智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于Transformer与技术词信息的知识产权实体识别方法

王宇晖, 杜军平, 邵蓥侠

引用本文:
王宇晖,杜军平,邵蓥侠. 基于Transformer与技术词信息的知识产权实体识别方法[J]. 智能系统学报, 2023, 18(1): 186–193.
WANG Yuhui,DU Junping,SHAO Yingxia. An intellectual property entity recognition method based on Transformer and technological word information[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(1): 186–193.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202203036

DOI: 10.11992/tis.202203036
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20220606.1515.002.html

基于 Transformer 与技术词信息的知识产权实体识别方法

王宇晖[1,2], 杜军平[1,2], 邵蓥侠[1,2]

（1. 北京邮电大学 计算机学院, 北京 100876; 2. 北京邮电大学 智能通信软件与多媒体北京市重点实验室, 北京 100876）

摘　要: 专利文本中包含了大量实体信息, 通过命名实体识别可以从中抽取包含关键信息的知识产权实体信息, 帮助研究人员更快了解专利内容。现有的命名实体提取方法难以充分利用专业词汇变化带来的词层面的语义信息。本文提出基于 Transformer 和技术词信息的知识产权实体提取方法, 结合 BERT 语言方法提供精准的字向量表示, 并在字向量生成过程中, 加入利用字向量经迭代膨胀卷积网络提取的技术词信息, 提高对知识产权实体的表征能力。最后使用引入相对位置编码的 Transformer 编码器, 从字向量序列中学习文本的深层语义信息, 并实现实体标签预测。在公开数据集和标注的专利数据集的实验结果表明, 该方法提升了实体识别的准确性。

关键词: 中文命名实体识别; 知识产权; Transformer 编码器; 信息融合; 向量表示; 科技大数据; 专利; 深度学习
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)01−0186−08

中文引用格式: 王宇晖, 杜军平, 邵蓥侠. 基于 Transformer 与技术词信息的知识产权实体识别方法 [J]. 智能系统学报, 2023, 18(1): 186–193.
英文引用格式: WANG Yuhui, DU Junping, SHAO Yingxia. An intellectual property entity recognition method based on Transformer and technological word information[J]. CAAI transactions on intelligent systems, 2023, 18(1): 186–193.

An intellectual property entity recognition method based on Transformer and technological word information

WANG Yuhui[1,2], DU Junping[1,2], SHAO Yingxia[1,2]

(1. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Patent text contains abundant entity information, from which the intellectual property (IP) entity information containing key information can be extracted through named entity recognition, which helps researchers understand patent content faster. For the existing named entity extraction method, the semantic information at the word level brought by a change in technical words is difficult to fully use. In this paper, the IP entity information extraction method based on Transformer and technical word information is proposed, which provides exact word vector representation based on the BERT language model. In the process of word vector generation, this method improves the representation ability of IP entities by adding the technical word information extracted by iterated dilated convolution neural network. Finally, the Transformer encoder with relative position coding is used to learn the deep semantic information of the text from the word vector sequence, realizing the prediction of the entity label. Experimental results on public and annotated patent datasets show that this method improves entity recognition accuracy.

Keywords: entity recognition named in Chinese; intellectual property; Transformer encoder; information fusion; vector representation; science and technology big data; patent; deep learning

随着科技的快速发展, 技术迭代速度也不断加快, 知识产权资源的数量呈现爆炸式增长。通过专利分析可以揭示技术的关系、技术发展的趋势等有价值的信息[1-3]。专利文献中包含了大量专业词汇, 具有专业背景的专利分析人员也需要投

收稿日期: 2022−03−21.　**网络出版日期:** 2022−11−17.
基金项目: 国家重点研发计划项目（2018YFB1402600）; 国家自然科学基金项目（61772083）.
通信作者: 杜军平. E-mail: junpingdu@126.com.

入大量的时间成本才能理解专利的内容。因此, 自动化地提取专利文本中的核心技术信息, 对帮助研究人员快速了解专利信息具有重要意义。

专利文献由专业的技术人员编写, 其中包含了大量专业词汇、技术术语, 具有语言精准、语义信息较为复杂、信息密度大等特点。词向量是自然语言处理的核心表征技术, 但传统的 word2vec 等词向量表示在一词多义等方面存在诸多局限性[4]。BERT 方法[5] 能捕捉整个句子中字序列和上下文关系等信息, 解决了一词多义问题, 具有较强的文本特征表示能力。目前实体识别常用的长短时记忆网络和条件随机场模型 (long shor-term memory and conditional random field, LSTM-CRF) 方法体系, 广泛应用于社交媒体[6] 等场景的实体提取, 能够更好利用文本的全局结构信息, 但是对词层面的信息不敏感, 不能充分利用专利文本中技术术语的语义信息。近年来 Transformer 方法[7] 在自然语言处理任务中得到了广泛应用并取得了很好的效果, Transformer 基于自注意力机制, 从而兼顾局部与上下文特征, 能避免文本语义特征的割裂, 具有出色的特征提取能力。

本文的贡献如下:

1) 本文提出基于 Transformer 与技术词信息的知识产权实体识别方法 (BWET), 通过引入膨胀卷积网络实现对专利文本中的技术词信息提取, 提高字向量对知识产权实体的表征能力。

2) 本文提出使用 Transformer 进行深层语义信息的提取, 减少语义信息损失, 并引入相对位置编码, 克服 Transformer 缺乏相对位置感知的问题, 提供出色的特征提取能力。

1　相关工作

知识产权实体是指专利文献中出现的包含技术信息的词, 主要包含反映科学技术、方法理论、所属领域的实体, 如自动驾驶、激光雷达等。知识产权实体可能由多个词语构成, 实体边界不易区分, 且专利文本中每个句子包含的实体数量较多, 这要求更准确更丰富的特征来描述知识产权实体。目前知识产权实体抽取主要分为两大类[8]: 一类是基于传统机器学习方法, 另一类是基于深度学习方法的方法。

基于传统机器学习的方法主要使用依存句法分析方法、领域词典等文本分析方法与传统统计机器学习方法[9] 相结合, 实现知识产权实体的抽取。Chen 等[10] 运用 Bootstrapping 算法并加入规则处理特殊情况, 减少语义漂移的影响, 实现低计算开销的专利中实体提取, 但是对长难句和由多个词组成复杂技术实体的抽取效果较差。

目前使用更多的方法是基于深度学习方法[11-12] 完成序列标注任务, 实现实体抽取, 在拥有较好性能的同时, 几乎不需要人工构造和选择特征[13]。目前应用最广泛的方法主要有字嵌入层、特征提取层和序列预测层三层结构, 文本序列先经过字嵌入层生成字向量, 再由特征提取层提取文本序列的语义特征, 最后由序列预测层生成序列标注的结果, 其中序列预测多使用条件随机场学习标签序列间的依存关系, 优化输出的标签序列, 针对方法的优化主要集中在字嵌入层与序列预测层。

针对专业领域实体, 多通过改进词嵌入层, 引入领域专业词特征提高识别准确率。Wang 等[14] 运用序列到序列方法提取通信领域专利文本的语义特征, Saad 等[15] 使用 BERT方法对生物领域语料生成动态字向量。在字向量中加入文本的词特征, 可以缓解数据稀疏和未登录词问题引起的字向量质量降低[8]。卷积神经网络 (convolution neural network, CNN)[16] 与中文分词词典[17] 是常用的词特征构建方式。但是知识产权实体长度跨度较大且可能由多个词构成, 以上方法可能导致词语义被割裂。Yan 等[18] 利用隐马尔可夫方法[19] 对分词词典进行动态更新, 提高了分词效果。目前特征提取层应用最广泛的方法结构是使用双向长短期记忆神经网络 (BiLSTM)[20] 提取文本的上下文特征。Jin 等[21] 提出通过增加注意力机制层, 进一步提取文本的隐藏语义关系, 但是无法并行加速训练速度慢、文本信息随递归会产生丢失的问题仍没有解决。近年来随着 Transformer 被广泛应用于自然语言处理的其他任务中并取得极佳的效果, 基于 Transformer 的改进方法[22] 也被越来越多地应用于命名实体提取中。

2　基于 Transformer 与技术词信息的实体抽取方法

本文提出基于 Transformer 与技术词信息的实体抽取方法, 其结构如图 1 所示, 该方法主要分为技术词信息融合层、知识产权语义抽取层和实体标签优化层。技术词信息融合层通过 BERT 方法, 根据输入专利文本生成包含上下文语义信息的动态字向量, 解决知识产权实体多义性问题。字向量经迭代膨胀卷积网络 (iterated dilated convolution neural network, IDCNN) 后生成的技术词信息与原始的字向量拼接, 得到融合技术词信息的字向量表示。Transformer 编码层由多个 Trans-

former 编码器构成，利用注意力机制提取字向量序列的深层语义特征。CRF 层负责学习标签之间的依赖关系，优化输出的标签序列。

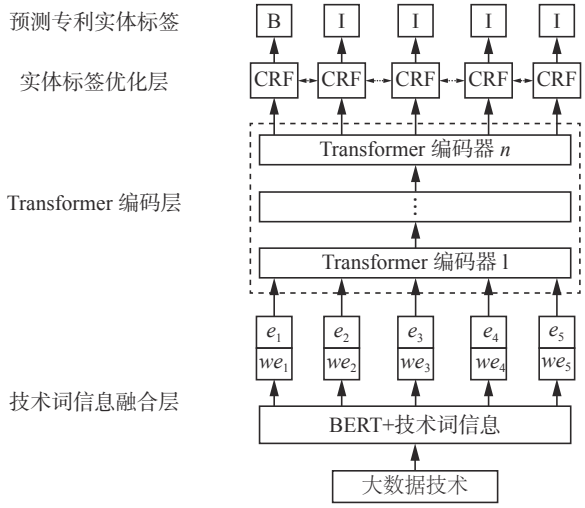

图 1　基于 Transformer 与技术词信息实体抽取方法

Fig. 1　Information entity extraction method based on Transformer and technical words

2.1　技术词信息融合层

技术词信息融合层的作用主要是将专利文本转化为语义向量表示，其结构如图 2 所示，主要由 BERT 方法和技术词词向量生成方法两部分组成，最终生成的字向量由 BERT 生成的初始字向量与词向量拼接而成。

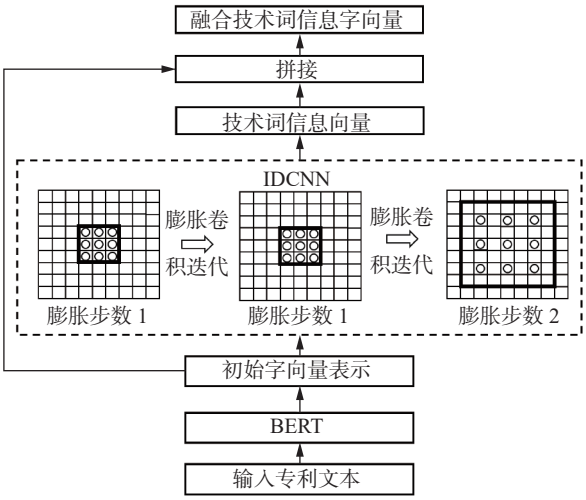

图 2　技术词信息融合层结构

Fig. 2　Structure of technical words information fusion layer

专利中的知识产权实体除了语义复杂、信息密度大外，其结构也比较复杂，可能由多个词组成，需要结合词信息才能很好地区分知识产权实体。BERT 虽然可以解决文字的多义性问题，但是无法提供词特征。传统的基于 CNN 或字典的词嵌入方式缺乏灵活性，面对长度与结构多变的知识产权实体反而容易造成词信息的割裂。

本文引入 IDCNN 来实现专利中技术词向量生成。膨胀卷积为滤波器设置了一个膨胀步长序列，它会忽略所有处于膨胀步长中的输入数据，同时保持卷积核的尺寸大小不变。随着卷积层数的增加，视野域指数扩散；如图 2 所示，步长为 1 时，膨胀卷积层的感受视野为加粗框中的 3×3 矩阵，步长为 2 时，使膨胀卷积的感受野由 3×3 的窗口扩展为 7×7 的窗口。生成词向量的膨胀卷积网络由 1 个膨胀卷积单元组成，膨胀卷积单元由膨胀步长为 1、1、2 的三层膨胀卷积层组成。前两层卷积网络使文本向量的每一个特征都会被膨胀卷积所提取，充分利用原始文本字向量的有效信息，最后一次卷积采用更大的膨胀步长，获得更广的输入矩阵数据。

2.2　知识产权语义抽取层

专利文本的语句较长，语义信息密度大，句子中包含的实体较多，对语义提取过程中的信息损失很敏感。因此，引入 Transformer 编码器完成语义信息的提取，其主要由多头注意力机制和前馈神经网络组成，通过残差连接将输入信息完整地传递给输出向量，有效地缓解方法中梯度消失的问题。但是由于自注意力机制没有卷积或递归网络结构，无法像 CNN 或 LSTM 网络在特征提取过程中依靠网络结构编码文本序列的位置信息，因此需要在输入的字向量中嵌入位置向量。直接在字向量中加入绝对位置编码，使用频率不同的正弦和余弦编码构建位置向量，具体计算过程为

$$E_{(p,2m)} = \sin\left(\frac{p}{10\,000^{2m/d}}\right)$$

$$E_{(p,2m+1)} = \cos\left(\frac{p}{10\,000^{2m/d}}\right)$$

式中：p表示字的位置；d表示位置向量的维度；$m \in [0, d/2-1]$表示向量的具体某一维度。但是这种编码方式在经过注意力机制后会丢失相对位置信息，仅能保留绝对位置信息[23]，而相对位置则是命名实体提取所需的关键特征。

根据自注意力机制与绝对位置嵌入编码方式，序列中第 i 个字与第 j 个字的注意力得分计算过程为

$$A_{i,j} = [(\boldsymbol{e}_i + \boldsymbol{p}_i)\boldsymbol{W}^q]\left[(\boldsymbol{e}_j + \boldsymbol{p}_j)\boldsymbol{W}^k\right]$$

$$A_{1,j} = \underbrace{\boldsymbol{e}_i\boldsymbol{W}^q\left(\boldsymbol{W}^k\right)^{\mathrm{T}}\boldsymbol{e}_j^{\mathrm{T}}}_{(a)} + \underbrace{\boldsymbol{p}_i\boldsymbol{W}^q\left(\boldsymbol{W}^k\right)^{\mathrm{T}}\boldsymbol{e}_j^{\mathrm{T}}}_{(b)} + \underbrace{\boldsymbol{e}_i\boldsymbol{W}^q\left(\boldsymbol{W}^k\right)^{\mathrm{T}}\boldsymbol{p}_j^{\mathrm{T}}}_{(c)} + \underbrace{\boldsymbol{p}_i\boldsymbol{W}^q\left(\boldsymbol{W}^k\right)^{\mathrm{T}}\boldsymbol{p}_j^{\mathrm{T}}}_{(d)}$$

式中：$\boldsymbol{e}$表示字向量；$\boldsymbol{p}$表示绝对位置嵌入向量；$\boldsymbol{W}^q$和$\boldsymbol{W}^k$分别是生成查询向量和键向量的权重矩阵。其中只有(d)部分包含了两个字的位置信息，但两

个权重矩阵的点积引起了相对位置的丢失[24]。

为解决此问题, 本文通过引入相对位置编码, 能够生成包含方向信息的特征, 其计算过程为

$$r_{i,j}=\left[\cdots \sin\left(\frac{i-j}{10\,000^{2m/d}}\right)\cos\left(\frac{i-j}{10\,000^{2m/d}}\right)\cdots\right]^{\mathrm{T}}$$

式中i和j表示输入字向量在文本中的位置。利用正弦函数$\sin(i-j)=-\sin(j-i)$的性质, 反映字符之间相对位置的变化, 利用余弦函数$\cos(i-j)=\cos(j-i)$的性质, 反映绝对位置关系, 从而在相对位置编码中同时保留相对位置和绝对位置信息。在引入相对位置后, 查询向量只需关注与键向量的相对位置即可, 键向量则不需要再计算查询向量的位置信息, 并且保留键向量与相对位置向量, 使注意得分能包含更多信息。修改后的注意力得分计算公式为

$$A_{i,j}=\boldsymbol{e}_i\boldsymbol{W}^q\left[\left(\boldsymbol{e}_j+\boldsymbol{r}_{i,j}\right)\boldsymbol{W}^k\right]+u\boldsymbol{e}_j\boldsymbol{W}^k+v\boldsymbol{r}_{i,j}$$

式中: $\boldsymbol{e}$表示字向量; $\boldsymbol{r}_{i,j}$表示相对位置嵌入向量; $\boldsymbol{W}^q$和$\boldsymbol{W}^k$分别是生成查询向量和键向量的权重矩阵; u和v为可学习参数。

2.3　实体标签优化层

专利文本经过技术词信息融合和 Transformer 编码器后, 可以根据文本的语义信息计算每个字属于各个类型标签的概率。但是由于没有考虑标签之间的依赖关系, 比如 I-X 标签必须出现在 B-X 标签之后, 导致生成无实体起始标记等无效的标签序列, 降低实体识别效果。

本文引入条件随机场学习标签之间的依赖关系, 为生成的标签之间的关系增加限制, 提高知识产权实体识别效果。对于长度为n的待预测的标签序列$y=\{y_1,y_2,\cdots,y_n\}$, 则通过式 (1) 表示的得分函数求得生成的序列标注为标签序列的概率, 为使其达到最大, 采用对数最大似然估计得到代价函数, 最后使用维特比算法求得最优标签序列。

$$\begin{aligned} s(x,y)&=\sum_{i=1}^{n}A_{y_{i-1},y_i}+\sum_{i=1}^{n}P_{i,y_i}\\ \log_2 P(y|x)&=s(x,y)-\log_2\sum_{y'}\exp\left(s(x,y')\right)\end{aligned}\tag{1}$$

式中: $s(x,y)$为标签序列y的得分; A_{y_{i-1},y_i}代表从第$i-1$个标签转移到第i个标签的概率; P_{i,y_i}为第i个词映射到第i个标签的概率; $P(y\mid x)$为生成序列标注为标签序列的概率。

BWET 方法的整体流程如下所示。

输入　专利文本序列

输出　知识产权实体标注序列

1) 对输入文本序列使用 BERT 生成字向量

2) 对字向量执行下列操作:

①使用 IDCNN 计算技术词信息

②将字向量与技术词信息拼接

③得到融合技术词信息的字向量

3) 对融合技术词信息的字向量使用 Transformer 提取深层语义信息

4) 根据语义信息生成每个字对应知识产权实体标签的概率

5) 使用 CRF 优化实体标签序列

6) 返回知识产权实体标注序列

3　实验结果与分析

3.1　评价指标与实验数据

实验使用准确率 P、召回率 R 以及 F_1 值作为基于 BERT 的融合特征文本分类方法对比实验的评价指标。

公开数据集选用 CLUENER 数据集[25], 训练集有 10748 条语句, 包含了地址、公司、电影名 10 种类型的实体。该数据集实体类别粒度较细, 且存在书名、电影名这种语义信息相近的实体, 比较接近知识产权实体识别的场景。专利实体识别数据集合共有 1972 条专利的标题及摘要, 通过手工标注了其中的实体, 按照 4:1 的比例划分为训练集和测试集, 其中实体类型的分布如表 1 所示。

表 1　知识产权实体类型及分布

Table 1　Types and distribution of intellectual property entities

编号	实体类型	实体含义	实体数量
1	DOM	所属领域	4925
2	TECH	技术术语	26027
3	USED	应用方向	4393
4	EFF	功效词	5594
5	INFO	数据源	6659
6	MAT	材料	3289

实验数据采用 BIO 标注: 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。比如, 将 X 表示为技术短语 (TECH), 则 BIO 的 3 个标记为: B-TECH 为技术短语的开头, I-TECH 为技术短语的中间部分, O 为不是技术短语的字符。

3.2　实验参数

实验中使用哈尔滨工业大学讯飞联合实验室训练的 BERT-wwm-ext, 共有 12 层 Transformer 结构, 12 个注意力头, 生成的字向量维度为 768 维,

并且方法不参与参数调节(不对 BERT 进行微调)。

IDCNN 的卷积核大小设置为 3, 卷积核数量设置为 128, 膨胀步长设置为(1、1、2), Transformer 的层数为 4 层, 有 16 个注意力头。单句最大长度为 128, batch_size 的大小设置为 32, epoch 的大小设置为 100, 学习率设置为 0.000 1, 训练时 dropout 设置为 0.5, 优化器选用 Adam。对比方法中, BiLSTM 的隐层神经元数量设置为 256, IDCNN 卷积核数量设置为 128, 学习率设置为 0.001 其余训练参数与 BWET 方法相同。

3.3　实验结果

实验 1　BWET 方法在公开数据集上的有效性验证

为了验证本文提出的方法的有效性, 对 BiLSTM-CRF、IDCNN-CRF 和加入相对位置的 Transformer-CRF 方法分别使用 word2Vec 和 BERT 生成字向量, 在公开数据集上进行了对比实验, 实验结果如表 2 所示。

表 2　各方法在 CLUENER 数据集上实体识别效果对比

Table 2　Comparison of entity recognition effects of various methods on CLUSTER dataset

方法	P	R	F_1
BiLSTM-CRF	0.727 5	0.686 7	0.706 5
IDCNN-CRF	0.643 9	0.677 6	0.660 3
Transformer-CRF	0.727 8	0.721 4	0.724 5
BERT	0.784 5	0.766 6	0.775 4
BERT-BILSTM-CRF	0.785 7	0.821 0	0.802 9
BERT-IDCNN-CRF	0.795 8	0.758 5	0.776 7
BERT-Transformer-CRF	0.786 3	0.835 0	0.809 9
BWET	**0.791 2**	**0.847 3**	**0.818 3**

从表 2 中可以看出, 提出的 BEWT 方法的 F_1 值达到了 0.818 3, 取得了最好的表现, 对实体识别性能有了巨大的提升。IDCNN 方法的表现相比其他方法有着较大差距, 说明在处理长文本时, IDCNN 网络较弱的上下文依赖特征获取能力对实体产生了很大制约。BERT-IDCNN 方法在准确率上高于其他方法, 表明词特征能够提高实体识别的准确率。BERT-BiLSTM 方法相比 BERT 召回率和 F_1 值有了 5% 和 3% 的提升, 表明 LSTM 方法能够很好学习到文本的上下文语义信息, 但是信息在传递过程中的损失影响了其表现。Transformer 方法通过注意力机制和残差连接显著减少了特征提取过程中的信息损失, 相比 BiLSTM 方法的召回率有了 4% 的提升, 在引入 BERT 后也有 1% 的提升。

实验 2　BWET 方法在专利数据集上的有效性验证

为了验证 BWET 方法针对专利数据具有更强的实体识别效果, 采用与实验 1 相同的对比方法, 在专利数据集上进行了对比实验, 实验结果如表 3 所示。

表 3　各方法在专利数据集上实体识别效果对比

Table 3　Comparison of entity recognition effects of various methods on patent dataset

方法	P	R	F_1
BiLSTM-CRF	0.591 4	0.516 6	0.551 5
IDCNN-CRF	0.460 7	0.437 8	0.448 9
Transformer-CRF	0.589 4	0.574 7	0.581 9
BERT	0.708 9	0.697 1	0.702 9
BERT-BILSTM-CRF	0.744 2	0.730 3	0.737 2
BERT-IDCNN-CRF	0.644 0	0.649 4	0.646 7
BERT-Transformer-CRF	0.769 2	0.788 4	0.778 7
BWET	**0.781 7**	**0.817 4**	**0.799 2**

从表 3 中可以看出, 在专利实体识别数据集上, 由于专利中技术实体, 因为所处的上下文不同而属于不同的实体类型, 如“人工智能”在“涉及人工智能领域”中表示所属领域, 而在“使用人工智能技术”中表示技术术语, 使得传统的静态字向量表现不佳。引入 BERT 后, 识别准确率、召回率和 F_1 值都有 15% 的提升, 很好地解决了多义性问题。由于专利文本中包含的实体数量和实体复杂程度都要高于一般文本, 因此在专利数据集上, BiLSTM 方法缺乏对局部特征的感知和信息损失的缺陷更为明显。BERT-Transformer 方法相比 BERT-BiLSTM 在专利数据集上, 识别的准确率提高了 2.5%, 召回率提高了 5%, 展现了比通用数据集更大的性能差异。本文提出的 BWET 方法, 由于在文本嵌入加入了词特征, 对技术术语带来的词级别语义变化更为敏感, 相比 BERT-Transformer 方法 F_1 值提高了 2%。

实验 3　BWET 方法训练时间对比

表 4 给出了各方法在公开数据集与专利数据集上平均每轮训练时间。BERT-Transformer 和 BWET 方法在 CLUE 数据集上训练时长相比 BERT-BiLSTM 方法更长, 但在专利数据集上则用时更少, 这是由于专利数据的文本长度更长, LSTM 网络不能并行计算的缺点更明显, 而 Transformer 多层编码器的结构, 在平均文本长度较短时, 反而需要更长时间。本文提出的 BWET 方法加入的技术词信息融合层并没有显著增加训练时间, 在

专利数据集上仅比 BERT-Transformer 方法增加约 1 s, 在 CLUE 数据集上仅增加约 2 s。

表 4　各方法平均训练每轮训练时间对比

Table 4　Comparison of average training time of each round of each method　　s

方法	CLUE数据集	专利数据集
BERT	73	42
BERT-BILSTM-CRF	76	55
BERT-IDCNN-CRF	77	43
BERT-Transformer-CRF	83	45
BWET	**85**	**46**

实验 4　BWET 方法 F_1 值随训练轮数变化

图 3 对比了 BWET、BERT-BiLSTM 和 BERT-Transformer 3 个方法在训练过程中 F_1 值的变化情况。从图 3 中可以看到, 在训练刚开始时, BERT-BiLSTMi 的 F_1 值提升速度远高于基于 Transformer 的方法, 这主要得益于 LSTM 拥有更少的参数。当训练到 10 轮左右时, Transformer 更加优异的特征提取能力得到了充分体现, F_1 值迅速超过 BERT-BiLSTM, 而 BWET 由于融合了技术词信息, F_1 的提升相比 BERT-Transformer 方法更早, 且提升幅度更高。在训练过程中期, BERT-BiLSTM 方法的 F_1 值经常出现比较大的波动, 其原因很可能是信息在递归传递过程中的损失, 而 BWET 方法则因为词信息与 Transformer 的共同作用, 能保留更丰富的语义信息, 从而减少了波动。

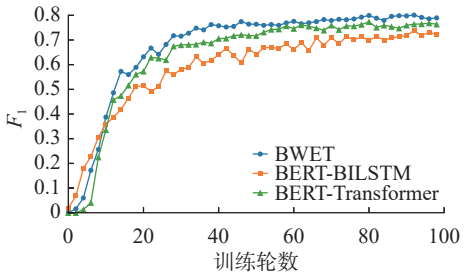

图 3　训练中方法 F_1 值变化趋势

Fig. 3　Change trend of method F_1 value in training

实验 5　BWET 方法不同实体的识别效果对比

本文对 BWET 针对各个类别实体的识别能力也做了探究, 图 4 展示了 BWET 方法对 6 种知识产权实体识别的准确率、召回率和 F_1 值。图 4 中可以看到, 所属领域和应用方向的实体识别效果显著好于其他类型的实体。这两类实体在专利文本中出现的位置与句子结构相对固定, 如所属领域主要在类似属于智慧交通领域、涉及汽车安全领域, 应用方向则常以用于全自动驾驶、基于深度学习的轨迹识别方法等形式出现, 对上下文的依赖相对较弱。技术实体虽然没有固定的句子结构体现, 所在句子的上下文也比较复杂, 但是依靠附加的词特征, 提升了技术实体语义信息的区分度, 获取了与结构相对稳定的功效词体相近的效果。材料类实体则受到训练数据不平衡的影响, 其识别效果较差。

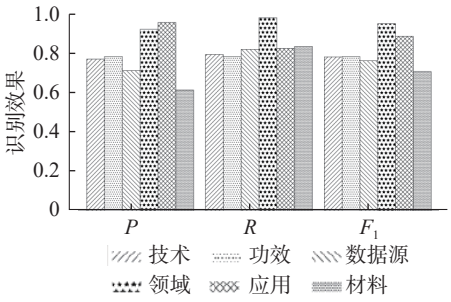

图 4　BWET 各类型实体识别表现

Fig. 4　Recognition performance of all types of entities in BWET

实验 6　技术词信息维度对 BWET 方法的影响

本文对技术词信息维度对实体识别效果的影响进行研究, 图 5 给出了 IDCNN 的卷积核数量分别为 64、128、256 和 512 时, BWET 方法在专利数据集上的识别效果。

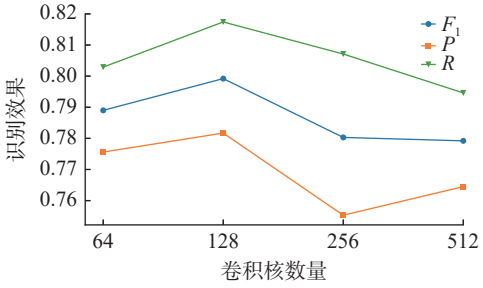

图 5　BWET 不同词信息维度实体识别表现

Fig. 5　Recognition performance of different word information dimension in BWET

从图 5 中可以看出, 卷积核数量为 64 和 128 时 BWET 方法的 F_1 值能够达到 0.79, 卷积核数量超过 128 时, F_1 值只有 0.78, 召回率出现了明显的下降。IDCNN 卷积核数量变化引起的识别表现差距达到了接近 2%, 表明方法对此参数比较敏感。当卷积核尺寸较小时, 无法充分挖掘文本的局部语义特征, 但是由于词信息密度较大, 仍能取得较好的识别效果。而当卷积核数量较大时, 虽然不会遗漏语义信息, 但是词向量中包含的语义信息更稀疏, 减弱了词向量在实体识别中的作用, 导致方法的识别效果降低。此外可以看出技术词信息质量下降时, 召回率产生了明显的下降而准确率则呈现出一定波动, 表明技术词信息的

主要作用是提高了方法的召回率。

4　结束语

针对专利文本中知识产权实体有较强的上下文依赖，且语义信息和实体词构成比较复杂，使得现有方法难以利用知识产权实体词层面语义信息的问题。

本文提出一种基于 Transformer 与技术词信息的知识产权实体识别方法。该方法利用 BERT 语言方法提供的动态字向量表示，将专利的上下文信息融入字向量中，解决实体的多义性问题。通过 IDCNN 提取技术词信息，并加入字向量中，加强对术语实体的感知。使用引入相对位置编码的 Transformer 编码器，使 Transformer 编码器能够感知文本序列相对位置和前后方向，提高语义信息提取能力。实验表明本文提出的方法在通用语料和专利数据集上能有效提高命名实体识别效果。

参考文献：

[1] 杨佳鑫, 杜军平, 邵蓥侠, 等. 面向知识产权的科技资源画像构建方法 [J]. 软件学报, 2022, 33(4): 1439–1450.
YANG Jiaxin, DU Junping, SHAO Yingxia, et al. Construction method of intellectual-property-oriented scientific and technological resources portrait[J]. Journal of software, 2022, 33(4): 1439–1450.

[2] WANG Yuhui, DU Junping, SHAO Yingxia, et al. A patent text classification method based on phrase-context fusion feature[C]//Proceedings of 2021 Chinese Intelligent Automation Conference. Singapore: Springer, 2022: 157−164.

[3] XU Mingying, DU Junping, XUE Zhe, et al. A scientific research topic trend prediction model based on multi-LSTM and graph convolutional network[J]. International journal of intelligent systems, 2022, 37(9): 6331–6353.

[4] KOWSARI K, JAFARI M K, MOJTABA H, et al. Text classification algorithms: A survey[J]. Information, 2019, 10(4): 150.

[5] DEVLIN J, CHANG MING-WEI, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018−10−11)[2022−05−23].https://arxiv.org/abs/1810.04805.

[6] KOU Feifei, DU Junping, HE Yijiang, et al. Social network search based on semantic analysis and learning[J]. CAAI transactions on intelligence technology, 2016, 1(4): 293–302.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017: 30.

[8] CHEN Hui, LIN Zijia, DING Guiguang, et al. GRN: gated relation network to enhance convolutional neural network for named entity recognition[J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(1): 6236–6243.

[9] SUN Bo, DU Junping, GAO Tian. Study on the improvement of K-nearest-neighbor algorithm[C]//2009 International Conference on Artificial Intelligence and Computational Intelligence. Shanghai: IEEE, 2009: 390−393.

[10] CHEN Tianci, LUO Mengfei, FU Hao, et al. Application of NER and association rules to traditional Chinese medicine patent mining[C]//2020 International Conferences on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data and IEEE Congress on Cybermatics. Rhodes: IEEE, 2020: 767−772.

[11] XUE Zhe, DU Junping, DU Dawei, et al. Deep low-rank subspace ensemble for multi-view clustering[J]. Information sciences, 2019, 482: 210–227.

[12] FANG Yuke, DENG Weihong, DU Junping, et al. Identity-aware CycleGAN for face photo-sketch synthesis and recognition[J]. Pattern recognition, 2020, 102: 107249.

[13] KRESTEL R, CHIKKAMATH R, HEWEL C, et al. A survey on deep learning for patent analysis[J]. World patent information, 2021, 65: 102035.

[14] WANG Yu, LI Yun, ZHU Ziye, et al. SC-NER: a sequence-to-sequence model with sentence classification for named entity recognition[M]//Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2019: 198−209.

[15] SAAD F, ARAS H, HACKL-SOMMER R. Improving named entity recognition for biomedical and patent data using Bi-LSTM deep neural network models[M]//Natural Language Processing and Information Systems. Cham: Springer International Publishing, 2020: 25−36.

[16] ZHAI Zenan, NGUYEN D Q, AKHONDI S A, et al. Improving chemical named entity recognition in patents with contextualized word embeddings[EB/OL]. (2019−07−05)[2022−05−23].https://arxiv.org/abs/1907.02679.

[17] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM[EB/OL]. (2018−05−05)[2022−05−23].https://arxiv.org/abs/1805.02023.

[18] YAN Xingyu, XIONG Xiaofan, CHENG Xiufeng, et al. HMM-BiMM: hidden Markov model-based word segmentation via improved bi-directional maximal matching algorithm[J]. Computers & electrical engineering, 2021, 94: 107354.

[19] ZHAO Hongke, LIU Qi, ZHU Hengshu, et al. A sequential approach to market state modeling and analysis in online P2P lending[J]. IEEE transactions on systems, man, and cybernetics:systems, 2018, 48(1): 21–33.

[20] ALZAIDY R, CARAGEA C, GILES C L. Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents[C]//WWW'19: The World Wide Web Conference. New York: ACM, 2019: 2551−2557.

[21] JIN Yanliang, XIE Jinfei, GUO Weisi, et al. LSTM-CRF neural network with gated self attention for Chinese NER[J]. IEEE access, 2019, 7: 136694–136703.

[22] LI Xiaonan, YAN Hang, QIU Xipeng, et al. FLAT: Chinese NER using flat-lattice transformer[EB/OL]. (2020−04−24)[2022−05−23].https://arxiv.org/abs/2004.11795.

[23] DAI Zihang, YANG Zhilin, YANG Yiming, et al. Transforme-xl: Attentive language models beyond a fixed-length context[EB/OL]. (2019−01−09)[2022−05−23].https://arxiv.org/abs/1901.02860.

[24] YAN Hang, DENG Bocao, LI Xiaonan, et al. TENER: adapting transformer encoder for named entity recognition[EB/OL]. (2019−11−10)[2022−05−23].https://arxiv.org/abs/1911.04474.

[25] YIN Xunwei, ZHENG Shuang, WANG Quanmin. Fine-grained Chinese named entity recognition based on RoBERTa-WWM-BiLSTM-CRF model[C]//2021 6th International Conference on Image, Vision and Computing. Qingdao: IEEE, 2021: 408−413.

作者简介：

王宇晖，硕士研究生，CCF 会员，主要研究方向为自然语言处理和数据挖掘。

杜军平，教授，CCF 会士，主要研究方向为人工智能、机器学习和模式识别。荣获吴文俊人工智能自然科学奖二等奖。

邵蓥侠，副教授，CCF 高级会员，主要研究方向为大规模图分析、并行计算框架和知识图谱分析。