 准确划分实体边界以及正确判断实体类别是提高系统性能的关键.

关键词 实体链接, 大语言模型, 知识增强, 适配器微调, 提示学习, 语境学习

引用格式 徐正斐, 辛欣. 基于大语言模型的中文实体链接实证研究. 自动化学报, 2025, **51**(2): 1−16

DOI 10.16383/j.aas.c240069 **CSTR** 32138.14.j.aas.c240069

An Empirical Study of Chinese Entity Linking Based on Large Language Model

XU Zheng-Fei[1, 2] XIN Xin[1, 2]

Abstract Large language models (LLMs) have recently made significant advancements in natural language processing. When scaled sufficiently, large language models exhibit reasoning capabilities that traditional pre-trained language models (PLMs) lack. In order to explore how to apply the emergent capabilities of large language models to the Chinese entity linking task, the following four methods are adapted: Knowledge augmentation, adapter fine-tuning, prompt learning, and in-context learning. Empirical studies on the Hansel and CLEEK datasets show that supervised learning methods based on Qwen-7B/ChatGLM3-6B outperform PLM-based methods. It achieves improvements ranging from 3.9% to 11.8% on the Hansel-FS dataset, 0.7% to 4.1% on the Hansel-ZS dataset, and 0.6% to 3.7% on the CLEEK dataset. When scaled to 72 billion parameters, Qwen-72B′s unsupervised methods yield results comparable to the supervised fine-tuning of Qwen-7B, with a performance range of −2.4% to +1.4%. Furthermore, the large language model Qwen has a clear advantage in the long-tail entity scenario (11.8%), and as the number of parameters increases, the advantage will become more obvious (13.2%). The analysis of the error cases (hereinafter referred to as error analysis) found that the errors related to entity granularity and entity type accounted for a high proportion, 36% and 25% respectively. This shows that in the entity linking task, accurately dividing entity boundaries and correctly judging entity types are the key to improving system performance.

Key words Entity linking, large language model (LLM), knowledge augmentation, adapter fine-tuning, prompt learning, in-context learning (ICL)

Citation Xu Zheng-Fei, Xin Xin. An empirical study of Chinese entity linking based on large language model. *Acta Automatica Sinica*, 2025, **51**(2): 1−16

在信息处理领域, 实体链接是将文本中的指称 (Mention) 与知识库中相应实体进行关联的任务, 主要应用于知识图谱扩充[1]、信息检索[2]、问答系统[3]等下游任务. 随着 Qwen[4]、ChatGLM (Chat general language model)[5]、GPT (Generative pre-trained transformer)[6] 等大语言模型的崛起, 实体链接有了更广泛的应用场景. 首先, 大语言模型有概率产生“幻觉”问题[7], 这限制了其可靠性; 其次, 大语言模型通过大规模预训练将知识隐式地存储于模型参数中[8], 这限制了知识更新和在特定领域的应用.

收稿日期 2024-01-31 录用日期 2024-08-07
Manuscript received January 31, 2024; accepted August 7, 2024
国家自然科学基金 (62172044) 资助
Supported by National Natural Science Foundation of China (62172044)
本文责任编委 刘洋
Recommended by Associate Editor LIU Yang
1. 北京理工大学计算机学院 北京 100081 2. 北京理工大学北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081
1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 2. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Institute of Technology, Beijing 100081

Kandpal 等[9] 的研究表明大语言模型难以通过持续训练纳入新的知识. 因此将大语言模型与外部知识库相结合, 成为缓解大语言模型的“幻觉”问题、扩展事实知识的有效方法[9]. 实体链接作为连接自然语言和结构化知识的桥梁, 可以为大语言模型提供上下文相关的事实知识, 将大语言模型与知识库相关联, 以提高大语言模型的能力[10].

当前的神经实体链接方法主要基于预训练 + 微调的范式. 为平衡效率和准确率, 目前的实体链接系统主要包含两个阶段: 候选实体生成和实体消歧, 本文主要关注实体消歧阶段. 实体消歧方法主要基于实体和指称的表示学习[11−13]. 随着预训练语言模型[14] 的发展, Yamada 等[15] 通过构建预训练任务, 在预训练阶段学习实体和指称的表示向量. Wu 等[16] 通过微调双编码器 (Dual encoder, DE), 将实体和指称映射到相同的向量空间, 并通过向量检索生成候选实体, 在消歧阶段, 通过交叉注意力编码器 (Cross-attention encoder, CA) 构建消歧向量实现消歧. 上述基于预训练 + 微调的实体消歧模型, 通过深度表示学习来构建指称和文本的语义向量, 利用这些向量的交互进行消歧.

相比于小规模预训练语言模型, 具备数十亿乃至上千亿参数的大语言模型涌现出一定的推理能力, 如逐步推理 (Multi-step reasoning)[17]、指令遵循 (Instruction following)[18] 和语境学习 (In-context learning, ICL)[6], 而人们通常认为小规模预训练语言模型缺少推理能力[17, 19]. 得益于这些涌现能力, 出现了很多新的模型交互方式. 例如通过自然语言形式的提问就能获得特定任务的答案. 以实体链接任务为例, 图 1 展示了大语言模型的三种交互方式[1]. 这些示例展现出大语言模型具有推理能力和内部知识, 这启发了将这些涌现能力应用在实体链接任务中, 以提高模型性能.

本文基于不同的涌现能力实例化了大语言模型的四种应用方法, 如图 1 所示. 1) 知识增强 (逐步推理): 利用思维链提示引导大语言模型生成推理步骤作为小模型的输入特征; 2) 适配器微调 (显式指令遵循): 通过固定指令并微调模型参数来实现指令遵循; 3) 提示学习 (隐式指令遵循): 通过固定模型参数并微调指令表示来实现指令遵循; 4) 语境学习: 检索相关演示示例, 构造小样本提示输入, 引导大模型进行语境学习, 从而实现推理.

本文主要研究大语言模型在中文实体链接任务中的改进效果. 构建了“检索−重排序”的两阶段中文实体链接基线系统, 在候选生成阶段使用双编码器方法检索实体; 在实体消歧阶段实现了知识增强、适配器微调、提示学习和语境学习四种应用方法. 在中文实体链接测试基准 Hansel[20] 和 CLEEK[21] 上的评估表明, 基于 BERT 等预训练模型的方法[14, 16] 存在“过拟合”和“长尾实体”问题; 而基于 Qwen-7B 和 ChatGLM3-6B 的知识增强、适配器微调和提示学习方法具有明显优于基线的效果; 基于语境学习和思维链推理的无监督方法在更大规模的大语言模型 Qwen-72B 上取得与小规模 Qwen-7B 监督微调相近的结果. 随着参数规模的增大, Qwen 模型在长尾实体上的优势进一步凸显. 此外, 本文对实验结果的错误案例进行了分析 (以下简称错误分析), 统计了实体粒度、实体类别、全局错误、局部错误、时间错误和地点错误六种主要错误类型的比例.

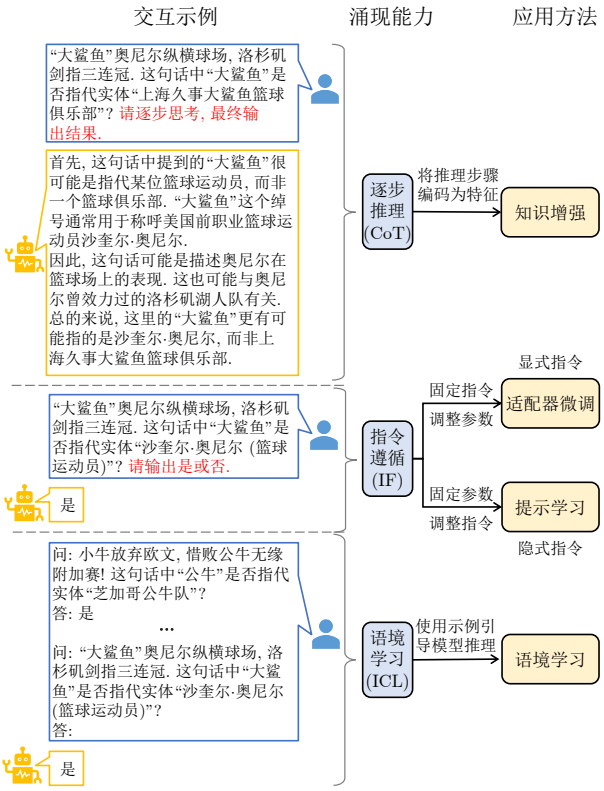


图 1 研究动机

Fig. 1 The research motivation

综上所述, 本文的主要贡献如下:

1) 实例化四种大语言模型应用方法, 分析了不同大语言模型在不同方法下的效果, 其中, Qwen-7B 和 ChatGLM3-6B 的监督学习方法在 Hansel-ZS、Hansel-FS 和 CLEEK 上均有提升, 准确率提升分别为 3.9% ~ 11.8%、0.7% ~ 4.1%、0.6% ~ 3.7%; 无监督方法在 Qwen-72B 上取得了与监督微调 Qwen-7B 模型相近的结果.

2) 实验发现, 大模型 Qwen 相比小模型

[1] 交互示例的输出内容来自: https://www.chatglm.cn

BERT 在长尾实体场景下有明显的优势, 且随着参数量的增加, 优势会更加明显. 此外, Qwen 大模型的参数高效微调方法相比 BERT 小模型微调和全参数微调更不容易出现“过拟合”问题.

3) 错误分析表明, 大语言模型在涉及粒度和类别的问题上存在较高的错误占比, 分别为 36% 和 25%. 最后通过定性分析给出了改进建议.

1　实体链接系统总体架构

1.1　任务定义

实体链接是给定一个知识库 E, 将一段文本中的指称 m 链接到知识库中对应实体 e^* 的任务. 由于知识库中实体规模庞大, 为平衡链接准确率和效率, 本文使用“检索−重排序”两阶段的方法完成实体链接任务, 如图 2 所示. 检索阶段进行候选实体生成, 从知识库 E 中选择指称 m 可能会链接到 k 个实体, 组成候选实体集合 $C(m)$. 重排序阶段进行实体消歧, 从候选实体集合中选择指称 m 真正指代的实体 e^*. 作为一般设置, 对于每个实体 $e \in C(m)$, 已知其名称和描述; 对于指称 m, 已知其上下文. 假设每个指称 m 都能从候选实体集合 $C(m)$ 中选出最相关的实体 $\hat{e}$ 作为链接实体, 本文将消歧任务建模为每个候选实体的打分问题, 同时不考虑目标实体不在知识库中的情况.

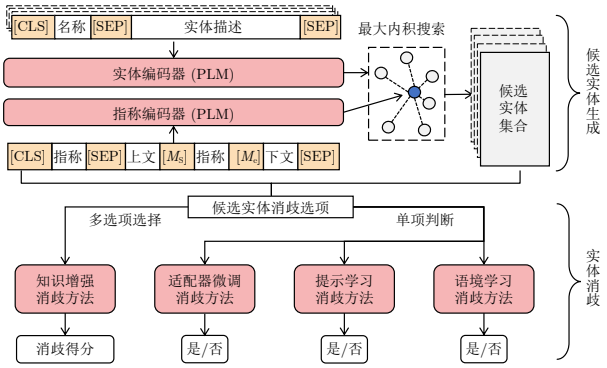


图 2　实体链接系统的整体架构

Fig. 2　The architecture of the entity linking system

1.2　候选实体生成

本文使用 Wu 等[16] 提出的双编码器进行候选实体生成, 如图 2 所示, 将指称和实体映射到相同的低维稠密向量空间, 通过最大内积搜索 (Maximum inner product search, MIPS) 实现基于向量检索的候选实体生成. 候选实体集合 $C(m)$ 的生成方法如下:

$$\begin{cases} s(m,\ e_i) = \boldsymbol{y}_m \cdot \boldsymbol{y}_{e_i} \\ C(m) = \arg \max\limits_{k,\ e_i \in E} s(m,\ e_i) \end{cases} \tag{1}$$

其中, “指称−实体”相似度得分 $s(m,\ e_i)$ 通过计算指称向量 $\boldsymbol{y}_m$ 和实体向量 $\boldsymbol{y}_{e_i}$ 的内积得到. 根据相似度得分, 从知识库中选择得分最高的 k 个实体构建候选实体集合, 其中指称向量 $\boldsymbol{y}_m$ 和实体向量 $\boldsymbol{y}_e$ 由下式给出:

$$\begin{cases} \boldsymbol{y}_m = \mathrm{red}(T_m(\tau_m)) \\ \boldsymbol{y}_e = \mathrm{red}(T_e(\tau_e)) \end{cases} \tag{2}$$

其中, T 表示可学习的 Transformer 编码器, τ 表示输入文本序列. $\mathrm{red}(\cdot)$ 是规约函数用于将输出的向量序列映射到表示向量空间, 例如在 BERT 模型中, 通常使用输入序列的首个特殊字符 [CLS] 对应的输出向量. 图 2 给出了指称和实体的输入文本建模, 不同于文献 [16, 20] 中的输入形式, 本文保持指称和实体的输入文本的对称性, 使得模型更易于在训练过程中收敛.

双编码器模型使用批次内负采样的方法进行训练, 对于随机采样的训练批次, 最大化同一批次中正确实体的分数, 在批次大小为 B 的情况下, 样本 $(m_i,\ e_i)$ 的损失函数表示为

$$\mathcal{L}(m_i,\ e_i) = -s(m_i,\ e_i) + \ln \sum_{j=1}^{B} \exp(s(m_i,\ e_j)) \tag{3}$$

1.3　实体消歧

本文研究了四种使用大语言模型进行实体消歧的方法, 分别为知识增强、适配器微调、提示学习和语境学习, 如图 2 下半部分所示. 其中知识增强方法中基于多选项选择的设置, 即对每个候选实体选项生成分数, 排序选出链接实体. 另外三种方法直接使用大语言模型进行实体消歧, 逐个判断每个候选实体, 输出“是”或“否”以及判断得分作为实体消歧依据. 上述四种实体消歧方法将在第 2 节进行具体的描述.

2　基于大语言模型的实体消歧方法

2.1　基于知识增强的实体消歧

经过大规模预训练的大语言模型积累了广泛的世界知识, 能够以文本补全或对话的形式输出信息. 其推理过程和结论作为外部知识, 可以增强基准模型解决具体问题的能力. 遵循 Wu 等[22] 提出的 CoT-KA 框架, 依赖大模型逐步推理的涌现能力, 本文采用图 3 中的思维链提示, 使用 Prompt 1 引导大语

言模型为“提及−实体”对 $(m,\ e_i)$, $e_i \in C(m)$ 生成判断其是否共指的思维链 cot (图 3 中的 CoT 给出了大语言模型输出的思维链示例), 然后将生成的思维链 cot 作为额外的知识加入到实体消歧的基准模型.

遵循先前的工作[16, 20, 23], 本文采用交叉注意力编码器作为基准模型. 如图 4(a) 左半部分所示. 具体来说, 将指称及其上下文和候选实体的文本拼接起来, 作为输入, 通过 Transformer 模型内部的交叉注意力模块进行语义融合, 获得交叉注意力特征向量, 如式 (4) 所示:

$$\boldsymbol{y}_{\mathrm{CA}} = \mathrm{red}(T_{\mathrm{CA}}(\tau_{m\text{-}e_i})) \tag{4}$$

其中, 输入文本 $\tau_{m\text{-}e_i}$ 具体表示为“带指称标记的上下文”、“指称”、“实体名称”、“实体描述”的顺序拼接, 使用特殊的字符对不同的字段进行分隔, 图 4(a) 给出了交叉注意力编码器的输入形式. T_{CA} 表示 Transformer 编码器通过交叉注意力模块进行表示向量的特征交互.

不同于 Wu 等[22] 直接将 cot 作为文本特征输入到小模型中, 考虑到小模型输入长度的限制, 本文通过额外的编码器将 cot 编码成特征向量 $\boldsymbol{y}_{\mathrm{CoT}}$, 与原本的交叉注意力模型的特征向量 $\boldsymbol{y}_{\mathrm{CA}}$ 拼接, 最后经过多层感知机 (Multilayer perceptron, MLP) 进行特征融合, 输出“指称−候选实体”的成对得分. 具体过程可用下式描述:

$$\begin{cases} cot = \mathrm{LLM}(\tau_{\mathrm{prompt}}(m,\ e_i)) \\ \boldsymbol{y}_{\mathrm{CoT}} = \mathrm{red}(T_{\mathrm{CoT}}(cot)) \\ s_{\text{CoT-KA}}(m,\ e_i) = \mathrm{MLP}(\boldsymbol{y}_{\mathrm{CA}} || \boldsymbol{y}_{\mathrm{CoT}}) \end{cases} \tag{5}$$

其中, $\boldsymbol{y}_{\mathrm{CA}}$ 由式 (4) 给出, $\tau_{\mathrm{prompt}}(\cdot)$ 表示图 3 中的提示词 Prompt 1, T_{CoT} 表示基于 Transformer 的思维链编码器.

为了优化 CoT-KA 模型, 本文使用二元交叉熵作为优化目标, 并通过平衡负样本挖掘的训练方法来缓解过拟合问题. 具体而言, 首先构建困难负样本数据集和随机负样本数据集. 由于双编码器模型检索出的候选实体相似性较高, 很多实体之间只存在细微差别, 因此将这些候选实体作为困难负样本训练集 D_{hard}. 同时随机采样 n 个实体作为负样本, 并添加上正确实体, 构建随机负样本训练集 D_{rand}. CoT-KA 模型首先在困难负样本 D_{hard} 上进行训练, 以提高模型对于相似实体的区分能力. 然后, 通过在随机负样本 D_{rand} 上训练缓解模型的过拟合问题, 使其更好地泛化到未见过的数据. 这样的训练策略可以有效地提高模型的性能和鲁棒性.

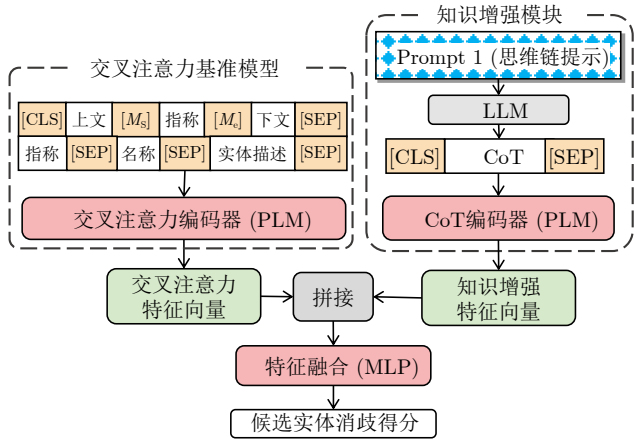
2.2 基于适配器微调的实体消歧

为了缓解大语言模型全参数微调成本高昂、灾难性遗忘等问题, Houlsby 等[24] 提出基于适配器微调的参数高效微调方法. 其在预训练神经网络之间插入一些小的神经网络模型, 称为适配器 (Adapter), 仅对适配器训练可以有效降低微调成本, 最小化对预训练参数的修改. 本文使用 Hu 等[25] 提出的 LoRA 微调方法对大语言模型进行指令微调, 以提高大模型遵循实体链接指令的能力, 如图 4(b) 所示, 以适配实体链接任务. 对大语言模型的监督微调主要包含以下三个步骤:

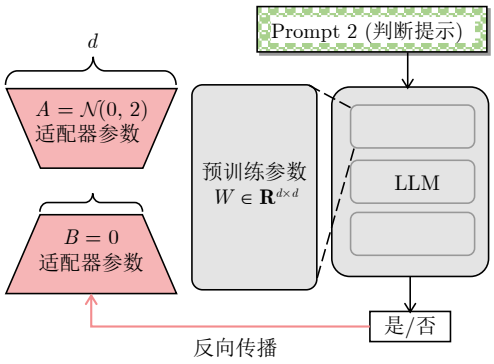
1) 构建指令格式化的微调数据集. 由于大模型

图 3 实体消歧方法中的提示语示例

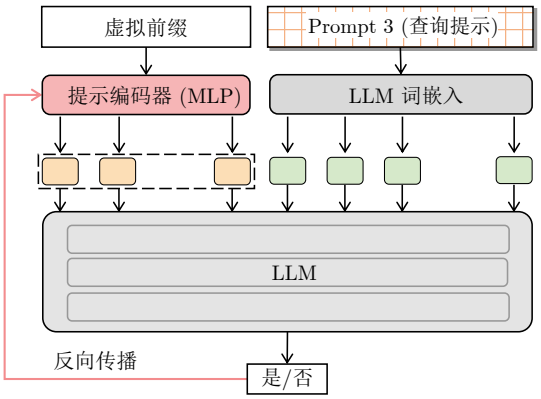
Fig. 3 Prompt examples in entity disambiguation



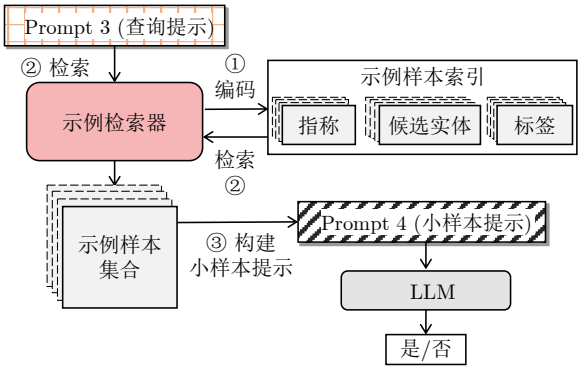
(a) 知识增强 (CoT-KA)
(a) Knowledge augmentation (CoT-KA)



(b) 适配器微调 (LoRA)
(b) Adapter fine-tuning (LoRA)



(c) 提示学习 (P-tuning)
(c) Prompt learning (P-tuning)



(d) 语境学习 (ICL)
(d) In-context learning (ICL)

图 4　基于大语言模型的四种实体消歧方法

Fig. 4　Four methods for LLM-based entity disambiguation

输入长度限制和候选集合规模之间的制约, 本文将实体消歧步骤建模为每个选项的判断问题, 即依次判断每个“指称–候选实体”对 $(m,\ e_i)$ 是否共指. 此外, 指令微调数据集中会给定一段任务描述 T 作为提示. 本文使用的输入样本示例如图 3 中的判断提示 Prompt 2 所示, 构建的微调数据集表示为

$$\{T,\ m,\ e_i,\ l_i | m \in M,\ e_i \in C(m)\} \tag{6}$$

其中, l_i 是经过标签映射器 (Verbalizer) 映射的标签词, 这里简单地使用“是”和“否”表示候选实体 e_i 是否为目标实体标签词.

$$l_i = \begin{cases} \text{是}, & e_i \text{ 是目标实体} \\ \text{否}, & \text{否则} \end{cases} \tag{7}$$

2) 低秩适配微调. 主要思想是冻结大语言模型中的参数 $W_\theta \in \mathbf{R}^{d\times d}$, 使用低秩分解矩阵 BA 模拟需要更新的参数矩阵的变化量 $W_\theta \leftarrow W_\theta + \Delta W = W_\theta + BA$, 其中, $B \in \mathbf{R}^{d\times r}$, $A \in \mathbf{R}^{r\times d}$. 由于一般设置 $r \ll d$, 因此可以大大降低训练的存储成本.

大语言模型微调的优化目标是标签词 l 在词表上的交叉熵损失, 以自回归语言模型为例, 输入样本表示为 $z = [x; y]$, 其中, x 表示模型已知信息 $x = T(m,\ e_i)$, 通过任务提示 T 将指称和候选实体信息组织起来, 其索引序列表示为 X_{idx}; $y = l_i$ 表示目标输出, 其索引序列表示为 Y_{idx}. 语言模型 LM_θ 表示为 $p_\theta(y \mid x)$, 其优化目标表示为

$$\max_\theta \ln p_\theta(y \mid x) = \max_\theta \sum_{i \in Y_{\text{idx}}} \ln p_\phi\left(z_i \mid h_{<i}\right) \tag{8}$$

3) 后处理. 由于候选实体集合中相似实体的存在, 大语言模型的推理结果在同一个候选实体集合中会出现冲突和重复的问题. 例如将多个相似的实体判断为链接实体, 或全部判断为非链接实体. 本文通过后处理解决上述问题. 对于候选实体 e_i 和指称 m 是否共指, 微调后的大语言模型输出“是”和

“否”以及该标签词 l_i 的得分 $p(l_i)$, 定义相关性得分 s_i 如下, 选择相关性得分最高的作为链接实体, 即

$$s_i = \begin{cases} 1 + p(l_i), & l_i = \text{“是”} \\ 1 - p(l_i), & l_i = \text{“否”} \end{cases} \tag{9}$$

2.3 基于提示学习的实体消歧

与调整大语言模型参数的适配器微调方法不同, 提示学习不改变模型参数, 而是通过学习提示词的向量表示实现对模型在特定任务或领域上的适配. 在提示学习方法中, 本文使用 P-tuning[26] 的方法对大语言模型进行微调. 如图 4(c) 所示, P-tuning 方法在 Transformer 的输入层中添加连续的提示向量序列作为输入序列前缀, 与任务相关的查询序列一同输入到大语言模型中. P-tuning 的虚拟前缀提示通过基于 MLP 的提示编码器 $PromptEncoder$ 进行重参数化, 这使得 P-tuning 方法能在保持参数高效性的同时提高微调的参数量, 同时使提示向量的学习更加稳定. 具体而言, P-tuning 方法中将需要学习的提示向量表示为向量序列 $[\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_l]$, 其中 l 为虚拟前缀长度, $\boldsymbol{p}_i = PromptEncoder(i) \in \mathbf{R}^d$, d 为模型中文本嵌入层的向量维度, 提示向量的索引序列表示为 P_{idx}. 此时模型的输入序列 I 表示为

$$I = \begin{cases} PromptEncoder(i), & i \in P_{\text{idx}} \\ Emb(w_i), & \text{否则} \end{cases} \tag{10}$$

其中, $Emb(w_i)$ 表示大语言模型的词嵌入层, w_i 表示输出文本的第 i 个词元. 训练时, 使用类似于适配器微调的指令格式化数据集, 区别在于使用虚拟前缀替代任务描述, 图 3 中的查询提示 Prompt 3 提供了提示学习输入的示例. 提示学习方法在训练过程中保持大语言模型参数不变, 使用式 (8) 作为优化目标训练提示编码器参数.

2.4 基于语境学习的实体消歧

语境学习方法依赖大语言模型的语境学习能力, 其核心思想是通过构建合适的演示示例来引导大语言模型解决特定任务. 根据文献 [6], 本文在给定任务描述 T 和相关演示示例 D_k 的情境下, 将它们组织成自然语言提示, 同时提供新问题 x 作为大语言模型的输入. 大语言模型在不进行任何参数更新的情况下, 通过模仿示例的输出来进行推理, 输出对新问题 x 的答案. 对于实体消歧任务, 图 4(d) 展示了语境学习方法的流程. 这个方法主要包含以下三个步骤:

1) 示例样本索引构建. 本文方法中演示示例包含“上下文”、“指称”、“候选实体”、“实体描述”等文本特征, 以及该示例的标签 l_i, 由式 (7) 给出. 图 3 中的查询提示 Prompt 3 给出演示示例的文本特征. 将演示示例形式化表示为 $d_i = [Prompt(m,\ e_i); l_i]$, 使用检索器为演示示例的文本特征构建索引, 索引的键值对为

$$\begin{cases} Key = Tokenize(Prompt(m,\ e_i)) \\ Value = [Prompt(m,\ e_i); l_i] \end{cases} \tag{11}$$

2) 演示示例检索. 根据文献 [27] 的观点, 演示示例的选择对大语言模型预测准确率有显著影响, 因此需要设计适当的策略构建演示示例. 本文采用三种不同的选择策略: 随机选择、稀疏检索 (BM25)、密集检索 (Sentence-bert, SBERT[28]). 这三种检索策略在给定新的待消歧查询 x 时, 通过检索器检索出最相关的演示示例集合 D_k.

3) 语境学习. 将待消歧查询 x 和检索出的演示示例 D_k 整合, 构建如图 3 中的 Prompt 4 所示的小样本提示, 将其输入到大语言模型, 引导其生成对应的推理结果. 对大语言模型的输出, 采用与 LoRA 和 P-tuning 相同的后处理步骤.

3 实验

3.1 数据

1) 知识库. 本文选用 Wikidata[2] 作为知识库, 并遵循 Xu 等[20] 的设置, 将 Wikidata 实体划分为 E_{known} 和 E_{new} 两组, 以真实反映知识库随时间演变的情况. 其中, E_{known} 来自 2018 年 8 月 13 日的 Wikidata 快照, 用于训练和验证时使用的知识库; E_{new} 包含 2018 年 8 月 13 日到 2021 年 3 月 15 日之间新增的 Wikidata 实体, 用于零样本设置. 对于中文任务, 本文将实体限定为出现在中文维基百科中的实体, 最终 E_{known} 中包含约 1 M 个实体, E_{new} 中包含约 57 K 个实体.

2) 数据集. 本文在 Hansel[20] 和 CLEEK[21] 数据集上进行实验. Hansel 的训练集 Hansel-Train 用于训练, 验证集 Hansel-Dev 用于验证, 三个测试集 Hansel-FS、Hansel-ZS 和 CLEEK 用于评测. 为了评测实体链接中存在的流行度偏差问题, Hansel-FS 主要包括流行度较低的尾部实体, Hansel-ZS 为训练和验证集中未登录的实体. CLEEK 是一个包含 100 篇不同难度设置的长文本数据集. 表 1 展示了 Hansel 和 CLEEK 数据集的统计信息. 为简化问题设置, 本文专注于 In-KB 的指称.

[2] https://dumps.wikimedia.org/wikidatawiki/

表 1　Hansel 和 CLEEK 数据集的统计信息
Table 1　Statistics of the Hansel and CLEEK datasets

数据集	# 指称	# 文档	# 实体		
			E_{known}	E_{new}	总计
Hansel-Train	9.89 M	1.05 M	541 K	—	541 K
Hansel-Dev	9 677	1 000	6 323	—	6 323
Hansel-FS	3 404	3 389	2 720	—	2 720
Hansel-ZS	4 208	4 200	1 054	2 992	4 046
CLEEK	2 412	100	1 100	—	1 100

3.2　实验设置

本文选择 ChatGLM3-6B[3] 系列模型和 Qwen-7B[4] 作为实体消歧任务中使用的基座大语言模型. 本文实验使用 Bert-base-Chinese 作为预训练的 Transformer 编码器. 如未加说明, 本文的实验在单张 RTX3090 上进行.

在候选实体生成阶段, 本文使用 AdamW 优化器训练双编码器模型进行候选实体生成, 学习率设为 1×10^{-5}, 预热比例为 10%, 训练批次大小为 64, 在 Hansel 训练集上训练 10 K 步. 指称上下文和候选实体描述的最大长度分别设为 64 和 128. 为说明候选生成模型的有效性, 本文使用下面两个基准模型进行比较, 在测试时选择的知识库为 $E=E_{\text{known}}\cup E_{\text{new}}$.

1) 别名表 (Alias table, AT). 遵循 Xu 等[20] 的工作, 使用由解析维基百科内部链接、重定向和页面标题获取的别名表生成指称 m 到实体 e 的先验概率 $p(e|m)$, 从而生成别名.

2) BM25. 基于 Gensim 代码库[5] 的 OkapiBM25 模型实现, 将指称和实体标题作为检索特征, 进行稀疏检索.

在实体消歧阶段, 为探究使用大语言模型增强实体链接系统的有效性, 本文在双编码器模型生成的 Top-10 候选实体集合上进行实体消歧, 并与六个中文实体链接的基线模型进行比较.

1) TyDE[20] 基于 Wikidata 类别体系, 将类别预测作为辅助监督任务, 用于增强实体链接.

2) Oops![29] 提出了一个从粗到细的基于词汇的检索器来分两层检索候选实体. 通过改进候选实体的召回率实现了最好的效果.

3) ITNLP[30] 同样使用两阶段算法, 第一阶段使用别名表生成候选实体, 第二阶段使用基于 ERNIE 的交叉注意力编码器进行细粒度交互.

4) YNU-HPCC[31] 使用 ElasticSearch 进行候选实体检索, 通过 Sentence-BERT[28] 建模指称和候选实体的相似度.

5) CA-tuned[20] 模型使用交叉注意力编码器对候选实体重排序, 本文重新训练了 CA 模型. 通过平衡负样本采样策略, 依次在 50 K 困难负样本 (双编码器模型生成的 Top-10 候选实体) 和 50 K 随机负样本 (随机检索生成的 Top-10 候选实体) 上进行训练. 模型输入长度限制为 256. 使用 AdamW 优化器先以 1×10^{-4} 的学习率在困难负样本上训练 3 轮, 然后以 1×10^{-5} 的学习率在随机负样本上训练 1 轮, 训练时批次大小为 6, 梯度累积 5 步.

6) mGENRE[32] 模型是当前最先进的实体链接模型 GENRE 的多语言版本, 采用端到端的文本生成的方法直接生成实体名称. 本文使用公开的模型参数检查点[6] 进行测试.

使用大语言模型的消歧模型实现细节如下:

1) 知识增强 (CoT-KA). CoT-KA 模型与 CA 模型训练相似, 不同之处在于使用大语言模型生成每个“指称–实体”对的思维链. 思维链编码器最大输入长度为 128. 在困难负样本上进行 4 轮次训练后收敛, 然后再通过随机负样本进行 1 轮次训练.

2) 适配器微调 (LoRA). 本文使用 PEFT 代码库[7] 对大语言模型进行 LoRA 微调, 在微调 ChatGLM3-6B 时, 更新的参数矩阵为 *query_key_value*; 在微调 Qwen-7B 时, 更新的参数矩阵包括 *c_attn*, *c_proj*, *w1*, *w2*. 更新矩阵的秩 r 设置为 64, 学习率设为 1×10^{-4}. 大语言模型在 Top-4 候选实体数据上进行微调, 批次大小为 1, 梯度累积 20 步更新一次参数, 使用 AdamW 优化器进行了 5 000 步参数更新.

3) 提示学习 (P-tuning). 同样由 PEFT 实现, 前缀长度设为 20, 梯度累积 20 步, 采用 AdamW 优化器进行 5 000 步的训练. 训练样本同样来源于双编码器模型检索生成的 Top-4 候选实体. 在微调 ChatGLM3-6B 时, 学习率设为 1×10^{-4}; 在微调 Qwen-7B 时, 学习率设为 1×10^{-2}.

4) 语境学习 (ICL). 在该实验中, 本文随机选择 50 K 个指称及其 Top-4 候选实体作为演示示例集合, 通过随机检索, BM25 检索和基于 SBERT 的密集检索各自检索出 10 个演示示例. 为保持小样本学习的设置, SBERT 使用预训练参数[8], 并且没有针对当前数据进行微调.

[3] https://github.com/THUDM/ChatGLM3
[4] https://github.com/QwenLM/Qwen
[5] https://github.com/piskvorky/gensim
[6] https://github.com/facebookresearch/GENRE
[7] https://github.com/huggingface/peft
[8] https://huggingface.co/DMetaSoul/sbert-chinese-general-v2

3.3 实验结果及分析

3.3.1 大语言模型与小模型整体性能对比

在候选实体生成实验中, 本文采用 Recall@k 指标对候选实体生成模型进行评估, 其中, k 分别取 1, 10, 100. 表 2 展示了候选实体生成模型在 Hansel 和 CLEEK 上的检索效果. 实验结果表明, 双编码器模型检索效果优于其他方法. 具体而言, 双编码器模型 (DE) 在 Top-10 候选实体设置下, 相对于别名表 AT 在 Hansel-FS 上提升了 20%, 相对于 BM25 方法在 Hansel-ZS 上提高了 9.1%, 相对于别名表 AT 在 CLEEK 上提升了 3.5%. 此外由于该模型不依赖别名表等外部资源, 因此在处理新实体时展现出较好的性能.

表 3 和表 4 展示了实体链接系统在 Hansel 和 CLEEK 上的链接准确率. 其中, $^\diamond$ 表示模型基于别名表 AT 生成候选实体; † 表示端到端的模型, 即不生成候选直接链接实体; ‡ 表示使用稀疏检索的方法生成候选实体; 其他模型都基于双编码器模型检索的 Top-10 候选实体.

1) 方法扩展性. 表 3 显示, 小模型方法在 Hansel-FS 和 Hansel-ZS 数据集上的 CA-tuned 最优准确率分别为 49.9% 和 83.5%, 在 CLEEK 上 mGENRE 最优为 73.7%; 表 4 显示, 大模型监督微调方法中, LoRA 微调表现最佳, 分别在 Hansel-FS、Hansel-ZS 和 CLEEK 上达到 61.7%, 87.6% 和 77.4%, 相比小模型最优方法提升了 11.8%, 4.1% 和 3.7%. 而性能最低的大模型方法 CoT-KA 在三个数据集上的成绩分别为 53.8%, 84.2% 和 74.3%, 相比小模型最优方法提升了 3.9%, 0.7% 和 0.6%. 大模型无监督的语境学习方法也接近 CA-tuned 模型.

2) 模型扩展性. 对比 Qwen-7B/ChatGLM3-6B 两个大语言模型和小模型基准的链接准确率发现, 大语言模型的应用方法在两者上均表现出有效性. 除语境学习方法外, 各种方法中 Qwen-7B 模型的实验结果整体上优于 ChatGLM3-6B, 对于语境学习方法, ChatGLM3-6B 模型在三个测试基准上都优于 Qwen-7B 模型.

3.3.2 中文实体链接任务中的“过拟合”问题

表 3 中, 基于 BERT、ERNIE 编码器的链接性能表明, 尽管交叉编码器 CA 的特征交互比双编码

表 2 候选实体生成模型在 Hansel 和 CLEEK 数据集上的召回率 (%)
Table 2 Recall of candidate entity generation model on Hansel and CLEEK (%)

方法	Hansel-FS			Hansel-ZS			CLEEK		
	R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
AT	0	61.1	63.0	70.6	78.5	78.8	**69.4**	77.8	79.1
BM25	13.1	41.9	71.1	69.7	84.1	90.9	34.9	46.8	57.2
DE	**46.8**	**81.1**	**92.6**	**78.2**	**93.2**	**97.2**	58.7	**81.3**	**92.2**

注: 加粗字体表示各列最优结果.

表 3 实体链接基线方法在 Hansel 和 CLEEK 数据集上的准确率 (%)
Table 3 Accuracy of entity linking baseline methods on Hansel and CLEEK (%)

数据集	TyDE$^\diamond$	Oops!$^{\diamond\ddagger}$	ITNLP$^\diamond$	YNU-HPCC‡	CA$^\diamond$	DE	CA-tuned	mGENRE†
Hansel-FS	11.7	44.6	30.7	21.1	46.2	46.8	**49.9**	36.6
Hansel-ZS	71.6	81.6	81.7	73.6	76.6	78.2	**83.5**	68.4
CLEEK	—	—	—	—	—	58.7	70.5	**73.7**

注: 加粗字体表示各行最优结果.

表 4 实体链接大语言模型方法在 Hansel 和 CLEEK 数据集上的准确率 (%)
Table 4 Accuracy of entity linking LLM methods on Hansel and CLEEK (%)

数据集	Qwen-7B				ChatGLM3-6B			
	CoT-KA	LoRA	P-tuning	ICL	CoT-KA	LoRA	P-tuning	ICL
Hansel-FS	53.8	**61.7**	56.7	49.2	52.5	51.6	47.2	50.6
Hansel-ZS	83.3	**87.6**	85.9	79.5	84.2	85.4	83.6	82.5
CLEEK	74.3	**77.4**	74.9	66.6	71.4	72.9	67.2	67.5

注: 加粗字体表示各行最优结果.

器 DE 更充分, 表达能力更强, 但其准确率低于 DE 模型. 同样对比本文微调的 DE 模型, 发现其效果 (46.8%) 在 Hansel-FS 数据集上也优于以往使用更复杂结构的小模型编码器的最好结果 (44.6%). 实际上, 本文只是通过约束 DE 模型的输入文本特征长度, 以及保持指称和实体文本输入的对称性实现的当前效果, 这本质上是一种正则化方法. 以上结果表明 DE 模型和 CA 模型及其变体在中文实体链接上存在一定程度"过拟合"问题.

为了解决小模型编码器方法的过拟合问题, 一方面改变模型的训练策略, 设计了平衡负采样策略, 这部分在第 3.4.4 节的消融实验中进行了详细分析; 另一方面研究分析了大语言模型在这一问题上的效果, 从表 4 展示的结果上可以发现, 经过充分预训练的大语言模型在参数高效的微调方法下, 没有出现性能下降至 DE 模型之下的情况, 这初步说明相比于小模型, 基于 LoRA 和 P-tuning 微调的大语言模型方法对过拟合问题有一定的缓解. 第 3.3.3 节将对这一问题进行更深入的分析.

3.3.3　大语言模型监督微调方法对比

表 4 的结果证明了基于适配器微调和提示学习的大模型监督指令微调方法的优势. 本文进一步使用 Qwen-7B 模型探讨了不同指令微调方法在实体链接任务上的表现. 如表 5 所示, 本文比较了 P-tuning、LoRA、AdaLoRA[33] 和全参数微调 (FT) 四种方法. 其中全参数微调使用 8×V100 进行实验, AdaLoRA 与 FT 方法都使用与 LoRA 微调相同的数据和参数设置. 结果显示, LoRA 微调的效果优于其他三种方法, 其中全参数微调的效果最差. 这是因为大语言模型在全参数微调过程中分类损失迅速下降, 显示出强大的拟合能力, 但在本研究的分类设置下, 全参数微调容易过拟合. 类似地, 表 3 展示的 DE 和 CA 模型效果也验证了上述发现. 相比之下, P-tuning、LoRA、AdaLoRA 三种方法都冻结了模型参数, 仅微调少量的额外参数相当于对模型进行了正则化, 因此没有出现明显的性能下降.

表 5　监督微调方法对准确率的影响 (%)
Table 5　Impact of supervised fine-tuning on accuracy (%)

微调方法	训练参数量	Top-1 准确率		
		Hansel-FS	Hansel-ZS	CLEEK
P-tuning	10 M	56.7	85.9	74.9
AdaLoRA	27 M	60.4	87.3	**77.5**
LoRA	286 M	**61.7**	**87.6**	77.4
FT	7 B	53.9	85.7	73.7

注: 加粗字体表示在不同数据集上的最优结果.

3.3.4　大语言模型无监督方法对比

为了研究大语言模型在无监督情况下的实体链接能力, 本文测试了四种不同的大语言模型: Qwen-7B、Qwen-14B、Qwen-72B 和 ChatGPT, 并对比了它们在语境学习 (ICL)、思维链推理 (CoT) 和思维链知识增强 (CoT-KA) 三种无监督方法下的表现, 如表 6 所示. 结果显示, 随着模型参数量的增加, 无监督推理能力也呈现出增长趋势, 其中 Qwen-72B 在所有方法中表现最佳. 并且在大部分场景中大语言模型在 ICL 到 CoT 以及 CoT 到 CoT-KA 的转变中都表现出了性能提升.

表 6　大语言模型的无监督推理能力 (%)
Table 6　Unsupervised reasoning capabilities of LLMs (%)

方法	大语言模型	Top-1 准确率		
		Hansel-FS	Hansel-ZS	CLEEK
ICL	Qwen-7B	49.2	79.5	66.6
	Qwen-14B	51.8	81.5	64.1
	Qwen-72B	**62.2**	**86.8**	**74.8**
	ChatGPT	52.7	79.4	66.4
CoT	Qwen-7B	49.8	78.9	67.6
	Qwen-14B	58.4	83.0	71.6
	Qwen-72B	**63.1**	**85.6**	**75.0**
	ChatGPT	55.4	78.7	67.9
CoT-KA	Qwen-7B	53.8	83.3	74.3
	Qwen-14B	60.1	86.3	75.6
	Qwen-72B	**61.8**	**87.2**	**77.4**
	ChatGPT	58.3	85.0	75.2

注: 加粗字体表示各组方法在不同数据集上的最优结果.

在长尾实体 (Hansel-FS) 场景下, 小模型最优准确率为 49.9%, Qwen-7B 监督方法提升至 61.7% (+11.8%), 而 Qwen-72B 无监督 CoT 方法进一步提升至 63.1%, 相比小模型提升 13.2%, 相比 Qwen-7B 提升 1.4%, 表明 Qwen 在长尾实体场景下具有显著优势, 且随着模型规模扩大, 优势更加明显.

在未登录实体 (Hansel-ZS) 场景下, 小模型最优准确率为 83.5%, Qwen-7B 监督方法提升至 87.6%, 而 Qwen-72B 无监督 CoT 方法为 85.6%, 相比小模型提升 2.1%, 但较 Qwen-7B 下降 2.0%.

在一般长文本场景 (CLEEK) 中, 小模型最优准确率为 73.7%, Qwen-7B 监督方法提升至 77.4%, 而 Qwen-72B 无监督 CoT 方法为 75.0%. 相比小模型提升 1.3%, 但较 Qwen-7B 下降 2.4%.

综上, 当模型规模从 7B 扩大到 72B 时, Qwen-

72B 无监督方法在性能上与 Qwen-7B 监督方法接近, 差异范围为-2.4%~+1.4.

3.4 消融实验

3.4.1 不同秩的大小对 LoRA 微调方法的影响

表 7 展示了在 Qwen-7B 模型下, LoRA 方法在设置不同大小的秩时的链接准确率. 实验结果显示, 随着秩 r 的增加, LoRA 方法在三个测试基准上的准确率逐渐提升, 但边际收益逐渐减小. 当 r 达到 64 时, 模型性能趋于稳定. 这一结果表明, 参数更新矩阵 ΔW 应该具有较小的内在秩, 过大的秩可能不会带来更多有效信息. 值得注意的是, 本文确定的秩 $r = 64$ 大于 Hu 等[25] 方法在 WikiSQL 和 MultiNLI 测试基准上测试得出的秩 $r = 4$, 这表明对于 LoRA 方法, 不同任务可能对应不同最佳秩的大小.

表 7 适配器的秩对准确率的影响 (%)
Table 7 Impact of adapter rank on accuracy (%)

秩	Top-1 准确率		
	Hansel-FS	Hansel-ZS	CLEEK
$r = 1$	53.8	85.4	72.4
$r = 2$	52.9	85.1	71.3
$r = 4$	54.5	86.0	73.0
$r = 8$	58.8	87.4	76.5
$r = 64$	61.4	87.3	**77.4**
$r = 128$	**61.7**	**87.6**	**77.4**

注: 加粗字体表示在不同数据集上的最优结果.

3.4.2 提示长度对提示学习方法的影响

表 8 展示了对于 Qwen-7B 模型使用 P-tuning 方法, 在设置不同提示长度时的链接准确率. 对于三个测试基准, 模型在提示长度为 20 时达到了最佳性能. 随着提示长度的增加, 模型的链接准确率波动降低. 通过观察训练过程, 发现随着提示长度的增加, 模型训练损失下降得更快. 这表明更长的提示长度对应更多的训练参数, 对大语言模型的影响能力更大, 更容易导致训练数据过度拟合. 因此, 在不同的任务中, 选择适当的提示长度是必要的, 以限制虚拟提示对模型的影响.

3.4.3 示例选择策略对语境学习方法的影响

本文测试了三种不同的示例选择策略以及演示数量的效果, 如图 5 所示. 结果显示: 1) 对于示例选择策略, 基于 SBERT 的密集检索在 CLEEK 和 Hansel-FS 数据集上优于其他的示例选择策略, 在 Hansel-ZS 数据集上也相对较好; 2) 对于演示示例数量, 不同的示例选择策略呈现出一致的趋势, 但在不同数据集和不同大语言模型上表现出一定的差异. 对于 Qwen-7B 模型, 在 Hansel-ZS 和 CLEEK 数据集上, 不指定演示示例能够达到最高的准确率; 而对于 ChatGLM3-6B 模型, 2 个演示示例足以得到最准确的输出. 与之前 Liu 等[27] 方法在生成式问答任务上的研究结论不同, 本文发现相比于示例选择策略, 示例数量对实验结果具有更显著的影响. 当上下文示例数量为 0 (Qwen-7B)/2 (ChatGLM3-6B) 时, 三种示例选择方法均能达到最高准确率.

表 8 虚拟提示长度对准确率的影响 (%)
Table 8 Impact of virtual prompt length on accuracy (%)

提示长度	Top-1 准确率		
	Hansel-FS	Hansel-ZS	CLEEK
10	54.6	85.7	74.1
20	**56.7**	**85.9**	**74.9**
40	55.4	85.6	74.6
60	51.4	85.1	71.3
80	53.0	85.2	72.4

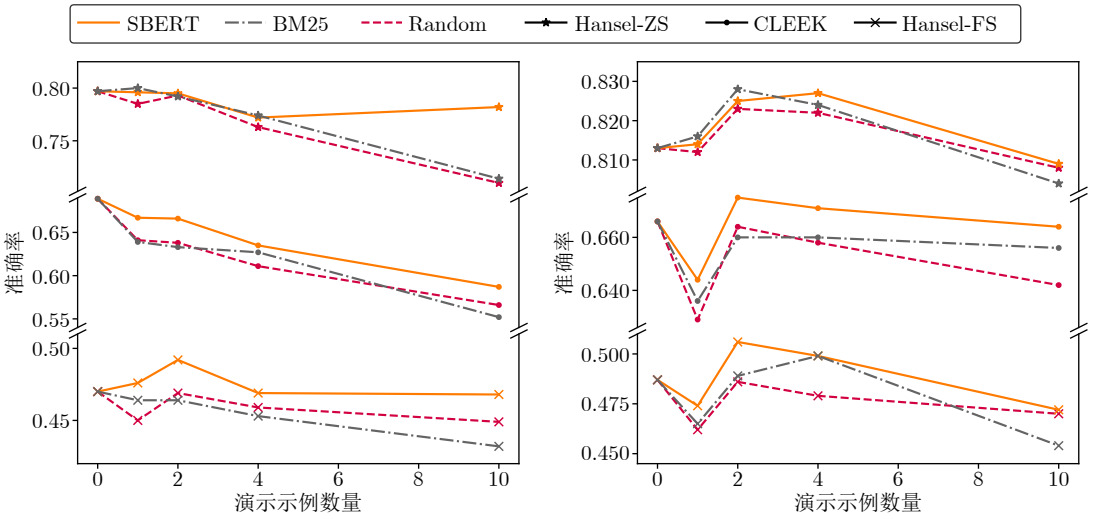
注: 加粗字体表示在不同数据集上的最优结果.

表 9 展示了示例数量为 2 时三种示例选择方法的准确率. 从整体上看, ChatGLM3-6B 模型的上下文学习能力强于 Qwen-7B 模型, 在 Hansel-FS 和 CLEEK 测试集上基于 SBERT 的密集检索策略效果最好, 在 Hansel-ZS 测试集上基于 BM25 的密集检索策略效果最好.

不同的示例选择策略在不同数据分布的数据集上表现不一致, 表明单一的示例选择策略难以适应复杂的实体链接环境. 在复杂场景下的实体链接中, 仅依赖字面形式或语义相似度检索演示示例效果不佳. 因此, 可以训练更强大的示例检索器, 根据问题类型和大语言模型的推理偏好, 为大语言模型检索更有价值的上下文示例, 从而提高实体链接的性能.

3.4.4 平衡负采样策略对知识增强方法的影响

表 10 展示了平衡负采样策略在知识增强方法中对实体消歧模型准确率的影响. 实验表明, 大语言模型思维链知识的引入, 提升了小模型的实体消歧性能, 引入平衡负采样策略则进一步提升了知识增强对小模型能力的正面影响. 具体分析如下: Qwen-7B 模型生成的思维链对基准模型的提升更大, 在 Hansel-FS 上提升了 5.2%, 在 CLEEK 上提升了 4.0%, 优于 ChatGLM3-6B 在 Hansel-FS 上的 1.7% 和 CLEEK 上的 0.8%, 这进一步表明知识增强方法在大语言模型基础能力上有良好的扩展性, 当使用更



(a) Qwen-7B　　(b) ChatGLM3-6B

图 5　示例数量和选择策略对语境学习的影响

Fig. 5　The impact of examples number and selection strategies on in-context learning method

表 9　不同示例选择策略下的语境学习的准确率 (%)

Table 9　Accuracy of ICL under different example selection strategies (%)

模型	选择策略	Top-1 准确率		
		Hansel-FS	Hansel-ZS	CLEEK
Qwen-7B	Random	46.9	79.3	63.8
	BM25	46.4	79.2	63.3
	SBERT	<u>49.2</u>	<u>79.5</u>	<u>66.6</u>
ChatGLM3-6B	Random	48.6	82.3	66.4
	BM25	48.9	**82.8**	66.0
	SBERT	**50.6**	82.5	**67.5**

注: 加粗字体表示各列最优结果; 下划线字体表示各列次优结果.

强大的大语言模型生成思维链知识时, 能更大程度地提高小模型的消歧能力.

平衡负采样策略能进一步提升小模型利用增强知识的能力. 相比单独的知识增强方法, 平衡负采样策略在 Hansel-FS 上提高 1.9% ~ 4.1%, 在 Hansel-ZS 上提高 0.8% ~ 1.7%, 在 CLEEK 上提高 0.2% ~ 0.5%. 原因如下: 首先, 表 2 中的召回率表明双编码器模型检索的候选实体在语义上与指称更相关, 导致消歧模型在偏差数据上过度拟合困难负样本. 相比之下, 随机负样本通常位于分类边界较远的类别内部, 不太可能引起分类边界的显著变化, 从而稳定分类边界, 缓解困难样本的噪声问题. 通过对随机负样本的再训练, 模型能更好地捕捉类别的典型特征, 有助于泛化到整个负样本分布, 降低过拟合风险. 对于知识增强的困难负样本, 其特征偏差更为显著, 因此平衡负采样策略能进一步提升模型性能.

3.5　错误分析

3.5.1　错误类型总结和统计

为进一步探究大语言模型在实体链接问题上的不足, 本文对四种大语言模型应用方法以及 DE,

表 10　知识增强方法的消融实验 (%)

Table 10　The ablation study on knowledge augmentation (%)

方法	Top-1 准确率		
	Hansel-FS	Hansel-ZS	CLEEK
CA	46.7	82.7	70.1
CA+平衡负采样	49.9 + 3.2	83.5 + 0.8	70.5 + 0.4
CA+知识增强 (Qwen-7B)	51.9 + 5.2	82.5 − 0.2	<u>74.1</u> + 4.0
CA+知识增强 (ChatGLM3-6B)	48.4 + 1.7	82.5 − 0.2	70.9 + 0.8
CA+平衡负采样+知识增强 (Qwen-7B)	**53.8** + 7.1	<u>83.3</u> + 0.6	**74.3** + 4.2
CA+平衡负采样+知识增强 (ChatGLM3-6B)	<u>52.5</u> + 5.8	**84.2** + 1.5	71.4 + 1.3

注: 加粗字体表示各列最优结果; 下划线字体表示各列次优结果.

mGENRE, CA-tuned 三个基准模型进行错误分析. 对于每种方法, 分别从 Hansel-FS, Hansel-ZS, CLEEK 三个测试基准的错误样本中随机采样, 最终构建了包含 1 320 个标注样本的错误样本集. 不考虑目标实体不在候选实体集合中的情况, 本文分析了 6 种主要的错误类型, 其覆盖了约 92% 的错误样本. 表 11 展示了这 6 种错误类型所占的比例. 下面对这些错误类型进行说明和分析.

1) 类别错误. 模型预测的实体与正确实体的类型不一致, 涉及借喻、比喻或是同名的两类事物等情况. 在 CA 基准模型中, 这类错误占比为 35%, 使用大语言模型的消歧方法有效地降低了该类错误的比例, 由表 11 统计可得大语言模型方法的类别错误占比平均为 25%, 其中知识增强方法在纠正这类错误时表现最好, 错误占比减少了约 13%. 这启发在知识增强方法中, 引导大语言模型生成更多类别相关的知识.

2) 粒度错误. 由表 11 统计可得大语言模型方法的粒度错误平均占比为 36%, 较 CA 基准模型的 28% 高出 7.6%, 表明其在指称粒度判断上存在不足. 具体统计错误类型包括: 特指实体预测为泛指实体 (53%)、泛指实体预测为特指实体 (16%)、整体预测为部分 (16%) 以及部分预测为整体 (15%).

3) 全局错误. 指需要考虑全局上下文信息才能预测正确的样本, 具体包括: a) 指代错误 (28%), 模型没有正确判断指称所指代的上下文内容或相关语义角色; b) 主题错误 (72%), 模型未能理解文章主题, 从而选择与文章主题无关而仅与指称相关的实体. 基于大语言模型的消歧方法的全局错误占比相比 CA 基准模型增加了 6%. 这表明大语言模型在全局实体链接方面仍有进一步提高性能的潜力.

4) 局部错误. 指称内部的信息与预测实体不一致. LoRA 微调方法和 Qwen-7B 模型在匹配局部信息方面效果最好, 相比 CA 基准模型降低了 3% 的错误比例. ChatGLM3-6B 该错误类型占比较多.

5) 时间错误. 预测实体的时间属性与上下文不一致. 此类错误占比约为 7%, 其中 ChatGLM3-6B 在该问题较优, 且基于提示微调和语境学习的方法在纠正时间错误的问题上表现更好, 优于 CA 基准模型.

6) 地点错误. 预测实体的地点信息与上下文不一致. 基于大语言模型的消歧方法整体优于 CA 基准模型, 此类错误占比约为 7%, 降低了 4%, 其中基于 LoRA 微调和提示微调和语境学习的方法最好.

3.5.2 典型错误样例分析

表 12 展示了错误样例. 针对类别和粒度两种占比较多的错误类型进行定性分析如下.

1) 类别错误. 在“除法兰西民族外, 边境地区还有阿尔萨斯, [科西嘉], 佛兰芒等少数民族, 大约占了总人口的 7.9%.”例子中, 指称“科西嘉”指代的实体是“科西嘉人”, 但 LoRA 微调的 Qwen-7B 模型错误地预测为“科西嘉岛”. 模型未能关注到“等少数民族”这样的同位语表述, 也未正确识别上下文的“民族”主题, 导致错误链接到更流行的实体.

2) 粒度错误. 在“跳水梦之队的两对 [世锦赛] 冠军组合都捍卫了自己的荣誉, 何冲–王峰赢得男子双人 3 米板金牌, 郭晶晶–李婷夺得女子双人 3 米板桂冠.”例子中, 正确实体为“世界游泳锦标赛”, 但模型错误链接到“世界锦标赛”. 模型未能从“跳水”及相关运动员的名字中推断出更细粒度的实体. 主要原因如下: a) 粗粒度实体在数据中更流行. 这导致“特指实体预测为泛指实体”的错误类型占比高达 56%. b) 分类损失函数难以区分实体粒度. 例如, 对于“世界游泳锦标赛”, “世界锦标赛”和“NBA”是同等负标签, 但“世界锦标赛”更接近正确答案. 相比之下, 通过对比损失进行训练的 DE 模型的粒度错误占比略低, 支持了这一分析.

3.5.3 未来改进方向

综合实验结果和错误分析表明, 引入大语言模型应用方法整体上提高了中文实体链接模型的准确率, 但仍存在类别预测和粒度推断的问题. 实例分析显示, 无论是类别错误还是粒度错误, 实体链接模型的性能都受到长尾问题的限制. 这些分析表明, 长尾实体问题是实体链接任务中的关键问题. 可以

表 11 六种错误类型在不同方法中所占比例 (%)

Table 11 Proportions of 6 error types across different methods (%)

错误类型	DE	mGENRE	CA-tuned	CoT-KA		LoRA		P-tuning		ICL	
				Qwen-7B	ChatGLM3-6B	Qwen-7B	ChatGLM3-6B	Qwen-7B	ChatGLM3-6B	Qwen-7B	ChatGLM3-6B
类别	24	26	35	22	21	31	29	30	21	25	19
粒度	23	32	28	36	33	41	29	42	36	36	33
全局	30	10	11	16	18	11	20	14	19	16	24
局部	4	14	6	8	11	3	10	6	14	9	11
时间	8	8	10	11	8	11	6	5	4	5	5
地点	11	9	11	7	9	4	7	4	6	9	8

表 12　错误样例: 包含上下文、预测实体、正确实体三个信息 (中括号内的内容表示指称)

Table 12　Error cases: Contains context, predicted entity, and correct entity, with content in brackets indicating mention

错误种类		上下文	预测实体	正确实体
类别		由猫腻的同名小说改编而成的 [《将夜》], 一播出就引起了网友们的关注.	将夜 (小说): 《将夜》为网络作家猫腻发布于起点中文网的玄幻网络小说.	将夜 (网络剧): Ever Night, 2018 年播出的玄幻古装剧.
粒度	“特指”预测为“泛指”	苹果提交的“通过动态属性而达到的 3D [用户界面] 显示效果”的专利就曾披露出其对眼部追踪技术的兴趣.	用户界面: User Interface, 简称 UI, 是系统和用户之间进行交互和信息交换的媒介, 它实现信息的内部形式与人类可以接受形式之间的转换.	图形用户界面: Graphical User Interface, 缩写: GUI, 是指采用图形方式显示的计算机操作界面.
	“泛指”预测为“特指”	新华社 11 月 8 日电 (记者: 肖世尧, 张华迎) 2019 [中国 (福州) 羽毛球公开赛] 8 日展开 1/4 决赛的较量, 赛会卫冕冠军陈雨菲直落两局战胜泰国名将.	2019 中国福州羽毛球公开赛: 第 2 届中国福州羽毛球公开赛, 是 2019 年世界羽联世界巡回赛的其中一站, 属于第三级别赛事.	中国福州羽毛球公开赛: 一项自 2018 年起成立、一年一度在中国福建省福州市仓山区举行的国际羽毛球公开锦标赛.
	“整体”预测为“部分”	古田会议的精髓就是思想建党、政治建军. 85 年前, 我军政治工作在古田奠基, 新型人民军队在 [古田] 定型.	古田会议会址: 位于福建省龙岩市上杭县古田镇, 1929 年 12 月, 毛泽东主持的中国共产党红军第四军第九次代表大会 (即古田会议) 在此召开, 通过了具有历史意义的《古田会议决议》.	古田镇: 古田镇是福建省上杭县下辖的一个镇, 位于上杭县境东北部, 是 2003 年评定的第一批中国历史文化名镇之一. 境内有古田会议纪念馆.
	“部分”预测为“整体”	在餐饮外卖行业, [美团] 强调更多的玩家进入餐饮外卖市场对行业是好事, 这意味着蛋糕会越做越大.	美团: 美团是一家面向本地消费产品和零售服务 (包括娱乐、餐饮、送货、旅行和其他服务) 的中文购物平台. 旗下经营美团网、美团外卖、大众点评网、摩拜单车等互联网平台.	美团外卖: 美团外卖是中国生活服务网站美团网旗下的互联网外卖订餐平台, 由北京三快在线科技有限公司运营, 创立于 2013 年, 目前合作商户数超过 200 万家, 覆盖 1300 多个城市.
全局	指代错误	奥地利选手梅尔泽以 7-6 和 6-1 击败克罗地亚卡洛维奇. [梅尔泽] 在半决赛中将对阵比利时名将奥–罗切斯.	莱昂纳多·梅耶尔: Leonardo Mayer, 出生于科连特斯, 是一位阿根廷男子职业网球运动员.	于尔根·梅尔策: 奥地利职业网球运动员. 于 1999 年转为职业选手. 单打最高世界排名是第 9 位.
	主题错误	摄影师从海南赶到茂名, 与当地一众天文爱好者一路 [追星], 终于拍下了这颗绿色彗星.	追星族: 崇拜明星, 积极追随并关注与其有关事物的爱好者.	天文摄影: 天文摄影为一特殊的摄影技术, 可记录各种天体和天象、月球、行星甚至遥远的深空天体.
	角色错误	双方签署协议共同成立“[新闻与传播学院] 院务委员会”. 北大将借助新华社的影响力, 建设国际传播研究智库, 打造教学实习和培养从业人员基地.	清华大学新闻与传播学院: 简称新闻学院、新传学院, 是清华大学直属的一个学院.	北京大学新闻与传播学院: 承担北京大学在新闻学和传播学领域教育与研究任务的一个直属学院.
局部		[《2019 MBC 演技大赏》] 于 12 月 30 日晚在首尔麻浦区上岩 MBC 举行, 由金成柱、韩惠珍主持.	2019 SBS 演技大奖: 《2019 SBS 演技大奖》为 SBS 于 2019 年度颁发的电视剧大奖.	2019 MBC 演技大奖:《2019 MBC 演技大奖》为 MBC 于 2019 年度颁发的电视剧大奖.
时间		当地时间 2018 年 9 月 15 日, 美国北卡罗来纳州, 飓风“[佛罗伦萨]”在美国北卡罗来纳州登陆.	2006 年飓风佛罗伦萨: 飓风佛罗伦萨是 2006 年大西洋飓风季形成的第 7 场热带风暴和第 2 场飓风.	飓风佛罗伦斯 (2018 年): 飓风佛罗伦斯为 2018 年大西洋飓风季第 6 个被命名的热带气旋.
地点		该段起于 11 号线 [左岭站] (不含), 终点位于葛店南站.	左岭站: 左岭站位于湖北省武汉市洪山区左岭镇, 是武黄城际铁路上的火车站, 武汉铁路局管辖.	左岭站 (武汉地铁): 左岭站是武汉地铁 11 号线的一座车站, 位于武汉市洪山区.

从以下四个方面进行改进.

1) 对长尾实体进行数据增强. 由第 3.3.4 节实验分析, 大语言模型在长尾实体理解问题上占明显优势, 且随着规模的增加优势会更加明显. 本文认为这得益于更大规模的预训练数据, 使得长尾实体也能有更多的相关语料. 由此启发可针对长尾实体进行数据增强, 以缓解长尾问题.

2) 进行在线困难样本挖掘 (Online hard example mining). 在训练过程中动态选择难以分类的长尾数据进行重点学习. 但是在此过程中, 可能需要注意解决模型的过拟合问题, 例如采用本文设计的平衡负采样策略.

3) 针对实体类别预测进行多任务学习. 这实际上是一种数据分层的策略, 由于类别数量远小于实体数量, 这有助于缓和类别上的长尾问题. 此外, 针对尾部实体的类别进行分层采样, 以确保长尾实体在训练过程中得到足够的关注.

4) 建模多层次实体标签体系. 基于分类损失的优化目标不能很好地区分实体的粒度, 对此可以应用层次化损失函数进行优化, 或使用图神经网络建模实体关系.

4　相关工作

4.1　大语言模型的涌现能力

Wei 等[34] 将大规模语言模型的涌现能力定义

为在小规模模型中不存在但在大规模模型中出现的能力. 涌现是指模型参数量、训练数据量和计算量达到一定规模后突然出现的能力. 典型的涌现能力包括: 1) 语境学习能力: Brown 等[6] 在 GPT-3 中报告, 通过在输入中放入几对输入–输出示例, 模型无需参数更新即可输出正确结果. 2) 逐步思考能力: Wei 等[17] 首次报告, 通过指导模型输出一系列中间推理步骤, 提高复杂推理能力. 3) 指令遵循能力: OpenAI 的 InstructGPT[18] 通过人类反馈的强化学习, 使模型输出与人类意图对齐. 文献 [26] 表明, 学习连续的虚拟提示指令也能使模型很好地完成自然语言理解任务.

4.2 大语言模型在自然语言处理中的应用方法

在自然语言处理领域, 使用大语言模型的涌现能力解决任务的方法整体可分为知识增强、适配器微调、提示学习和语境学习四类方法, 下面将简要介绍每种方法的相关工作.

1) 知识增强. 现有研究通过知识增强方法利用大语言模型提升小模型能力, 主要分为两类: a) 知识蒸馏. 大模型作为教师模型, 训练小模型模仿其输出. 例如, Bonifacio 等[35] 提出的 InPairs 利用大语言模型的小样本学习能力生成信息检索数据集, Ferraretto 等[36] 使用大语言模型生成数据集样本标签的解释, 再微调小模型生成标签和解释. b) 知识源. 大语言模型生成推理所需的相关知识作为小模型输入, 增强其推理能力. 例如, Wu 等[22] 提出的 CoT-KA 框架, 通过提示大语言模型进行链式思考生成思维链, 并将其作为附加知识输入小模型.

2) 适配器微调. 为实现参数高效的微调, Houlsby 等[24] 提出基于适配器的微调方法. 对于大语言模型, Hu 等[25] 提出了低秩适配器微调方法 (Low-rank adaptation, LoRA), 将可训练的低秩分解矩阵注入到 Transformer 的每一层中, 模拟训练过程中权重矩阵的参数变化量, 进一步降低了训练的参数量. Dettmers 等[37] 在 LoRA 的基础上通过对大语言模型预训练参数进行 4bit-NormalFloat 量化提出了 QLoRA 方法, 在有限的性能损失下, 进一步减少了显存需求.

3) 提示学习. 根据提示词在输入序列中的位置, 模型是否每层都插入提示词以及重参数化方法不同发展出多种提示微调方法, 包括 Li 等[38] 提出的 Prefix-tuning, Lester 等[39] 提出的 Prompt tuning 以及 P-Tuning[26] 和 P-Tuning v2[40] 等.

4) 语境学习. 语境学习通过将演示示例作为条件输入引导大语言模型进行任务推理. 与提示学习相似, 语境学习通过离散地构造演示示例来提高大语言模型在下游任务上的能力. 由于语境学习无需参数更新, 因此在数据稀缺的领域和任务中表现出较强的适应性. 该方法最早由 GPT-3[6] 引入, 之后 Liu 等[27] 和 Lu 等[41] 分别探讨了示例选择策略和样本排列顺序对语境学习效果的影响. 后续的研究[42−44] 通过校准消除大语言模型偏见或元学习等方法进一步提高了大语言模型的语境学习能力, 从而改进语境学习应用的效果.

4.3 大语言模型在实体链接中的应用

使用大语言模型进行实体链接的研究目前仍处于初步阶段, 目前仅有两项相关工作在该领域进行了尝试. Cho 等[45] 将实体消歧建模为多项选择的问题, 基于零样本提示的方法引导大语言模型作出选择. 此外, 他们还利用大语言模型对指称上下文进行文本摘要, 并将其作为全局文本添加到大语言模型的提示输入中. Shi 等[46] 在多模态实体链接方面应用大语言模型, 提出 GEMEL 模型, 该工作基于语境学习方法, 并借鉴 Cao 等[47] 在 2021 年提出的 GENRE 模型. GEMEL 将视觉特征映射为软提示, 结合样本提示引导参数冻结的大语言模型在预定义的前缀树上进行受限波束搜索, 以直接生成实体名称作为消歧结果.

5 结束语

本文研究了大语言模型的涌现能力对中文实体链接系统的影响, 主要分析了 Qwen-7B 和 ChatGLM3-6B 模型的四种方法 (知识增强、适配器微调、提示学习和语境学习) 对中文实体链接任务中的改进效果和不足. 实验结果表明, 基于 Qwen-7B/ChatGLM3-6B 的监督学习方法在 Hansel-FS 数据集上提升 3.9% ~ 11.8%, 在 Hansel-ZS 数据集上提升 0.7% ~ 4.1%, 在 CLEEK 数据集上提升 0.6% ~ 3.7%. 这些方法在一定程度上改善了基于 BERT 等预训练编码器方法的“过拟合”问题. 经过不同规模的 Qwen 模型的对比实验, 可以发现无监督的语境学习方法在模型参数量达到 720 亿时也取得了与监督微调 70 亿参数模型相近的效果 (−2.4% ~ +1.4%), 并且更大参数的 Qwen 模型在长尾问题上表现出优势. 通过消融实验, 本文深入分析了不同设置对大语言模型实体链接性能的影响. 错误样本的定量分析显示, 大语言模型在粒度范围和实体类别方面仍有 36% 和 25% 的错误占比. 通过错误样例的定性分析, 本文进一步定位了问题原因, 并给出相关改进方向. 这些研究为进一步完善大语言模型在实体链接任务中的应用提供了启示.

References

1 Guo Hao, Li Xin-Yi, Tang Jiu-Yang, Guo Yan-Ming, Zhao Xiang. Adaptive feature fusion for multi-modal entity alignment. *Acta Automatica Sinica*, 2024, **50**(4): 758−770
(郭浩, 李欣奕, 唐九阳, 郭延明, 赵翔. 自适应特征融合的多模态实体对齐研究. 自动化学报, 2024, **50**(4): 758−770)

2 Hasibi F, Balog K, Bratsberg S E. Exploiting entity linking in queries for entity retrieval. In: Proceedings of the ACM International Conference on the Theory of Information Retrieval. Newark, Delaware, USA: Association for Computing Machinery, 2016. 209−218

3 Liu Qiong-Xin, Wang Ya-Nan, Long Hang, Wang Jia-Sheng, Lu Shi-Shuai. Generative knowledge question answering technology based on global coverage mechanism and representation learning. *Acta Automatica Sinica*, 2022, **48**(10): 2392−2405
(刘琼昕, 王亚男, 龙航, 王佳升, 卢士帅. 基于全局覆盖机制与表示学习的生成式知识问答技术. 自动化学报, 2022, **48**(10): 2392−2405)

4 Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. Qwen technical report. arXiv preprint arXiv: 2309.16609, 2023.

5 Zeng A H, Xu B, Wang B W, Zhang C H, Yin D, Zhang D, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv preprint arXiv: 2406.12793, 2024.

6 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv: 2005.14165, 2020.

7 Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy S M T I, Chadha A, et al. The troubling emergence of hallucination in large language models——An extensive definition, quantification, and prescriptive remediations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 2541−2573

8 Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual Event: Association for Computational Linguistics, 2020. 5418−5426

9 Kandpal N, Deng H K, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. arXiv preprint arXiv: 2211.08411, 2023.

10 Wang X T, Yang Q W, Qiu Y T, Liang J Q, He Q Y, Gu Z H, et al. KnowledGPT: Enhancing large language models with retrieval and storage access on knowledge bases. arXiv preprint arXiv: 2308.11761, 2023.

11 Ganea O-E, Hofmann T. Deep joint entity disambiguation with local neural attention. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 2619−2629

12 Le P, Titov I. Improving entity linking by modeling latent relations between mentions. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018. 1595−1604

13 Jiang Sheng-Chen, Wang Hong-Bin, Yu Zheng-Tao, Xian Yan-Tuan, Wang Hong-Tao. Domain integrated-entity links based on relationship indices and representation learning. *Acta Automatica Sinica*, 2021, **47**(10): 2376−2385
(蒋胜臣, 王红斌, 余正涛, 线岩团, 王红涛. 基于关系指数和表示学习的领域集成实体链接. 自动化学报, 2021, **47**(10): 2376−2385)

14 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 4171−4186

15 Yamada I, Washio K, Shindo H, Matsumoto Y. Global entity disambiguation with BERT. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022. 3264−3271

16 Wu L, Petroni F, Josifoski M, Riedel S, Zettlemoyer L. Scalable zero-shot entity linking with dense entity retrieval. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual Event: Association for Computational Linguistics, 2020. 6397−6407

17 Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022, **35**: 24824−24837

18 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv: 2203.02155, 2022.

19 Talmor A, Tafjord O, Clark P, Goldberg Y, Berant J. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 2020, **33**: 20227−20237

20 Xu Z R, Shan Z F, Li Y X, Hu B T, Qin B. Hansel: A Chinese few-shot and zero-shot entity linking benchmark. In: Proceedings of the 16th ACM International Conference on Web Search and Data Mining. Singapore: Association for Computing Machinery, 2023. 832−840

21 Zeng W X, Zhao X, Tang J Y, Tan Z, Huang X Q. CLEEK: A Chinese long-text corpus for entity linking. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020. 2026−2035

22 Wu D J, Zhang J, Huang X M. Chain of thought prompting elicits knowledge augmentation. arXiv preprint arXiv: 2307.1640, 2023.

23 Humeau S, Shuster K, Lachaux M A, Weston J. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv: 1905.01969, 2020.

24 Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, Laroussilhe Q D, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. arXiv preprint arXiv: 1902.00751, 2019.

25 Hu E J, Shen Y L, Wallis P, Allen-Zhu Z Y, Li Y Z, Wang S, et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685, 2022.

26 Liu X, Zheng Y N, Du Z X, Ding M, Qian Y J, Yang Z L, et al. GPT understands, too. *AI Open*, 2024, **5**: 208−215

27 Liu J C, Shen D H, Zhang Y Z, Dolan B, Carin L, Chen W Z. What makes good in-context examples for GPT-3? In: Proceedings of the Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Dublin, Ireland: Association for Computational Linguistics, 2022. 100−114

28 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv preprint arXiv: 1908.10084, 2019.

29 Huang S J, Wang B B, Qin L B, Zhao Q, Xu R F. Improving few-shot and zero-shot entity linking with coarse-to-fine lexicon-based retriever. In: Processings of the Natural Language Processing and Chinese Computing: 12th National CCF Conference, Foshan, China: 2023. 245−256

30 Zhou H Y, Sun C J, Lin L, Shan L L. ERNIE-AT-CEL: A Chinese few-shot emerging entity linking model based on ERNIE and adversarial training. In: Processings of the Natural Language Processing and Chinese Computing: 12th National CCF Conference. Foshan, China: 2023. 48−56

31 Xu Z, Shan Z, Hu B, Zhang M. Overview of the NLPCC 2023

shared task 6: Chinese few-shot and zero-shot entity linking. In: Proceedings of Natural Language Processing and Chinese Computing: 12th National CCF Conference. Foshan, China: 2023. 257−265

32 de Cao N, Wu L, Popat K, Artetxe M, Goyal N, Plekhanov M, et al. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 2022, **10**: 274−290

33 Zhang Q R, Chen M S, Bukharin A, He P C, Cheng Y, Chen W Z, et al. Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv: 2303.10512v1, 2023.

34 Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models [Online], available: https://openreview.net/forum?id=yzkSU5zdwD, January 15, 2025

35 Bonifacio L, Abonizio H, Fadaee M, Nogueira R. InPars: Unsupervised dataset generation for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain: Association for Computing Machinery, 2022. 2387−2392

36 Ferraretto F, Laitz T, Lotufo R, Nogueira R. ExaRanker: Synthetic explanations improve neural rankers. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China: Association for Computing Machinery, 2023. 2409−2414

37 Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv: 2305.14314, 2023.

38 Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. VirtualEvent:AssociationforComputationalLinguistics,2021.4582−4597

39 Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Virtual Event: Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 3045−3059

40 Liu X, Ji K X, Fu Y C, Tam W L, Du Z X, Yang Z L, et al. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, 2022. 61−68

41 Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics, 2022. 8086−8098

42 Zhao T, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: Improving few-shot performance of language models. arXiv preprint arXiv: 2102.09690, 2021.

43 Rubin O, Herzig J, Berant J. Learning to retrieve prompts for in-context learning. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022. 2655−2671

44 Min S, Lewis M, Zettlemoyer L, Hajishirzi H. MetaICL: Learning to learn in context. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022. 2791−2809

45 Cho Y M, Zhang L, Callison-Burch C. Unsupervised entity linking with guided summarization and multiple-choice selection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. 9394−9401

46 Shi S B, Xu Z R, Hu B T, Zhang M. Generative multimodal entity linking. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy: ELRA and ICCL, 2024. 7654−7665

47 de Cao N, Izacard G, Riedel S, Petroni F. Autoregressive entity retrieval. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021.

徐正斐 北京理工大学计算机学院硕士研究生. 主要研究方向为知识工程. E-mail: zhengfei@bit.edu.cn

(**XU Zheng-Fei** Master student at the School of Computer Science and Technology, Beijing Institute of Technology. His main research interest is knowledge engineering.)

辛 欣 北京理工大学计算机学院副教授. 主要研究方向为自然语言处理, 知识工程, 信息检索. 本文通信作者. E-mail: xxin@bit.edu.cn

(**XIN Xin** Associate professor at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers natural language processing, knowledge engineering, and information retrieval. Corresponding author of this paper.)