

面向领域知识图谱的实体关系抽取模型仿真

何山, 肖晰, 张嘉玲

(西南石油大学 计算机与软件学院, 成都 610599)

摘要: 针对目前领域知识图谱实体关系抽取效果不佳的问题, 提出一种面向领域知识图谱的实体关系抽取模型研究方法. 先建立由编解码模块、实体识别模块和实体关系抽取模块组成的实体关系抽取模型, 在实体关系抽取模型中, 通过双向长短期记忆神经网络对文本句子进行编码处理, 将编码后文本句子特征表示向量输入至基于深度神经网络的实体识别模块中进行文本句子的实体识别, 并将识别结果输入至基于卷积神经网络的实体关系抽取模块中进行实体关系抽取, 然后将实体关系抽取获取的实体关系三元组输入至编解码模块中进行解码操作, 实现最终的面向领域知识图谱的实体关系抽取. 实验结果表明, 该方法的实体关系抽取效果和整体应用效果较好.

关键词: 知识图谱; 实体关系抽取; 实体识别; 卷积神经网络

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5489(2025)02-0465-07

Simulation of Entity Relationship Extraction Model for Domain Knowledge Graph

HE Shan, XIAO Xi, ZHANG Jialing

(School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610599, China)

Abstract: Aiming at the problem of poor performance of entity relationship extraction in current domain knowledge graphs, we proposed a research method for entity relationship extraction models oriented towards domain knowledge graphs. Firstly, we established an entity relationship extraction model consisting of an encoding and decoding module, an entity recognition module, and an entity relationship extraction module. In the entity relationship extraction model, a bidirectional long short-term memory neural network was used to encode text sentences, and the feature representation vectors of the encoded text sentences were input into a deep neural network-based entity recognition module for entity recognition of text sentences, and the recognition results were input into the entity relationship extraction module based on convolutional neural networks for entity relationship extraction. Secondly, the entity relationship triplet obtained from entity relationship extraction was input into the encoding and decoding module for decoding operation, achieving the final entity relationship extraction for domain oriented knowledge graph. The experimental results show that the proposed method has better entity relationship extraction effect and overall application effect.

Keywords: knowledge graph; entity relationship extraction; entity recognition; convolutional neural network

收稿日期: 2024-02-07.

第一作者简介: 何山(1972—), 男, 汉族, 硕士, 副教授, 从事大数据挖掘和机器学习的研究, E-mail: heshanwjz@163.com.

基金项目: 国家自然科学基金面上项目(批准号: 62276099).

领域知识图谱是一种将某一特定领域知识可视化的知识表示方法,是目前人工智能领域的研究热点.知识图谱实体关系抽取是建立知识图谱的重要步骤,但其在面临复杂语义关系时,常存在实体关系抽取效果不佳的问题,因此,进行面向领域知识图谱的实体关系抽取方法研究有一定的现实意义^[1-2].

夏鸿斌等^[3]建立了由基于 BERT(bidirectional encoder representations from Transformers)编码和链式解码模块、实体抽取和关系抽取模块、关系修正模块组成的实体关系抽取模型,但该方法的计算复杂度较高,且依赖于训练数据的质量,实体关系抽取结果质量较低;张亮等^[4]建立了由基于 PATB(position and attention based Booster)信息聚合器的编解码模块、基于实体抽取器的实体识别模块等模块组成的实体关系抽取模型,但该方法在构建模型时未考虑实体之间存在的多种关系,导致实体关系抽取效果不佳;Huang 等^[5]先进行实体识别,然后将识别出的实体输入至长短期记忆(LSTM)网络完成实体关系抽取,但该方法存在因数据稀疏性导致 LSTM 网络处理实体关系时性能下降的问题,从而导致实体关系抽取中文本编解码效果不佳;Sun 等^[6]通过选择门网络实现实体关系抽取,但该方法对特定任务具有依赖性,且需要大量标注数据,导致实体关系抽取模型的可解释性不足.为解决上述方法中存在的问题,本文设计一种面向领域知识图谱的实体关系抽取模型.

1 实体关系抽取模型构建

实体关系抽取模型是一种从文本中识别实体并建立实体之间关系的模型,本文建立基于深度神经网络的实体识别模块、基于双向长短期记忆神经网络的编解码模块和基于卷积神经网络的实体关系抽取模块三部分组成的实体关系抽取模型^[7-8].其中,深度学习神经网络可通过层叠多个隐藏层提取更高级别的特征表示,使实体识别模块能更好地理解复杂的文本语义信息;双向长短期记忆神经网络能捕捉上下文依赖关系,允许模型同时考虑过去和未来的上下文,从而提高编解码的精度和效果;卷积神经网络在局部区域上进行参数共享和池化操作,能有效捕获局部特征并保持空间位置信息,因此适用于实体关系的抽取.综合使用这些神经网络模型,可有效提取文本中的实体和实体关系,实现精准实体关系抽取任务.本文实体关系抽取模型结构如图 1 所示.

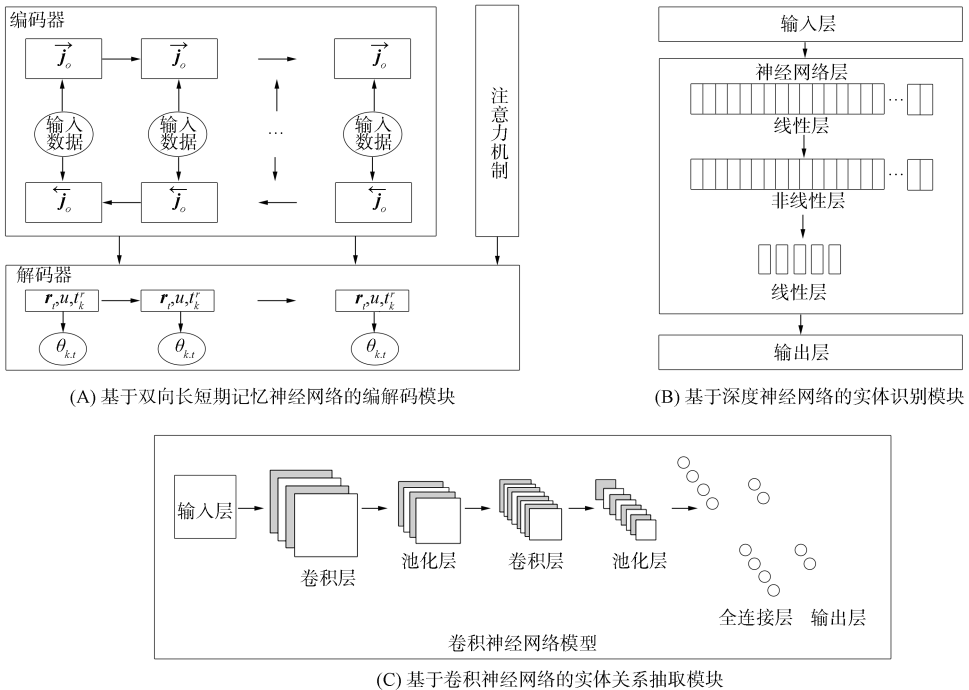


图 1 实体关系抽取模型示意图

Fig. 1 Schematic diagram of entity relationship extraction model

在建立的实体关系抽取模型中,实体识别模块可获取文本句子的实体,为后续的实体关系抽取提供基础数据;编解码模块中的编码用于实体关系抽取前的文本句子特征向量表示,解码用于实体关系抽取后的原始文本形式转换;实体关系抽取模块用于实现实体关系抽取^[9-10].

实体关系抽取模型步骤如下:

- 1) 利用基于双向长短期记忆神经网络的编解码器转换文本句子至适用于实体识别和实体关系抽取的表示形式;
- 2) 将通过编码器编码获取的文本句子特征表示向量输入至实体识别模块中,获取文本句子中的实体;
- 3) 将获取的文本句子实体实行编码处理后,继续输入至实体关系抽取模型中,实现实体关系抽取;
- 4) 将获取的实体关系抽取三元组输入至编解码器中实现解码处理,恢复实体关系的原始表达.

1.1 实体编码处理

由于文本句子和实体通常以字符或词的形式存在,而神经网络^[11]对字符或词序列的直接处理较困难,因此在进行实体识别与实体关系抽取前,需先对文本句子和实体进行编码操作,通过编码转换文本句子和实体为特征表示向量形式^[12-13].通过双向长短期记忆网络模型对文本句子进行编码处理,文本句子/实体的特征表示向量可通过双向长短期记忆网络编码器的前向传播输出与后向传播输出结果拼接得到.

编码器前向传播输出值 \vec{j}_o 用公式表示为

$$\vec{j}_o = \vec{E}_{\text{LSTM}}(\vec{j}_{o+1}, r_o), \quad (1)$$

其中 \vec{E}_{LSTM} 表示长短期记忆神经网络前向编码函数, \vec{j}_{o+1} 表示第 $(o+1)$ 个文本句子/实体的双向特征向量, r_o 表示第 o 个文本句子/实体的嵌入向量.编码器后向传播输出值 \vec{j}_o 用公式表示为

$$\vec{j}_o = \vec{E}_{\text{LSTM}}(\vec{j}_{o-1}, r_o), \quad (2)$$

其中 \vec{E}_{LSTM} 表示长短期记忆神经网络后向编码函数, \vec{j}_{o-1} 表示第 $(o-1)$ 个文本句子/实体的双向特征向量.

拼接前向传播输出和后向传播输出,得到用于实体识别和实体关系抽取的文本句子/实体特征表示向量 j_o^R 为

$$j_o^R = (\vec{j}_o; \vec{j}_o). \quad (3)$$

1.2 实体识别模块

将获取的文本句子特征表示向量输入至基于深度神经网络中完成实体识别,获取实体数据,为领域知识图谱的实体关系抽取提供基础数据.实体识别即通过实体识别模型获取文本中具有特定意义的实体,通过对输入文本句子的特征表示向量进行标签预测,选择文本最优描述标签的过程^[14-15].

神经网络分为输入层、神经网络层和输出层,输入层用于输入数据的低维映射,神经网络层由两个线性层夹杂一个非线性层组成,其用于输入数据的特征表示,在输出层获取文本句子的标签预测结果,即实体识别结果.将文本句子特征表示向量 j_o^R 作为深度神经网络的输入,通过神经网络层得到输入文本句子特征表示向量 j_o^R 的特征表示 E_{hid}^v 为

$$E_{\text{hid}}^v = \omega_{\text{hid}} j_o^R + b_{\text{hid}}, \quad (4)$$

其中 ω_{hid} 表示权重, b_{hid} 表示偏置.

在输出层得到文本句子特征表示向量 j_o^R 的标签预测 $\text{score}^v(\theta, j_o^R)$,用公式表示为

$$\text{score}^v(\theta, j_o^R) = E_{\text{out}}^v \times h(E_{\text{hid}}^v \times g_1^v(j_o^R) + n_{\text{hid}}^v) + n_{\text{out}}^v, \quad (5)$$

其中 E_{out}^v 表示输出层的文字特征向量, n_{hid}^v 和 n_{out}^v 分别表示隐含层和输出层的训练参数, h 表示激活函数, θ 表示输入的真实标签.

1.3 实体关系抽取模块

实体关系抽取即从句子中获取头实体、尾实体、关系组成的实体关系三元组,将识别得到的文本句子实体输入至基于卷积神经网络的实体关系抽取模块中,然后通过卷积层的卷积操作获取特征图,

再通过池化层优化卷积层获取的特征图, 获取对应的特征映射图, 其中分别设置 2 层卷积层与池化层, 最后经全连接层得到实体关系抽取结果.

经卷积层特征处理输出的特征图 k_l 表示为

$$k_l = \tanh(e_v \cdot \text{score}^v(\theta, \mathbf{j}_o^R)), \quad (6)$$

其中 e_v 表示权重. 在池化层进行最大池采样, 以进一步获取卷积层中有用的局部特征信息 $A(o)$, 用公式表示为

$$A(o) = \max_{l=1,2,\dots,L} \{k_l(o)\}. \quad (7)$$

选择表示关系分类参数的五元组 $\theta = (c; M; E_1; E_2; E_3)$ 作为卷积神经网络的参数, 其中 c 表示输入文本, M 表示网络结构, E_1, E_2, E_3 表示网络的权重参数. 最终得到输出概率分布结果 $A(o|c, \theta)$, 即实体关系抽取结果为

$$A(o|c, \theta) = \frac{e^{p_k}}{\sum_{l=1}^n e^{p_l}}, \quad (8)$$

其中 e^{p_k}, e^{p_l} 分别表示第 k, l 个分量包含关系 p 的权重, n 表示所有关系分类的数量.

对卷积神经网络进行训练, 以优化神经网络的关系分类参数 θ , 通过优化网络的参数逐渐调整到更准确的状态, 使网络能更好地识别和抽取文本中的实体关系. 关系分类参数 θ 的优化可利用随机梯度下降法获取参数的最大对数似然值实现, 对一个句子中的实体训练数据对 (c, u) , 可得参数的对数似然值 $K(\theta)$ 为

$$K(\theta) = \sum_{c=1}^Y \log p(u|c, \theta). \quad (9)$$

通过随机梯度下降法最大化对数似然值, 更新网络参数 θ :

$$\theta' = \theta + \mu \frac{\partial \log p(u|c, \theta)}{K(\theta)}, \quad (10)$$

其中 μ 表示关系索引, ∂ 表示偏导率.

利用卷积神经网络参数 θ 的不断优化, 即可获取最佳输出概率分布结果, 实现实体关系抽取.

1.4 实体解码处理

在利用实体关系抽取获取实体关系三元组后, 仍需通过双向长短期记忆神经网络完成解码操作. 将上下文向量 \mathbf{r}_t 、实体关系三元组 u 、关系嵌入 t_k^r 作为解码器的输入, 通过解码层和映射层获取解码结果^[16-17]. 关系嵌入即将结构化数据中的关系转变为向量表示, 可将关系嵌入描述为词嵌入, 根据关系索引 μ 查询关系的向量表示.

在解码器中引入注意力机制以获取解码器隐含向量的上下文向量 \mathbf{r}_t , 用公式表示为

$$\mathbf{r}_t = \text{Attention}(\mathbf{j}_{k,t}^F, \mathbf{j}_{k,t}^R) \times \theta', \quad (11)$$

其中 $\mathbf{j}_{k,t}^F$ 表示解码层隐含向量, $\mathbf{j}_{k,t}^R$ 表示编码层隐含向量. 在解码层中, 可得隐含向量 $\mathbf{j}_{k,t}^F$ 如下:

$$\mathbf{j}_{k,t}^F = \text{LSTM}^F(\mathbf{r}_{t-1} \parallel t_k^r \parallel u, \mathbf{j}_{k,t-1}^F), \quad (12)$$

其中 LSTM^F 表示解码器的计算单元, \mathbf{r}_{t-1} 表示上一时刻的上下文向量, $\mathbf{j}_{k,t-1}^F$ 表示上一时刻的隐含向量. 最后将解码层获取的信息输入至映射层中获取原始文本形式 $\theta_{k,t}$ 如下:

$$\theta_{k,t} = \text{Softmax}\{\hat{\theta}_{k,t}, \mathbf{j}_{k,t}^F\}. \quad (13)$$

通过将实体关系抽取的结果输入至编解码器进行解码处理, 可实现对实体关系的还原和重建, 从而更好地理解文本中实体之间的关联性.

2 实验与结果分析

为验证面向领域知识图谱实体关系抽取模型的有效性, 下面对其进行测试.

2.1 实体关系与实验指标选取

选择维基百科的地质领域公共数据集作为实验对象, 根据地质句子实例, 将实体关系类型分为

实例、子类、属于、可分为、位于、用途、形状、形成原因、地质年代、颜色、组成和其他关系, 地质句子的实体关系抽取表示实例如图 2 所示。

采用本文方法、文献[3]方法和文献[4]方法, 分别在编解码效果、实体识别效果和实体关系抽取效果三方面进行对比。实验采用编解码效果和 P-R (precision-recall) 曲线两个评价指标。

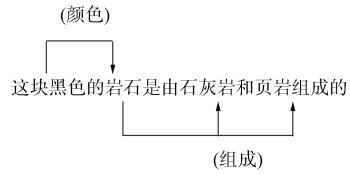


图 2 实例样本

Fig. 2 Example sample

1) 编解码效果: 文本的编解码效果直接影响实体关系抽取效果的好坏, 故要进行编解码效果的测试。引入 ROUGE (recall-oriented understudy for gisting evaluation) 指标评价文本编解码效果, ROUGE 可用于评估模型生成的文本特征表示与人工文本特征表示之间的相似度。

2) P-R 曲线: 引入 P-R 曲线评价实体识别效果与实体关系抽取效果。P-R 曲线是一种描述查准率与查全率关系的曲线, P-R 曲线下面积越大, 表明实体识别、实体关系抽取效果越好。

2.2 抽取效果分析

2.2.1 编解码效果

采用文献[3]方法、文献[4]方法和本文方法完成文本编解码处理, 记录 3 种方法的 ROUGE-1, ROUGE-2, ROUGE-L 结果, 其中 ROUGE-1, ROUGE-2, ROUGE-L 分别表示基于词的评估效果、基于词对的评估效果和基于最长公共子序列的评估效果。实验结果如图 3 所示。

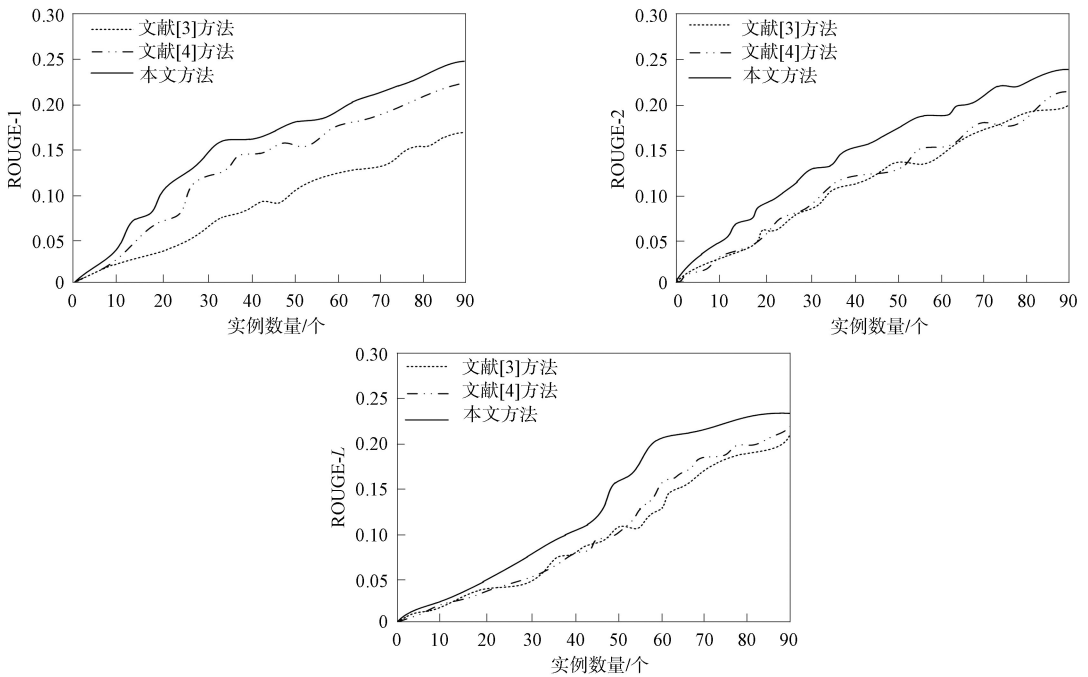


图 3 不同方法的编解码效果对比

Fig. 3 Comparison of encoding and decoding effects of different methods

由图 3 可见, 本文方法在 ROUGE-1, ROUGE-2, ROUGE-L 上的结果均高于文献[3]方法和文献[4]方法, 表明本文方法生成的文本特征表示与人工文本特征表示之间的相似度更高, 编解码效果更好, 更利于后续实现实体关系抽取模型中的实体关系抽取。

2.2.2 P-R 曲线

采用本文方法、文献[3]方法和文献[4]方法完成实体识别和实体关系抽取, 记录 3 种方法的 P-R 曲线, 结果如图 4 所示。由图 4 可见, 在实体识别的 P-R 曲线上, 本文方法 P-R 曲线下面积明显大于文献[3]方法和文献[4]方法, 表明本文方法的实体识别效果更好; 在实体关系抽取的 P-R 曲线上, 本文方法的 P-R 曲线下面积同样大于文献[3]方法和文献[4]方法, 表明本文方法的实体关系抽取效果更

佳. 综合结果表明, 本文方法构建的实体关系抽取模型应用效果更好.

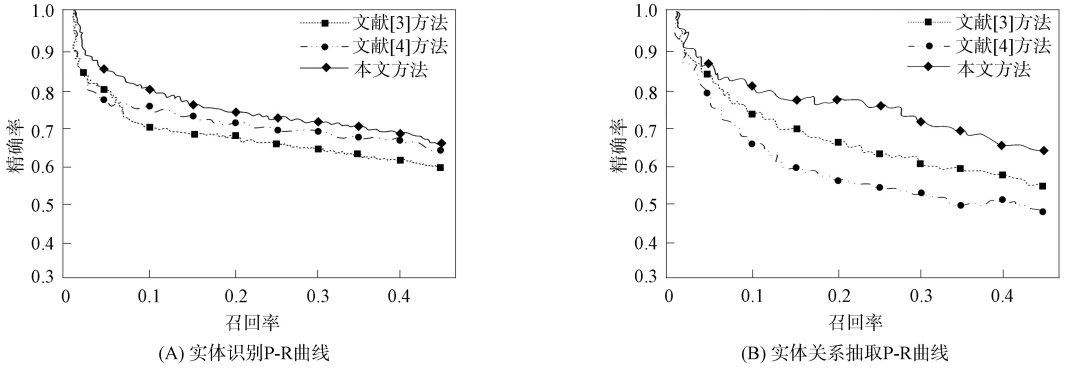


图 4 不同方法的 P-R 曲线对比

Fig. 4 Comparison of P-R curves of different methods

上述实验结果表明, 本文方法的实体关系抽取效果较好, 这是因为该方法建立了由基于深度神经网络的实体识别模块、基于双向长短期记忆网络的编解码模块和基于卷积神经网络的实体关系抽取模块组成的面向领域知识图谱的实体关系抽取模型, 编解码模块过程准确且流畅, 结果一致性高, 实体识别模块准确获得了实体数据, 提高了后续实体关系抽取的准确性, 实体关系抽取模型也具有好的性能表达能力, 最终获得了实体关系抽取效果良好的实体关系抽取模型.

综上所述, 针对目前领域知识图谱中实体关系抽取效果不佳的问题, 为提高实体关系抽取任务的准确性和应用能力, 本文提出了一种面向领域知识图谱实体关系抽取模型的方法. 该方法将文本句子经过双向长短期记忆神经网络进行编码处理, 捕捉上下文依赖关系, 提高了实体识别模块对文本语义的理解能力. 通过神经网络模型实现实体识别, 可以更好地理解文本中的实体信息. 采用卷积神经网络模型进行实体关系抽取, 捕捉局部特征并保持空间位置信息, 提高了抽取的准确性. 实验结果表明, 该方法在实体关系抽取任务中具有更好的性能和应用效果.

参 考 文 献

- [1] 赵丹丹, 张俊朋, 孟佳娜, 等. 基于预训练模型和混合神经网络的医疗实体关系抽取 [J]. 北京大学学报(自然科学版), 2023, 59(1): 65-75. (ZHAO D D, ZHANG J P, MENG J N, et al. Medical Entity Relation Extraction Based on Pre-trained Model and Hybrid Neural Network [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2023, 59(1): 65-75.)
- [2] 胡代旺, 焦一源, 李雁妮. 一种新型高效的文库知识图谱实体关系抽取算法 [J]. 西安电子科技大学学报, 2021, 48(6): 75-83. (HU D W, JIAO Y Y, LI Y N. Novel and Efficient Algorithm for Entity Relation Extraction with the Corpus Knowledge Graph [J]. Journal of Xidian University, 2021, 48(6): 75-83.)
- [3] 夏鸿斌, 沈健, 刘渊. 基于过滤机制的链式实体关系抽取模型 [J]. 模式识别与人工智能, 2023, 36(7): 590-601. (XIA H B, SHEN J, LIU Y. Chain Entity Relation Extraction Model with Filtering Mechanism [J]. Pattern Recognition and Artificial Intelligence, 2023, 36(7): 590-601.)
- [4] 张亮, 卢玲, 王爱娟, 等. PATB: 一种面向联合实体和关系抽取的信息聚合器 [J]. 小型微型计算机系统, 2023, 44(10): 2338-2345. (ZHANG L, LU L, WANG A J, et al. PATB: An Information Booster for Joint Entity and Relationship Extraction [J]. Journal of Chinese Computer Systems, 2023, 44(10): 2338-2345.)
- [5] HUANG H Y, LEI M, FENG C. Graph-Based Reasoning Model for Multiple Relation Extraction [J]. Neurocomputing, 2021, 420(8): 162-170.
- [6] SUN J, LI Y, SHEN Y T, et al. Selection Gate-Based Networks for Semantic Relation Extraction [J]. International Journal of Embedded Systems, 2021, 14(3): 211-217.
- [7] 丁相国, 桑基韬. 基于关系自适应解码的实体关系联合抽取 [J]. 计算机应用, 2021, 41(1): 29-35. (DING X G, SANG J T. Joint Extraction of Entities and Relations Based on Relation-Adaptive Decoding [J]. Journal of Computer Applications, 2021, 41(1): 29-35.)

- [8] 乔勇鹏,于亚新,刘树越,等. 图卷积增强多路解码的实体关系联合抽取模型 [J]. 计算机研究与发展, 2023, 60(1): 153-166. (QIAO Y P, YU Y X, LIU S Y, et al. Graph Convolution-Enhanced Multi-channel Decoding Joint Entity and Relation Extraction Model [J]. Journal of Computer Research and Development, 2023, 60(1): 153-166.)
- [9] 李晓林,潘治霖,邓庆康,等. 基于融合关系信息编码的法律文书实体关系抽取方法 [J]. 中文信息学报, 2023, 37(4): 90-97. (LI X L, PAN Z L, DENG Q K, et al. Relation Enhanced Embedding Based Entities Relation Extraction from Legal Documents [J]. Journal of Chinese Information Processing, 2023, 37(4): 90-97.)
- [10] 廖开际,邹珂欣,席运江. 一种在线医疗社区问答文本实体识别方法——基于卷积神经网络和双向长短期记忆神经网络 [J]. 科技管理研究, 2021, 41(8): 173-179. (LIAO K J, ZOU K X, XI Y J. An Online Medical Community Q&A Text Entity Recognition Method: Based on CNN and BiLSTM [J]. Science and Technology Management Research, 2021, 41(8): 173-179.)
- [11] 汤志康,武毓琦,李春英,等. 基于知识图谱卷积网络的学习资源推荐 [J]. 计算机工程, 2024, 50(9): 153-160. (TANG Z K, WU Y Q, LI C Y, et al. Recommendation of Learning Resource Based on Knowledge Graph Convolutional Network [J]. Computer Engineering, 2024, 50(9): 153-160.)
- [12] 张洪程,李林育,杨莉,等. 基于对比学习与语言模型增强嵌入的知识图谱补全 [J]. 计算机工程, 2024, 50(4): 168-176. (ZHANG H C, LI L Y, YANG L, et al. Knowledge Graph Completion Based on Contrastive Learning and Language Model-Enhanced Embedding [J]. Computer Engineering, 2024, 50(4): 168-176.)
- [13] 景鹏,袁代标,杜刘洋,等. 基于科学知识图谱的自动驾驶技术接受度研究综述 [J]. 江苏大学学报(自然科学版), 2023, 44(1): 14-21. (JING P, YUAN D B, DU L Y, et al. Research of Acceptance of Autonomous Vehicles Technology Based on Mapping Knowledge Domain [J]. Journal of Jiangsu University (Natural Science Edition), 2023, 44(1): 14-21.)
- [14] 王明常,丁文,赵竞争,等. 基于知识图谱与随机森林的落叶松毛虫害遥感识别 [J]. 吉林大学学报(地球科学版), 2023, 53(6): 2006-2017. (WANG M C, DING W, ZHAO J Z, et al. Remote Sensing Identification of Dendrolimus Superans Infestation Based on Knowledge Graph and Random Forest [J]. Journal of Jilin University (Earth Science Edition), 2023, 53(6): 2006-2017.)
- [15] 刘琼昕,牛文涛,王佳升. 融合知识和约束图的远程监督关系抽取方法 [J]. 北京理工大学学报, 2024, 44(7): 731-739. (LIU Q X, NIU W T, WANG J S. Extracting Method of Distant Supervised Relation Based on Fusion of Knowledge and Constraint Graph [J]. Transactions of Beijing Institute of Technology, 2024, 44(7): 731-739.)
- [16] 邓亮,齐攀虎,刘振龙,等. BGP NRE: 一种基于 BERT 的全局指针网络实体关系联合抽取方法 [J]. 计算机科学, 2023, 50(3): 42-48. (DENG L, QI P H, LIU Z L, et al. BGP NRE: A BERT-Based Global Pointer Network for Named Entity-Relation Joint Extraction Method [J]. Computer Science, 2023, 50(3): 42-48.)
- [17] 任安琪,柳林,王海龙,等. 面向文本实体关系抽取研究综述 [J]. 计算机科学与探索, 2024, 18(11): 2848-2871. (REN A Q, LIU L, WANG H L, et al. Review of Text-Oriented Entity Relation Extraction Research [J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(11): 2848-2871.)

(责任编辑:韩 啸)