文章编号:1674-2974(2024)06-0108-11 DOI:10.16339/j.cnki.hdxbzkb.2024271

基于学术知识图谱的增强语义表示与检索

沈思[1†],严大钰[1],卞嘉欣[1],何宏旭[2]
(1. 南京理工大学 经济管理学院, 江苏 南京 210094;
2. 南京农业大学 信息管理学院, 江苏 南京 210095)

摘 要:知识图谱作为一个巨大的知识网络图,其中包含着实体概念、关系等信息.基于深度学习的语义表示虽然泛化性强,但对于一些专有知识的敏感度不高,所以许多研究尝试将知识图谱与神经网络结合.目前大部分知识图谱语义表示的方法是围绕通用领域知识图谱展开的,没有针对学术领域的知识图谱语义表示方法的研究.本文以学术文献的全文本数据为研究对象,从基于学术知识图谱的语义表示方法切入研究,在构建学术知识图谱的基础上,对通用领域的研究方法(K-BERT)进行领域化改进(KEBERT),进一步使用实体知识增强文本的语义信息.通过开展下游任务的对比实验,在学术检索数据集上验证KEBERT、K-BERT和ERNIE的性能.实验采用检索任务中常用的NDCG评价指标对结果进行评价,实验结果表明改进后的KEBERT在检索任务上的效果优于其他模型.

关键词:知识图谱;语义表示;增强语义;学术检索
中图分类号:G255.1 **文献标志码**:A

Enhanced Semantic Representation and Retrieval Based on Academic Knowledge Graph

SHEN Si[1†], YAN Dayu[1], BIAN Jiaxin[1], HE Hongxu[2]
(1. School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China;
2. College of Information Management, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: As a huge knowledge network diagram, the knowledge graph contains entity concepts, relationships, and other information. Although the semantic representation based on deep learning has strong generalization, it is not sensitive to some proprietary knowledge, so many researchers try to combine knowledge graphs with neural network. At present, most of the methods of semantic representation of knowledge graphs are based on general domain knowledge graphs, and there is no research on the semantic representation of knowledge graphs in the academic field. In this paper, the full-text data of academic literature is taken as the research object, and the semantic representation method based on an academic knowledge graphs is studied. On the basis of constructing academic knowledge graph, the research method of the general field (K-BERT) is improved (KEBERT), and entity knowledge is further used to enhance the semantic information of the text. By conducting comparative experiments on downstream tasks, the performance of KEBERT, K-BERT, and ERNIE is verified on academic

* **收稿日期**:2023-06-01
基金项目:国家自然科学基金资助项目(71974094),National Naturel Science Foundation of China (71974094)
作者简介:沈思(1983—),女,江苏南京人,南京理工大学教授,博士
† 通信联系人,E-mail:shensi@njust.edu.cn

retrieval datasets. The experiment uses the NDCG evaluation index commonly used in the retrieval task to evaluate the results. The experimental results show that the improved KEBERT is superior to other models in the retrieval task.

Key words: knowledge graph; semantic representation; enhanced semantics; academic retrieval

随着学术论文数量日益增加,如何有效利用文献成为新的挑战.特别对于检索而言,在进入一个研究领域时,需要快速找到该研究领域的热点等,这就要求检索算法可以准确理解用户检索意图和文献内容,对检索词和文献语义表示进行深入挖掘,然后快速、准确地找到相关文献,提高检索匹配的精确度.面对大量的学术领域数据,如何在数据中挖掘出隐含在学术论文中的知识,为学术论文的检索提供更为丰富的语料数据,对学术搜索和知识搜索的发展具有重要意义的研究问题.知识图谱技术可以在一定程度上解决上述问题.在学术领域,知识图谱可以将学术论文表示为知识对象,利用自然语言处理、知识抽取等技术,实现对学术论文中具体实体、概念等信息的抽取,对学术论文全文本进行细粒度标注,丰富学术图谱的语义信息,从而将这些语义信息融入检索、推荐等下游任务中,其可以有效提高模型的效果,改善用户的使用体验.

同时相关研究一直在尝试通过自然语言处理等技术增强检索系统的智能化程度和对自然语言的理解深度.ArnetMiner、Microsoft Academic Search(MAS)等文献检索系统都在利用实体抽取、数据挖掘等多种技术来提升用户的检索体验.比如MAS检索系统通过建立学术知识图谱,为用户提供更丰富多样的检索结果,可以在检索结果页面给出检索词相关的研究主题和领域相关作者.除此之外,部分研究也在尝试利用知识图谱中的实体信息优化检索模型,现有的方法主要有基于扩展的方法、潜在因素建模的方法、语言建模的方法和深度学习方法等.知识图谱中的实体信息优化检索主要是利用实体信息丰富检索词和文档的语义表示,没有利用知识图谱中的实体关系信息.但是知识图谱中的关系关联着头实体、尾实体,是头尾实体连接的桥梁,包含着大量的语义信息,是知识图谱中非常关键的部分,因此对于关系信息的利用显得尤为重要.

在上述研究背景下,本文先从学术文献全文本入手,对文献全文进行实体相关知识的抽取并构建一个学术领域的知识图谱.随后将现有的知识图谱增强语义表示方法应用到学术领域,并在学术知识图谱数据上进行了实验验证.后续将上述语义表示方法应用于检索任务中,提出了一个基于知识图谱增强语义表示的学术检索模型,并在学术检索数据集上进行了实验验证.

1　相关工作

实体增强的语义表示主要是指将实体本身或实体相关的描述信息、属性信息等作为知识来源,对文本的语义表示进行扩展,为检索、问答等任务提供更好的语义表示.实体增强的语义表示主要由实体融合、文本融合和规则融合模型构成.

现有的知识增强语义表示方法通常将实体视为有监督的数据,以学习其语义信息或者从中获取额外的关键特征.稀有或歧义的实体信息使模型能够更好地学习其语义信息,并在命名实体识别、情感分析、词义消歧以及问答任务中取得了良好的性能效果.Li等[1]提出用字词层面对齐的注意力机制来拓展基于字符层的预训练模型,并且提出了一种池化方法来说明字符层的注意力对齐到字词层,以缓解多源分词器的分词误差传播的问题.Ke等[2]对语言知识加入预训练模型中的有效性展开研究,揭示通过在预训练任务中注入文中提到的知识,来提高情感分析的效果.Roberts等[3]基于T5的text-to-text范式,以closed-book的形式在没有对应答案的上下文的情况下直接向模型输入问题以获取对应的答案,这样的问题设定可以考验模型在微调的过程中能学到什么.SenseBERT[4]模型的优势在于不仅能够预测遮掩掉的词汇,还能够预测当前词汇在给定语境下的实际含义.SKEP[5]是百度提出的一种情感任务上的知识增强方法,SKEP对原始输入句子中的部分词语进行遮掩.Yamada[6]提出了一种新的专门用于处理实

体相关任务的语义表示方法LUKE,是对transformer原有的注意机制的有效扩展.

文本融合的方法在问答任务中有着良好的表现.RAG[7](Retrieval Augmented Generation)是一种将信息检索和seq2seq生成器结合在一起的模型,其中信息检索系统是DPR[8],seq2seq生成器则使用BART[9]模型.采用标准的seq2seq模型就意味着RAG不需要考虑序列的长度和顺序,每一个输入序列都对应一个输出序列.TEK[10]结合检索到的百科全书知识与输入的文本表示进行联合学习,以捕获和保存实体相关的事实知识.REALM[11]通过从大型语料库检索文档来增强语言表示,其中有代表性的为使用transformer实现知识增强的编码器,用于向文档中注入实体并进行预测.

规则可以表示为非正式的约束或是严格的逻辑表达式,它的优点是可解释性强.这些规则不仅可以作为监督信号,而且可以作为独立的推理系统来进行决策.Gangopadhyay等[12]提出了一种结构化的深度学习和符号推理结合的方法来解决图像分类问题.SimpleNet[13]将半词汇法字符串识别为字母表的元素,最后利用规则对歧义标记进行判断,提供了一个基于领域知识的推理框架来解释复杂问题.Deng等[14]提出了一个允许对标签之间的灵活关系进行编码新的模型.Amizadeh等[15]提出将视觉问题问答任务的推理和感知部分分开并进行评估,提出了一种新颖的自上而下的校准技术.PTR[16]根据逻辑规则将与任务和类别相关的先验知识进行表征,以关系分类任务为例,预测的关系通常与句子的关系表达和标记实体的类型相关.

上述各个知识层面的融合策略为本研究的探究一方面提供了方法上的可借鉴之处,另一方面也提供了可操作性的流程和规范,同时也为本研究的算法改进提供了相应的算法理论基础.

基于知识图谱增强的语义表示主要有句法树融合和知识图谱融合两种.

句法知识呈现了句子的关键组成部分,不仅在自然语言推理和理解任务中有着重要作用,而且在句法分析、语义标注和共指消解任务中也起着重要作用.Bai等[17]首次尝试证明语法树在预训练语言模型中的作用,提出将自注意力机制分解为不同的方面.Zhou等[18]提出通过多任务学习来学习跨语言任务的语言表征.Sachan等[19]系统地研究了在语义角色标注、命名实体识别和关系抽取三个有代表性地信息提取任务中将语法树融入transformer中的作用.

知识图谱融合与句法树融合类似,将适当的知识表示学习来作为语义嵌入,如Zhang等[20]提出一种利用富含信息的实体来增强语言表征的方法.Peters等[21]在输入文本中显式建模实体跨度,并使用实体链接从知识图谱中检索相关的实体嵌入,以形成知识增强的实体跨度表示.Wang等[22]提出了一种通过知识库进行视觉问答的方法,这种方法回答用自然语言表述的问题,也可以以自然语言的形式进行回答.Wang等[23]将检索出的视觉概念与知识图谱对齐并连接到子图中,再根据上述的信息构建一个特殊的查询去请求上一步中构建好的子图,最后进行关键词筛选得出与最初问题相对应的答案.Xiong等[24]通过替换的策略引入外部知识,根据文本中的实体,提出了弱监督的训练目标来判断实体是否被替换.Lauscher等[25]提出的LIBERT使用外部的词汇知识来补充原始文本,将满足语义相似度约束的实体对作为训练实例.Shen等[26]提出了一种新的模型,利用知识图谱来帮助选择信息量大且有学习价值的实体进行遮掩,并且使用知识图谱来构造被遮掩掉实体的负样本,从而计算损失函数.COMET[27]模型使用现有的三元组作为种子集进行训练,这种方法可以用来自动构建知识图谱或进行知识图谱补全.GRF是Ji等[28]提出的一个多跳推理框架,在GPT-2的基础上引入了ConceptNet作为外部知识,并且基于文本的上下文和知识图谱进行结尾的生成.Liu等人[29]提出了KG-BART模型,通过融合知识图谱信息和文本信息,生成高质量的句子,纯文本作为结构化知识的补充.KT-NET[30]采用注意力机制从知识库中自适应地选择所需知识,然后将所选知识与BERT进行融合,以实现上下文感知和知识感知预测.

以往语义表示方法的研究对象都局限在文本本身,随着知识图谱这类包含着大量知识信息的知识库的发展,越来越多的学者开始研究如何利用这些知识库中的外部知识对文本的语义表示进行扩展,使模型能更好地理解语言.本研究在综合比较分析前人研究的基础上,提出了一种基于学术知识图谱的实体增强语义表示方法,将实体增强和实体关系增强整合到一个框架当中.

2　增强语义表示研究框架及模型

2.1　研究框架

首先根据用户需求和学术文献特征制定学术实体标注规范,征集标注人员对学术文献数据进行人工标注和校对,利用标注的实体和文献题录数据构建学术知识图谱.其次根据学术数据特点对现有的实体增强语义表示方法进行改进,并利用上述构建的学术知识图谱数据在检索任务上进行实验验证,并且将本研究的语义表示方法与其他两个语义表示方法进行实验对比,分析其实验结果.整体研究流程如图1所示.

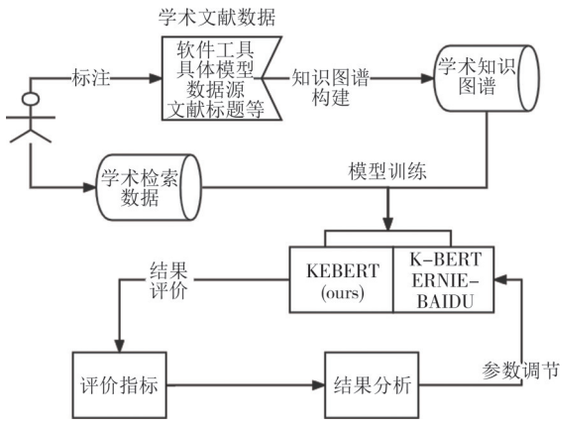


图1　基于学术知识图谱的实体增强语义表示方法研究流程

Fig.1　Research process of entity enhanced semantic representation method based on academic knowledge graph

2.2　KEBERT模型构建

K-BERT模型是一个通用的、基于知识图谱的语义表示模型,本研究对K-BERT模型进行改进,在原有融入知识图谱的基础上,强化实体的整体性与语义性,将实体整体作为一个token进行语义特征提取.以学术图谱为实体和关系的数据支撑,使其能够适应学术领域知识图谱数据.

2.2.1　K-BERT结构

K-BERT模型主要分为四个模块:知识层、嵌入层、可视层和Mask-Transformer编码器.

1)知识层.知识层主要有两个任务,一是将输入文本中的实体链接到知识图谱中,主要是通过知识图谱中的实体与输入文本进行匹配实现的;二是将匹配到的三元组知识插入到输入文本中.如图2所示,以“Tim Cook is visiting Beijing now”为例,输入句子在经过链接后会将“Beijing”链接到知识图谱中的“Beijing”实体,然后就会得到(Beijing,capital,China)三元组,随后可以将三元组中的关系和尾实体插入到句子中,头实体即句子中与三元组相匹配的实体.经过知识层处理,输入的句子会变成一个句子树.

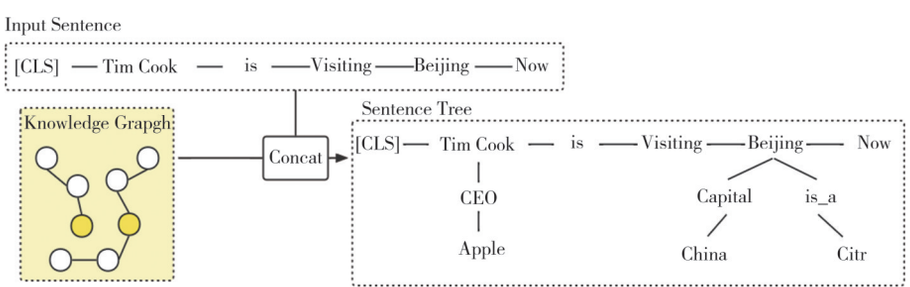


图2　知识层结构示意图

Fig.2　Schematic diagram of knowledge layer structure

$$M_{ij}\begin{cases}0 & w w_j \quad \text{same branch}\\ -\infty & \text{else}\end{cases} \tag{1}$$

M_{ij}本质上是一个掩码矩阵,它的目的是用来避免不相关词之间的交互.式中:W_i表示输入句子中的第i个词;W_j表示输入句子中的第j个词.当两个词同时出现在一个三元组中或者同在一个句子中时,M_{ij}等于0,即W_i需要关注W_j;其余情况下,M_{ij}设置为负无穷,即W_i不需要关注W_j.以图2中的句子树为例,“is”和“visiting”这种树的主干上的实体和词只能与主干上的词交互,“Tim Cook”和“Beijing”这种交叉的实体节点与主干上的词和对应分支上的关系节点、尾实体节点都是可以交互的,但是分属不同分支的节点之间不能交互,分支上的节点只能与当前分支上的节点交互.

4)Transformer层.从某种程度上说,可视矩阵$\boldsymbol{M}$中包含了句子树的结构信息,为了利用这些信息,避免插入的知识对原始句子的语义造成负面影响,本文对Transformer结构进行了调整,使其能够根据矩阵$\boldsymbol{M}$来限制自注意力机制的可见区域.Mask-Transformer由多个mask-self-attention堆叠而成,与BERT结构一样,其层数为L,隐藏层输出的大小为H,mask-self-attention的头数量为A.mask-self-attention的自

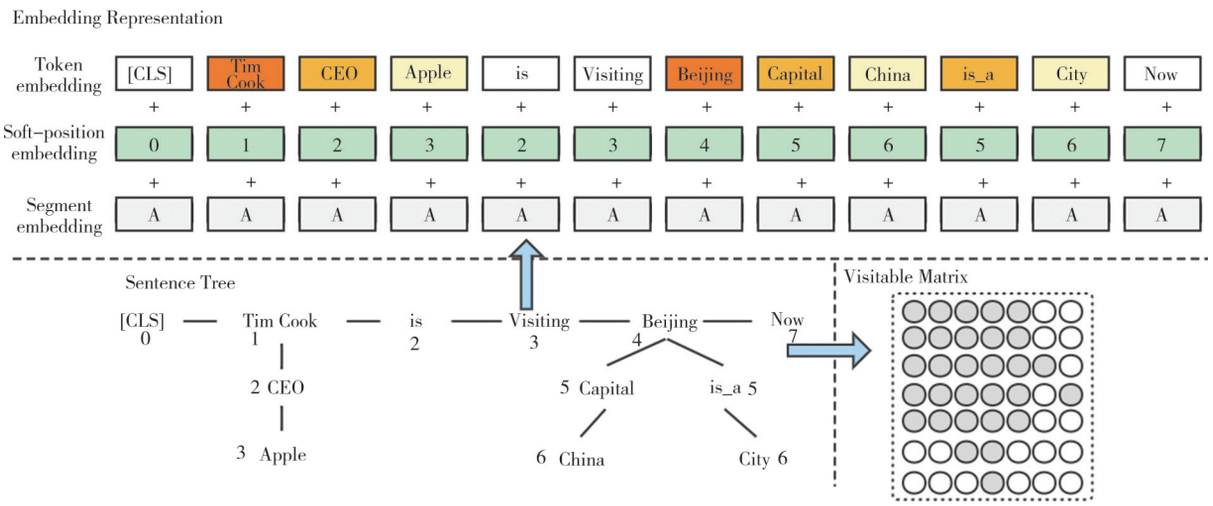


图3 句子树转化成嵌入表示和可视矩阵

Fig.3 Transform sentence trees into embedded representations and visual matrices

注意力计算过程如式(2)~式(4).

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v \tag{2}$$

$$S^{i+1} = \text{soft max}\left(\frac{Q^{i+1}K^{i+1^{\mathrm{T}}} + M}{\sqrt{\mathrm{d}_k}}\right) \tag{3}$$

$$h^{i+1} = S^{i+1}V^{i+1} \tag{4}$$

式中:W_q、W_k、W_v是可训练参数;h_i是第i个mask-self-attention块的隐藏状态,$\boldsymbol{M}$是可视层输出的可视矩阵.

2.2.2 KEBERT模型结构

基于K-BERT方法,本研究针对学术知识图谱数据的特点对上述方法进行了改进,改进后称为KEBERT,具体流程示意如图4所示.改进后的方法主要有四个模块.在输入层将学术文本句子作为输入,抽取出其中的学术实体并进行实体链接.以“support vector machine model uses Twitter data for classification”句子为例,通过实体链接将句子中的“support vector machine”和“Twitter”实体抽取出来并链接到学术知识图谱中,检索到与之相关的三元组.与原始的K-BERT的不同,本文的KEBERT在进行处理时,将句子中的实体作为一个整体进行操作.在知识层,将链接的三元组和原始的学术句子进行组合构造一个句子树.将知识层输出的句子树作为嵌入层的输入进行嵌入表示.嵌入表示主要有三个部分:token embedding、soft position embedding和segment embedding.相比于原始的K-BERT模型在进行嵌入表示时将句子中的每个单词都转化为一个单独的token来进行表示,本文在KEBERT模型中将同一个实体的单词作为一个整体转化为一个token进行表示,在后续的掩码过程中也可以对整个实体同时进行掩码,保留了实体信息的完整性,使模型能够更完整地学习知识图谱中的三元组知识.soft position embedding和segment embedding的表示方法保持不变.在可视层,由于在学术文献中,每个句子都可能包含多个实体.因此可能出现大量的三元组被链接到句子中,但是在这些三元组中只有一部分是对句子的语义理解有效的三元组.因此如果将全部的三元组和句子一起通过自注意力机制,就会使模型对原本的语义理解出现偏差,导致知识噪声问题的出现.因此在可视层中依旧延续前文中的方法,Transformer层也与前文中的处理方法保持一致.

3 增强语义表示及检索实验

3.1 数据集构建

为了验证上述方法的可行性和有效性,本研究通过网络爬虫爬取了JASIST期刊的全文本文献数据,经过一系列处理、标注等操作(图5)构建了学术知识图谱,作为后续实验中的知识图谱数据.

首先确定需要爬取的数据来源及格式,以此为根据编写爬虫获取JASIST期刊的文献数据以及对应文献的题录数据,从而形成知识图谱构建的数据源.随后对数据进行清洗,利用规则化抽取的方法从清洗后的数据中抽取出文献相关的题录数据;除此之外还要对数据进行知识实体的人工标注,根据制定好的标注规范进行细粒度实体标注,并进行多次校对,保证标注的准确度.最后将题录信息和标注的实

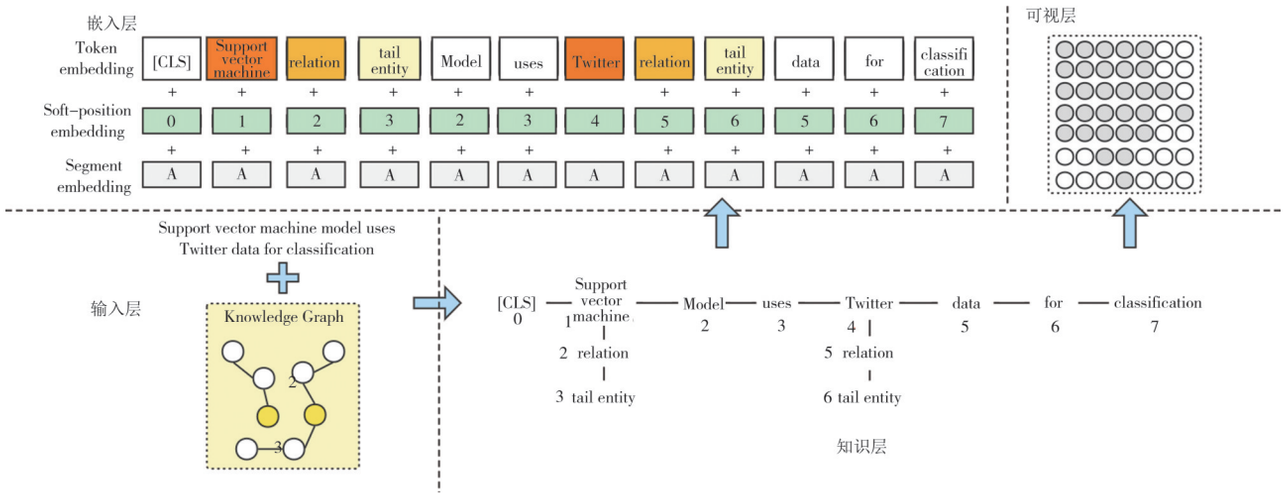


图 4　学术知识图谱增强的语义表示方法示意图

Fig.4　Schematic diagram of semantic representation method for enhancing academic knowledge graph

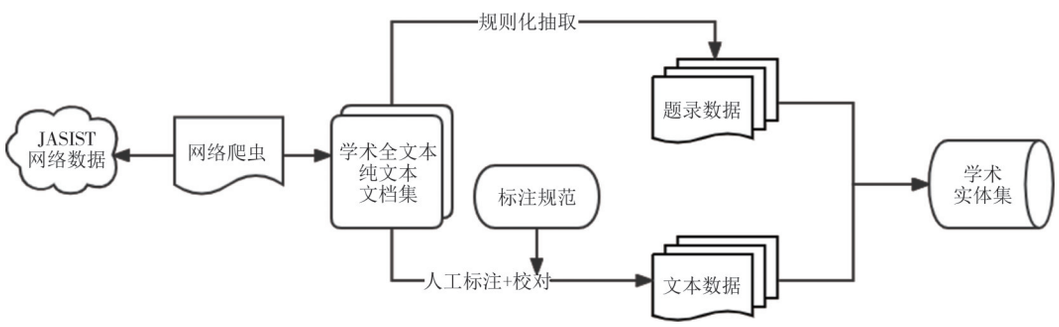


图 5　学术文献数据处理与标注流程图

Fig.5　Academic literature data processing and annotation flowchart

体信息共同作为学术知识图谱的实体来源数据.

本研究采用的知识图谱数据主要来源于2016—2020年JASIST全文本文献,共502篇(表1).在获得数据后首先对文本数据进行清洗,将文献全文按照句子进行划分,文献数据中共包含138 197个句子,平均每个句子有31.88个单词.将每篇文献单独保存为一个txt格式文件,供后续标注使用.

3.2　学术知识图谱构建

根据所构建学术知识图谱的需要,本研究确定了需要标注的三个实体类别,并制定详细的标注规范,进而根据标注规范对文献全文进行实体标注.需要人工标注的实体包括数据源、软件工具、具体模型三类,具体的标注规范如表2所示.标注平台使用BRAT,它是一款应用于webserver端的文本标注工具,可以用于标注实体、关系、事件、属性等信息,可以满足学术知识图谱实体的标注需求.

表 1　学术知识图谱数据源数据分布表

Tab.1　Academic knowledge graph data source data distribution table

项目	数量
文献数量	502
句子总数	138 197
每个句子的平均单词数	31.88

表 2　学术实体标注规范

Tab.2　Specification for academic entity labeling

编号	实体类型	标注规则
1	具体模型	文献中主要描述的模型、方法等,尽量以“model、method”等词语开头或结尾,当出现具体模型时则无须标注这些词语,例如“support vector machine model”仅需标记“support vector machine”.
2	软件工具	“SPSS、VOSviewer、Matlab”等常见的软件工具.
3	数据源	“Wikipedia、Google Scholar”等常见的数据源,或者是文献中自行构建的数据集.

由于人工标注可能会带有主观判断,为了保证标注的准确性,本研究在进行一轮标注后对标注后的实体进行了二次校对,由不同的标注人员再次根据标注规范进行检查,如果两位标注人员的标注结果一致,则认为标注正确,否则将交给第三位标注人员再次进行标注检查,以此确保标注数据的准确性.随后使用SCI-BERT和BERT模型对上述实体进行实体抽取.最后将所有的实体和关系以及对应文献的题录数据存储到Neo4j图数据库中进行知识图谱的构建.总计标注了1 501个实体,5 069个三元组,具体统计数据如表3所示.

表3 学术知识图谱实体分布表

Tab.3 Academic knowledge graph entity distribution table

实体类型	数量
软件工具	367
具体模型	587
数据源	329
文献标题	218
总计	1 501

本研究将构建好的学术知识图谱存储在Neo4j图数据库中,Neo4j是一个高性能的NOSQL图形数据库,可以作为知识图谱的存储工具,可以提供知识图谱、存储、检索和可视化等功能.不同于关系数据库将结构化数据存储在表中,Neo4j将结构化数据存储在图网络中,它不存在表的概念,只有节点和关系,可以理解为图中的点和边,而且节点和关系都可以具备属性,但是Neo4j中的关系为单向关系,不具备双向或者无向关系.构建好的学术知识图谱示例如图6所示.

绿色节点为文献节点,粉色节点代表数据源节点,蓝色节点代表具体模型,黄色节点代表软件工具.绿色文献节点在知识图谱中为中心节点,各个文献的周围有多个子节点,不同颜色的子节点代表不同的实体.多篇文献通过相同的子节点互相连接,这个使文献之间互相连接的子节点包括数据源、具体模型、软件工具.搜索某数据源,查看该对象所有的相关文献.以数据源“Web of Science”为例,查询语句为:MATCH q=()-[]->(n:Datasource {entity: "Web of Science"}) RETURN q.查询后的可视化结果如图7左图所示,可以看出,与“Web of Science”存在关系的有大约20个相关文献,说明这些文献在研究过程中均使用了Web of Science数据.搜索某软件工具,查看该对象所有的相关文献.以软件工具“SPSS”为例,查询语句为:MATCH q=()-[]->(n:Software {entity:"SPSS"}) RETURN q.查询后的可视化结果如图7右图所示,可以看出,与“SPSS”存在关系的有大约7个相关文献,说明这些文献在研究过程中均使用了SPSS软件.

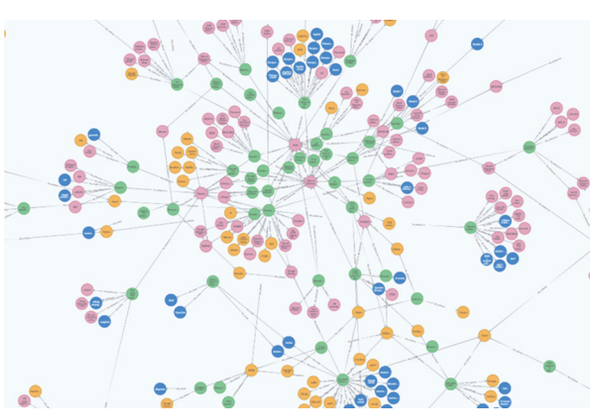


图6 学术知识图谱可视化

Fig.6 Visualization of academic knowledge graph

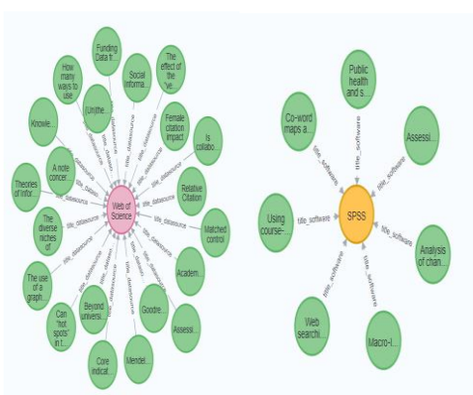


图7 学术知识图谱检索子图可视化

Fig.7 Visualization of academic knowledge graph retrieval subgraphs

3.3 三元组处理与实体链接

本实验中主要使用的实体类型有软件工具、模型、数据源以及对应的实体关系,因此需要对上述知识图谱数据进行预处理,将(头实体-关系-尾实体)转化为相应的文本表示,尤其是实体关系需要表示为相应的文本.以(模型-关系-文献)三元组为例,经过处理将其表示为(模型-a tool used by-文献),以便在后续的嵌入表示过程中得到更准确的语义表示.预处理后的数据示例如图8所示,头实体-关系-尾实体之间采用制表符分隔.

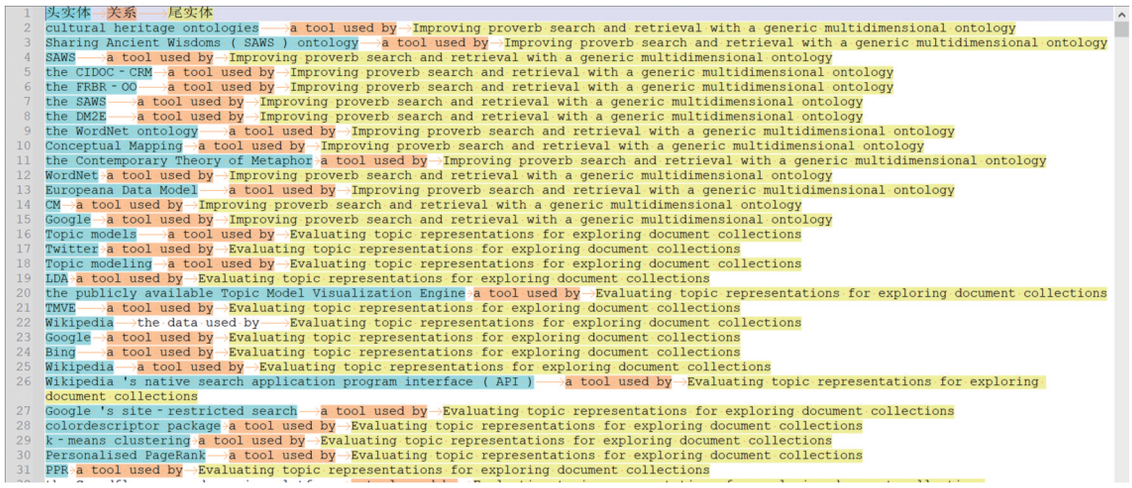


图 8　学术知识图谱数据样例

Fig.8　Sample academic knowledge graph data

在实验过程中需要通过实体链接对输入的检索词和文档进行句子树构建,以检索词“tf-idf similarity”为例,通过实体链接技术在知识图谱中查询到与当前检索词相关的三元组“TF-IDF-a tool used by-A survey on tag recommendation methods”和“TF-IDF-a tool used by-Indexing biomedical documents with a possibilistic network”,因此可以将检索词的树结构设计为如图 9 所示的样式.

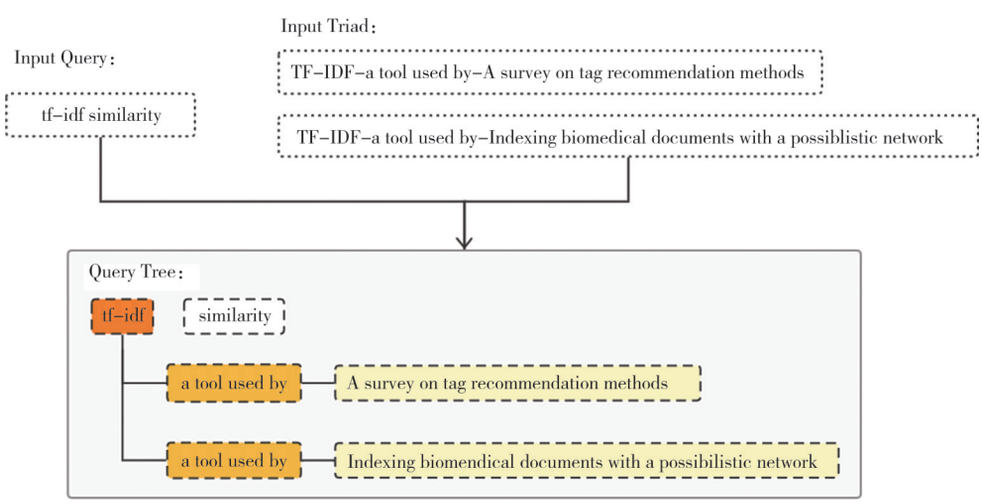


图 9　检索词树结构示意

Fig.9　Schematic diagram of search word tree structure

3.4　检索实验设计与结果分析

为了验证本文提出的知识图谱增强的语义表示方法的有效性,本文在检索任务上进行了初步验证.并且实验采用 NDCG@k 值作为评价指标,对检索结果进行评价与分析.本实验中具体实验细节如图 10 所示.首先本文将检索词和文档用[SEP]进行拼接,输入样例如下:“[CLS] One Sentence containing entity knowledge in JASIST [SEP] One Query in Google [SEP].”其中[CLS]表示包含实体知识句子的开头,第一个 SEP 表示包含实体知识句子与 Google 检索词的分隔符,第二个表示 Google 检索词的结束符.然后通过知识层,链接知识图谱中的三元组,对检索词和文档进行句子树扩展,然后将扩展后的句子树输入向量层进行表示,得到向量表示和对应的可视化矩阵,随后输入 Mask-Transformer 层,随后接入一个池化层构建文档与检索词之间的语义表示,并将此语义表示输入回归层,计算包含实体知识的文档与检索词之间的语义相关度.

本模型采用均方差的方式计算损失,并利用梯度下降的方式对模型的参数进行更新.其中模型参数分别为:热身比率(Warmup Proportion)设置为 0.1,学习率(Learning Rate)为 5.0E-5,最大序列长度

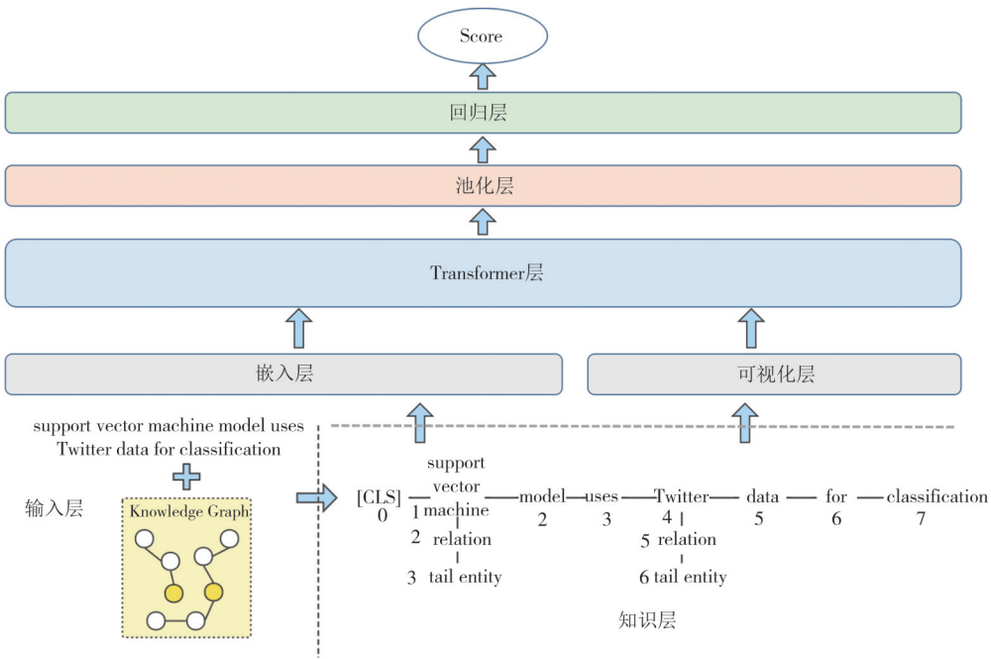


图 10 学术知识图谱增强的语义表示实验图

Fig.10 Experimental diagram of semantic representation enhanced by academic knowledge graph

(Max Sequence Length)为 512,训练周期(Epoch)设置为 3.本研究利用 NDCG@1,NDCG@3,NDCG@5,NDCG@10 对检索性能进行评估,@1 表示只看第一个文档,@3 表示只看前三个文档,以此类推.NDCG 表示从第一个检索文档到第 k 个检索文档的"归一化累计折扣信息收益".然后,对每个检索词下的 NDCG 值进行平均,得到排序模型的整体 NDCG.NDCG 是使用 IDCG 的规范化过程,它表示当前 DCG 值与 IDCG 值的比例.具体公式如公式(5)和公式(6)所示.

$$\mathrm{NDCG}_{@K} = \frac{\mathrm{DCG}_{@K}}{\mathrm{IDCG}_{@K}} \tag{5}$$

$$\mathrm{NDCG}_{@K} = \sum_{i=1}^{k} \frac{\mathrm{rel}_i}{\log_2(i+1)} \tag{6}$$

式中:rel_i 表示当前排序列表中结果 i 的相关性得分;i 是指结果 i 当前的位置序号;IDCG(Ideal DCG)则是将相关性打分由大到小排列之后,按照 DCG 的公式进行计算得到的.

在对上述模型进行实验验证时,将基础的 K-BERT 模型和 ERNIE 模型作为本实验的对比模型,采用 NDCG@1、NDCG@3、NDCG@5、NDCG@10 四个评价指标分别对比实验效果,具体实验结果如表 4 所示.

由上表可以看出,KEBERT 模型和 K-BERT 模型在四个指标上都略高于 ERNIE 模型,在整体性能上更有优势.具体来说,本文提出的 KEBERT 模型的

表 4 知识图谱增强语义表示方法实验结果

Tab.4 Experimental results of knowledge graph enhanced semantic representation method

评价指标	本文方法	对比模型	
		K-BERT	ERNIE
NDCG@1	**29.9%**	29.7%	29.7%
NDCG@3	**44.2%**	42.8%	37.6%
NDCG@5	52.8%	**53.1%**	48.0%
NDCG@10	71.5%	**71.6%**	69.9%

NDCG@1、NDCG@3 分别为 29.9% 和 44.2%,均大于 K-BERT 的 29.7% 和 42.8%,在这两个指标上的效果都优于其余两个模型,说明在检索词和文档的匹配程度上,本文提出的 KEBERT 模型更优.这一结果的原因主要为改进后的模型利用知识图谱中的实体在检索词和文档的表示中起到了扩展优化的作用,使得原始文本中的知识实体获得了拓展,在进行相似度计算时提高了匹配精度,更容易检索到与检索词最相关的文档.在 NDCG@5、NDCG@10 两个指标上,本文提出的 KEBERT 和原始模型的表现效果相比略低,但是整体来看两个模型的表现效果相当.

4 结 语

针对当前学术知识图谱增强检索应用研究中的不足,本文从知识图谱的角度出发,结合学术全文本数据和学术检索数据,构建了学术知识图谱,并探究

了如何将学术领域知识图谱融入到文本的语义表示中.在语义表示方法部分,本文结合学术领域数据的特点,将知识图谱增强的语义表示方法应用到学术文本中,并与多个实体增强语义表示方法进行了实验对比,证明了本研究在学术文本数据中的有效性.主要工作和研究成果如下.一方面,本文将502篇JASIST学术文献全文本作为数据源,构建学术领域知识图谱,并储存在Neo4j图数据库中,供后续实验使用.另一方面,将知识图谱增强的语义表示方法应用于学术领域数据.本文基于K-BERT模型,对实体链接部分进行了修改,将改进后的语义表示方法命名为KEBERT.后续本研究利用学术知识图谱数据进行了实验,将KEBERT与其他两种实体增强语义表示方法进行了实验对比,对比分析后发现,KEBERT方法在语义表示方面要优于其他两种方法,证明了本研究实验的有效性.

由于学术文本数据的特殊性和多样性,以及实验的研究广度有限,本文还存在以下不足以及要进一步完善的地方.一方面,数据量较小,在后续的研究中可以扩大数据量和数据的学科类别,构建更为全面的面向全文本的学术知识图谱.另一方面,本文是在K-BERT模型的基础上进行了改进,使该模型能够更好地适用于学术数据.在今后的研究中,可以将此类模型作为对比,选择更优的适用于学术领域数据的语义表示模型.

参考文献

[1] LI Y Z, YU B W, XUE M G, et al. Enhancing pre-trained Chinese character representation with word-aligned attention[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 3442-3448.

[2] KE P, JI H Z, LIU S Y, et al. SentiLARE: sentiment-aware language representation learning with linguistic knowledge[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6975-6988.

[3] ROBERTS A, RAFFEL C, SHAZEER N. How much knowledge can you pack into the parameters of a language model?[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 5418-5426.

[4] LEVINE Y, LENZ B, DAGAN O, et al. SenseBERT: driving some sense into BERT[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 4656-4667.

[5] TIAN H, GAO C, XIAO X Y, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 4067-4076.

[6] YAMADA I, ASAI A, SHINDO H, et al. LUKE: deep contextualized entity representations with entity-aware self-attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6442-6454.

[7] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, Vancouver, BC, Canada: ACM, 2020: 9459-9474.

[8] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6769-6781.

[9] LEWIS M, LIU Y H, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 7871-7880.

[10] JOSHI M, LEE K, LUAN Y, et al. Contextualized representations using textual encyclopedic knowledge[J]. ArXiv e-Prints, 2020: arXiv: 2004.12006.

[11] GUU K, LEE K, TUNG Z, et al. REALM: retrieval-augmented language model pre-training [C]//Proceedings of the 37th International Conference on Machine Learning. ACM, 2020: 3929-3938.

[12] GANGOPADHYAY B, HAZRA S, DASGUPTA P. Semi-lexical languages: a formal basis for using domain knowledge to resolve ambiguities in deep-learning based computer vision[J]. Pattern Recognition Letters, 2021, 152: 143-149.

[13] HASANPOUR S H, ROUHANI M, FAYYAZ M, et al. Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures[J]. ArXiv e-Prints, 2016: arXiv: 1608.06037.

[14] DENG J, DING N, JIA Y Q, et al. Large-scale object

classification using label relation graphs [C]//European Conference on Computer Vision. Cham:Springer,2014:48-64.

[15] AMIZADEH S, PALANGI H, POLOZOV O, et al. Neuro-symbolic visual reasoning: disentangling "visual" from "reasoning"[C]// Proceedings of the International Conference on Machine Learning (ICML). 2020.

[16] HAN X, ZHAO W L, DING N, et al. PTR: prompt tuning with rules for text classification[J]. AI Open,2022,3:182-192.

[17] BAI J G, WANG Y J, CHEN Y R, et al. Syntax-BERT: improving pre-trained transformers with syntax trees[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 3011-3020.

[18] ZHOU J R, ZHANG Z S, ZHAO H, et al. LIMIT-BERT: linguistics informed multi-task BERT [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online. Stroudsburg, PA, USA: Association for Computational Linguistics,2020:4450-4461.

[19] SACHAN D S, ZHANG Y, QI P, et al. Do syntax trees help pre-trained transformers extract information? [C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021: 2647-2661.

[20] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics,2019:1441-1451.

[21] PETERS M E, NEUMANN M, LOGAN R, et al. Knowledge enhanced contextual word representations[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019:43-54.

[22] WANG P, WU Q, SHEN C H, et al. Explicit knowledge-based reasoning for visual question answering[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: ACM, 2017: 1290-1296.

[23] WANG P, WU Q, SHEN C H, et al. FVQA: fact-based visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(10):2413-2427.

[24] XIONG W, DU J, WANG W Y, et al.Pretrained encyclopedia: weakly supervised knowledge-pretrained language model [C]// International Conference on Learning Representations.2020.

[25] LAUSCHER A, VULIĆ I, PONTI E M, et al. Specializing unsupervised pretraining models for word-level semantic similarity [C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online). Stroudsburg, PA, USA: International Committee on Computational Linguistics,2020:1371-1383.

[26] SHEN T, MAO Y, HE P C, et al. Exploiting structured knowledge in text via graph-guided representation learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics,2020:8980-8994.

[27] BOSSELUT A, RASHKIN H, SAP M, et al. COMET: commonsense transformers for automatic knowledge graph construction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019:4762-4779.

[28] JI H Z, KE P, HUANG S H, et al. Language generation with multi-hop reasoning on commonsense knowledge graph [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 725-736.

[29] LIU Y, WAN Y, HE L F, et al. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021,35(7): 6418-6425.

[30] YANG A, WANG Q, LIU J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019:2346-2357.