软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1439−1450 [doi: 10.13328/j.cnki.jos.006483]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向知识产权的科技资源画像构建方法*

杨佳鑫[1], 杜军平[1], 邵蓥侠[1], 李 昂[1], 奚军庆[2]

[1](智能通信软件与多媒体北京市重点实验室(北京邮电大学), 北京 100876)

[2](中华人民共和国司法部信息中心, 北京 100020)

通信作者: 杜军平, E-mail: junpingdu@126.com

摘 要: 大数据时代, 面向知识产权的科技资源呈现数据规模大、时效性高和价值密度较低等趋势, 为有效利用知识产权资源带来严峻的挑战. 同时, 各个国家对知识产权中隐匿信息挖掘的需求日益增加, 使得面向知识产权的科技资源画像构建成为当下的研究热点. 目标是通过智能化的数据获取、实体识别以及可视化的方式对知识产权进行画像构建. 然而, 现有的科技资源画像构建方法只适用于结构化数据, 忽略了词语的词性对句子语义理解的影响. 因此, 提出了一种新颖的面向知识产权的科技资源画像构建算法, 针对自动获取的知识产权资源, 通过引入词性级别的注意力机制提高实体识别准确率, 并以可视化的形式构建知识产权科技资源画像. 相比于现有方法, 所提出的面向知识产权的科技资源画像构建方法具有以下优势: 1) 该算法利用词语的词性信息学习句子语义层面的含义, 并融合注意力机制, 以有监督的方式避免语义理解中的歧义; 2) 该模型能够智能自动地完成科技数据获取、命名实体识别、科技资源画像构建; 3) 大量实验结果表明, 所提方法利用词语的词性进行有监督学习, 在命名实体识别任务中综合性能优于对比算法.

关键词: 科技资源画像; 实体识别; 数据获取; 知识产权

中图法分类号: TP18

中文引用格式: 杨佳鑫, 杜军平, 邵蓥侠, 李昂, 奚军庆. 面向知识产权的科技资源画像构建方法. 软件学报, 2022, 33(4): 1439−1450. http://www.jos.org.cn/1000-9825/6483.htm

英文引用格式: Yang JX, Du JP, Shao YX, Li A, Xi JQ. Construction Method of Intellectual-property-oriented Scientific and Technological Resources Portrait. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1439−1450 (in Chinese). http://www.jos.org.cn/1000-9825/6483.htm

Construction Method of Intellectual-property-oriented Scientific and Technological Resources Portrait

YANG Jia-Xin[1], DU Jun-Ping[1], SHAO Ying-Xia[1], LI Ang[1], XI Jun-Qing[2]

[1](Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia (Beijing University of Posts and Telecommunications), Beijing 100876, China)

[2](Information Center, Ministry of Justice of the People's Republic of China, Beijing 100020, China)

Abstract: In the era of big data, intellectual-property-oriented scientific and technological resources show trends such as large data scale, high timeliness, and low value density, which poses severe challenges for the effective use of intellectual property resources. At the same time, the demand for the mining of hidden information in intellectual property rights is increasing in various countries, making the construction of intellectual-property-oriented scientific and technological resource portraits a current research hotspot. This study aim at building a portrait of intellectual property through intelligent data acquisition, entity recognition and visualization. However, the existing methods for constructing scientific and technological resource portraits are only suitable for structured data and ignore the impact of

* 基金项目: 国家重点研发计划(2018YFB1402600); 国家自然科学基金(61772083, 61802028); 广西科技重大专项(桂科AA18118054)

本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.

收稿时间: 2021-08-27; 修改时间: 2021-07-16, 2021-11-07, 2021-12-27; 采用时间: 2022-01-29; jos 在线出版时间: 2022-01-29

words' part of speech on the semantic understanding of sentences. Therefore, a novel algorithm is proposed for the construction of intellectual-property-oriented portraits of scientific and technological resources. Regarding the automatically acquired intellectual property resources, attention mechanism of part-of-speech level is introduced to improve the accuracy of entity recognition, and intellectual-property-oriented scientific and technological resource portraits are visually constructed. Compared with the existing methods, the proposed method has the following advantages: 1) This utilizes the part-of-speech information of words to learn the semantic meaning of sentences, and integrates the attention mechanism to avoid ambiguities in semantic understanding in a supervised way. 2) This model can intelligently and automatically complete sci-tech data acquisition, named entity recognition, and construction of scientific and technological resource portraits. 3) Extensive experiments demonstrate that our method performs better than baselines in named entity recognition by utilizing the part of speech of words for supervised learning.

Key words: scientific and technological resource portrait; entity recognition; data acquisition; intellectual property

知识产权是保护和激励创新的制度基石, 且隐匿着一个国家背后科技创新的实情. 面向知识产权的科技资源画像构建已成为当下的研究热点, 它是对科技资源实体的抽象和概括, 主要由知识产权资源实体的各种属性、实体间的关联关系以及由属性和关系挖掘得到的高维信息特征构成. 通过精准画像, 可以从海量知识产权数据中获取隐匿、时变的信息, 掌握知识产权中蕴含的发展规律. 借助这些信息能够进一步设计满足用户需求的应用, 比如面向政府部门的管理决策支撑服务[1]、知识产权推荐系统[2]等. 本文的目标是通过智能化的数据获取、实体识别以及可视化的方式对知识产权进行画像构建.

针对科技资源画像的构建, 现有方法主要包括基于本体的构建方法[3−6]、基于主题的构建方法[7]、基于剖面矩阵的构建方法[8]、基于语义挖掘的构建方法[9−11]等. 然而, 上述方法仅适合处理结构化的科技资源信息. 在现实的开放互联网环境中, 海量的科技资源通常以非结构化的文本形式存在, 因此, 从非结构化文本中识别出实体, 是科技资源画像构建的重要环节. 针对科技资源的实体识别, 现有方法主要包括基于文本观测序列的识别方法[12]、基于上下文语境信息的识别方法[13]、结合文本观察序列和上下文语境信息优势的实体识别方法[14]、对左右上下文进行联合调节的方法[15]等. 然而, 这些方法忽略了词语词性对于语义理解的重要影响.

本文充分考虑词语词性在自然语言理解中的重要性, 引入中文词性级别的注意力机制, 以有监督的方式对知识产权数据集进行学习与训练, 提出了适用于中文知识产权资源的实体识别算法, 用于构建面向知识产权的科技资源画像. 在画像构建过程中, 笔者首先对海量的科技数据进行高效地处理, 过滤无关信息, 获取知识产权数据, 并以恰当的形式存储到合适的位置. 然后, 针对专利摘要等非结构化的知识产权数据中, 笔者运用本文提出的结合分词词性级别注意力机制的命名实体识别算法, 抽取文本中的命名实体; 对于专利信息等结构化数据, 使用基于规则匹配的抽取方法, 获得实体的属性信息. 接着, 笔者对知识产权数据进行知识融合, 构建专利申请单位或个人与抽取实体的关系, 利用 neo4j 图数据库存储已经获取的实体以及实体关系, 为知识图谱的创建提供必要的技术支持. 最后, 在知识图谱的基础上, 笔者运用词频分析、共词分析、网络中心度、聚类分析等方法, 结合数据统计与挖掘方法, 通过各项统计指标反映技术的热度、企业的影响力等科技实体隐藏在知识图谱中的知识, 完成科技资源画像的构建.

本文实现了服务于国家知识产权局和专家学者的面向知识产权的科技资源画像构建系统. 通过本文的面向知识产权的科技资源画像构建方法, 可以获取大量科技资源; 然后结合本文提出的实体识别方法, 规则化科技资源信息, 最终可以得到科技资源的知识图谱、科技资源的发展趋势和科技资源的热点词云.

本文的主要贡献如下:

1) 提出了一种针对科技资源的实体识别算法(entity recognition based on word segmentation attention, ERWSA), 通过引入中文词性级别注意力机制, 利用词语的词性对句子语义进行有监督学习, 能够精准识别出科技资源中的实体;

2) 提出了一种面向知识产权的科技资源画像构建方法, 通过结合本文提出的科技资源获取方法、ERWSA 和可视化技术, 实现了面向知识产权的智能自动化科技资源获取、命名实体识别、科技资源画像构建;

3) 大量的实验结果表明: 本文所提的 ERWSA 算法在知识产权相关实体识别的准确率、召回率、$F1$ 值上的整体表现优于基于双向长短时循环网络(BLSTM)和双向编码器表征模型(BERT)的对比方法; 但在时间、组

织等具体实体类别的识别准确率、召回率、*F*1 值上稍差于基于 BERT 的命名实体识别方法.

1 相关工作

近年来, 越来越多的学者注意到科技资源信息的重要性[16,17]. 然而, 科技资源画像相比于常见的用户画像[18]、学者画像[19]而言面临更多挑战: 一方面, 用户画像、学者画像可以使用公开数据集进行构建. 在社交媒体分析场景[20], Liang 等人[18]利用 Twitter 数据进行用户画像构建, Tang 等人[21]建立了以科研人员、科技文献、学术活动这 3 大类数据为基础的 AMiner 数据平台, 为进行学者画像研究提供了数据基础. 相比之下, 科技资源画像缺乏公共数据集, 且数据广泛分布于互联网中, 因此需要一套从海量数据中获取科技资源并构建画像的自动化方法. 另一方面, 科技资源画像的构建需要从获取到的科技资源中精准识别出有效信息[22], 然而现有的用户画像、学者画像方法的准确率有限, 为科技资源的画像构建增加了难度.

在科技资源画像构建方面, 现有方法主要包括基于本体的构建方法、基于主题的构建方法、基于剖面矩阵的构建方法、基于语义挖掘的构建方法. 欧洲科学家采用系统 Euro-CRIS 构建了统一的描述模型 CERIF[3], 构建多类科技资源的画像. Wang 等人[4]提出了一种基于知识图谱的本体模型构建科技资源画像的方法. 然而知识图谱的质量管理仍需花费大量的人力, 实体对齐的准确度仍需提高. 葛胤池等人[6]提出了一种基于领域本体的科技资源聚类方法, 通过构建科技资源领域本体树和概念语义关系矩阵, 并对其运用主成分分析方法进行降维处理以构建科技资源向量空间, 最终通过聚类构建科技资源画像. 但是该方法依赖于基于领域本体的数据预处理, 而且大规模领域本体构建目前仍然是一个困难的工作. Zha 等人[7]利用主题模型解决科技文本的多标签分类问题, 提出了一种种子引导的多标签主题模型, 无需使用任何标记数据或外部资源. 然而该方法仅适用于训练数据规模小或数据质量无法保证的场景. Kuan 等人[8]采用二维剖面矩阵, 更全面地捕捉专利组合的技术内容, 对专利资源在两个或多个符号的共同转让中可能“隐藏”的信息进行科技资源画像. 然而实验只在小规模专利信息上进行, 尚未验证该方法在海量科技资源中的有效性. Wang 等人[9]采用语义分析方法来构建科技资源标签体系. Kozerenko 等人[10]和 Qin 等人[11]从科研专家的参与基金项目、发表文章、专著和授权专利中挖掘科研成果信息, 运用 k 近邻[23]等算法构建了基于语义的科研成果画像. 然而, 上述方法只适用于结构化的科技资源信息, 难以处理现实场景中的以非结构化的形式存在的科技资源.

在开放的网络资源中, 文本信息难以规则化, 不利于构建科技资源画像, 因此基于文本信息进行实体识别变得尤为重要. Lafferty 等人[12]提出了条件随机场(CRF), 结合最大熵模型和隐马尔可夫模型[24]的特点, 在词性标注和命名实体识别等序列标注任务中取得了不错的效果. 但 CRF 方法需要手动提取序列特征, 增加了对数据的人为干扰. 近年来, 基于深度学习的方法能够从原始数据中自主地学习, 不需要人为设定特征, 因此可以减少对数据的人为干扰. Graves [13]提出了 BLSTM 模型, 可以从前向和后向 2 个方向对句子进行建模, 既能保存前面的上下文信息, 又能考虑到句子未来的上下文信息, 使其在中文命名实体识别任务中取得更好的效果. 然而, 这样的模型无法学习到输出的标注之间的转移依赖关系以及序列标注的约束条件. Huang 等人[14]结合 CRF 和 BLSTM, 在 LSTM 后面再加一层 CRF, 以获得两者的优点, 使得该模型既可以学习到句子的约束条件, 又可以捕捉句子的上下文语境. Devlin 等人[15]提出了一种来自变换器的双向编码器表征模型(BERT), 通过在所有神经网络层中对左右上下文进行联合调节, 从未标记文本中预训练深层双向表示. 最近的模型 BERT+CRF 在预训练语言模型 BERT 基础上引入 CRF 模块, 学习句子的约束条件. 但是上述模型忽略了词语的词性在语义理解中的重要性. 本文针对专利文本, 在 BLSTM 与条件随机场序列标注模型(CRF)的基础上引入了分词后词性标注结果的注意力机制进行有监督学习, 通过学习中文词性避免句子语义可能存在的歧义, 进一步提升了实体识别的效果.

2 面向知识产权的科技资源画像构建方法

本文提出了面向知识产权的科技资源画像构建方法, 其框架如图 1 所示, 该框架主要包括科技资源获取模块、实体识别和关系构建模块、科技资源画像构建模块.

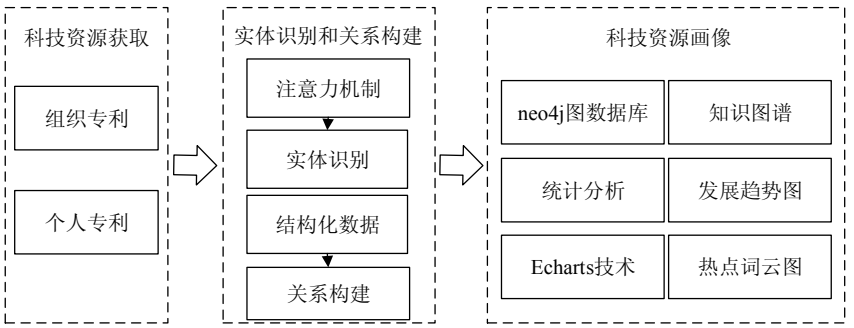


图 1 面向知识产权的科技资源画像构建方法框架

科技资源获取模块负责爬取机构、组织、个人发表的专利等知识产权数据. 然后, 在实体识别和关系构建模块中, 框架结合注意力机制对专利数据进行命名实体识别, 利用爬取到的结构化数据完成实体间的关系构建. 最后, 框架利用得到的实体关系和统计分析方法对科技资源进行画像, 展示科技资源的知识图谱、专利的发展趋势和专利的热点词云.

2.1 科技资源数据获取

科技资源数据获取模块主要使用 scrapy 框架对 Innojoy 专利网站(http://www.innojoy.com/search/index.html)进行数据抓取, 并配置代理池, 分多线程抓取页面; 使用布隆过滤器判断页面是否进行访问, 并用正则匹配出抓取的下一个站点. 针对专利文本, 科技资源数据获取模块采用正则截取方法对组织或个人与专利的关系进行提取, 获取科技资源数据. 科技资源数据获取流程如图 2 所示.

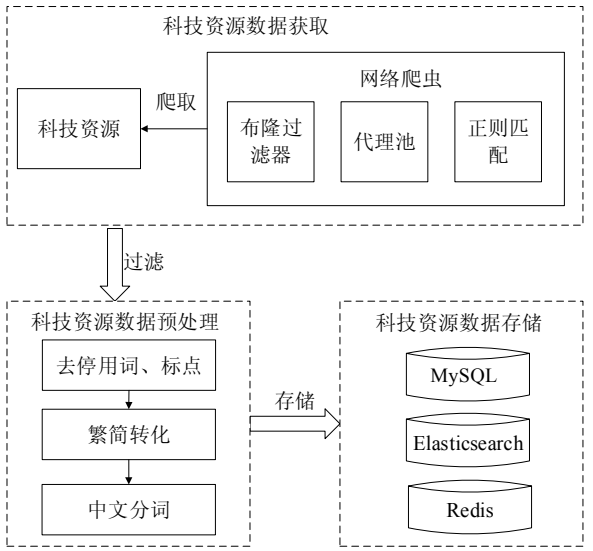


图 2 科技资源数据获取流程

此模块在获取科技资源原始数据后, 通过构建信任网络对科技资源大数据进行评估, 去除无关和虚假的内容, 将有效的内容进行进一步的处理. 对于提取到的文本数据, 去除内容过短、重复的专利文本数据. 然后对有效数据进行去停用词、繁简转化等操作, 减少同义文本的影响. 最后, 将预处理后的数据进行分词, 对单词进行 ID 化, 并构造字典, 得到一系列由 ID 符号构成的文档.

2.2 实体识别与关系构建

为了提升专利文本数据的实体识别准确性, 有效支持知识图谱可视化, 本文提出了一种结合分词词性的

注意力机制的命名实体识别算法 ERWSA. 此算法在 BLSTM 与 CRF 的基础上引入了分词后词性标注结果的注意力机制, 提高了科技大数据的实体识别准确率. 具体而言, 首先利用 jieba 分词方法将专利文本数据进行文本分词, 然后利用 Word2Vec 模型得到的分词向量和字符向量表示作为 BLSTM 初始输入, 经过字符级别的注意力机制后, 融合分词词性注意力机制, 得到输出为从专利文本中识别出的实体向量. 结合分词词性的注意力机制的命名实体识别方法如图 3 所示.

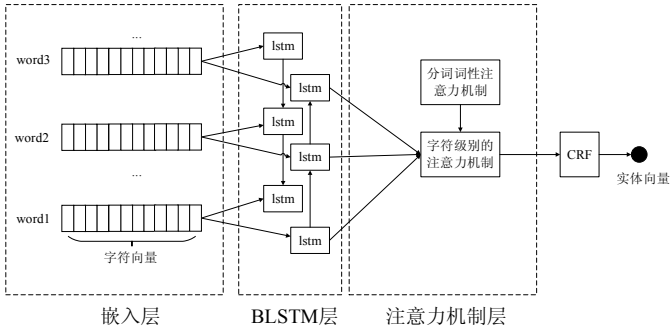


图 3 结合分词词性的注意力机制的命名实体识别方法

ERWSA 算法的输入为专利文本的字向量, 其中, 字向量是从一个均匀分布中随机采样的值. 假设一个句子存在 n 个汉字, 且每个汉字的特征向量为 m, 则句子的输入向量为 $Sen_{n,m}$. 之后需要经过 BLSTM, 得到输出结果 $h_{n,k}$, 其中, n 指句子中汉字的个数, k 指 BLSTM 隐藏层输出的维度.

将 $h_{n,k}$ 融入引入分词词性的注意力机制层中, 得到融入中文分词词性特征的注意力机制层的输出结果. 其中, 注意力机制层的原理如图 4 所示.

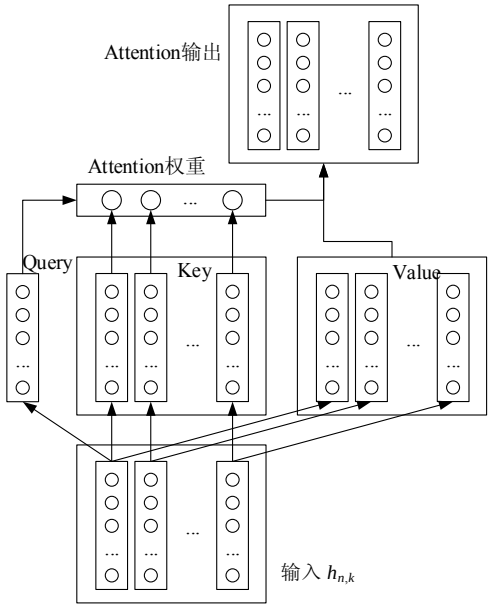


图 4 引入分词词性的注意力机制层结构图

注意力机制层中的 Value 特征矩阵为输入的原始矩阵, 即 Value 矩阵与输入向量矩阵是一样的. 而注意力机制层中向量矩阵 Key 是由原始输入文本的分词词性特征生成而来的. 利用均匀分布进行每个词性特征向量的初始化, 其中, 词性特征向量的维数应与 BLSTM 隐藏层输出的维度一样. 利用注意力机制层中的 $Q(query)$, $K(key)$, $V(value)$的关系, 可以得到如下公式(1)、公式(2):

$$att\_w_{n,1}=softmax(key_{n,k}*query_{k,1}) \tag{1}$$

其中, *softmax* 是用来进行归一化处理的; $att\_w_{n,1}$ 为注意力机制层的权重值, 表示为对于一个句子的 n 个汉字进行实体名分类的权值.

$$att\_r_{n,k}=att\_w_{n,1}\cdot value_{n,k} \tag{2}$$

其中, ·表示注意力机制层的权重值与 *Value* 矩阵进行乘积运算, $att\_r_{n,k}$ 是注意力机制层的输出值.

在输入 CRF 之前需要进行一次全连接层操作, 其计算如公式(3)所示:

$$crf\_i_{c,n} = softmax(W_{c,k} * att\_r_{n,k}^{T} + B_{c,1}) \tag{3}$$

其中, c 为实体识别的类别数量, $B_{c,1}$ 为全连接层的偏移向量, $crf\_i_{c,n}$ 为 CRF 层的输入特征矩阵, $W_{c,k}$ 为全连接层的权重矩阵.

将全连接层的输出特征再输入到 CRF 层, 计算如公式(4)所示:

$$S(X,Y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=0}^{n}P_{i,y_i} \tag{4}$$

其中, A 是状态转移矩阵, 代表从第 i 个 tag 转移到第 j 个 tag 的概率. 利用极大似然的方法求得 $S(X,Y)$最大值, 得到最佳的输出标签序列. 最后预测标签转换为单标签向量生成期望的标签结果. 利用梯度下降算法训练实体识别模型, 从而得到专利文本数据的实体识别模型.

通过本文提出的 ERWSA 方法对现有的知识产权数据进行实体抽取, 为科技资源画像的构建提供重要的技术支持. 根据爬取到的每个专利文本对应的申请单位或个人, 构建申请单位或个人与抽取实体的关系, 然后利用 neo4j 图数据库存储已经获取的实体以及实体关系, 为知识图谱的创建提供必要的技术支持.

2.3 面向知识产权的科技资源画像

通过科技资源数据获取与预处理(第 2.1 节)和实体识别与关系构建(第 2.2 节), 得到了专利文本知识产权数据的实体和关系. 在此基础上, 笔者进一步实现面向知识产权的科技资源画像, 整体流程如图 5 所示.

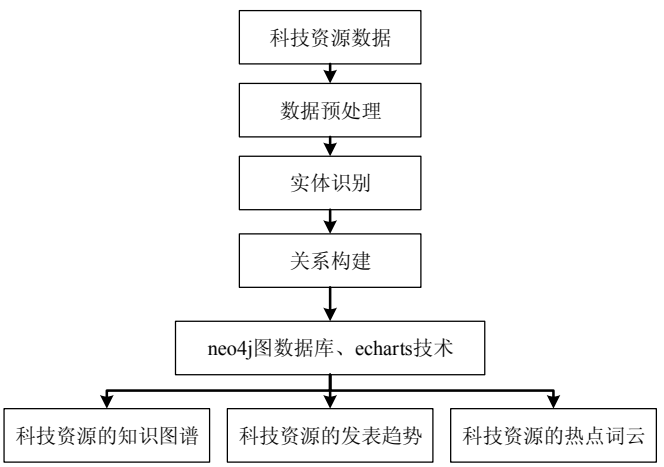


图 5 科技资源画像流程

笔者利用数据统计与挖掘方法, 通过计算各项统计指标发现能够反映技术热度、组织或个人影响力的科技实体, 挖掘隐藏在知识图谱中的知识, 完成科技大数据知识图谱的构建. 例如, 统计科技资源历年发表数量, 绘制出每年科技资源(专利)的发表趋势图, 从而方便用户跟踪科技资源发表热点. 统计数据库中识别的实体出现次数, 按照词频出现次数对识别的实体进行排名, 取 top25 用于构造热点词云, 完成科技资源画像.

2.4 实验结果及分析

1) 数据集

为了验证本文提出的结合分词词性的注意力机制的命名实体识别算法(ERWSA), 构建了一个面向知识产权的科技资源数据集. 通过开发基于 scrapy 框架的分布式抓取系统, 对科技资源数据进行抓取, 同时抓取组织或个人与专利发表的关系. 抓取的科技资源数据集统计见表 1.

表 1 面向知识产权的科技资源数据集

数据种类	数量(个)
专利文本	272 636
组织发表专利	215 880
个人发表专利	58 756

利用本文开发的爬虫系统, 构造含有 272 636 个专利文本的科技资源数据集, 其中, 组织发表专利数 215 880 个, 个人发表专利数 58 756 个. 手动标记其中部分专利文本数据, 用于验证本文提出的结合分词词性注意力机制的实体识别方法的有效性. 标记得到的专利文本数据见表 2.

表 2 用于实体识别的数据集

	训练集	测试集
标记的专利文本数	6 884	1 720
字符数	388 019	97 210
带标记字符数	66 740	18 042
标记占比	17.20%	18.56%
带词性标注数	204 400	51 162
实体总数	21 524	4 025
实体类别	6	6

用于实体识别的专利数据集包括训练集与测试集两部分, 其中, 专利文本训练数据集有 6 884 条, 专利文本测试数据集有 1 720 条. 按照训练集大小和测试集大小比例为 8:2 设定. 数据集的详细分布如表 2 所示. 训练集与测试集中标签占比基本一致, 保证数据的一致性.

在实验中将所有的专利文本数据的标签分为 B_LOC, M_LOC, E_LOC, B_ORG, M_ORG, E_ORG, B_PER, M_PER, E_PER, B_TIME, M_TIME, E_TIME, B_TECH, M_TECH, E_TECH 以及 O. 其中: B_XXX 表示实体名的开始, M_XXX 表示实体名中间数据, E_XXX 表示实体名的结尾; LOC 表示地点; ORG 表示组织; PER 表示人名; TIME 表示时间; TECH 表示技术实体; O 表示其他实体.

2) 对比算法

本文将所提的 ERWSA 算法与如下基准算法进行对比.

- CRF[12]: 结合最大熵模型和隐马尔可夫模型的特点, 利用状态序列下观测序列的条件概率分布, 学习到句子的约束条件;
- BLSTM[13]: 考虑到当前时刻的输出不仅和之前的状态有关, 还与未来的状态有关系. BLSTM 既能保存句子先前的上下文信息, 也能同时考虑到句子未来的上下文信息;
- BLSTM+CRF[14]: 结合 CRF 和 BLSTM 两者的优点, 既可以学习到句子的约束条件, 又可以捕捉句子的上下文语境;
- BERT[15]: 在所有神经网络层中对左右上下文进行联合调节, 从未标记文本中预训练深层双向表示;
- BERT+CRF: 在预训练语言模型基础上引入 CRF 模块, 学习句子的约束条件.

3) 实验 1: ERWSA 的有效性验证

为验证本文提出的 ERWSA 方法的有效性, 本文采用准确率、召回率和 $F1$ 值作为结果的评价指标. 分别使用 CRF, BLSTM, BLSTM+CRF, BERT, BERT+CRF 这 5 种算法进行对比实验, 实验结果如表 3 所示.

表 3 ERWSA 方法与对比算法的性能比较

对比算法名	准确率	召回率	*F*1 值
CRF	0.795 8	0.632 3	0.704 7
BLSTM	0.796 0	0.634 2	0.705 9
BLSTM+CRF	0.841 8	0.752 6	0.794 7
BERT	0.684 3	0.776 7	0.727 6
BERT+CRF	0.734 9	0.793 1	0.760 3
ERWSA	**0.850 2**	**0.798 4**	**0.823 5**

从表 3 可以得到以下发现: 1) 本文提出的 ERWSA 方法在准确率、召回率、*F*1 值这 3 个性能指标上均优于其他的对比方法, 验证了本文所提算法的有效性; 2) BLSTM 的性能略微优于 CRF, 表明考虑句子的上下文语境比学习句子的约束条件更有利于进行实体识别; 3) BLSTM+CRF 的性能优于 BLSTM, 验证了通过引入 CRF 层可以加入一些约束来保证最终预测结果的有效性, 提升 BLSTM 在实体识别上的性能; 4) ERWSA 的性能优于 BLSTM+CRF 的性能, 表明考虑词性在句子语义理解中的重要性以及融合注意力机制有助于进一步提升实体识别的性能; 5) 基于预训练语言模型 BERT 的整体效果弱于 ERWSA, 一个原因是缺乏大规模的专利相关文本进行预训练模型的微调, 另一个原因是对特定的实体识别效果不佳, 详见实验 2~实验 4 的结果分析.

为了进一步论证本文提出的 ERWSA 方法在多类实体识别上的有效性, 分别统计 CRF, BLSTM, BLSTM+CRF, BERT, BERT+CRF 这 5 种对比算法在组织、时间、地点、人名、技术实体识别任务中的准确率、召回率和 *F*1 值.

4) 实验 2: ERWSA 在不同实体识别的准确率对比

为验证本文所提的 ERWSA 算法针对不同类别的实体在识别准确率上均有提升, 将 ERWSA 与对比算法在组织、时间、地点、人名、技术这 5 类实体识别任务上进行了准确率的比较, 实验结果如表 4、图 6 所示.

表 4 ERWSA 方法与对比方法的准确率比较

	组织	时间	地点	人名	技术
CRF	0.773 8	0.727 2	0.832 2	0.850 9	0.846 9
BLSTM	0.671 2	0.669 1	0.834 4	0.851 2	0.851 2
BLSTM+CRF	0.798 3	0.783 6	0.864 2	0.880 6	0.878 6
BERT	**0.912 1**	0.990 5	0.314 6	0.714 2	0.645 4
BERT+CRF	0.911 1	**0.992 6**	0.615 4	0.842 7	0.694 4
ERWSA	0.799 3	0.814 3	**0.896 5**	**0.885 2**	**0.895 4**

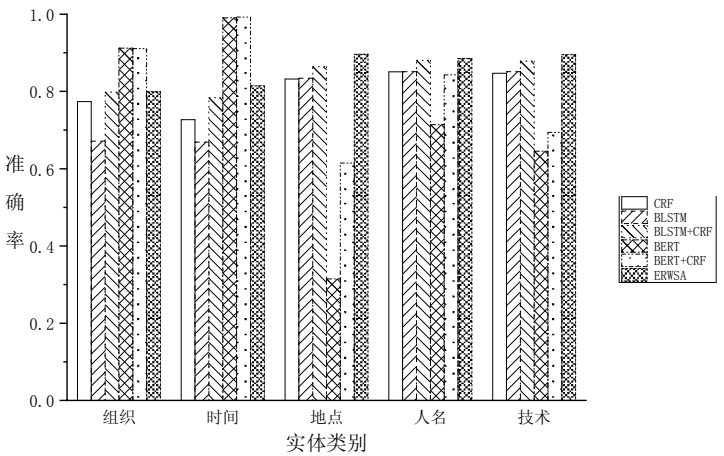


图 6 ERWSA 方法与对比方法在各类实体的识别精确率

从表 4 和图 6 可以看出, 本文提出的 ERWSA 方法在组织、时间、地点、人名、技术这 5 个实体识别任务中的准确率都要好于 CRF, BLSTM, BLSTM+CRF 算法, 其中, 在地点实体识别任务提升最多, 准确率相较于目前最优算法 BLSTM+CRF 提高了 3.23%; 在组织实体识别任务提升最少, 但准确率相较于目前最优算法

BLSTM+CRF 也提高了 0.1%, 从而进一步验证了本文提出的 ERWSA 方法的有效性. 与 BERT 和 BERT+CRF 等预训练语言模型相比, ERWSA 算法在地点、人名、技术等 3 类实体识别的准确率更优, 在组织和时间等两类实体的准确率上表现较差. 这说明 ERWSA 算法在组织和时间实体方面的识别准确率有待进一步提升. 总体结果表明: ERWSA 方法在 BLSTM+CRF 的基础上引入分词词性级别的注意力机制, 能够进一步提升实体识别准确率.

5) 实验 3: ERWSA 在不同实体识别的召回率对比

为验证本文所提的 ERWSA 算法针对不同类别的实体在识别召回率上均有提升, 将 ERWSA 与对比算法在组织、时间、地点、人名、技术这 5 类实体识别任务上进行了召回率的比较, 实验结果如表 5、图 7 所示.

表 5 ERWSA 方法与对比方法的召回率比较

	组织	时间	地点	人名	技术
CRF	0.521 3	0.526 5	0.577 6	0.767 1	0.767 4
BLSTM	0.547 8	0.525 5	0.586 9	0.755 5	0.755 5
BLSTM+CRF	0.634 3	0.738 8	0.744 0	0.831 5	0.820 5
BERT	0.929 0	**0.994 2**	0.235 2	0.746 4	0.748 6
BERT+CRF	**0.911 1**	0.993 1	0.134 4	0.725 1	0.770 8
ERWSA	0.750 6	0.700 2	**0.795 7**	**0.868 7**	**0.871 2**

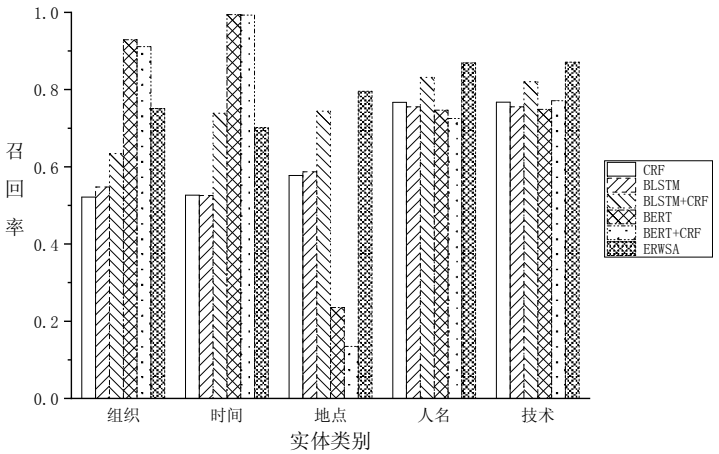


图 7 ERWSA 方法与对比方法在各类实体的识别召回率

从表 5 和图 7 可以看出, 本文提出的 ERWSA 方法在地点、人名、技术这 3 个实体识别任务中的召回率都要好于 CRF, BLSTM, BLSTM+CRF, BERT, BERT+CRF 算法, 其中: ERWSA 算法在地点实体识别任务提升最为明显, 召回率相较于目前最优算法 BLSTM+CRF 提高了 5.17%; 在人名实体识别任务提升最少, 召回率比目前最优算法 BLSTM+CRF 提高了 3.72%, 从而进一步验证了算法的有效性. 与 BERT 和 BERT+CRF 相比, 在召回率方面, ERWSA 算法在不同实体类别识别性能上的表现与准确率上的表现相似. ERWSA 算法在 BLSTM+CRF 算法的基础上融入分词词性注意力机制, 使得具有词性标注的词语进一步辅助语义理解, 达到提升命名实体识别的效果. ERWSA 算法在 5 个实体识别任务中的平均召回率要高于其他对比算法, 也说明算法的有效性.

6) 实验 4: ERWSA 在不同实体识别的 $F1$ 值对比

为了验证本文所提的 ERWSA 算法针对不同类别的实体在识别 $F1$ 值上均有提升, 将 ERWSA 与对比算法在组织、时间、地点、人名、技术这 5 类实体识别任务上进行了 $F1$ 值的比较, 实验结果如表 6、图 8 所示.

表 6 ERWSA 方法与对比方法的 *F*1 值比较

	组织	时间	地点	人名	技术
CRF	0.622 9	0.610 8	0.681 9	0.806 8	0.805 2
BLSTM	0.603 3	0.588 7	0.689 1	0.800 5	0.800 5
BLSTM+CRF	0.706 9	0.760 5	0.799 6	0.855 3	0.848 6
BERT	**0.920 4**	0.992 4	0.269 2	0.730 0	0.693 2
BERT+CRF	0.911 1	**0.992 9**	0.220 6	0.771 7	0.730 6
ERWSA	0.774 2	0.753 0	**0.843 1**	**0.876 9**	**0.883 1**

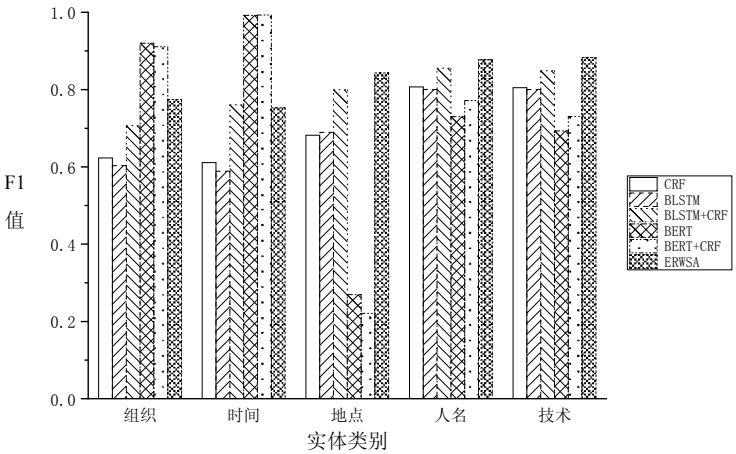


图 8 ERWSA 方法与对比方法在各类实体的识别 *F*1 值

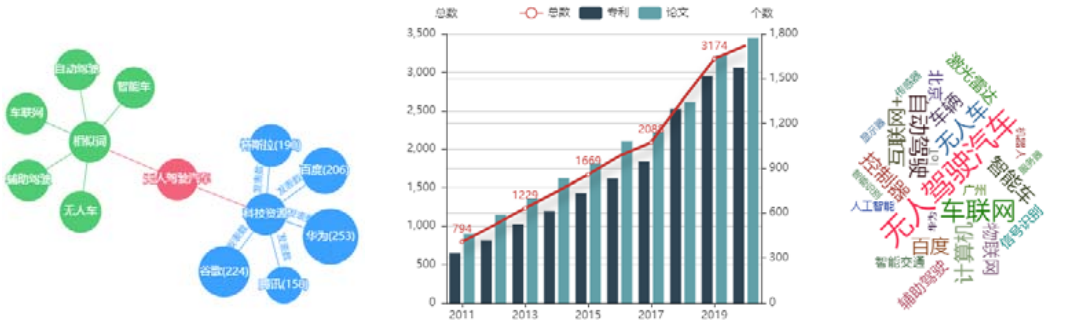
从表 6 和图 8 可以看出, 本文提出的 ERWSA 方法在地点、人名、技术这 3 个实体识别任务中的 *F*1 值都要好于 CRF, BLSTM, BLSTM+CRF, BERT, BERT+CRF 算法, 其中: ERWSA 算法在地点实体识别任务提升最为明显, *F*1 值相较于目前最优算法 BLSTM+CRF 提高了 4.35%; 在人名实体识别任务提升最少, 召回率比目前最优算法 BLSTM+CRF 提高了 2.16%, 进一步表明了该方法的有效性. 由于 *F*1 值是通过召回率和准确率综合计算得到, ERWSA 算法与 BERT 和 BERT+CRF 相比在 *F*1 值上的性能表现与实验 2 和实验 3 分析一致. 实验表明: ERWSA 算法在 BLSTM+CRF 算法的基础上利用分词词性标注的特点引入注意力机制, 对于实体名的边界进行重点识别, 能够提升命名实体识别的效果.

3 面向知识产权的科技资源画像展示

在知识图谱的基础上, 运用聚类分析、词频分析等方法, 结合数据统计与挖掘方法, 通过各项统计指标反映技术的热度、企业的影响力等科技实体, 完成科技资源画像的构建. 以“无人驾驶汽车”相关的专利为例, 构建面向知识产权的科技资源画像, 如图 9(a)所示.

统计科技资源的历年发表数量, 绘制出每年科技资源的发表趋势图, 从而方便用户跟踪科技资源的发展趋势. 以“无人驾驶汽车”相关的专利为例, 绘制出其 2011–2020 年科技资源发展趋势画像, 如图 9(b)所示.

得到实体识别模型后, 可以对所有的专利科技资源进行实体识别. 对与关键词相关的专利文本中的实体进行排序, 取实体词频的 top25 用于构建科技资源的热点词云画像. 用户通过科技资源的热点词云图画像, 可以追踪到最近的热门技术名词, 方便用户获取科技信息. 以关键词“无人驾驶汽车”为例, 构建相关的科技资源热点词云画像, 如图 9(c)所示.



(a) 科技资源的知识图谱 (b) 科技资源发表历年趋势图 (c) 科技资源的热点词云

图 9 面向知识产权的科技资源画像系统

4 总 结

构建面向知识产权的科技资源画像, 有助于挖掘其中隐匿的深层科研价值, 进一步助力国家的科技自主创新. 在科技资源画像构建任务中, 针对现有的画像构建方法只适用于结构化数据、忽略了词语词性在句子语义理解中的作用等不足, 本文提出了一种面向知识产权的科技资源画像构建方法, 主要包括科技资源数据获取、实体识别与关系构建、面向知识产权的科技资源画像、面向知识产权的科技资源画像系统构建. 通过数据爬虫技术、数据过滤、数据预处理等技术, 完成对科技资源的高效获取; 结合分词词性、双向长短时循环网络和注意力机制, 利用有监督方法对知识产权相关数据集进行学习与训练, 得到针对科技资源的实体识别; 利用爬取到的申请组织或个人结构化数据来构建与抽取实体之间的关系; 采用 Neo4j 图数据库、Echarts 技术完成对科技资源画像的构建, 得到科技资源的知识图谱、科技资源的发展趋势和科技资源的热点词云. 本文提出的方法在知识产权相关实体识别的准确率、召回率、*F*1 值上的表现均优于对比方法.

References:

[1] Gao X. Intellectual property powers scientific and technological innovation to drive high-quality development. Management and Administration, 2021, 39(4): 82–85 (in Chinese with English abstract).

[2] He XJ, Dong YB, Zhen Z, *et al*. Weighted meta paths and networking embedding for patent technology trade recommendations among subjects. Knowledge-Based Systems, 2019, 184(22): 104899.

[3] Kremenjaš D, Udovičić P, Orel O. Adapting CERIF for a national CRIS: A case study. In: Proc. of the 43rd Int'l Conf. on Information, Communication and Electronic Technology (MIPRO). 2020. 1633–1638.

[4] Wang Y, Qian L, Xie J, *et al*. Building knowledge graph with sci-tech big data. Data Analysis and Knowledge Discovery, 2019, 3(1): 15–26 (in Chinese with English abstract).

[5] Logesh R, Subramaniyaswamy V, Vijayakumar V, *et al*. Efficient user profiling based intelligent travel recommender system for individual and group of users. Mobile Networks and Applications, 2019, 24(3): 1018–1033.

[6] Ge YC, Zhang H, Song WY, *et al*. Scientific and technology resource clustering based on domain ontology. Frontiers of Data & Computing, 2020, 2(5): 13–22 (in Chinese with English abstract).

[7] Zha D, Li C. Multi-label dataless text classification with topic modeling. Knowledge and Information Systems, 2019, 61(1): 137–160.

[8] Kuan CH, Tu WM, Chen DZ. Two-dimensional technology profiling of patent portfolio. In: Proc. of the IEEE Int'l Conf. on Industrial Engineering and Engineering Management (IEEM). 2018. 1342–1347.

[9] Wang T, Liu L, Liu N, *et al*. A multi-label text classification method via dynamic semantic representation model and deep neural network. Applied Intelligence, 2020, 50(8): 2339–2351.

[10] Kozerenko E, Sinyaghina Y, Somin N, *et al*. Problem domain ontology mining based on distributional semantics. In: Proc. of the Int'l Conf. on Computational Science and Computational Intelligence (CSCI). 2019. 439–444.

[11] Qin X, Zhang H, Zheng H. Construction of agricultural knowledge service platform for scientific and technological innovation. In: Proc. of the Int'l Conf. on Computer Engineering and Intelligent Control (ICCEIC). 2020. 18–21.

[12] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML). 2001. 282–289.

[13] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5-6): 602−610.
[14] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015.
[15] Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the NAACL-HLT. 2019. 4171−4186.
[16] Wu L, Wu C, Wu E. Research on safety information resources management mode from the perspective of big data. Science and Technology Management Research, 2020, 40(9): 156−162 (in Chinese with English abstract).
[17] Yang JX. Research on knowledge service and interactive visualization component of cross_media big data of science and technology [MS. Thesis]. 2021 (in Chinese with English abstract).
[18] Liang SS, Zhang XL, Ren ZC, *et al*. Dynamic embeddings for user profiling in Twitter. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2018. 1764−1773.
[19] Yuan S, Tang J, Gu XT. A survey on scholar profiling techniques in the open Internet. Journal of Computer Research and Development, 2018, 55(9): 1903−1919 (in Chinese with English abstract).
[20] Kou FF, Du JP, He YJ, *et al*. Social network search based on semantic analysis and learning. CAAI Trans. on Intelligence Technology, 2016, 1(4): 293−302,
[21] Tang J, Zhang J, Yao LM, *et al*. Arnetminer: Extraction and mining of academic social networks. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 990−998.
[22] Li J, Sun AX, Han JL, *et al*. A survey on deep learning for named entity recognition. IEEE Trans. on Knowledge and Data Engineering, 2020, 2981314.
[23] Sun B, Du JP, Gao T. Study on the improvement of *K*-nearest-neighbor algorithm. In: Proc. of the 2009 Int'l Conf. on Artificial Intelligence and Computational Intelligence, Vol.4. 2009. 390−393.
[24] Zhao HK, Liu Q, Zhu HS, *et al*. A sequential approach to market state modeling and analysis in online p2p lending. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2017, 48(1): 21−33.

附中文参考文献:

[1] 高翔. 发挥知识产权支撑保障作用助力科技创新驱动高质量发展. 经营与管理, 2021, 39(4): 82−85.
[4] 王颖, 钱力, 谢靖, 等. 科技大数据知识图谱构建模型与方法研究. 数据分析与知识发现, 2019, 3(1): 15−26.
[6] 葛胤池, 张辉, 宋文燕, 等. 基于领域本体的科技资源聚类方法研究. 数据与计算发展前沿, 2020, 2(5): 13−22.
[16] 吴林, 吴超, 吴娥. 大数据视域下安全信息资源管理模式研究. 科技管理研究, 2020, 40(9): 156−162.
[17] 杨佳鑫. 跨媒体科技大数据的知识服务与交互可视化构件研究[硕士学位论文]. 北京: 北京邮电大学, 2021.
[19] 袁莎, 唐杰, 顾晓韬. 开放互联网中的学者画像技术综述. 计算机研究与发展, 2018, 55(9): 1903−1919.

杨佳鑫(1996－), 男, 硕士, 主要研究领域为自然语言处理, 数据挖掘.

李昂(1993－), 男, 博士, 主要研究领域为信息检索, 数据挖掘, 机器学习.

杜军平(1963－), 女, 教授, CCF 会士, 主要研究领域为人工智能, 机器学习, 模式识别.

奚军庆(1977－), 男, 硕士, 主要研究领域为科技信息化建设应用, 大数据分析.

邵蓥侠(1988－), 男, 副教授, CCF 高级会员, 主要研究领域为大规模图分析, 并行计算框架, 知识图谱分析.