doi:10.3969/j.issn.1672-1292.2023.01.005

基于 FastBert 的水稻病虫害实体关系抽取研究

周　烨[1],徐向英[1,2],章永龙[1],陈佳云[1],汪洪江[1]

(1.扬州大学信息工程学院,江苏 扬州 225012)
(2.扬州大学教育部农业与农产品安全国际合作联合实验室,江苏 扬州 225127)

[**摘要**] 针对水稻病虫害知识图谱构建所需实体和关系,提出了一种基于 FastBert 模型的中文实体关系抽取方法. 首先,在中文语料收集的基础上,使用 Hanlp 工具和农业词典提取了与水稻病虫害相关的领域实体,并依据实体间关系的特点定义了病虫害别名、为害部位、为害地区、防治方法等 7 种类型. 然后,在词嵌入和句子嵌入的基础上通过 FastBert 模型实现水稻病虫害关系的抽取. 该模型与 Robert、Electra、Distilbert 等其它 Bert 相关模型的关系抽取结果比较显示,基于 FastBert 模型的中文水稻病虫害关系抽取效果更好,模型获得的实体间关系 *F*1 值达 0.72,模型精度达 0.69. 该方法为中文农业病虫害知识图谱的自动化构建提供了参考.

[**关键词**] 水稻病虫害,知识图谱,关系抽取
[**中图分类号**]TP391.1　[**文献标志码**]A　[**文章编号**]1672-1292(2023)01-0033-06

Relationship Extraction of Entities About Rice Diseases and Insect Pests Based on FastBert

Zhou Ye[1],Xu Xiangying[1,2],Zhang Yonglong[1],Chen Jiayun[1],Wang Hongjiang[1]

(1.College of Information Engineering,Yangzhou University,Yangzhou 225012 China)
(2.Joint International Research Laboratory of Agriculture and Agri-Product Safety,the Ministry of Education of China,Yangzhou 225127 China)

Abstract:A FastBert model based Chinese entity relationship extraction method is proposed to extract the entities and relationships required for rice pest and disease knowledge graph. First of all,on the basis of Chinese corpus collected,a tool named Hanlp and a agricultural dictionary are used to extract the domain entities related to rice diseases and insect pests. According to the characteristics of the relationship between entities,seven types of diseases and pests are defined,such as alias,harm parts,suffer region,prevention and treatment,etc. Based on word embedding and sentence embedding,the extraction of the relation of rice diseases and insect pests is realized through the FastBert model. And the results are compared with those of other Bert related models. It shows that the FastBert model is better than other Bert related models in the relationship extraction task of entities in the Chinese corpus of rice diseases and insect pest. The *F*1 value obtained by the FastBert model is 0.72,and the accuracy of the model is 0.69. This method provides a reference for automated construction of Chinese knowledge map of agricultural pests and diseases.

Key words:rice diseases and insect pests,knowledge graph,relationship extraction

水稻作为我国的主要粮食作物,其病虫害防治一直备受关注. 近年来,随着智慧农业[1-2]相关技术的发展,物联网、遥感监测[3]、知识图谱、数字孪生[4]等信息技术在水稻种植过程中发挥了越来越重要的作用. 水稻病虫害实体关系抽取技术是构建水稻病虫害知识图谱的关键技术,是进行水稻病虫害自动化诊断和问答系统的基础,对保障我国的水稻安全生产具有积极意义.

实体关系抽取是通过识别句子中两个实体之间存在的一种或多种关系,将非结构化或半结构化数据转化为结构化数据,为构建知识图谱和知识图谱的下游应用,如知识问答,智能化推荐等自然语言处理任务提供丰富的数据[5].

收稿日期:2022-09-15.
基金项目:教育部农业与农产品安全国际合作联合实验室开放课题项目(JILAR-KF202007)、扬州大学交叉学科基金项目(yzuxk202008)、扬州市市校合作专项项目(YZ2021150).
通讯作者:徐向英,博士,研究方向:农业信息化. E-mail:xuxy@yzu.edu.cn

知识图谱是谷歌公司首次提出并广泛应用于搜索引擎的一项技术. 目前,它已然成为人工智能领域的关键技术. 知识图谱是具有图结构的知识数据库,将知识以图的形式呈现出来,在图结构上展示实体与实体之间的关系. 知识图谱构建[6]过程中存在两大步骤:实体的抽取和关系的抽取[7]. 关系抽取建立在实体抽取基础之上,通过对句子中两个实体之间的关系识别和分类,形成关系三元组,从而构建知识图谱.

在农业领域,目前还没有开放的水稻病虫害知识图谱以支持水稻病虫害的自动诊断,因此构建水稻病虫害知识图谱成为业界的热点研究.

传统方法对于病虫害的关系抽取较多地依赖人工分类,需要大量的人力开销. 本研究使用预训练模型 FastBert[8]自动化地对句子中的实体关系进行分类的方法,极大地减轻了实体关系分类的时间和人力成本开销,能够为构建大规模领域知识图谱提供所需的知识三元组.

1 相关工作

为构建知识图谱,通常需要将关系三元组表示为:($e1$,R,$e2$),其中 $e1$ 是实体 1,$e2$ 是实体 2,R 是两个实体之间的关系. 目前,关系抽取方法可以分为 3 类,第一类是传统的模板定义的方法,这类方法需要人工设置模板规则,通常需要丰富的领域经验. 第二类是基于机器学习的方法,包括基于特征的、基于核函数的方法等,如 Culotta 等[9]通过学习上下文和关系模式构建了一个条件随机场模型来提取实体间关系. Mooney 等[10]提出了一种核函数的方法,使用自然语言中的 3 种类型的子序列模式来识别两个实体之间的关系. 第三类是基于深度学习的方法,包括有监督和无监督的学习方式. 如 Zeng 等[11]提出了一种卷积深度神经网络来提取词汇和句子级别的特征,只需将单词标记作为输入,不需要复杂的预处理,将句子级特征与词汇级特征连接起来作为最终提取的特征向量. Zhou 等[12]提出了基于注意力的双向长短期记忆网络来捕获句子中任意位置的重要语义信息. Lin 等[13]针对句子噪声引入带来的关系抽取性能下降问题,提出了一种基于句子级的注意力关系提取模型,通过在多个实例上建立句子级别的注意力,减少噪声的权重从而使用卷积神经网络来嵌入句子的语义信息. 随着深度学习模型的不断推出和模型性能的不断提升,基于深度学习[14]的关系抽取方式将逐步取代机器学习的方式,成为关系抽取领域主流的方法.

本研究采用了基于预训练模型 FastBert 的实体关系抽取方法,通过对维基百科中文语料进行清洗和预处理,筛选其中的农业语料并结合人工收集的语料构建水稻病虫害文本数据集,在人工划分实体间类型的基础上,通过 FastBert 模型,实现水稻病虫害的关系分类和三元组的构建.

2 水稻病虫害关系数据集

2.1 数据获取

本研究的数据来源主要来自国家农业科学数据共享中心(http://crop.agridata.cn)以及维基百科中文语料库. 国家农业科学数据共享中心的半结构化数据采用 Scrapy 爬虫的方式对水稻病虫害数据文本进行爬取并保存. 对于维基百科中文语料库采用 wikiextractor 抽取工具,先从 wikidump 下载维基百科中文词条压缩文件,并根据词条文章目录抽取维基百科中文文本,由于抽取到的文本均为繁体中文,先使用繁简转换程序将繁体中文转换为简体中文,然后在简体维基百科中文语料中抽取与农业病虫害相关语料. 最终语料数据集中的句子数量达 6 868 句,其中从国家农业科学数据中心获得语料 1 200 句,从维基百科获得语料 5 668 句.

2.2 关系分类

知识图谱用实体表示现实世界中的事实概念等,用实体之间的关系表示事实之间存在的联系. 本研究在构建水稻病虫害关系抽取模型时,首先对病虫害相关实体的类别进行了探索. 例如,水稻病虫害总体可以分为“病害”和“虫害”[15]. “病害”以病原为划分依据又可以细分为“真菌病害”“细菌病害”“病毒病害”“线虫病害”“种传病害”等. “真菌病害”可分为“稻瘟病”“纹枯病”“胡麻斑病”“稻曲病”“恶苗病”等具体的水稻病害类型.

因此依据水稻病虫害实体的特点,将水稻病虫害实体关系类型定义为 7 种类型(如表 1 所示),各关系类别的举例如表 2 所示,将所有关系数据集按照 8∶2 的比例分为训练集和测试集.

表 1　水稻病虫害实体关系类型

Table 1　The entity relationship types of rice pests and diseases

序 号	关系名称	含义
1	属(Is a)	表示属于关系
2	别名(alias)	表示别称
3	受害地域(Suffer region)	表示水稻病虫害与发生灾害的地区之间的关系
4	发病时期(Sick period)	表示水稻病虫害发生病虫害的物候期
5	病原(Pathogen)	表示水稻病虫害与病原体的关系
6	为害部位(Harm parts)	表示水稻病虫害与侵害的水稻部位的关系
7	防治(prevention and treatment)	表示病虫害与治理措施对应关系

表 2　实体关系示例

Table 2　Examples of relation between entities

实体 1	实体 2	关系	实体 1	实体 2	关系
稻梨孢菌	半知菌	属	水稻稻苗疫病	秧苗	发病时期
恶苗病	徒长病	别称	稻苗疫病	串珠镰孢	病原
恶苗病	江苏	受害地区	倍式波尔多液	稻苗疫病	防治
霜霉病	叶片	为害部位			

3　水稻病虫害关系抽取模型构建

3.1　实体识别

使用汉语自然语言处理工具 Hanlp[16],对语料数据集中的句子进行命名实体识别. 为了识别出句子中与水稻病虫害相关的实体,从数据堂(http://www.datatang.com/)获取农业专业词典,并添加入 Hanlp 工具,从而实现对水稻病虫害相关实体的提取.

3.2　句子嵌入

给定一个长度为 n 的句子,$\boldsymbol{S}=[w_0,w_1,\cdots,w_n]$,将其转化为嵌入向量 $\boldsymbol{e}$,

$$\boldsymbol{e}=\text{Embedding}(\boldsymbol{S}). \tag{1}$$

通过分词器为每句的开头添加标记[CLS],每句末尾添加标记[SEP],通过对句子的嵌入表示获取句中每个词的嵌入向量以及整个句子的嵌入表示,如图 1 所示,[CLS]标记的向量表示即为整个句子的特征表示.

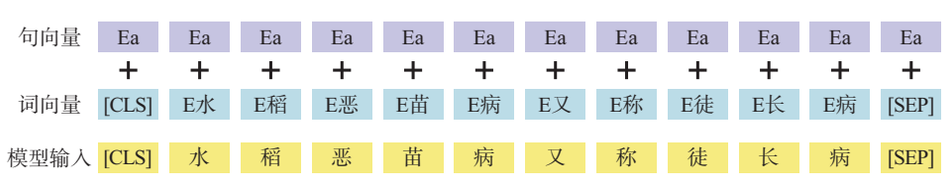

图 1　嵌入格式

Fig. 1　Format of embedding

3.3　关系抽取模型构建

本研究使用 Bert 模型的发展版本 FastBert0 模型进行关系抽取,FastBert 在 Bert 模型的基础上添加了自蒸馏功能,即通过一个教师分类器和多个学生分类器进行自适应的推理[8]. 教师分类器和学生分类器的蒸馏过程如图 2 所示,学生分类器在分类概率较高时能够提前输出预测标签,因此该模型减少了大量训练数据给模型带来的沉重压力,在时间和空间上,运行效率都比 Bert 模型提高很多.

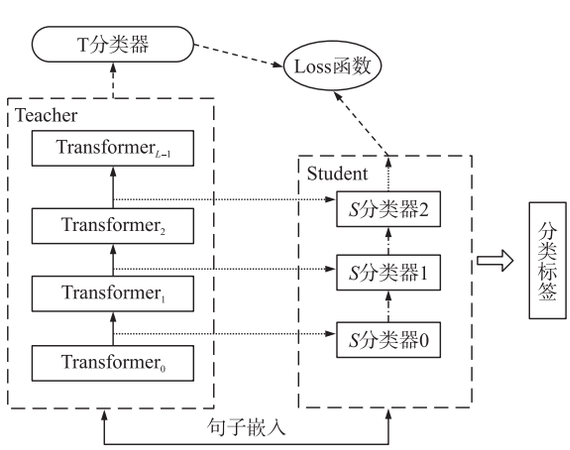

图 2　FastBert 模型

Fig. 2　FastBert model

3.3.1　Transformer 层

FastBert 的主干网络中使用了多个双向 Transformer 模块,很好地融合了句子上下文信息. Transformer 由一个自注意力层网络和一个前馈神经网络层(Feed Forward)组成[17]. 当句子通过嵌入表示输入时,首先由

注意力层通过特定的编码为每个单词添加权重信息,之后会被送入到前馈神经网络层中.

如图 3 所示,我们首先使用嵌入算法将每个输入单词变成一个 768 维度的向量,每个向量都会经过自注意力层和前馈神经网络,假设输入单词为 $\boldsymbol{X}_1$,在经过自注意力层编码后得到聚合特征向量 $\boldsymbol{Z}_1$,送到前馈神经网络编码后最终输出向量 $\boldsymbol{R}_1$. Transformer$_0$ 的输出 $\boldsymbol{R}_1$ 继续被送入到 Transformer$_1$ 中进行编码,最终 Transformer$_L$ 得到的特征向量为 $\boldsymbol{R}_{L-1}$.

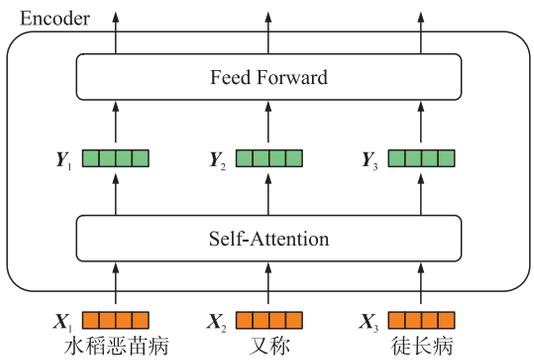

图 3　Encoder 过程

Fig. 3　Encoder process

3.3.2　注意力层

注意力层[18]可以融合当前单词和整个句子中各单词的权重信息,将句子与当前单词关联起来. 自注意力值的计算过程是,先对每个输入的单词 $\boldsymbol{X}$ 分别创建三个向量 $\boldsymbol{Q}$(query)、$\boldsymbol{K}$(key)、$\boldsymbol{V}$(value):

$$\boldsymbol{Q}=\boldsymbol{W}_Q\cdot\boldsymbol{X}, \tag{2}$$

$$\boldsymbol{K}=\boldsymbol{W}_K\cdot\boldsymbol{X}, \tag{3}$$

$$\boldsymbol{V}=\boldsymbol{W}_V\cdot\boldsymbol{X}. \tag{4}$$

式中,$\boldsymbol{W}_Q$,$\boldsymbol{W}_K$,$\boldsymbol{W}_V$ 是三个权重矩阵. 之后计算当前单词 $\boldsymbol{X}$ 相对于输入句子中的每一个词进行评分,这个得分反映了当前词与句子中其他单词的关联程度,句子中所有词的两两关联度得分即为句子的注意力矩阵 $\boldsymbol{A}$,通常用 $\boldsymbol{K}$ 和 $\boldsymbol{Q}$ 点乘的方式计算:

$$\boldsymbol{A}=\boldsymbol{K}^{\mathrm{T}}\cdot\boldsymbol{Q}. \tag{5}$$

对 $\boldsymbol{A}$ 除以当前向量维度的平方根,这样可以使得注意力的梯度更加的稳定,然后将得分送入 softmax 函数进行归一化操作. 将每个单词的 $\boldsymbol{V}$ 值乘以它们的 softmax 分数,然后对所有的输出向量求和得到输出 $\boldsymbol{Z}$. 对于单词 $\boldsymbol{X}_1$ 而言,它的自注意力最终输出为 $\boldsymbol{Z}_1=\mathrm{softmax}*v_1+\mathrm{softmax}*v_2\cdots$.

$$\boldsymbol{Z}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\times\boldsymbol{K}}{\sqrt{dk}}\right)\times\boldsymbol{V}. \tag{6}$$

式中,$\boldsymbol{V}$ 是单词的合集,v_1,$v_2\cdots$代表每一个单词,dk 是 $\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$ 的长度. 多头注意力机制是在自注意力的基础上进,通过添加位置编码和多表示子空间的形式提高了注意力层的性能. Transformer 使用了 8 个注意力头,每个头都含有 $\boldsymbol{W}_Q$,$\boldsymbol{W}_K$,$\boldsymbol{W}_V$ 的权重矩阵用来计算不同的 $\boldsymbol{Q}$ 值、$\boldsymbol{K}$ 值和 $\boldsymbol{V}$ 值. 因此每个编码器和解码器都会有 8 个集合,每组集合都会在训练之后将输入嵌入投影到下一层编码器的表示子空间中. 由于使用了 8 个注意力头,最终的注意力输出将会产生 8 个不同的 $\boldsymbol{Z}$ 向量,将这些 $\boldsymbol{Z}$ 向量做横向拼接操作再乘以一个额外的权重矩阵 $\boldsymbol{Wo}$,得到最后的结果.

4　关系分类结果与分析

4.1　分类预测结果

利用查准率 P 和 $F1$ 值对 7 种实体间关系的预测结果进行评价,如图 4 所示. 其中 P 值最高的关系是:发病时期,达到了 0.75,而 $F1$ 值最高的关系是受害地区,达到了 0.79. 不同关系的预测准确率存在明显差异,反映了不同的关系在语料和算法识别上的不平衡.

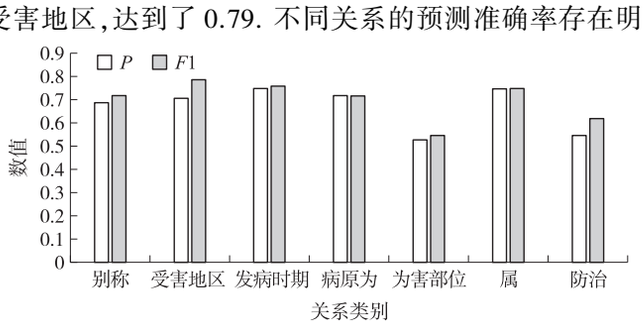

图 4　FastBert 模型对不同关系的处理结果

Fig. 4　FastBert model processing results for different relationships

4.2　不同模型预测结果对比分析

将 FastBert 模型与其它 Bert 相关模型进行对比,实体关系预测的结果显示,FastBert 与其它 5 种模型相比查准率 P 值最高,达到了 0.69,$F1$ 值达到 0.72(如表 3 所示),反映了 FastBert 模型对于中文水稻病虫害语料的处理能力相比于其它模型更优,提取实体关系的准确率较高.

由于 FastBert 模型使用了自蒸馏技术,其模型运行的时间效率也是本研究关注的重点. 比较上述 6 个模型在相同硬件环境下运行的时间,如图 5 所示. 操作系统为 Window10,CPU 为 Intel-i5 12,GPU 为 NVIDIA RTX 3060,设置学习率为 5e-5. 测试结果显示 FastBert 模型所需时间最少,为 18 min. 所需时间最多的模型是 RoBert 模型,为 34 min,表明 FastBert 模型更适合于对时间要求较高的应用.

表 3　不同模型关系提取数

Table 3　Values extracted from different model relationships

模型	*P*	*F*1
FastBert	**0.69**	**0.72**
Bert	0.54	0.62
Camembert	0.31	0.34
Robert	0.57	0.60
Electra	0.35	0.37
Distilbert	0.41	0.52

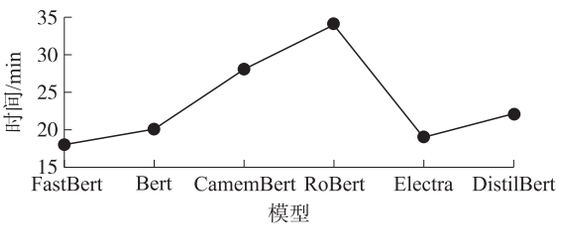

图 5　不同模型运行结果的时间对比

Fig. 5　Time comparison of running results of different models

4.3　模型参数的影响分析

对 FastBert 模型微调可获得不同的模型性能. 为了达到模型的最佳性能,本研究对 FastBert 模型的超参数进行了微调,并根据测试结果确定了最终的参数设置(如表 4 所示). 对 FastBert 模型的 learning_rate、epoch、batch_size 参数的微调及对应模型的 *F*1 值,如表 5 所示. 微调的结果表明,learning_rate 设置为 5e-5、epoc 设置为 500、batch_size 设置为 12 时模型性能最优,*F*1 值达 0.72.

表 4　FastBert 模型参数设置

Table 4　Parameter settings of FastBert model

参数	值	参数	值
attention_probs_dropout_prob	0.1	num_attention_heads	12
directionality	bidi	num_hidden_layers	12
gradient_checkpointing	false	pad_token_id	0
gradient_checkpointing	gelu	pooler_fc_size	768
hidden_dropout_prob	0.1	pooler_num_attention_heads	12
hidden_size	768	pooler_num_fc_layers	3
initializer_range	0.02	pooler_size_per_head	128
intermediate_size	3 072	pooler_type	transform
layer_norm_eps	1e-12	type_vocab_size	2
max_position_embeddings	512	vocab_size	21 128
model_type	bert		

表 5　FastBert 模型参数结果对比

Table 5　Comparison of model parameter results of FastBert model

学习率	Epoch	Batch_size	*F*1	学习率	Epoch	Batch_size	*F*1
2e-5	500	12	0.59	5e-5	1 000	12	0.67
3e-5	500	12	0.61	5e-5	500	4	0.65
4e-5	500	12	0.57	5e-5	500	8	0.58
5e-5	500	12	0.72	5e-5	500	24	0.57
5e-5	100	12	0.32	5e-5	500	48	0.43

5　结论

针对专业语料进行实体关系抽取仍然是目前自然语言处理研究中较为困难的内容. 本文研究了基于 FastBert 模型的水稻病虫害实体关系抽取方法,并与多种 Bert 相关模型进行了对比分析. 结果表明,FastBert 模型的关系抽取效果在同类模型中表现最佳,可能是由于本研究中使用的 FastBert 模型采用了中文语料进行预训练,对中文实体关系的提取更加有利. 另外,从时间上看,FastBert 模型比其它模型更快获得结果.

[参考文献](References)

[1] STINNER D H,PAOLETTI M G,STINNER B R. In search of traditional farm wisdom for a more sustainable agriculture:a study of Amish farming and society[J]. Agriculture,Ecology and Environment,1989,27:77-90.

[2] 闫靖昆,黄毓贤,秦伟森,等. 棉田复杂背景下棉花黄萎病病斑分割算法研究[J]. 南京师大学报(自然科学版),2021,44(4):127-134.

[3] ROGAN J,CHEN D M. Remote sensing technology for mapping and monitoring land-cover and land-use change[J]. Progress in Planning,2004,61(4):301-325.

[4] JONES D,SNIDER C,NASSEHI A,et al. Characterising the digital twin:a systematic literature review[J]. CIRP Journal of Manufacturing Science and Technology,2020,29:36-52.

[5] PUJARA J, MIAO H, GETOOR L, et al. Knowledge graph identification[C]//International Semantic Web Conference. Berlin,Heidelberg:Springer,2013:542-557.

[6] MARTINEZ R J L,LÓPEZ A I,RIOS A A B. Openie-based approach for knowledge graph construction from text[J]. Expert Systems with Applications,2018,113:339-355.

[7] BHATIA P,CELIKKAYA B,KHALILIA M,et al. Comprehend medical:a named entity recognition and relationship extraction web service[C]//2019 18th IEEE International Conference on Machine Learning and Applications. Boca Raton,Florida,USA,2019:1844-1851.

[8] LIU W J,ZHOU P,ZHAO Z,et al. Fastbert:a self-distilling bert with adaptive inference time[J]. arXiv Preprint arXiv:2004.02178,2020.

[9] CULOTTA A,MCCALLUM A,BETZ J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text[C]//Proceedings of the Human Language Technology Conference of the NAACL. New York,NY,USA,2006:296-303.

[10] MOONEY R, BUNESCU R. Subsequence kernels for relation extraction[J]. Advances in Neural Information Processing Systems,2005,18.

[11] ZENG D J,LIU K,LAI S,et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014,the 25th International Conference on Computational Linguistics:Technical Papers. Dublin,Ireland,2014:2335-2344.

[12] ZHOU P,SHI W,TIAN J,et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,Berlin,Germany,2016:207-212.

[13] LIN Y K,SHEN S,LIU Z,et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers). Berlin,Germany,2016:2124-2133.

[14] 施寒瑜,曲维光,魏庭新,等. 基于组合深度模型的现代汉语数量名短语识别[J]. 南京师大学报(自然科学版),2022,45(1):127-135.

[15] 夏迎春. 基于知识图谱的农业知识服务系统研究.[D]. 合肥:安徽农业大学,2018.

[16] YANG Y X,REN G C. HanLP-based technology function matrix construction on Chinese process patents[J]. International Journal of Mobile Computing and Multimedia Communications,2020,11(3):48-64.

[17] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document transformer[J]. arXiv Preprint arXiv: 2004.05150,2020.

[18] ZHANG H,GOODFELLOW I,METAXAS D,et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning. Long Beach,California,USA,2019:7354-7363.

[责任编辑:陈　庆]