面向标注数据稀缺专利文献的科技实体抽取

原之安[1,2,3]，彭甫镕[1,2,3]，谷　波[1,2,3]，钱宇华[1,2,3]

(1. 山西大学 大数据科学与产业研究院　山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室　山西 太原 030006；
3. 山西大学 计算机与信息技术学院　山西 太原 030006)

摘要：专利中的科技实体是指专利文献中富含科技信息的词汇,抽取专利中的科技实体对科研工作者提高科研效率、企业布局专利体系都至关重要。提出一种基于半监督学习框架与命名实体识别模型相结合的科技实体抽取方法,半监督学习能够利用无标记数据的优势弥补标注数据稀缺的缺陷,利用大量的专利语料在通用领域的 BERT 模型基础上进行预训练,得到适用于专利领域的 BERT 模型 BERT-Patent,有效提升模型对专利中科技实体的抽取性能。在专利数据集上的实验结果表明,提出的方法在准确率、召回率、$F1$ 值指标上分别提高了 6.37%、2.99%、4.63%;在人民日报数据集上准确率、召回率、$F1$ 值分别提高了 2.87%、1.24%、2.07%。

关键词：科技实体；专利挖掘；数据稀缺；BERT；半监督学习

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2021)04-0061-08

DOI：10.13705/j.issn.1671-6841.2021196

0　引言

专利是指专有的权利和利益,囊括了全球 95%以上的最新技术情报[1]。通常科技类文献可以通过关键字快速掌握文献信息。但在专利文献中,没有关键词这一重要属性,读者想要快速获取专利的科技信息,需要了解专利分类号的含义。我国采用国际专利分类号对专利进行标注,即便是专利工作者,也很难了解每一条分类号的分类信息,导致不能像读论文一样快速了解文献中的信息。针对这一问题,本文将抽取专利文献中富含科技信息的词,即科技实体,这将有利于读者通过科技实体快速了解专利文献中蕴含的科技信息、提高专利检索的精度、构建专利知识图谱、提升专利自动分类的效果、梳理专利发展趋势等。

科技实体具体指在专利文献中能够反映某种科技成就、经验与方法、科学理论等的实体。反映科技成就的实体有物联网、存储器等;反映经验与方法的实体有人工智能、汉字识别等;反映科学理论的实体有微分方程、波函数等。科技实体的定义包括但不限于以上描述。

在类似的科技实体抽取研究中,Jang 等[2]通过句法分析处理专利数据,对描述一项技术的单词 TechWord 进行依赖关系分析并识别其词性,从名词短语和 SAO(subject-action-object)结构中提取 TechWord 候选词,通过中心性指标将每个词之间的关系解释为一个网络,评估 TechWord 候选词的重要性,进而抽取出专利中的 TechWord。该方法速度较快,且不需要对专利中的科技词进行标注,但没有涵盖所有词性,而是通过句法关系提取科技词,且对专利进行句法标注也是一大难题。

自然语言处理(natural language processing, NLP)中的命名实体识别(named entity recognition, NER)[3]技术是指识别文本中具有特定意义的实体。命名实体识别问题通常被抽象为序列标注任务。该方法适用于本文的科技实体抽取。因此本文将使用命名实体识别技术对专利文本进行深度挖掘,抽取专利中的科技实体。

早期的命名实体识别研究大都是基于领域专家建立手工规则对实体进行抽取[4-6],该方法成本较高,且速度较慢。随着深度学习在自然语言处理上的应用[7-8],基于神经网络的命名实体识别方法[9-12]针对训练

收稿日期:2021-05-18

基金项目:国家自然科学基金项目(61672332);山西省重点研发计划项目(201903D421003);山西省教育厅科技成果转化培育项目(2020CG001)。

作者简介:原之安(1998—),男,硕士研究生,主要从事数据挖掘、机器学习研究,E-mail:jczhianyuan@163.com;通信作者:钱宇华(1976—),男,教授,主要从事大数据、数据挖掘与机器学习研究,E-mail:jinchengqyh@126.com。

数据开展基于词语级或字符级的实体语义特征学习,进而实现对测试数据的实体抽取。李建等[13]使用BERT-BiLSTM-CRF结构对专利中的实体进行抽取,由于使用了BiLSTM结构,训练速度慢,并且这些神经网络模型的最大局限性是依赖大量的标注数据。

专利数据规模较大,需要大量具有领域知识的标注人员进行标注,这对数据标注造成了一定的阻碍。本文考虑在有限的标注数据上,建立一套可靠的命名实体识别模型。

在有限的标注数据上,一些研究人员使用迁移学习的方法进行突破[14-16],该方法需要有标记的平行语料库,对专利数据来说,该语料库难以获得,无法用迁移学习解决专利中科技实体抽取任务。远程监督学习也是解决低资源问题的常用方法[17-19]。该方法需要特定任务领域的词典进行监督,且性能会受到词典中数据量的影响,若词典仅包含待标注语料所属的领域,将大幅提升该方法针对特定领域命名实体识别的精确率。以上方法都需要带标注的语料库或外部词典。Liang等[20]使用半监督学习的方法从科学文献中提取生物医学实体,选择BiLSTM-CRF作为命名实体识别模型,然后执行Bootstrapping过程,该方法使用了循环神经网络结构,无法并行计算,使得训练速度下降。综上,现有方法中存在一些不足:1)专利中标记数据和外部语料稀缺,且构建字典代价昂贵,导致传统监督学习过拟合问题。2)现有的BERT模型在预训练时使用通用领域的语料库,语料库中没有专利方面的数据,对专利分析问题有一定的影响。3)BiLSTM网络结构无法并行计算,训练速度较慢。

1 基于半监督学习和预训练语言模型的科技实体抽取方法

近年来,在自然语言处理领域,使用预训练语言模型的方法在多项NLP任务上都取得了突破性的进展。在众多预训练语言模型中比较具有代表性的模型有ELMo(embeddings from language models)[21]、OpenAi GPT(generative pre-training)[22]和BERT(bidirectional encoder representations from transformers)[23],它们都是神经网络模型。因神经网络语言模型在当前研究中表现最佳,且使用神经网络的方法可以结合上下文信息学到一些语义信息有利于翻译系统性能的提升,所以本文也采用了该方法预训练语言模型。

本文将采用半监督学习与命名实体识别模型相结合的方法解决标注数据稀缺的问题,命名实体识别模型选择BERT模型学习融合上下文关系的字的表征信息,在BERT输出部分添加条件随机场(conditional random field, CRF)[24],学习标签之间的依赖关系,进而获得全局最优的实体标记序列,并根据输出序列的规范化条件概率用于半监督学习算法中判断输出标记序列的可靠性。由于BERT模型的性能会受到训练时选择的语料库的直接影响,本文还将使用大量的专利数据在原有的BERT模型基础上进行预训练,使其适应专利领域的应用。模型框架如图1所示。其中:$\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_N$表示词向量,即模型的输入;*Trm*表示Transformer编码器结构;$\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_N$表示模型的输出。

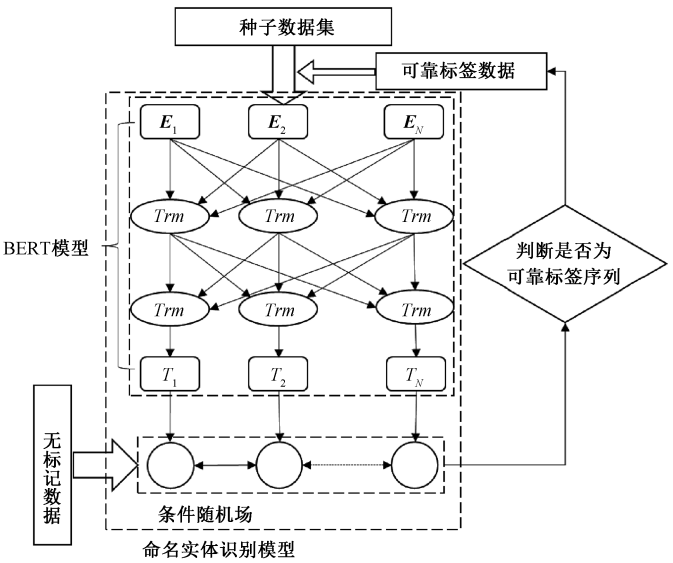

图1 基于半监督学习和预训练语言模型的科技实体抽取方法

Figure 1 Technology entity extraction method based on semi supervised learning and pre training language model

1.1 预训练语言模型

BERT是一个预训练语言模型,它在大规模语料库上采用多层双向Transformer[25]编码结构,基于自注意力机制对文本建模。由于采用了双向编码器结构,训练出的结果反映了词所在句子的上下文语义关系,能够学到词的多义性,可以获得更好的词的分布表示。大规模语料的预训练语言模型能够为某些数据规模较小的自然语言处理任务提供模型参数,进而提升模型的性能。

Transformer是一个基于多头自注意力机制的深度网络,自注意力机制主要是通过同一个句子中词与词之间的关联程度调整权重系数矩阵来获取词的表征,所用公式为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}, \tag{1}$$

其中:$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 是字向量矩阵,d_k 是词向量的维度。

在 Transformer 结构中无法和循环神经网络(recurrent neural network,RNN)一样获取句子中字的序列信息,为了解决这个问题,Transformer 在数据预处理之前加入了位置编码,并与输入向量的数据进行相加,得到句子中每个字的相对位置。

我们在 Google 发布的 BERT-Base Chinese 基础上进行专利数据上的预训练,得到一个专利领域的 BERT 预训练语言模型,该模型对我们的任务能够有所帮助。

1.2　条件随机场

在命名实体识别任务中,通常采用 BIO 标注方法,将每个元素标注为“B-X”、“I-X”或者“O”。其中,“B-X”表示此元素所在的片段属于 X 类型,并且此元素在此片段的开头;“I-X”表示此元素所在的片段属于 X 类型,并且此元素在此片段的中间位置,“O”表示不属于任何类型。BERT 模型能够得到融合输入上下文关系的字的表征信息,但是无法处理 BIO 标签模式之间的依赖关系,如:“O”标签的下一个标签可以是“B”标签,但是不可以是“I”标签;“I”标签一定在“B”标签之后才出现。所以条件随机场能够通过相邻的标签获得一个最优的预测序列,从而弥补 BERT 模型的不足。条件随机场模型是基于概率图模型的分类学习方法,是条件概率分布模型 $P(Y \mid X)$。对于任一个文本序列 $X = \{x_1, x_2, \cdots, x_n\}$,根据 BERT 模型的输出预测序列 $Y = \{y_1, y_2, \cdots, y_n\}$,通过条件概率 $P(y \mid x)$ 进行建模:

$$P(y \mid x) = \frac{1}{Z(x)}\exp\Big(\sum_{i,m}\lambda_m\mu_m(y_{i-1}, x, i) \cdot \sum_{i,n}\beta_n t_n(y_n, x, i)\Big), \tag{2}$$

其中:i 表示当前节点在 x 中的索引;m、n 分别表示当前节点上的特征函数总个数;μ_m 表示定义在边上的特征函数,称为转移特征,描述从前一个标签到当前标签的条件概率,只与当前位置和前一个节点位置有关;t_n 表示定义在节点上的特征函数,称为状态特征,描述一个观测序列标记为某标签的条件概率,只与当前位置有关;β_n、λ_m 分别表示特征函数 t_n 和 μ_m 对应的权重系数,用于衡量特征函数的信任度;$Z(x)$ 为归一化因子,在所有可能的输出序列上进行求和,以便获得全局归一化,所用公式为

$$Z(x) = \sum_{y}\exp\Big(\sum_{i,m}\lambda_m\mu_m(y_{i-1}, x, i) \cdot \sum_{i,n}\beta_n t_n(y_n, x, i)\Big)。 \tag{3}$$

在定义特征函数阶段,可以将节点处和边上的特征函数的初始值都设置为 1,由于每个特征函数都分配有权重参数,因此在训练的时候,如果节点之间不存在依赖关系,则该特征函数的权重参数会在训练结束后趋近于 0,以此学习标签之间的相互依赖关系。

1.3　半监督学习

首先人为地标注一部分数据作为种子数据集 *Data-Seed*,使用种子数据集训练一个初始的命名实体识别模型 M_0,使用该模型在未做标记的数据集 *Data-Unlabeled* 上预测样本的标签序列,本研究设置一个阈值选取出较高概率的标签序列,用于评估标签序列的可靠性。将可靠标签序列的样本集 $Data^*$ 与 *Data-Seed* 融合得到一个增广数据集,用于迭代命名实体识别模型,直到达到最大迭代次数或所有标签序列均达到可靠标签序列的水准。其具体步骤如算法 1 所述。

算法 1　半监督学习的命名实体识别算法。

输入:在专利数据集上预训练的 BERT 模型 *BERT-Patent*,带有标记的种子数据集 *Data-Seed*,不带有标记的数据集 *Data-Unlabeled*。

输出:命名实体识别模型 M_s。

1)　$M_0 = BERT\text{-}Patent(Data\text{-}seed)$

2)　for $s \leftarrow 1$ to K

3)　if $Data\text{-}unlabeled = \varnothing$

4)　　return M_s

5)　end if

6)　$sentence = M_{s-1}(Data\text{-}unlabeled)$

7) $p = M_{s-1}(Data\text{-}unlabeled)$

8) if $p > \theta$

9) $Data^* \leftarrow sentence$

10) end if

11) $Data\text{-}seed += D^*$

12) $M_s = M_{s-1}(Data\text{-}seed)$

13) end for

14) return M_s

2 实验设计

2.1 数据描述和实验设置

本文收集了6 080 130条专利的摘要数据用来对BERT模型进行预训练,从而得到适应于专利领域的BERT模型,数据来源于山西大学大数据科学与产业研究院自主研发的"山西大学知识产权大数据分析平台"。

以"计算机"为检索词收集了1 691篇发明专利的摘要数据,共计4 025个样本(一句话为一个样本),三个人使用brat标注工具标注了18 100个实体,并进行人工校验。为了验证种子数据集的规模对模型学习效率的影响,将实验数据按比例划分有标记的数据和无标记的数据,划分结果如表1所示。

表1 数据集划分及规模

Table 1 Dataset splitting and size 单位:个

划分比例	有标记句子数	有标记实体数	未标记句子数	未标记实体数
0.1	403	1 966	3 622	16 134
0.2	805	3 578	3 220	14 522
0.3	1 207	5 278	2 818	12 822
0.5	2 013	9 151	2 012	8 949

本文按照BIO标注格式对专利中的科技实体进行标注:B-TEC,即Begin Technology,表示当前科技实体的开始;I-TEC,即Intermediate Technology,表示当前科技实体的延续或结束;O,即Other,表示其他字符,用于标记无关字符。标注示例见表2。

表2 数据标注示例

Table 2 Example of data annotation

字	一	种	自	动	驾	驶	的	传	感	器	模	块
标注	O	O	B-TEC	I-TEC	I-TEC	I-TEC	O	B-TEC	I-TEC	I-TEC	O	O

本文选择通用领域的BERT模型作为基线模型。其他模型如下:

1) 半监督学习结合通用领域BERT模型(BERT with SS);

2) 专利领域的BERT模型(BERT on Patent);

3) 半监督学习结合专利领域的BERT模型(BERT on Patent with SS)。

为了验证本文方法在通用领域数据集上的有效性,选择人民日报命名实体识别数据集进行对比实验。

2.2 评价指标

本文采用命名实体识别任务中常用的实体准确率pre、实体召回率rec,$F1$值以及标签正确率acc作为科技实体抽取模型性能的评价指标,计算公式分别为

$$pre = \text{识别出的正确实体数} / \text{识别出的实体数},$$

$$rec = \text{识别出的正确实体数} / \text{样本的实体数},$$

$$F1 = (2 \times pre \times rec)/(pre + rec),$$

$$acc = \text{识别出的正确标签数} / \text{样本的标签数}。$$

2.3 超参数设置

实验采用Google发布的BERT-Base,共有12层Transformer结构,字向量维度为768,12个自注意力头,

参数量共 110 M。最大序列长度为 202,batch_size 为 64,学习率为 1e-5,dropout 为 0.5。LSTM 隐藏层单元数为 128。

2.4 BiLSTM 层对命名实体识别模型的影响

表 3 显示了 BERT-CRF 结构和 BERT-BiLSTM-CRF 结构在不同划分比例专利数据集上的性能对比。通过实验发现,加入了 BiLSTM 层的网络结构对命名实体识别的效果提升并不显著,但比不加 BiLSTM 层的网络结构更耗时。这反映了 BiLSTM 层不能并行加速的不足,同时,BERT 模型通过 Self-Attention 机制能够学习到上下文的词汇表示。相比之下,本文选择了 BERT-CRF 结构作为命名实体识别模型的网络结构。

表 3　BiLSTM 层对命名实体识别性能的影响

Table 3　Influence of BiLSTM layer on named entity recognition performance

划分比例	训练时间/s		实体 *F*1 值	
	BERT-CRF	BERT-BiLSTM-CRF	BERT-CRF	BERT-BiLSTM-CRF
0.1	282.766	352.075	0.798 9	0.780 4
0.2	530.257	648.131	0.837 4	0.840 2
0.3	619.859	772.758	0.888 7	0.883 2
0.5	1 078.408	1 301.786	0.919 9	0.922 5

2.5 阈值的选择

图 2 显示了具有不同标签概率阈值的开发集上 *F*1 值的曲线。本实验是为不同比例的训练数据寻找最优阈值。从理论上讲,提高阈值可以降低误报率,从而提高精确度,相反,若阈值设置较低会得到较高的召回率。通过实验发现慢慢提高标签概率会使 *F*1 值升高,但是当标签概率高于阈值时,*F*1 值开始下降。这说明随着标签概率的升高,实体精确度上升的同时,召回率在下降,通过实验可以确定不同数据集划分下标签概率阈值的选择。

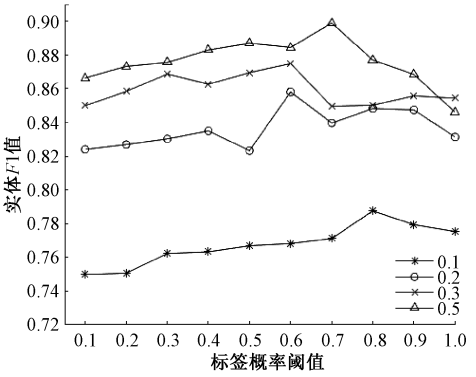

图 2　不同标签概率阈值下的实体 *F*1 值

Figure 2　Entity F1 values under different label probability thresholds

2.6 实验结果及分析

专利数据集在不同模型上的实验结果如表 4 所示,人民日报数据集在不同模型上的实验结果如表 5 所示。

表 4　不同模型在专利数据集上的实验效果

Table 4　Experimental Performance of different models on patent dataset

评价指标	划分比例	BERT on patent with SS	BERT with SS	BERT on Patent	BERT
实体准确率	0.1	0.857 9	0.839 2	0.794 9	0.754 2
	0.2	0.872 3	0.856 4	0.822 6	0.799 3
	0.3	0.910 6	0.897 7	0.886 4	0.863 0
	0.5	0.933 3	0.933 1	0.916 6	0.902 7
实体召回率	0.1	0.856 3	0.854 9	0.822 2	0.813 5
	0.2	0.880 0	0.855 3	0.859 3	0.846 4
	0.3	0.900 0	0.892 6	0.889 5	0.881 8
	0.5	0.915 4	0.905 8	0.913 9	0.890 5
实体 *F*1 值	0.1	0.852 1	0.847 0	0.808 3	0.782 7
	0.2	0.877 4	0.855 8	0.840 6	0.822 2
	0.3	0.905 3	0.895 2	0.887 9	0.872 3
	0.5	0.924 3	0.919 3	0.915 3	0.896 5
标签正确率	0.1	0.967 1	0.964 9	0.957 2	0.952 2
	0.2	0.971 9	0.967 8	0.965 7	0.963 1
	0.3	0.978 8	0.977 6	0.975 7	0.972 8
	0.5	0.984 8	0.982 3	0.983 6	0.978 9

表5 不同模型在人民日报数据集上的实验效果

Table 5 Experimental performance of different models on People's Daily dataset

评价指标	划分比例	BERT on Patent with SS	BERT with SS	BERT on Patent	BERT
实体准确率	0.1	0.910 2	0.921 2	0.865 2	0.898 3
	0.2	0.915 3	0.923 7	0.899 9	0.909 5
	0.3	0.925 9	0.929 2	0.913 6	0.917 5
	0.5	0.936 5	0.943 3	0.923 9	0.931 2
实体召回率	0.1	0.916 4	0.922 7	0.895 8	0.910 7
	0.2	0.926 8	0.932 7	0.919 4	0.922 9
	0.3	0.935 5	0.935 2	0.927 3	0.925 7
	0.5	0.951 5	0.946 6	0.944 9	0.943 4
实体 $F1$ 值	0.1	0.913 3	0.922 0	0.880 2	0.904 5
	0.2	0.921 0	0.928 2	0.909 5	0.916 1
	0.3	0.930 7	0.932 2	0.920 4	0.921 6
	0.5	0.944 0	0.944 9	0.934 3	0.937 2
标签正确率	0.1	0.989 7	0.990 8	0.988 0	0.989 3
	0.2	0.990 2	0.990 9	0.989 7	0.990 3
	0.3	0.991 3	0.991 4	0.991 0	0.990 9
	0.5	0.993 0	0.993 2	0.992 6	0.992 7

通过实验发现,本文的方法和传统的 BERT-CRF 方法相比,准确率、召回率、$F1$ 值指标分别平均提高了 6.37%、2.99%、4.63%;在人民日报数据集上分别平均提高了 2.87%、1.24%、2.07%。

半监督学习方法对科技实体抽取性能的影响由表 4 可以得出,采用了半监督学习框架后,通用领域的 BERT 和专利领域的 BERT 模型性能在准确率上分别平均提升了 5.18%和 3.84%。出现这一现象的原因是半监督学习在训练过程中加入许多无标签数据和伪标签信息,使模型的准确率性能上升。

预训练语言模型对科技实体抽取性能的影响由表 4 可以得出,专利领域的 BERT 模型相比通用领域的 BERT 能够在更少的数据集上找到更多的科技实体。使用了专利领域的 BERT 模型之后,使用半监督学习框架和未使用半监督学习框架的实体召回率分别平均提升 1.32%和 1.08%。与此同时,使用了专利领域的 BERT 模型之后,随着标记数据集的扩大,模型性能上升较稳定。这说明在使用了大量的专利数据之后,BERT 模型能够更高效地抽取专利文本中内容。

本文方法在不同数据集上的性能对比由表 5 可以得出,在半监督学习框架下,对人民日报数据集上的命名实体识别模型依旧表现出了比较好的性能。同时发现,专利领域的 BERT 模型在人民日报数据集上的性能不如通用领域的 BERT 模型,这进一步说明了 BERT 模型的性能与预训练时语料库的选择有密切关系。

由表 4 和表 5 标签正确率发现,标签正确率要远远高于实体的评价指标,出现这个现象的原因是本文的方法不能准确地找到实体边界,导致实体的评价指标降低。

3 结束语

本文针对专利文本中的科技实体抽取任务,研究了基于半监督学习和预训练语言模型的专利科技实体抽取模型。通过在大量专利上对 BERT 模型进行预训练,使其能够更好地抽取专利中的特征信息,并结合 CRF 构建命名实体识别模型,使用半监督学习的思想解决专利科技实体抽取问题。通过实验证明了本文的方法能够在标注数据不足的情况下抽取专利中的科技实体。本文工作在专利检索、专利价值评估等专利挖掘工作上将发挥重要的作用。

本文接下来的工作,将考虑进一步提高科技实体抽取的准确率,准确识别实体边界,并考虑利用科技实体信息对专利文本进行更深层次的挖掘。

参考文献:

[1] 冯岭, 彭智勇, 刘斌, 等. 一种基于潜在引用网络的专利价值评估方法[J]. 计算机研究与发展, 2015, 52(3): 649-

660.

FENG L, PENG Z Y, LIU B, et al. A latent-citation-network based patent value evaluation method[J]. Journal of computer research and development, 2015, 52(3): 649-660.

[2] JANG H, JEONG Y, YOON B. TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing[J]. Expert systems with applications, 2021, 164: 114042.

[3] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae investigationes, 2007, 30(1): 3-26.

[4] 昝红英, 关同峰, 张坤丽, 等. 面向医学文本的实体关系抽取研究综述[J]. 郑州大学学报(理学版), 2020, 52(4): 1-15.

ZAN H Y, GUAN T F, ZHANG K L, et al. Entity and relation extraction for medical text: a survey [J]. Journal of Zhengzhou university (natural science edition), 2020, 52(4): 1-15.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12:2493-2537.

[6] 闫丹辉, 毕玉德. 基于规则的越南语命名实体识别研究[J]. 中文信息学报, 2014, 28(5): 198-205, 214.

YAN D H, BI Y D. Rule-based recognition of Vietnamese named entities[J]. Journal of Chinese information processing, 2014, 28(5): 198-205, 214.

[7] 陈观林, 侍晓龙, 周梁, 等. 基于深度强化学习的文本相似语义计算模型[J]. 郑州大学学报(理学版), 2020, 52(3): 1-8.

CHEN G L, SHI X L, ZHOU L, et al. A text similarity semantic computing model based on deep reinforcement learning[J]. Journal of Zhengzhou university (natural science edition), 2020, 52(3): 1-8.

[8] 邵恒, 冯兴乐, 包芬. 基于深度学习的文本相似度计算[J]. 郑州大学学报(理学版), 2020, 52(1): 66-71, 78.

SHAO H, FENG X L, BAO F. Text similarity computation based on deep-learning[J]. Journal of Zhengzhou university (natural science edition), 2020, 52(1): 66-71, 78.

[9] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 2670-2680.

[10] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE transactions on knowledge and data engineering, 2020,99: 1.

[11] LV H, NING Y S, NING K. ALBERT-based Chinese named entity recognition[M]. Cham: Springer, 2020: 79-87.

[12] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 1554-1564.

[13] 李建, 靖富营, 刘军. 基于改进 BERT 算法的专利实体抽取研究: 以石墨烯为例[J]. 电子科技大学学报, 2020, 49(6): 883-890.

LI J, JING F Y, LIU J. Study on patent entity extraction based on improved bert algorithms—a case study of graphene[J]. Journal of university of electronic science and technology of China, 2020, 49(6): 883-890.

[14] CHAUDHARY A, XIE J T, SHEIKH Z, et al. A little annotation does a lot of good: a study in bootstrapping low-resource named entity recognizers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong:ACL Press, 2019: 5164-5174.

[15] PENG D L, WANG Y R, LIU C, et al. TL-NER: a transfer learning model for Chinese named entity recognition[J]. Information systems frontiers, 2020, 22(6): 1291-1304.

[16] GLIGIC L, KORMILITZIN A, GOLDBERG P, et al. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks[J]. Neural networks, 2020, 121: 132-139.

[17] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. New York:Curran Associates Inc, 2013: 3111-3119.

[18] SHANG J B, LIU L Y, GU X T, et al. Learning named entity tagger using domain-specific dictionary[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2054-2064.

[19] LIU A L, DU J F, STOYANOV V. Knowledge-augmented language model and its application to unsupervised named-entity recognition[C]//Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics Press,

2019: 1142-1150.

[20] LIANG C, YU Y, JIANG H M, et al. BOND: BERT-assisted open-domain named entity recognition with distant supervision [C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1054-1064.

[21] MATHEW J, FAKHRAEI S, AMBITE J L. Biomedical named entity recognition via reference-set augmented bootstrapping[EB/OL]. [2020-05-17]. https://arxiv.org/abs/1906.00282.

[22] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics Press, 2018: 2227-2237.

[23] RADFORD A, KARTHIK N. Improving language understanding by generative pre-training[EB-OL]. [2021-03-16]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[24] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// North American Association for Computational Linguistics (NAACL). Minneapolis: ACL Press, 2019: 4171-4186.

[25] JOHN L, ANDREW M, PEREIRA FERNANDO C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[J]. Proceedings of the eighteenth international conference on machine learning. ACM Press, 2001: 282-289.

Technology Entity Extraction of Patent Literature with Limited Annotated Data

YUAN Zhi'an[1,2,3], PENG Furong[1,2,3], GU Bo[1,2,3], QIAN Yuhua[1,2,3]

(1. *Research Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006, *China*; 2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006, *China*; 3. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006, *China*)

Abstract: Technological information contained in patent documents was in the form of vocabulary. These vocabulary was called patent technology entity. Extracting the entity accurately from the patent was crucial for scientists to improve the efficiency of scientific research, and for enterprises to deploy the patent system. A method of extracting scientific and technological entity was proposed based on semi-supervised learning framework and named entity recognition model. It took advantage of semi-supervised learning to make up for the insufficiency of annotated data. At the same time, BERT-Patent model was pre-trained from the generic BERT model over a large patent corpus, in order to improve the feature extraction performance effectively in patent context. The proposed method had superior performance in terms of accuracy, recall rate, and $F1$ measure; specifically, it was scored 6.37%, 2.99%, and 4.63% higher respectively on the patent dataset, and 2.87%, 1.24%, and 2.07% higher respectively on People's Daily dataset.

Key words: technology entity; patent mining; data scarcity; BERT; semi supervised learning

(责任编辑:方惠敏 孔 薇)