计算机科学
COMPUTER SCIENCE

利用语义增强提示和结构信息的知识图谱补全模型

蔡启航，徐彬，董晓迪

引用本文

蔡启航，徐彬，董晓迪. 利用语义增强提示和结构信息的知识图谱补全模型[J]. 计算机科学，2025，52(9): 282-293.

CAI Qihang, XU Bin, DONG Xiaodi. Knowledge Graph Completion Model Using Semantically Enhanced Prompts and Structural Information [J]. Computer Science, 2025, 52(9): 282-293.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700201

利用语义增强提示和结构信息的知识图谱补全模型

蔡启航　徐　彬　董晓迪
东北大学计算机科学与工程学院　沈阳 110169
(qducqh@163.com)

摘　要　知识图谱补全旨在根据已有事实推断新事实,增强知识图谱的全面性和可靠性,从而提升其实用价值。为了解决现有基于预训练语言模型的方法对头实体和尾实体预测效果差异大、连续提示初始化随机性强导致训练过程波动较大以及未充分利用知识图谱结构信息的问题,提出了利用语义增强提示和结构信息的知识图谱补全模型(SEPS-KGC)。该模型遵循多任务学习框架,联合知识图谱补全任务与实体预测任务。首先,设计了基于示例引导的关系模板生成方法,针对预测头实体和预测尾实体的不同任务,利用大语言模型生成两种更具针对性的关系提示模板,并结合语义辅助信息,使模型更好地理解实体间的语义关联。其次,设计了基于有效初始化的提示学习方法,使用关系标签的预训练嵌入进行初始化。最后,设计了结构信息提取模块,利用卷积和池化操作提取知识图谱结构信息,提升模型的稳定性和关系理解能力。在两个公开数据集上进行实验,证明了 SEPS-KGC 的有效性。

关键词:知识图谱;知识图谱补全;预训练语言模型;大语言模型;提示学习;结构信息
中图分类号　TP391

Knowledge Graph Completion Model Using Semantically Enhanced Prompts and Structural Information

CAI Qihang,XU Bin and DONG Xiaodi
School of Computer Science and Engineering,Northeastern University,Shenyang 110169,China

Abstract　Knowledge graph completion aims to infer new facts based on existing facts,enhance the comprehensiveness and reliability of the knowledge graph,and thus improve its practical value. In order to solve the problems that existing methods based on pre-trained language models have large differences in the prediction effects of head and tail entities,large fluctuations in the training process due to the stochastic initialization of consecutive prompts,and under-utilization of structural information of the knowledge graph,this paper proposes the knowledge graph completion model using semantically enhanced prompts and structural information(SEPS-KGC). The model follows a multi-task learning framework that unites the knowledge graph completion task with the entity prediction task. Firstly,an example-guided relationship templates generation method is designed to generate two more targeted relationship prompt templates using a large language model for the different tasks of predicting head entities and predicting tail entities,and incorporating semantic auxiliary information to enable the model to better understand the semantic associations between entities. Secondly,a prompt learning method based on effective initialization is designed,using pre-trained embeddings of relational labels for initialization. Finally,a structural information extraction module is designed to extract knowledge graph structural information using convolution and pooling operations to improve the stability and relationship understanding of the model. The effectiveness of SEPS-KGC is demonstrated on two public datasets.

Keywords　Knowledge graph,Knowledge graph completion,Pre-trained language model,Large language model,Prompt learning,Structural information

1　引言

大数据时代的到来推动了以专家知识为核心的知识工程向以数据驱动为核心的大数据知识工程的转变。知识图谱(Knowledge Graph,KG)正是这一新型知识工程的典型代表。知识图谱以结构化的三元组形式描述概念、实体和关系,提供了更好的组织、管理和理解世界上大量信息的能力。因此,知识图谱被广泛应用于信息检索[1]、问答系统[2]和推荐系统[3]

到稿日期:2024-07-31　返修日期:2024-10-21
基金项目:国家自然科学基金(62137001);辽宁省自然科学基金面上项目(2022-MS-119)
This work was supported by the National Natural Science Foundation of China(62137001) and Liaoning Natural Science Foundation(2022-MS-119).
通信作者:徐彬(xubin@mail.neu.edu.cn)

等众多领域。目前，较为常用的知识图谱有 FreeBase[4]，Wikidata[5]，DBpedia[6]，YAGO[7]等，但它们都存在不同程度的信息缺失。以 FreeBase 为例，其中 70%的人物实体缺失出生地信息，99%的人物实体缺失种族信息[8-9]。知识图谱补全(Knowledge Graph Completion，KGC)任务旨在根据知识图谱中已有事实推断出新的事实，使知识图谱更完整。

知识图谱嵌入(Knowledge Graph Embedding，KGE)模型是早期研究使用的主流模型。KGE 模型可分为 3 类：基于翻译的模型、基于张量分解的模型和基于神经网络的模型。近年来，越来越多的人开始发现预训练语言模型(Pre-trained Language Model，PLM)在解决自然语言处理(Natural Language Processing，NLP)问题时的潜力，设计了许多基于 PLM 的模型，包括 KG-BERT[10]，PKGC[11]，MEM-KGC[12]等。

尽管上述基于 PLM 的模型已经取得令人满意的结果，但这些方法还存在一些不足。1)手工定义提示模板不仅耗费大量人力资源，而且由于知识图谱中的关系具有方向性，按照人类的语言表达习惯，通常将尾实体放在关系提示模板的主语位置，这使得其上下文信息更加丰富，从而更容易预测。因此，模型预测尾实体时的性能明显优于预测头实体时的性能。2)只使用离散提示无法充分捕捉和利用丰富的语义信息，而 P-tuning[13]等方法虽然在一定程度上缓解了这一问题，但初始化时的随机性较强，可能导致训练过程出现较大的波动，并且需要更多的训练数据来达到理想的性能。3)现有的基于 PLM 的方法大多仅利用了知识图谱中的语义信息，而忽略了其重要的结构信息。尽管 CSProm-KG[14]模型利用了结构信息，但是由于过度依赖软提示，学习过程具有不稳定性，模型偶尔会在 WN18RR 等小型数据集上崩溃。

为了解决上述问题，本文提出了 SEPS-KGC 模型，遵循多任务深度神经网络模型(Multi-task Deep Neural Networks，MT-DNN)中的多任务学习框架[15]。该框架是微软的开源框架，结合了多任务学习(Multi-task Learning，MTL)和 BERT 的优点，并在多个流行的自然语言理解基准测试中得到了最优的结果，具有公认的优越性。因此，本文 SEPS-KGC 模型遵循该框架，联合 KGC 任务与实体预测任务，共享 PLM 和分类器，使得 PLM 和分类器可以从实体预测任务中学习更多的实体属性，更好地区分相似实体，从而有利于完成 KGC 任务。具体来说，首先，充分发掘大语言模型[16-18](Large Language Model，LLM)的潜能，通过基于示例引导的关系模板生成方法，结合语义辅助信息，为 PLM 设计更符合 KGC 任务的提示模板，能够在提高模型预测尾实体的能力的同时，有效提高模型预测头实体的能力。其次，设计了基于有效初始化的提示学习方法(Prompt Learning with Effective Initialization，PLEI)，将离散提示与连续提示结合并进行有效的初始化，解决了 P-tuning 等方法初始化时随机性强的问题。最后，设计了结构信息提取模块，提取知识图谱的结构信息，弥补了基于 PLM 的方法的不足。与同样使用结构信息的 CSProm-KG[14]模型相比，该模块使用卷积神经网络(Convolutional Neural Networks，CNN)和池化操作，在处理结构化数据时更具稳定性，同时能够提取多尺度特征，捕捉不同层次的知识图谱信息。

本文的主要贡献如下。

1)设计了基于示例引导的关系模板生成方法，针对预测头实体和预测尾实体的不同任务，利用 LLM 为 PLM 生成两种更有针对性的关系提示模板，并结合语义辅助信息，使 PLM 更好地理解实体之间的语义关联。

2)设计了 PLEI 方法，使用关系标签的预训练嵌入进行有效的初始化，增强提示的语义一致性。

3)设计了一个结构信息提取模块，提取知识图谱的结构信息，弥补以往大多数基于 PLM 的模型忽略知识图谱结构信息的不足，且更具稳定性。

4)在两个常用数据集上进行了一系列实验，验证了 SEPS-KGC 模型在 KGC 任务中的有效性和可行性，并通过消融实验展示了不同模块的效果，验证了 SEPS-KGC 模型的架构合理性。

2 相关工作

执行 KGC 任务的方法的共同点是，建立一个有效的机制来衡量三元组的合理性。主流的方法有基于嵌入的方法、基于 PLM 的方法以及将它们相结合的方法。近年来，LLM 取得了飞速进展，并在文本相关任务中显示出了强大的能力[19]，一些研究者也开始使用 LLM 完成 KGC 任务。

2.1 基于嵌入的知识图谱补全方法

基于嵌入的方法将实体和关系表示为嵌入向量，并在向量空间中维护它们的语义关系。Bordes 等[20]将三元组($\boldsymbol{h}$，$\boldsymbol{r}$，$\boldsymbol{t}$)表示为头实体向量 $\boldsymbol{h}$ 到尾实体向量 $\boldsymbol{t}$ 的变换过程，令头实体向量 $\boldsymbol{h}$ 与关系向量 $\boldsymbol{r}$ 之和尽可能靠近尾实体向量 $\boldsymbol{t}$。Yany 等[21]将双线性模型中的所有关系嵌入转换为对角矩阵。Sun 等[22]将每个关系嵌入表示为复向量空间中从头实体到尾实体的旋转，令三元组的映射操作等价于实体嵌入沿坐标轴旋转的过程。然而，考虑到早期模型的低效性与复杂性，无法满足对大规模知识图谱补全的要求，越来越多的研究人员开始转向探索神经网络来识别三元组之间的关联和重要模式。Dettmers 等[23]和 Nguyen 等[24]提出了使用 CNN 的嵌入模型来捕获实体和关系的复杂语义。然而，CNN 捕捉的交互关系依然有限。为了令实体向量与关系向量更充分地进行交互，Nathani 等[25]使用图注意力网络(Graph Attention Network，GAT)通过加权的优化显著增强了模型的表达能力，构建出 KBGAT 模型。

2.2 基于 PLM 的知识图谱补全方法

基于嵌入的方法通过学习结构嵌入来进行三元组识别，但忽略了文本信息。同时，由于预训练语言模型，如 BERT[26]等，显著提高了几种 NLP 任务的性能，一些研究试图从 PLM 中获取知识，完成 KGC 任务。Yao 等[10]提出的 KG-BERT 模型是第一个使用 PLM 执行 KGC 任务的模型。该方法连接每个三元组的头实体、关系和尾实体构成输入序列，将 KGC 转换为具有二元交叉熵损失函数的序列分类问题。但是，简单的连接会导致句子不连贯，不能充分利用 PLM 中的隐含知识。为了解决这个问题，Lyu 等[11]提出了 PKGC，使用手动设计的三元组提示和精心选择的支持提示作为 PLM 的输入，为了使三元组提示更具表现力，使用了 P-

tuning 方法，在离散提示间添加了一些连续提示。然而，这些连续提示是随机初始化的，可能导致训练过程中出现较大的波动。Choi 等[12]利用多任务学习框架[15]和掩码语言模型的分类过程，将实体预测和超类预测两个额外任务与 KGC 任务结合起来。

2.3 嵌入与 PLM 相结合的知识图谱补全方法

嵌入与 PLM 相结合的知识图谱补全方法大多遵循编码器-解码器框架，首先使用图神经网络（Graph Neural Network，GNN）等将知识图谱中的实体和关系嵌入低维向量空间。然后，利用 PLM 对实体描述和关系描述进行编码，获取语义向量。接着，将两种表示拼接在一起。最后，经过解码器解码完成 KGC 任务。Ju 等[27]提出了一种常识知识库补全模型，该模型通过 GAT 和 PLM 分别学习常识知识图谱节点和关系的结构表示和上下文表示。使用解码器对给定的三元组解码，从而实现更准确的预测。Sun 等[28]提出了 SS-KGC 模型，该模型使用 PLM 得到知识图谱语义向量，使用多头注意力机制得到知识图谱实体和关系的嵌入向量，然后将这两种向量进行拼接，利用 CNN 提取特征完成 KGC 任务。

2.4 基于 LLM 的知识图谱补全方法

LLM 通常以自回归的方式进行预训练，通过预测序列中的下一个元素来训练模型，在文本理解和生成方面表现出较强的能力。在执行 KGC 任务方面，Zhu 等[29]直接将 LLM 应用于 KGC 任务。然而，实验结果证明，单纯使用 LLM 无法达到最先进的性能。Yao 等[30]提出 KGLLaMA 模型，该模型使用三元组的实体和关系描述作为提示，利用 LLM 的响应进行预测。Wei 等[31]构建了一个集成 LLM 和基于三元组的 KGC 检索器的框架，减轻了长尾问题，且不会产生额外的训练开销。

本文提出的 SEPS-KGC 模型是 PLM 和 LLM 相结合的方法。首先，基于嵌入的知识图谱补全方法主要依赖于实体和关系的低维向量表示，学习结构嵌入进行三元组识别，可能无法充分捕捉实体之间复杂的语义关联。因此，本文选择了使用 PLM 完成 KGC 任务。现有的基于 PLM 的方法的一个重要限制是，使训练过程变成了基于文本的学习，难以在知识图谱中捕获复杂的结构信息。因此本文设计了结构信息提取模块，以提取知识图谱的结构信息。与 2.4 节中基于 LLM 的知识图谱补全方法不同的是，SEPS-KGC 模型不是直接使用 LLM 完成 KGC 任务，而是将 LLM 作为工具，释放 LLM 自身的力量为 PLM 提供更好的关系提示模板和语义辅助信息。此外，考虑到 LLM 中普遍存在幻觉现象，在生成文本时，输出内容中可能包含不真实、不准确或完全虚构的信息。本文使用最大边界相关（Maximal Marginal Relevance，MMR）算法对 LLM 的输出进行筛选，可以有效地降低幻觉现象产生的影响。

3 SEPS-KGC 模型

3.1 问题形式化

知识图谱是一种以图的形式组织的多关系数据，其中节点表示实体，边表示实体之间的关系。一个知识图谱可以表示为 $G=\{E,R,T,D\}$，其中 E 和 R 分别表示实体集和关系集。$T=\{(h,r,t)\mid h,t\in E,r\in R\}$ 是三元组集合，h,r,t 分别表示三元组 (h,r,t) 中的头实体、关系和尾实体，例如（Crazy Heart，/film/film/language，English）是一个描述电影 *Crazy Heart* 使用的语言的三元组。D 是每个实体和关系的文本描述集；$D(e)$，$D(r)$ 分别表示实体 $e\in E$ 和关系 $r\in R$ 的文本描述。例如，实体/m/0d2kt 的文本描述是 D（'/m/0d2kt'）= 'River Thames'。KGC 的目的是在给定头实体和关系的情况下预测尾实体，例如（Crazy Heart，/film/film/language，?）；或者在给定尾实体和关系的情况下预测头实体，例如（?，/film/film/language，English）。

3.2 模型框架

本文提出的 SEPS-KGC 模型分为知识图谱补全模块、实体预测模块、结构信息提取模块，模型结构如图 1 所示。

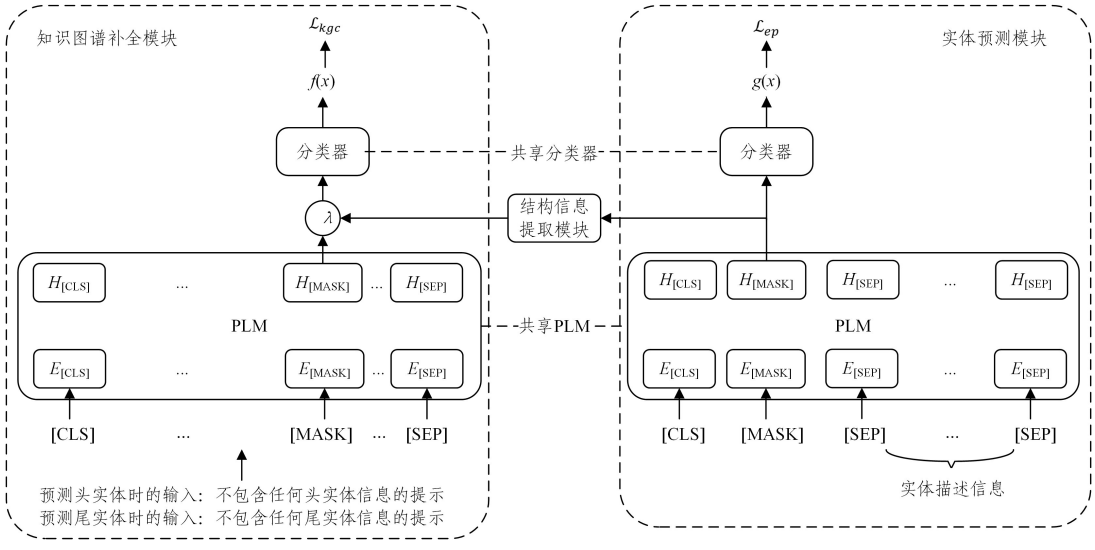

图 1 SEPS-KGC 模型的架构

Fig. 1 Architecture of SEPS-KGC model

知识图谱补全模块中，利用 LLM 赋能，将三元组的头实体和关系信息（预测尾实体）或者尾实体和关系信息（预测头

实体）转换成一个提示句，并融合更多语义辅助信息，以提高提示句的表达能力。设计了连续提示优化策略与 KGC 过程共同优化，充分挖掘 PLM 中蕴含的知识。最后将提取到的语义信息与结构信息提取模块提取的结构信息以 λ 作为加权系数进行加权融合，进行知识图谱补全。

实体预测模块中，将实体描述视为实体的上下文信息，根据实体的描述预测实体。该模块有两个作用：1）与知识图谱补全模块一起遵循 MT-DNN 中的多任务学习框架[15]，使得 PLM 和分类器可以学习更多的实体属性，更好地区分相似实体，从而有利于完成 KGC 任务；2）将学习到的实体信息提供给结构信息提取模块，帮助结构信息提取模块提取知识图谱结构信息。

结构信息提取模块提取知识图谱结构信息，弥补了以往大多数基于 PLM 的方法没有充分利用知识图谱结构信息的不足。知识图谱语义信息与结构信息加权之后的数据和知识图谱补全模块中 PLM 输出的数据维度一致，因此可以将结构信息提取模块删除而不影响模型的运行。

3.3 知识图谱补全模块

3.3.1 基于示例引导的关系模板生成方法

知识图谱中原始的关系描述通常是模糊的，例如 FB15k-237-N 数据集中的关系“/film/film/written_by”。现有的大多数方法通过去除特殊符号（例如将其转换为“film film written by”）获得关系标签，并简单地拼接三元组中实体和关系的标签，并将其作为 PLM 的输入。这导致句子不连贯，与预训练的任务有差距，因此不能充分利用 PLM 中的知识。为了解决这个问题，Lyu 等[11]为每个关系手动定义关系提示模板，将每个三元组转换为自然提示句，以更好地表达三元组的语义。但是手工定义关系提示模板不仅会耗费昂贵的人力资源，而且由于关系的方向性以及人类的语言表达习惯，在手工定义关系提示模板时，尾实体通常作为主语，并且位于句子的开头或其他特定位置，具有的上下文信息更丰富，这使得模型更容易预测尾实体，而预测头实体的效果较差。

本文设计了一种基于示例引导的关系模板生成方法，使用 LLM 针对预测头实体和预测尾实体的不同任务生成更有针对性的关系提示模板。首先，采用了关系分组的方法，对于关系集合中的每个关系 r，将训练集和验证集中所有具有该关系的三元组聚成一组 T_r。设 G_{train} 和 G_{valid} 分别为知识图谱的训练集和验证集，E 为实体集合，R 为关系集合，则 $T_r=\{(h,r,t)\in G_{\text{train}}\cup G_{\text{valid}}\mid h,t\in E,r\in R\}$。接着，将关系 r 的三元组集合 T_r 作为示例集提供给 LLM，引导 LLM 针对预测头实体和预测尾实体的不同任务生成两种更有针对性的关系提示模板。在接下来的 KGC 任务中，预测头实体时使用其中一个模板，预测尾实体时使用另一个模板。

在将具有相同关系的三元组作为示例集提供给 LLM 生成关系提示模板时，示例集的多样性非常重要。多样性的示例可以帮助模型更好地理解和泛化关系，从而生成更准确和通用的模板。然而，由于具有相同关系的三元组数量通常很多，如何有效选择具有代表性和多样性的三元组成为关键问题。

为了实现质量控制机制，本文使用了 MMR 算法，结合内容的相关性和多样性来对这些三元组进行排序，有效地挑选出多样性较高的示例集。综合分析本文使用的数据集的数据规模，规定每个关系的示例集的大小为 30。具体来说，MMR 通过同时考虑每个三元组与查询的相关性以及与已经选择的三元组之间的差异性，来选择下一个最优的三元组。其计算式如式（1）所示：

$$MMR=\arg\max_{D_i\in T_r\setminus T}[\beta Rel(D_i,r)-(1-\beta)\max_{D_j\in T}Sim(D_i,D_j)] \tag{1}$$

$$Rel(D_i,r)=\frac{\boldsymbol{D}_i\cdot\boldsymbol{r}}{\|\boldsymbol{D}_i\|\ \|\boldsymbol{r}\|} \tag{2}$$

$$Sim(D_i,D_j)=\frac{\boldsymbol{D}_i\cdot\boldsymbol{D}_j}{\|\boldsymbol{D}_i\|\ \|\boldsymbol{D}_j\|} \tag{3}$$

其中，T_r 是所有候选三元组的集合；T 是已选择的三元组集合；D_i 是候选三元组；$Rel(D_i,r)$ 表示三元组 D_i 与关系 r 的相关性，计算式如式（2）所示；$Sim(D_i,D_j)$ 表示三元组 D_i 与已选三元组集合 T 中的某个三元组 D_j 的相似性，计算式如式（3）所示；β 是一个权重参数，用于平衡相关性和多样性。算法的具体流程如算法 1 所示。

算法 1 MMR 算法

输入：(r, T_r)
输出：T
1. 初始化：T←Ø / * T 为选择后的三元组集合 * /；
2. while len(T)＜30 and T_r
3. max_mmr_score←-float('inf')
4. best_candidate←None
5. / * 计算 D_i 与关系 r 的相关性 * /
6. for $D_i\in T_r$ do
7. relevance←Rel(D_i, r)
8. max_similarity←0
9. / * 计算 D_i 与已选集合 T 中 D_j 的相似性 * /
10. for $D_j\in T$ do
11. sim←Sim(D_i, D_j)
12. if sim＞max_similarity then
13. max_similarity←sim
14. end if
15. end for
16. / * 计算 MMR 分数 * /
17. mrr_score←β×relevance－(1－β)×sim
18. if mrr_score＞ max_mmr_score then
19. max_mmr_score←mmr_score
20. best_candidate←D_i
21. end if
22. end for
23. if best_candidate then
24. T.add(best_candidate)
25. T_r.remove(best_candidate)
26. end if
27. end while

基于示例引导的关系模板生成方法使用的提示如表 1 所列。将 LLM 的响应信息进行简单处理，得到了关系“/film/film/written_by”的两个关系模板：预测头实体时使用的模板“[MASK] is a film written by [Y].”和预测尾实体时使用的

模板“[MASK] is the writer of film [X].”。其中[Y]和[X] 分别是尾实体和头实体的占位符。

表 1 基于示例引导的关系模板生成方法使用的提示

Table 1 Prompt for example-guided relationship templates generation method

提示	LLM 响应
Your task is to understand and summarize. Here are some triples for you: The Sound of Music /film/film/written_by Ernest Lehman My Fair Lady /film/film/written_by Alan Jay Lerner Family Plot /film/film/written_by Ernest Lehman Into the Wild /film/film/written_by Sean Penn The Conversation /film/film/written_by Francis Ford Coppola In above examples, What do you think “A /film/film/written_by B” mean? Please summarize and describe the meaning using A and B as subjects, respectively.	The relationship “A /film/film/written_by B” indicates that B is the writer of the film A. Using A and B as subjects, this can be summarized as: “A is a film written by B.” “B is the writer of film A.”

本文通过 3 种方式提高模板生成过程的可重复性:1)在参考范围内,调整生成过程的参数(如温度、核采样等),评估输出;2)保存上下文和对话状态;3)使用固定的模型版本。

3.3.2 语义辅助信息

除了知识图谱包含的三元组信息外,语义辅助信息也能显著提高 KGC 任务的性能。与基于嵌入的知识图谱补全方法相比,基于 PLM 的方法很容易在不改变模型结构的情况下灵活地整合语义辅助信息。这种灵活性不仅简化了模型的设计和实现过程,而且通过在输入中显式地加入语义辅助信息,PLM 可以更好地理解实体之间的语义关联,从而提高知识图谱补全的效果。本文使用了实体类型信息和近义词信息两种语义辅助信息。

实体类型信息可以减小候选实体的搜索空间,例如如果已知头实体的类型是“作家”,那么与其关联的尾实体更可能是“书籍”类型,而不是“运动”类型。

近义词信息可以通过扩展实体的语义范围来增强模型对实体间关系的理解,例如“enjoyment”可以扩展为“delight”“pleasure”等近义词,进一步帮助模型做出更准确的预测。数据集中并没有提供实体的近义词,本文设计了简单的提示,利用 LLM 获得实体的近义词。具体来说,本文使用的提示为“Give synonyms for [Entity] based on the definition of [Entity]:[Entity Definition],and answer in the format {[Entity]:[your answer]}.”,其中,[Entity]表示实体标签,[Entity Definition]表示实体定义。表 2 列出了一个具体的例子。

表 2 近义词信息提示举例

Table 2 Examples of synonym information prompts

提示	LLM 响应
Give synonyms for enjoyment based on the definition of enjoyment: the pleasure felt when having a good time, and answer in the format {enjoyment:[your answer]}	{enjoyment:[pleasure, delight, amusement, satisfaction, gratification, joy, fun, happiness, contentment, recreation]}

LLM 会生成多个近义词,但是由于 LLM 存在幻觉现象,并不是所有生成的近义词都能作为有效的近义词信息使用。为了确保近义词在具体应用中的准确性和适用性,通过 BERT 模型获取实体和近义词的向量表示,计算它们之间的余弦相似度,精确地评估每个近义词在特定上下文中的语义价值,从而筛选出那些真正具有相同或相近意义的近义词。相似度计算式如式(4)所示:

$$cosine\_similarity(e,e_i)=\frac{\boldsymbol{e}\cdot\boldsymbol{e}_i}{\|\boldsymbol{e}\|\ \|\boldsymbol{e}_i\|},e_i\in S \tag{4}$$

其中,S 为 LLM 生成的近义词集合,e 为原始实体,e_i 为原始实体的近义词。

利用上述方法,得到了实体类型信息和近义词信息两种辅助信息。本文定义了模板来将辅助信息转换为相应的句子,具体内容如表 3 所列。

表 3 语义辅助信息模板

Table 3 Semantic auxiliary information templates

语义辅助信息类型	模板
实体类型信息	“[*Entity*] is a [$type_1$], a [$type_2$], …, a [$type_m$].”
近义词信息	“The synonyms of [*Entity*]: [$synonyms_1$], [$synonyms_2$], …, [$synonyms_p$].”

表 3 中,[*Entity*]为实体标签,[$type_i$]为实体的第 i 种类型,[$synonyms_i$]为实体的第 i 个近义词,m 为实体类型的数量,p 为近义词的数量。

3.3.3 基于有效初始化的提示学习方法

Liu 等[13]观察到离散提示具有很大程度的不稳定性。对于冻结的语言模型,更改提示中的单个单词可能会导致性能大幅下降。离散提示是人类可读的,因此更容易解释,但是无法直接进行梯度优化,模型只能基于预训练时学到的知识来解释这些提示。而连续提示可以向模型引入可学习的嵌入向量,并通过反向传播进行优化,使模型逐渐调整这些嵌入向量,以更好地适应特定任务,提高整体性能。现有的一些方法虽然在一定程度上缓解了这一问题,但初始化时随机性较强,可能导致训练过程中出现较大的波动。

因此,本文提出了 PLEI 方法,将连续提示嵌入与离散提示连接起来,并将它们作为新的提示输入 PLM 中。在训练过程中,连续提示嵌入作为可学习参数,与模型的其他参数一起通过梯度下降进行更新。模型会调整这些向量,使得在给定任务下的损失函数最小化。与 PKGC 使用的 P-tuning 不同的是,PLEI 不是在提示的每个分割位置插入一个连续提示嵌入,而是在固定位置连续插入多个连续提示嵌入。单个连续提示嵌入可能在不同位置上有不同的语义,而多个连续提示嵌入在特定位置上共同作用,可以形成更加丰富和准确的语义表示。考虑到关系标签中单词可能会对关系的语义起到概括作用,本文使用关系标签的预训练嵌入来初始化每个 $[V]_i$,以预测尾实体为例,具体形式如式(5)所示:

$$PROMPT=[e_{desp}][e_s][\boldsymbol{V}]_1[\boldsymbol{V}]_2\cdots[\boldsymbol{V}]_n[template] \tag{5}$$

其中,[e_{desp}]表示头实体描述;[e_s]表示头实体语义辅助信息;每个[$\boldsymbol{V}$]$_i\in\mathbb{R}^H$是与 PLM 的输入嵌入具有相同维度的

密集向量；[template]表示预测尾实体时的关系提示；n 为插入的连续提示的数量，根据不同数据集的关系标签长度而定，对于 FB15k-237-N 数据集，n 设置为 6，对于 WN18RR 数据集，n 设置为 4。

3.3.4 知识图谱补全任务

图 2 给出了执行 KGC 任务时的例子。

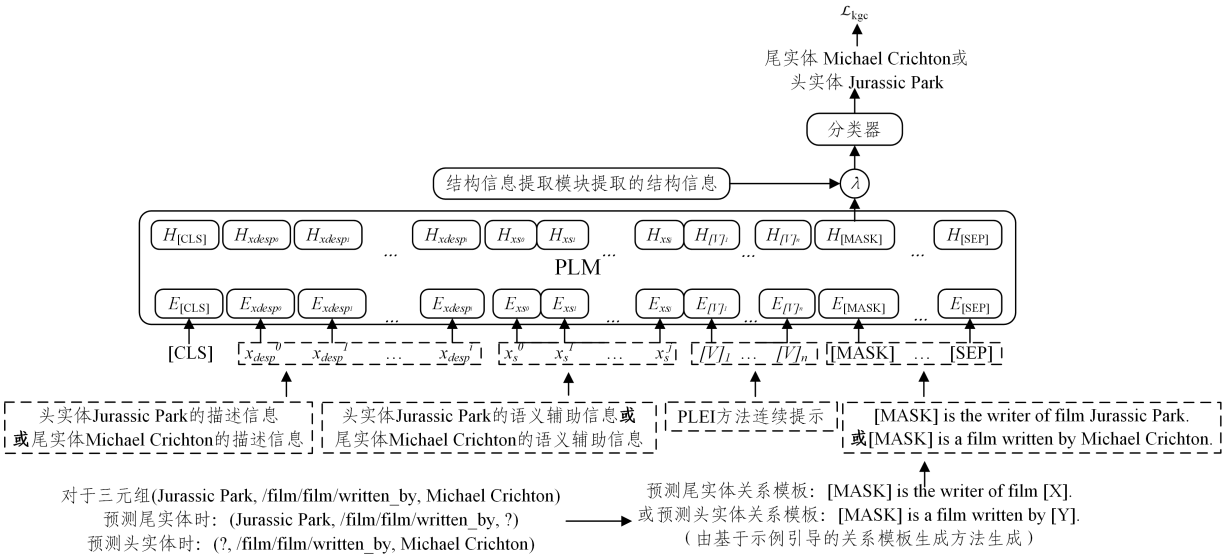

图 2 知识图谱补全模块的结构

Fig. 2 Structure of knowledge graph completion module

具体来说，给定一个三元组(h,r,t)，其中 h 表示头实体，r 表示关系，t 表示尾实体。预测尾实体时，把前文所述的预测尾实体的提示模板中的[X]替换为头实体 h，并送入 PLM 中，将[MASK]标记处的最后一层隐藏状态的向量与结构信息向量加权融合后预测尾实体 t。同理，预测头实体时，把前文所述的预测头实体的提示模板中的[Y]替换为尾实体，并送入 PLM 中，将[MASK]标记处的最后一层隐藏状态的向量与结构信息向量加权融合后预测头实体。分类过程如式(6)和式(7)所示：

$$\boldsymbol{H}_{\mathrm{MASK}}^{kgc}=PLM(x_{\mathrm{kgc}})_{\mathrm{MASK}} \tag{6}$$

$$f(x)=\mathrm{softmax}[(\lambda\boldsymbol{H}_{\mathrm{MASK}}^{kgc}+(1-\lambda)\boldsymbol{H}_s)\boldsymbol{W}^{\mathrm{T}}] \tag{7}$$

其中，x_{kgc}表示知识图谱补全模块的输入，$\mathrm{PLM}(\cdot)_{\mathrm{MASK}}$表示 PLM 的输出中[MASK]标记处的最后一层隐藏状态的向量，$\boldsymbol{H}_{\mathrm{MASK}}^{kgc}\in\mathbb{R}^H$，$H$ 为隐藏层大小，对于 BERT 模型来说是 768；$\boldsymbol{H}_s\in\mathbb{R}^H$表示结构信息提取模块提取的知识图谱结构信息向量，λ 表示加权系数，$\boldsymbol{W}\in\mathbb{R}^{K\times H}$是分类层，$f(x)\in\mathbb{R}^K$表示所有实体的输出概率分布，$K$ 为实体数量。为了训练模型，使用交叉熵损失函数计算知识图谱补全损失：

$$\mathcal{L}_{\mathrm{kgc}}=-\frac{1}{K}\sum_{i=1}^{K}y_i\ln f(x_i) \tag{8}$$

其中，y_i为真实实体类别。

3.4 实体预测模块

实体描述通常包含丰富的上下文信息，这些信息可以帮助模型更好地理解和区分不同的实体。例如，实体描述可以包括实体的属性、关系、历史背景等，提供更多的细节来辅助预测。早期的方法 KG-BERT 很难从词汇相似的实体中挑选出答案实体。例如，给定头实体和关系为(take a breather, derivationally related for)，正确的尾实体为 breathing time，而 KG-BERT 却将与 breathing time 词汇相似的 snorkel breather和 breath 判定为高分实体，导致模型的 MRR 和 Hits@K 指标值降低。因此，受到 Choi 等[12]的启发，本文借助实体预测模块，将实体预测作为 KGC 的辅助任务，模块结构如图 1 右侧部分所示。

将实体描述视为实体的上下文信息，与[MASK]拼接起来作为 PLM 的输入，如式(9)所示：

$$x_{\mathrm{ep}}=[\mathrm{CLS}][\mathrm{MASK}][\mathrm{SEP}]e_{\mathrm{desp}}[\mathrm{SEP}] \tag{9}$$

其中，e_{desp}表示实体 e 的实体描述信息。由于实体预测与 KGC 具有相似的分类过程，因此共享分类器，分类过程如式(10)和式(11)所示：

$$\boldsymbol{H}_{\mathrm{MASK}}^{\mathrm{ep}}=PLM(x_{\mathrm{ep}})_{\mathrm{MASK}} \tag{10}$$

$$g(x)=\mathrm{softmax}(\boldsymbol{H}_{\mathrm{MASK}}^{\mathrm{ep}}\boldsymbol{W}^{\mathrm{T}}) \tag{11}$$

其中，x_{ep}表示实体预测模块的输入，$\mathrm{PLM}(\cdot)_{\mathrm{MASK}}$表示 PLM 的输出中[MASK]标记处的最后一层隐藏状态的向量，$\boldsymbol{H}_{\mathrm{MASK}}^{\mathrm{ep}}\in\mathbb{R}^{\mathrm{H}}$，$H$ 为隐藏层大小，对于 BERT 模型来说是 768；$g(x)\in\mathbb{R}^K$表示所有实体的输出概率分布，K 为实体数量。$\boldsymbol{W}\in\mathbb{R}^{K\times H}$是与知识图谱补全模块共享的分类层。为了训练模型，使用交叉熵损失函数计算实体预测损失：

$$\mathcal{L}_{\mathrm{ep}}=-\frac{1}{K}\sum_{i=1}^{K}y_i\ln g(x_i) \tag{12}$$

其中，y_i为真实实体类别。

3.5 结构信息提取模块

基于 PLM 的方法利用了 PLM 的功能，但使训练过程变成了基于文本的学习，难以在知识图谱中捕获复杂的结构信息。本文设计了一个结构信息提取模块，在预测尾实体时融合同一头实体和关系的其他尾实体信息，在预测头实体时融合同一尾实体和关系的其他头实体信息，从而提取知识图谱的结构信息。模块结构如图 3 所示。下面以预测尾实体为例进行说明，预测头实体时与之类似。

首先，采用了分组的方法，对于每个(头实体，关系)对，将训练集和验证集中所有具有相同(头实体，关系)的三元组聚成一组 T_{er}，计算式如式(13)所示：

$$T_{er}=\{(h,r,t')\in G_{train}\cup G_{valid}\mid(h,r,t)\in G_{train}\cup G_{valid},t\neq t'\} \tag{13}$$

其中，G_{train} 和 G_{valid} 分别为知识图谱的训练集和验证集，E 为实体集合，R 为关系集合，t 为要预测的尾实体。

接着，对 T_{er} 中三元组的尾实体从实体预测模块得到的信息进行特征提取和信息融合，图 3 中的 $desp_i$ 表示第 i 个尾实体描述。将实体的状态表示堆叠成一个张量序列，使用 CNN 对这些张量进行处理，提取出不同卷积核下的关键特征。池化操作进一步将每个实体的多维特征映射到一个单一的特征向量中，该向量能够更好地表示实体在整个序列中的重要性和信息丰富度。最后，将上述得到的结构信息送到知识图谱补全模块，知识图谱补全模块会将原尾实体的语义信息和结构信息进行加权融合。

知识图谱中的结构信息不仅包括拓扑结构，还包含丰富的语义。GNN 通过消息传递机制聚合节点邻居的信息，但通常不直接处理邻居节点之间的关系。本文利用 CNN 通过卷积核捕捉同一头/尾实体及其关系下的多个尾/头实体的局部交互，并将其与预训练语言模型提取的中心节点语义信息进行聚合。这不仅考虑了中心节点与邻居节点之间的关系，还涵盖了邻居节点之间的关系。此外，CNN 在计算效率上优于 GNN。

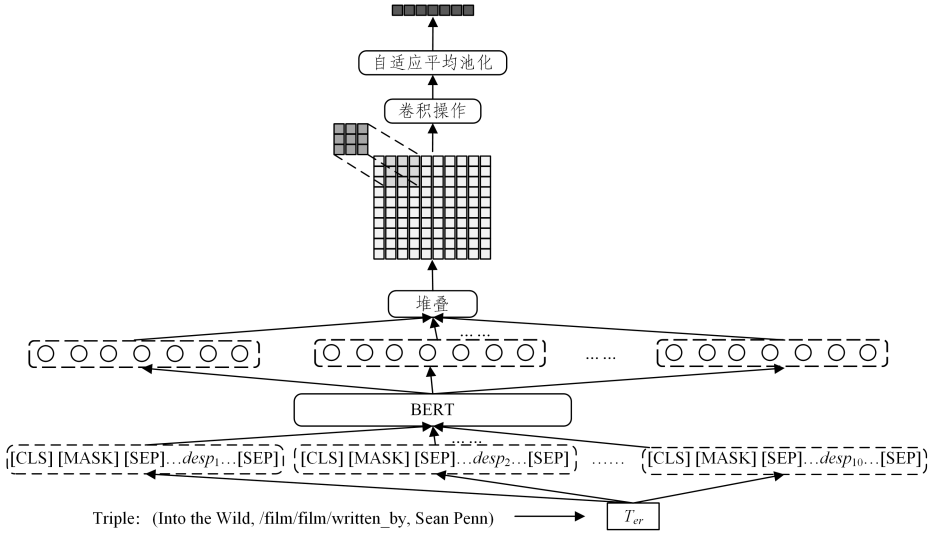

图 3 结构信息提取模块结构

Fig. 3 Structure of structural information extraction module

4 实验

4.1 数据集

本文在 KGC 任务常用的两个数据集 FB15k-237-N 和 WN18RR 上评估了模型性能。Akrami 等[32]表明 FB15k-237 数据集不恰当地提高了模型的准确性。为了提高任务难度，更接近真实场景中的 KGC 任务，Lyu 等[11]创建了 FB15k-237-N 数据集。WN18RR 数据集是 WordNet 的一个子集，主要涉及词义、词类和单词之间的各种关系，如近义词、反义词、上位词、下位词等，由约 41 000 个实体和 11 个关系组成。两个数据集的统计数据如表 4 所列。

表 4 数据集的统计信息

Table 4 Statistics of datasets

Datasets	$\lvert E\rvert$	$\lvert R\rvert$	Train	Valid	Test
FB15k-237-N	13 104	93	87 282	14 082	16 452
WN18RR	40 943	11	86 835	3 034	3 134

4.2 评价指标

为了评估模型的性能，本文使用了平均倒数排名(Mean Reciprocal Rank，MRR)和结果在排名的前 K 位命中率 Hits@K 作为评价指标，其中 $K\in\{1,3,10\}$。

MRR 是对每个 KGC 任务所对应真实答案在预测结果中排名的倒数取平均值，侧重于评测模型的整体预测效果。其计算式如式(14)所示：

$$MRR=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{14}$$

其中，Q 表示三元组查询集合，$|Q|$ 表示三元组查询集合的大小；$rank_i$ 表示第 i 个查询的正确答案的排名。$MRR=[0,1]$，其值越大，预测效果就越好。

Hits@K 表示预测中正确答案的排名结果等于或小于阈值 K 的比率，其计算式如式(15)所示：

$$\text{Hits@}K=\frac{1}{|Q|}Count(rank_i\leqslant K),1<i\leqslant|Q| \tag{15}$$

其中，Q 表示三元组查询集合，$|Q|$ 表示三元组查询集合的大小；$rank_i$ 表示第 i 个查询的正确答案的排名；$Count(\cdot)$ 表示计数函数。Hits@K 的值越高，模型的预测效果就越好。

4.3 基线模型

为了评估 SEPS-KGC 在 KGC 任务中的性能，将其与以下方法进行了比较，并将这些方法根据所使用的技术分为 4 类：基于嵌入的模型、基于 PLM 的模型、嵌入与 PLM 相结合的模型和基于 LLM 的模型。

4.3.1 基于嵌入的模型

1) TransE[20]：该方法对头实体进行关系特定的平移，然后与尾实体进行距离度量。

2）DistMult[21]：该模型通过双线性函数计算实体和关系的嵌入向量间的相互作用，用于知识图谱补全。

3）ConvE[23]：该方法将知识图谱中的实体和关系嵌入二维矩阵中，并使用 CNN 来捕捉它们之间的复杂交互关系，从而进行链接预测。

4）TuckER[33]：该方法依赖于 Tucker 分解，将一个三元组张量分解为一个共享核心张量与一系列矩阵的乘积。

5）RotatE[22]：该方法将三元组投影到复数向量空间，并将关系嵌入定义为旋转矢量，令三元组的映射操作等价于实体嵌入沿坐标轴旋转的过程。

6）CompGCN[34]：该方法利用图卷积网络结合实体和关系嵌入，来捕捉知识图谱中节点和边的复杂交互关系，从而进行图结构化数据的学习和推理。

7）Att-HousE[35]：该方法设计了一个带注意力机制的规则生成器和一个带 HousE 嵌入的规则预测器，可以更全面地抓取和形成多边关系。

8）GANPUL[36]（该方法没有提供模型的英文简写，在本文中暂时使用 GANPUL 表示 Hu 等[36]的模型）：该方法是一种基于生成式对抗网络和正类无标签学习的知识图谱补全算法，利用生成式对抗网络生成无标签样本，并使用正类无标签学习缓解假阴性标签问题。

4.3.2 基于 PLM 的模型

1）KG-BERT[10]：第一个使用 PLM 执行 KGC 任务的模型，将头实体、关系和尾实体拼接后进行序列分类。

2）MTL-KGC[37]：该方法结合链接预测、关系预测和相关性排序 3 个任务进行多任务学习。

3）MEM-KGC[12]：该方法结合链接预测、实体预测和超类预测 3 个任务进行多任务学习。

4）PKGC[11]：手工设计关系提示模板并使用提示调优技术来增强 PLM 在 KGC 任务中的性能。

5）KG-S2S[38]：该方法通过序列到序列架构生成知识图谱中的三元组，从而实现知识图谱补全。

6）CSProm-KG[14]：该模型提出了在结构信息和文本知识之间保持平衡的条件软提示，通过优化 PLM 以增强 KGC 任务的性能。

4.3.3 嵌入与 PLM 相结合的模型

1）SS-KGC[28]：该模型将知识图谱的图嵌入信息和语义信息进行拼接，利用 CNN 提取特征。

2）ISA-KGC[39]：该模型将 GNN 与基于 Transformer 的模型相结合，并使用适配器将图嵌入空间与文本空间对齐，然后将这两种嵌入连接起来，从 Transformer 的最后一层提取编码，再进行卷积操作。

4.3.4 基于 LLM 的模型

1）$\text{ChatGPT}_{\text{zero-shot}}$[29]，$\text{ChatGPT}_{\text{one-shot}}$[29]：该模型直接使用 ChatGPT 执行 KGC 任务。

2）KICGPT[31]：该方法构建了一个集成 LLM 和基于三元组的 KGC 检索器的框架。

4.4 实验设置

本文使用的 PLM 为 BERT-base，LLM 为 GPT-3.5-turbo，对 30 个 epoch 的多任务设置进行微调。设置 mini-batch 为 32，使用 AdamW 优化器，模型学习率为 0.00005，PLM 学习率为 0.000001。在使用 LLM 生成关系模板的过程中，设置 temperature 为 0，top-p 为 1，max-token 为 300。模型中的语义信息与结构信息的加权系数 λ 设置为 0.8，这是本文经过实验得出的最优加权系数。

4.5 实验结果

表 5 列出了基线模型以及提出的 SEPS-KGC 模型的性能。

表 5 知识图谱补全结果

Table 5 Knowledge graph completion results

Model		年份	FB15k-237-N				WN18RR			
			MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
KGE Models	TransE[20]	2013	25.5	15.2	30.1	45.9	24.3	4.3	44.1	53.2
	DistMult[21]	2015	20.9	14.3	23.4	33.0	44.4	41.2	47.0	50.4
	ConvE[23]	2018	27.3	19.2	30.5	42.9	46.0	39.0	43.0	48.0
	TuckER[33]	2019	31.2	22.8	34.6	48.6	47.0	44.3	48.2	52.6
	RotatE[22]	2019	27.9	17.7	32.0	48.1	47.6	42.8	49.2	57.1
	CompGCN[34]	2019	31.6	23.1	34.9	48.0	47.9	44.3	49.4	54.6
	Att-HousE[35]	2024	—	—	—	—	41.8	41.3	49.4	55.8
	GANPUL[36]	2024	—	—	—	—	49.2	43.5	49.8	59.4
PLM-based Models	KG-BERT[10]	2019	20.3	13.9	20.1	40.3	21.6	4.1	30.2	52.4
	MTL-KGC[37]	2020	24.8	15.5	25.6	43.6	33.1	20.3	38.3	59.7
	MEM-KGC[12]	2021	—	—	—	—	<u>57.2</u>	48.9	**62.0**	<u>72.3</u>
	PKGC[11]	2022	33.2	26.1	34.6	48.7	—	—	—	—
	KG-S2S[38]	2022	35.4	<u>28.5</u>	38.8	49.3	57.0	**52.5**	59.7	65.4
	CSProm-KG[14]	2023	<u>36.0</u>	28.1	<u>39.5</u>	<u>51.1</u>	55.2	50.0	57.2	65.7
KGE-PLM Integration Model	SS-KGC[28]	2023	30.1	22.7	32.7	46.3	—	—	—	—
	ISA-KGC[39]	2024	—	—	—	—	47.8	41.2	51.6	59.5
LLM-based Models	$\text{ChatGPT}_{\text{zero-shot}}$[29]	2023	—	—	—	—	—	19.0	—	—
	$\text{ChatGPT}_{\text{one-shot}}$[29]	2023	—	—	—	—	—	21.2	—	—
	KICGPT[31]	2023	—	—	—	—	54.9	47.4	58.5	64.1
ours	SEPS-KGC	2024	**37.3**	**28.6**	**40.8**	**53.7**	**58.9**	<u>51.9</u>	<u>61.8</u>	**72.4**

注：粗体表示最优结果，下划线表示次优结果。

可以得到以下结论：SEPS-KGC 模型在 2 个数据集上优于其他前沿方法，尤其是在 FB15k-237-N 数据集上始终优于

其他所有模型。具体来说，与基于嵌入的方法相比，SEPS-KGC 表现更好，说明了 PLM 在处理 KGC 任务方面的潜力。与基于 PLM 的方法相比，SEPS-KGC 比简单拼接三元组各部分和使用手工定义关系提示模板的模型表现更好，这表明了基于示例引导的关系模板生成方法、语义辅助信息和 PLEI 方法的重要性和有效性。此外，MEM-KGC，KG-S2S 等基于 PLM 的模型仅考虑知识图谱语义信息，在同类型模型中取得了良好表现，但是在 Hits@10 和 MRR 指标上仍落后于 SEPS-KGC 模型，这体现了结构信息提取模块的可行性。与同样结合了结构信息的 CSProm-KG 相比，SEPS-KGC 使用 CNN 和池化操作，使得模型更具稳定性，且补全效果更好。尽管 SEPS-KGC 模型在 WN18RR 数据集的 Hits@1 和 Hits@3 指标上表现欠佳，但它在 Hits@10 和 MRR 指标方面优于所有基线模型。上述结果证明了 SEPS-KGC 的有效性和可行性。

4.6 消融实验

为了更具体地探究本文方法对模型性能是否有提升，以仅由知识图谱补全模块和实体预测模块组成的模型作为基础模型。知识图谱补全模块的输入为三元组元素的简单拼接，其中要预测的实体用[MASK]令牌代替，如预测尾实体时输入为头实体、关系和[MASK]令牌的简单拼接。实体预测模块的输入与本文方法相同，即[MASK]令牌与实体描述的拼接。消融实验的实验结果如表 6 所列。从第二行开始，每一行表示在上一行的基础上添加新的模块或使用新的方法获得的实验结果。

表 6 知识图谱补全结果

Table 6 Knowledge graph completion results

对比方案	FB15k-237-N				WN18RR			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
基础模型	36.3	27.8	39.6	52.9	56.7	48.7	60.9	71.2
+基于示例引导的关系模板生成方法	36.7	28.3	40.2	53.3	57.5	49.9	61.1	71.8
+语义辅助信息	36.9	28.3	40.7	**53.8**	57.7	50.3	61.2	71.7
+PLEI	37.1	28.6	40.7	53.6	58.9	51.8	**62.1**	72.2
+结构信息提取模块	**37.3**	**28.6**	**40.8**	53.7	**58.9**	**51.9**	61.8	**72.4**

注：粗体表示最优结果。

4.6.1 基于示例引导的关系模板生成方法的影响

Lyu 等[11]提供了 FB15k-237-N 数据集的关系提示模板，Wei 等[31]提供了 WN18RR 数据集的关系提示模板，但是对于每个关系，仅有一个关系提示模板。为了更具体地探究使用基于示例引导的关系模板生成方法生成两种更有针对性的关系提示模板对模型性能是否有提升，本文仅将基础模型的知识图谱补全模块输入改为使用一个关系提示模板，可以得到表 7 中第二行的结果。仅将基础模型的知识图谱补全模块输入改为使用基于示例引导的关系模板生成方法生成的两种关系提示模板，可以得到表 7 中第三行的结果。

表 7 基于示例引导的关系模板生成方法的消融实验

Table 7 Ablation experiment of example-guided relationship templates generation method

对比方案	FB15k-237-N								WN18RR							
	head				avg				head				avg			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
基础模型	22.3	15.0	24.1	37.1	36.3	27.8	39.6	52.9	47.7	39.9	51.0	62.3	56.7	48.7	60.9	71.2
P-one	22.2	14.7	24.3	36.8	36.6	28.2	40.2	52.9	48.6	41.3	50.8	**62.7**	57.4	49.7	61.0	**71.9**
P-two	**23.0**	**15.2**	**24.6**	**38.0**	**36.7**	**28.3**	**40.2**	**53.3**	**48.6**	**41.5**	**51.1**	62.4	**57.5**	**49.9**	**61.1**	71.8

注：P-one 表示使用一个关系提示模板，P-two 表示使用基于示例引导的关系模板生成方法。

可以发现，无论是使用一个关系提示模板，还是使用基于示例引导的关系模板生成方法改进基础模型的知识图谱补全模块输入，均对模型性能有一定提升。其中，使用基于示例引导的关系模板生成方法的提升更为显著，尤其是在预测头实体时，较基础模型和使用一个关系提示模板的方法在绝大多数的指标上有所提升，这表明该方法能够弥补单个提示模板不能很好地捕捉头实体语义的缺陷，在利用关系信息方面具有更强的效果。同时，关系提示模板的有效性在于能否准确地表达关系的语义，而与关系类型不直接相关，因此关系属于哪种类型，并不会影响关系提示模板的使用。然而，值得注意的是，虽然大部分指标有提升，但在 WN18RR 数据集上的 Hits@10 指标略有下降，这可能是由于 WN18RR 数据集表示的是词义、词类和单词之间的各种关系，对关系提示模板的简练性和准确性要求较高，而对所预测词汇在句子中的位置的要求较低。

4.6.2 语义辅助信息的影响

在 KGC 任务中，语义辅助信息能够提供更丰富的语义信息，支持模型进行更深入的推理和语义理解。为了验证语义辅助信息的有效性，本文在两个数据集上分别比较了模型在添加(表 6 第 3 行)和不添加(表 6 第 2 行)语义辅助信息时的性能表现。可以发现，在任何数据集上，语义辅助信息的引入都有效地提升了 MRR，Hits@3 的指标值，这验证了添加语义辅助信息的有效性。

此外，本文在 WN18RR 的数据集上分析了近义词的添加数量对模型性能的影响，如图 4 所示，并非所有近义词均能够给模型带来正面效果，近义词添加数量过多会降低模型性能。模型在选取 2 个最为相似近义词的情况下取得了最好的性能。这是因为 LLM 存在幻觉现象，直接将 LLM 生成的近义词添加到文本中，可能会为当前上下文引入噪声，所以本文综合考虑近义词与当前语境是否匹配，通过筛选并保留合适

近义词来进一步提高模型性能。

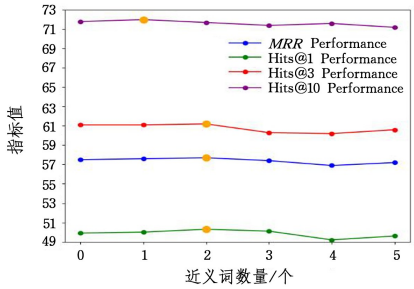

图 4 近义词数量的影响

Fig. 4 Impact of the number of synonyms

4.6.3 PLEI 方法的影响

为了验证该方法的有效性，本文在 2 个数据集上分别对比了不使用连续提示（表 8 第 1 行）、随机初始化连续提示（表 8 第 2 行）和使用 PLEI 方法（表 8 第 3 行）时的结果。

可以观察到，在一些指标上，使用随机初始化连续提示的方法相比于不加入连续提示的方法有所提升。这表明，连续提示与离散提示相结合的方法不仅能够利用离散提示提供的显性语义信息，还能够通过连续提示逐步细化和强化模型对复杂关系的捕捉和推理能力。同时，对比表 8 的第 2 和第 3 行可以发现，使用 PLEI 的方法能够进一步弥补随机初始化的不足，使得模型在 KGC 任务中表现出更高的准确性和鲁棒性。

表 8 PLEI 方法的消融实验

Table 8 Ablation experiment of prompt learning with effective initialization

对比方案	FB15k-237-N				WN18RR			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
不加入连续提示	36.9	28.3	40.7	**53.8**	57.7	50.3	61.2	71.7
随机初始化连续提示	36.9	28.4	40.3	53.6	57.8	50.5	61.4	71.7
PLEI	**37.1**	**28.6**	**40.7**	53.6	**58.9**	**51.8**	**62.1**	**72.2**

4.6.4 结构信息提取模块的影响

结构信息提取模块通过提取知识图谱的结构信息，进一步提升了模型的性能。实验结果显示（表 6 中第 4 行和第 5 行对比），加入结构信息提取模块后，模型在 FB15k-237-N 和 WN18RR 两个数据集上的 MRR 和排名指标有显著提升。具体来说，在 FB15k-237-N 数据集上，MRR 从 37.1 提升到 37.3，Hits@1 保持在 28.6，Hits@3 和 Hits@10 也有所提高；在 WN18RR 数据集上，MRR 保持在 58.9，Hits@1 从 51.8 提升到 51.9，Hits@10 从 72.2 提升到 72.4。这表明，首先，引入其他尾实体（或头实体）信息可以丰富模型对头实体（或尾实体）和关系的理解，通过多样化的信息输入，模型能够更全面地捕捉实体之间的复杂关联和语义。其次，加权融合确保了原始尾实体（或头实体）信息的重要性，同时通过引入其他实体信息，使得模型能够更好地处理样本偏差，提高对实体的泛化能力。

4.6.5 加权系数 λ 的影响

本文在 FB15k-237-N 数据集上分析了知识图谱语义信息与结构信息的加权系数 λ 的影响，如图 5 所示。

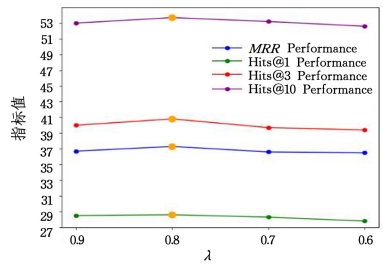

图 5 加权系数 λ 的影响

Fig. 5 Impact of the weighting coefficient λ

总体来看，当 λ 设置为 0.8 时，模型在所有指标上的表现均达到最优，这表明在此加权系数下，知识图谱的语义信息与结构信息的结合效果最好。较低或较高的 λ 值则会导致性能下降，说明平衡知识图谱的语义信息与结构信息对模型性能至关重要。

结束语 本文分析了现有模型在 KGC 任务上的不足，为了更加全面地利用知识图谱的语义信息和结构信息，本文提出了 SEPS-KGC 模型。具体来说，本文使用基于示例引导的关系模板生成方法，引导 LLM 针对预测头实体和预测尾实体的不同任务生成两种更有针对性的关系提示模板；同时，加入语义辅助信息，使 PLM 更好地理解实体之间的语义关联。此外，设计了 PLEI 方法，并使用关系标签的预训练嵌入进行有效初始化，增强了提示的语义一致性。最后，设计了一个结构信息提取模块，弥补了以往基于 PLM 的模型忽略知识图谱结构信息的不足。在两大公开数据集上进行的详细实验，证明了本文 SEPS-KGC 的有效性和可行性。消融实验展示了不同模块的效果，从而验证了 SEPS-KGC 模型的架构合理性。在今后的研究中，将在更多数据集上验证本文方法的性能，并在现阶段工作的基础上进行更多尝试，如调整大模型的提示，为每个关系生成更多而不是 2 个关系提示模板，实现多方面的质量控制。并考虑在结构信息提取模块中尝试更多信息提取方式，如在 GNN 的基础上加入邻居节点间的信息聚合。

参 考 文 献

[1] XIONG C Y, POWER R, CALLAN J. Explicit semantic ranking for academic search via knowledge graph embedding[C]// Proceedings of the 26th International Conference on World Wide Web. ACM, 2017: 1271-1279.

[2] SAXENA A, TRIPATHI A, TALUKDAR P. Improving multi-hop question answering overknowledge graphs using knowledge base embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 4498-4507.

[3] WANG H W, ZHANG F Z, ZHANG M D, et al. Knowledge-aware graph neural networks with label smoothness regulariza-

tion for recommender systems[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. New York: ACM, 2019: 968-977.

[4] BOLLACKER K, EVANS C, PARITOSH P, et al. FreeBase: a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, 2008: 1247-1250.

[5] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[6] LEHMANN J, ISELE R, JAKBO M, et al. DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia [J]. Semantic Web, 2015, 6(2): 167-195.

[7] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]// Proceedings of the 16th International World Wide Web Conference. New York: ACM, 2007: 697-706.

[8] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion [C]// Proceedings of the 20th ACM SI-GKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601-610.

[9] WEST R, GABRILOVICH E, MURPHY K, et al. Knowledge base completion via search-based question answering[C]// Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 515-526.

[10] YAO L, MAO C S, LUO Y. KG-BERT: BERT for knowledge graph completion[J]. arXiv: 1909. 03193, 2019.

[11] LYU X, LIN Y K, CAO Y X, et al. Do pre-trained models benefit knowledge graph completion? a reliable evaluation and a reasonable approach[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 3570-3581.

[12] CHOI B, JANG D, KO Y. MEM-KGC: masked entity model for knowledge graph completion with pre-trained language model [J]. IEEE Access, 2021, 9: 132025-132032.

[13] LIU X, ZHENG Y N, DU Z X, et al. Gpt understands, too[J]. arXiv: 2103. 10385, 2021.

[14] CHEN C, WANG Y F, SUN A X, et al. Dipping PLMs sauce: bridging structure and text for effectiveknowledge graph completion via conditional soft prompting[C]// Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 11489-11503.

[15] LIU X D, HE P C, CHEN W Z, et al. Multi-task deep neural networks for natural language understanding[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4487-4496.

[16] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report [J]. arXiv: 2303. 08774, 2023.

[17] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models [J]. arXiv: 2302. 13971, 2023.

[18] ZENG A H, LIU X, DU Z X, et al. GLM-130B: an open bilingual pre-trained model[J]. arXiv: 2210. 02414, 2022.

[19] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[J]. arXiv: 2303. 18223, 2023.

[20] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Advances in Neural Information ProcessingSystems. 2013: 2787-2795.

[21] YANY B, HE X. Embedding entities and relations for learning and inference in knowledge bases[C]// Proceedings of the 2015 International Conference on Learning Representations. 2015: 1412-1420.

[22] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: knowledge graph embedding by relational rotation in complex space[C]// Proceedings of the 7th International Conference on Learning Representations. 2019: 1-18.

[23] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. Menlo Park: AAAI, 2018: 1811-1818.

[24] NGUYEN D Q, NGUYEN T D, NEUYEN D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network[J]. arXiv: 1712. 02121, 2017.

[25] NATHANI D, CHAUHAN J, SHARMA C, et al. Learning attention-based embeddings for relation prediction in knowledge graphs[J]. arXiv: 1906. 01195, 2019.

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

[27] JU J H, YANY D Q, LIU J P. Commonsense knowledge base completion with relational graph attention network and pre-trained language model[C]// Proceedings of the 31st ACM International Conference on Information and Knowledge Management Association for Computing Machinery. New York: ACM, 2022: 4104-4108.

[28] SUN W, LI Y F, YAO J F, et al. Combining structure embedding and text semantics for efficient knowledge graph completion[C]// Proceedings of the 35th International Conference on Software Engineering and Knowledge Engineering. 2023: 317-322.

[29] ZHU Y Q, WANG X H, CHEN J, et al. Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities[J]. arXiv: 2305. 13168, 2023.

[30] YAO L, PENG J Z, MAO C S, et al. Exploring large language models for knowledge graph completion[J]. arXiv: 2308. 13916, 2023.

[31] WEI Y B, HUANG Q S, ZHANG Y, et al. KICGPT: large language model with knowledge in context for knowledge graph

completion[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL,2023:8667-8683.

[32] AKRAMI F,SAEEF M S,ZHANG Q,et al. Realistic re-evaluation of knowledge graph completion methods: an experimental study[J]. arXiv:2003.08001,2020.

[33] BALAŽEVIC I, ALLEN C, HOSPEDALES T. Tucker: tensor factorization for knowledge graph completion[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. 2019:5185-5194.

[34] VASHISHTH S,SANYAL S,NITIN N,et al. Composition based multi-relational graph convolutional networks[C]// Proceedings of the 8th International Conference on Learning Representations. 2020.

[35] ZHU Y L, LIU J T, RAO Z Y, et al. Knowledge Reasoning Model Combining HousE with Attention Mechanism[J]. Computer Science,2024,51(S1):147-154.

[36] HU B H,ZHANG J P,CHEN H C. Knowledge Graph Completion Algorithm Based on Generative Adversarial Network and Positiveand Unlabeled Learning[J]. Computer Science, 2024, 51(1):310-315.

[37] KIM B, HONG T, KO Y, et al. Multi-task learning for knowledge graph completion with pre-trained language models[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:1737-1743.

[38] CHEN C,WANG Y,LI B,et al. Knowledge is flat:a seq2seq generative framework for various knowledge graph completion [C]// Proceedings of the 29th International Conference on Computational Linguistics. Stroudsburg: ACL,2022:4005-4017.

[39] LIU X Y, WANG Z X, SUN Y, et al. ISA-KGC: integrated semantics-structure analysis in knowledge graph completion[J]. IEEE Access,2024,12:57250-57260.

CAI Qihang, born in 1999, postgraduate, is a member of CCF(No. U9520G). Her main research interest includes knowledge graph completion.

XU Bin, born in 1980, Ph. D, associate professor, is a member of CCF (No. 21664S). His main research interests include artificial intelligence and smart education.

(责任编辑:喻藜)