计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):260−265 [doi: 10.15888/j.cnki.csa.007657]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BiLSTM_ATT_CNN 中文专利文本分类①

杜恒欣, 朱习军

(青岛科技大学 信息技术学院, 青岛 266061)
通讯作者: 杜恒欣, E-mail: 1220626623@qq.com

摘　要: 随着大数据和人工智能的发展, 将人工处理专利的方式转换为自动化处理成为可能. 本文结合卷积神经网络 (CNN) 提取局部特征和双向长短记忆神经网络 (BiLSTM) 序列化提取全局特征的优势, 在 BiLSTM 隐藏层引入注意力机制 (Attention 机制), 提出了针对中文专利文本数据的 BiLSTM_ATT_CNN 组合模型. 通过设计多组对比实验, 验证了 BiLSTM_ATT_CNN 组合模型提升了中文专利文本分类的准确率.

关键词: 专利文本; 卷积神经网络; 长短记忆神经网络; 注意力机制

引用格式: 杜恒欣,朱习军.基于 BiLSTM_ATT_CNN 中文专利文本分类.计算机系统应用,2020,29(11):260–265. http://www.c-s-a.org.cn/1003-3254/7657.html

Chinese Patent Text Classification Based on BiLSTM_ATT _CNN Model

DU Heng-Xin, ZHU Xi-Jun

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: With the development of big data and artificial intelligence, it is possible to transform the manual processing of patents into automated processing. In this study, combined with the advantages of Convolutional Neural Network (CNN) to extract local features and Two-way Long and Short Term Memory neural network (BiLSTM) to serialize and extract global features, the attention mechanism is introduced in the hidden layer of BiLSTM, and a BiLSTM_ATT_CNN combination model for Chinese patent text data is proposed. The BiLSTM_ATT_CNN combined model improves the accuracy of Chinese patent text classification by designing multiple comparison experiments.

Key words: patent text; Convolution Neural Network (CNN); long short memory neural network; attention mechanism

如今科技成为衡量一个国家实力水平的重要标准, 而专利文献中包含大量的创新发明, 企业可以通过利用专利中的创新发明, 减少研发周期. 为了有效利用专利文献中的关键信息, 易于管理专利, 对专利分类成为必不可少的一步. 目前, 我国使用国际专利分类 (IPC) 体系, IPC 分类体系共有 5 个级别和 7 万类, 是一个多层次多标签的分类系统[1]. 专利作为科技类的文献, 有极强的专业性, 但是在底层类别之间的相似度很高, 需要对该领域知识较为熟悉, 才能给专利赋予较为合理的分类号[2], 随着计算机技术和人工智能的发展, 人们开始关注如何使用计算机协助人工进行半自动化或自动化的专利文本分类, 本文采用深度学习的方法进行中文专利文本分类.

1　相关工作

传统的文本分类多采用机器学习算法, 如 K 近邻算法 (KNN)、朴素贝叶斯 (Naive Bayesian) 和支持向量机 (SVM) 等, 但这些机器学习算法需要人工提取文本特征, 而文本特征提取的质量和文本分类的准确度有密切关系[3], 另外人工提取特征费时费力, 还不能确

① 收稿时间: 2020-03-11; 修改时间: 2020-04-12, 2020-04-29; 采用时间: 2020-05-10; csa 在线出版时间: 2020-10-29

保所提取特征的准确性和全面性.

目前, 随着深度学习的发展, 专家学者们十分重视深度学习在文本分类领域上的应用, 主要方法有卷积神经网络 (CNN)、循环神经网络 (RNN) 和长短时记忆网络 (LSTM) 等. 其中 CNN 可以通过卷积和池化过程对文本进行局部特征提取, 由人工设置卷积核大小和个数, 实现权值共享. Kim[4] 采用预先训练好的词向量作为 CNN 模型的输入, 使用 CNN 的卷积和池化过程对句子信息进行 n-gram 特征抽取, 结果表明 CNN 对文本的特征提取效果良好, 得到了较优的文本分类结果. RNN 是一种序列模型, 能够提取上下文数据, 但是在长序列或复杂的文本中易出现梯度消失和梯度爆炸问题, 专家们在 RNN 的基础上提出改进算法, 如 LSTM. Rao 等[5] 通过使用 LSTM 捕捉上下文中的依赖关系, 获得较好的文本特征信息, 提高了分类准确率. 而 GRU 是基于 LSTM 的一种改进, 方炯焜等[6] 在使用 GloVe 词向量化的基础上, 利用 GRU 神经网络模型进行训练, 结果证明该算法对提高文本分类性能有较明显的作用. 另外为了更好的提取文本特征, 引入注意力机制 (Attention 机制), 张冲[7] 提出 Attention-Based LSTM 模型用于提取特征, 其中 LSTM 模型解决了传统 RNN 的梯度消失的问题, 同时通过 Attention-Based 减少特征向量提取过程中的信息丢失和信息冗余. 还有专家学者提出, 将多个学习模型结合起来, 利用模型优势互补原则, 提升分类性能, 例如 Lai 等[8] 通过结合 CNN 和 RNN 模型, 提出了 RCNN 混合模型, 获得了非常好的分类效果. 李洋等[9] 将 CNN 提取的文本局部特征和双向 LSTM 提取的文本全局特征进行特征融合, 解决了 CNN 忽略上下文语义信息, 又避免了 RNN 梯度消失和梯度爆炸问题.

在专利文本分类上, 马建红等[10] 在进行专利文本分类时, 从挖掘专利与效应对应关系的角度出发, 提出利用基于 Attention 机制的双向 LSTM 模型训练专利语料, 得到专类所属的效应. 余本功等[11] 提出一种双通道特征融合的专利文本自动分类, 将文本专利分别映射为 Word2Vec 词向量序列和 POS 词性序列, 分别使用这两种特征通道训练 WPOS-GRU 模型, 该方法节省了大量的人力成本, 并提高了专利文本分类的准确度. 胡杰等[12] 利用 CNN 进行专利文本特征提取, 结合随机森林作为分类器, 相对对于单一算法模型, 提高了专利文本的分类准确度. 通过研究, 本文针对计算机领域的中文专利文本, 提出将基于注意力机制的双向长短期记忆网络 (BiLSTM) 和卷积神经网络 (CNN) 组合的方法, 设计了 BiLSTM_ATT_CNN 模型, 并对该模型在专利文本数据进行训练学习. 实验结果表明, 该模型在一定程度上提高了中文专利文本分类的准确率.

2 方法

2.1 专利文本预处理

文本预处理过程主要工作是对数据集进行分词操作. 目前, 国内一些专家开发出效果较好的现代分词系统, 主要有结巴分词、盘古分词 NLPIR、语言云、搜狗分词、Boson NLP 分词系统、IKAnalyzer、中国科学院计算所汉语词法分析系统 ICTCLAS 等[13].

本文使用基于隐马尔科夫算法的结巴分词器进行分词, 目前结巴分词有 4 种模式, 精确模式、全模式、搜索引擎模式和 paddle 模式, 本文采用结巴分词的全模式. 在专利摘要文本中, 相比于其他文本数据, 语言较为领域化专业化, 而且专业术语设计的较多, 传统方法对专业术语词不能进行很好的覆盖, 当出现新的专业术语时, 需要重新计算特征向量. 再者专利中专业术语较多, 在调用已有的分词系统进行分词时, 专业术语词往往是会被分开的. 为了尽量避免这些情况, 本文将专利文本摘要中的关键字提取出来, 建立一个领域词典, 添加到已有的分词系统的词典中, 来减少分词不准确带来的误差. 分词后的词集中, 有很多像“我们”、“这样”、“之后”等这样没有实际意义的词, 这些词对分类没有贡献, 甚至影响文本分类的准确性, 在此利用停用词表来去除文本的停用词.

2.2 专利文本表征

词语是人类的抽象总结, 是符号形式的 (比如中文、英文、拉丁文等), 在文本处理时需要把词语转换成数值形式. 2013 年 Mikolov 等[14] 提出词向量概念, 使得自然语言处理方向有了 word embedding, 即词嵌入. Word2Vec 就是词嵌入的一种, 相较于传统 NLP 的高维、稀疏的表示法 (One-hot Representation), Word2Vec 利用了词的上下文信息, 语义信息更加丰富. Word2Vec 可以将初始特征词映射到实数低维向量, 避免了传统词袋模型词向量维度过大的问题, 并且在词向量生成过程中, 用向量空间里的向量运算来代替对文本内容的处理, 结合了词的上下文内容, 提供含有语义信息的词向量. 一般分为 CBOW (Continuous Bag-Of-Words)

与 Skip-gram 两种模型. CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量, 而输出就是这特定的一个词的词向量. Skip-gram 模型和 CBOW 的思路是相反的, 即输入是特定的一个词的词向量, 而输出是特定词对应的上下文词向量. CBOW 对小型数据库比较合适, 而 Skip-gram 在大型语料中表现更好. 本文利用 Google 开源推出的 Word2Vec 工具包, 选用 Skip-gram 模型, 使用 Python 语言实现了词向量化. 由于数据集是由专利文本的题目、摘要、主权项和分类号组成, 不能保证各样本数据的长度统一, 采用 padding 机制, 经实验验证, 将词向量长度设为 100, 样本数据长度固定为 400 时, 分类效果较好.

2.3 分类模型

2.3.1 BiLSTM_ATT 模型

循环神经网络 (RNN) 可以获取全局的特征信息, 但是存在梯度消失和梯度爆炸问题. 而长短期记忆神经网络 (LSTM) 可以避免这个问题, 作为 RNN 的一种改进模型, LSTM 不仅涵盖了 RNN 的优点, 还具有更强的记忆能力. LSTM 模型的基本神经元是由记忆单元和遗忘门f_t, 记忆门i_t和输出门o_t这 3 种门组成, 记忆单元是自连接单元, 能够记忆远距离上下文信息, 而这 3 种门共同决定如何更新当前记忆单元c_t和当前隐藏状态h_t. 遗忘门f_t, 可以看作是一个控制来自旧记忆细胞的信息会被丢弃到什么程度的函数; 记忆门i_t, 控制有多少新信息要存储在当前的存储单元中; 输出门o_t, 根据存储单元c_t控制要输出什么.

LSTM 转换函数定义如下:

输入数据为通过使用 Word2Vec 得到的词向量x_t, 维度为 K=100, 记忆门由隐藏状态h_{t-1}和输入x_t计算得到, 其中σ是逻辑函数 Sigmoid, 输出值在 0 到 1 之间, W_i是权重矩阵, b_i是偏置项, 计算公式如式 (1):

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{1}$$

遗忘门也是由隐藏状态h_{t-1}和输入x_t计算得到, 其中W_f是权重矩阵, b_f是偏置项, 计算公式如式 (2):

$$f_t = \sigma\left(W_f.[h_{t-1}, x_t] + b_f\right) \tag{2}$$

式 (3) 中, q_t为临时记忆状态, 记忆单元更新计算公式如式 (4), tanh表示输出为 [–1, 1] 的双曲切线函数, ⊙表示元素间的乘法:

$$q_t = \tanh(W_a \cdot [h_{t-1}, x_t] + b_a) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot q_t \tag{4}$$

输出门计算公式如式 (5), 其中W_o是权重, b_o是偏置项:

$$o_t = \sigma(W_O[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_{t-1} = o_t * \tanh(C_t) \tag{6}$$

传统的 LSTM 只在一个方向上读取样本数据, 从前往后或者从后往前, 为了更好提取文本语义, 需要考虑到文本词语的语境含义, 本文采用基于 LSTM 的改进的双向长短时记忆网络 (BiLSTM), 即同时从前往后和从后往前读取文本数据, 可以进一步增强语义对上下文的依赖程度.

在专利文本数据中, 有很多词是无足轻重的, 但是在前期的去停用词时没有将其全部去掉, 使得对分类影响大的词语融汇在大量词汇中, 减弱了其对分类的影响程度, 另外, BiLSTM 模型不能将专利文本中对分类重要的词语标记出来, 而且其隐藏层会损失一定的前文信息, 而注意力 (Attention) 机制能够很好的改善这个问题[15,16]. Attention 机制可以根据该词包含的语义信息和对文本分类的重要程度, 进行分配不同的权值, 进而减弱数据稀疏性, 提高文本分类的准确性. 本文在 BiLSTM 模型的隐藏层添加 Attention 机制, 对隐藏层输出的特征向量赋予不同注意力分配值, 把注意力集中到对分类任务较重要的词语上, 进一步提高专利文本分类的准确率. BiLSTM_ATT 结构如图 1 所示.

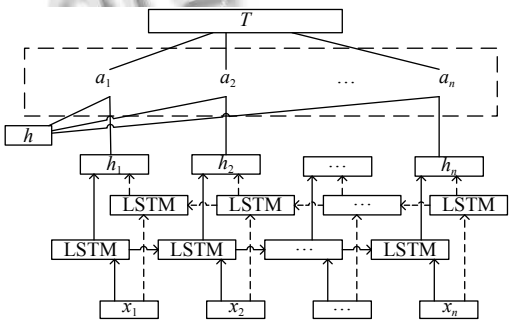

图 1 BiLSTM_ATT 模型结构图

由图 1 可知, BiLSTM_ATT 模型主要由输入层、前向 LSTM、后向 LSTM、注意力计算层组成. 中文专利文本经过预处理后, 将词语输入到词向量模型中, 本文使用 Word2Vec 工具, 输入层的中文专利文本数据表示为$X = \{x_1, x_2, \cdots, x_n\}$, 专利文本有 n 个词语, 分别作为输入数据, 进入前向 LSTM 和后向 LSTM, 得到

前向 LSTM 隐藏层的输出$\overrightarrow{h_i}$和后向 LSTM 隐藏层的输出$\overleftarrow{h_i}$, BiLSTM 模型隐藏层的输出为这两者之和, 即$H=\{h_1,h_2,\cdots,h_n\}$, 隐藏层的输出向量维度为 d, 在此 BiLSTM 的隐藏层上引入 Attention 机制, 注意力计算公式如式 (7), 得到注意力分配值分别为$a_1,a_2,\cdots,a_n$.

$$score(\bar{h},h_i)=w^{\mathrm{T}}\tanh(A\bar{h}+Bh_i+b_i) \tag{7}$$

$$a_i=\frac{\exp(score(\bar{h},h_i))}{\sum_j \exp(score(\bar{h},h_j))} \tag{8}$$

$$T=\sum_{i=0}^{n} a_i h_i \tag{9}$$

其中, h_i为第 i 个隐藏层的输出值, h为专利文本向量, 式 (7) 表示h_i所占专利文本向量的注意力权重, w, A, B 为权值矩阵, b_i为偏置项. 再将注意力权重通过 Softmax 函数进行概率化, 如式 (8), 得到注意力分布值. 最后如式 (9) 将隐藏层输出值和注意力分布值进行点乘、累加, 得到中文专利文本的特征向量矩阵 T.

2.3.2 CNN 模型

本文利用 BiLSTM_ ATT 模型可以获取文本上下文的全局特征矩阵, 但是无法体现专利文本局部特征, 故采用卷积神经网络 (CNN) 和 BiLSTM_ ATT 模型进行组合. 将 BiLSTM_ ATT 模型获得的文本信息特征向量和原始文本的词向量x进行首尾连接作为 CNN 的输入, 该输入包含原始的专利文本信息, 又包含经过 BiLSTM_ ATT 模型提取的全局特征. 利用 CNN 对其进行进一步的局部特征提取, 既解决了 BiLSTM_ ATT 模型无法获取专利文本局部特征的问题, 又避免了 CNN 无法提取专利文本上下文语义信息的问题.

具体计算过程如下:

卷积层的输入为$M\in R^{L\times d}$, M 表示由 BiLSTM_ ATT 模型获得的专利文本特征向量和原始文本的向量 x 连接得到的新专利文本向量, R 表示新专利文本向量的集合, d 是词向量维度大小为 K+K_1, L 是新专利文本向量的长度.

使用大小为$m\times d$的卷积核 w 对输入的新专利文本向量 M 进行卷积操作, 计算公式如式 (10).

$$c_i\in f(w\times M_{i,i+m+1}+b) \tag{10}$$

式中, $M_{i,i+m+1}$表示在 M 的第 i 行到第$i+m+1$行之间进行提取局部特征, f 是一个非线性转换函数, 又称激活函数, 本文采用 ReLU 作为激活函数, 表示卷积过程, b 表示偏置项.

将特征映射进行列连接, 得到对该样本的特征矩阵 W, c_i是由第 i 个卷积核生成的特征映射, 计算公式如式 (11):

$$W=[c_1,c_2,\cdots,c_{L-m+1}] \tag{11}$$

在卷积神经网络中一般会采用 max-pooling 或者动态的 k-max-pooling 等池化函数对卷积后的特征映射进行池化, 相当于降采样, 池化层可以减小数据空间大小, 在一定程度上控制过拟合. 本文采用 k-max-pooling, 从卷积层提取的文本特征中选取 k 个对分类最重要的文本特征, 并进行拼接, 计算公式为式 (12), 式 (13) 为全连接层.

$$H=\max(c_1,c_2,\cdots,c_{L-m+1})=\max\{C_1,C_2,\cdots,C_k\} \tag{12}$$

$$V=\{H_1,H_2,\cdots,H_L\} \tag{13}$$

2.3.3 BiLSTM_ATT_CNN 模型

本文设计了 BiLSTM_ATT_CNN 模型用来实现专利文本分类, 其中, BiLSTM_ATT 模型不仅可以考虑上下文信息, 在隐藏层添加的注意力机制, 还可以为词语分配不同的注意力分布概率值, 有效地防止信息丢失. 同时, 利用 CNN 模型能够捕获空间或时间结构的局部关系的优势, 进一步提取专利文本的局部特征. 该模型共分为 5 层, 输入层、BiLSTM_ATT层、CNN 层、全连接层、分类层. BiLSTM_ATT_CNN组合模型结构如图 2 所示.

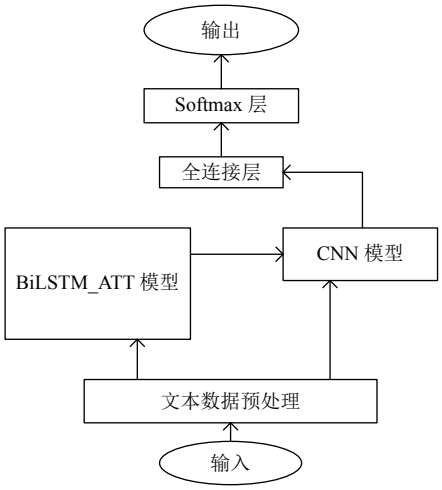

图 2 BiLSTM_ATT_CNN 模型结构图

第 1 层为输入层, 主要功能是将专利文本数据进行分词和词向量化, 首先利用结巴分词系统将专利文本数据进行分词, 并去除停用词, 然后利用 Word2Vec 进行词向量化. 专利文本格式: “专利题目”+“摘要”+“主权项”, “分类号”.

第 2 层为 BiLSTM_ATT 层, LSTM 具有学习序列数据的能力, 还避免了 RNN 模型的梯度爆炸和消失问题, 而 BiLSTM 由正向 LSTM 和反向 LSTM 组合而成, 可以更好的捕捉双向的语义依赖. 并在隐藏层加入 Attention 机制, 减弱数据稀疏性而造成重要信息的损失. 另外, 采用 L2 正则化方法防止过拟合.

第 3 层为连接层, 将 BiLSTM_ATT 层的输出和输入层的输出进行首尾连接, 作为 CNN 层的输入, 可以使得文本特征信息更为丰富.

第 4 层为 CNN 层, CNN 层可以从大量的训练数据中自动学习, 捕获空间、时间结构的局部特征, 将连接层的输出作为输入, 通过 CNN 的卷积层进行特征提取, 并在池化层采用 *k*-max-pooling 进行下采样, 并添加全连接层.

第 5 层为分类层, 采用 Softmax 函数实现专利文本分类, 为了防止模型过拟合, 在 Softmax 层采用 Dropout 策略和 L2 正则化策略, 最后将结果输出, 并保存于数据库中.

3 实验

3.1 实验数据

本文数据来源于 SOOPAT 的公开数据, 针对较为熟悉的计算机领域, 采用 Python 编程语言, 利用 requests、beautifulsoup 和 selenium 等库, 进行爬虫获取样本数据, 从中整理得到训练数据和测试数据. 同时考虑到数据的均衡性, 计算机领域专利主要集中在 G 部, 爬取的专利文本数据为 G06K3/02、G07B1/02、G08G1/01、G09B5/08 和 G11B5/012 这 5 类别, 各 2000 条专利文本数据. 其中 1800 条作为训练数据, 200 条作为测试数据, 训练数据和测试数据不重合.

专利文本数据包括专利题目、摘要、主权项、正文、主分类号等文本, 其中摘要和主权项中包含了专利的核心内容, 阅读者通过阅读专利摘要和主权项就可以对该专利的类别有所把握, 所以本文采用专利的题目、摘要、主权项和主分类号这 4 部分内容作为样本数据, 数据格式图 3 所示.

计算机 | 本发明涉及一种计算机。所述计算机包括显示器、耳机与收容筒体，所述显示器上设置有显示屏，所述显示器的侧壁上开设有耳机插孔，所述耳机包括耳塞与耳机线缆，所述耳塞连接于所述耳机线缆的端部，所述耳机线缆远离所述耳塞的一端插设于所述耳机插孔中，所述显示屏为矩形块状，所述收容筒体设置于所述显示器的侧壁上并邻近所述耳机插孔。所述计算机收容耳机较为方便。| 1.一种计算机，其特征在于，包括显示器、耳机与收容筒体，所述显示器上设置有显示屏，所述显示器的侧壁上开设有耳机插孔，所述耳机包括耳塞与耳机线缆，所述耳塞连接于所述耳机线缆的端部，所述耳机线缆远离所述耳塞的一端插设于所述耳机插孔中，所述显示屏为矩形块状，所述收容筒体设置于所述显示器的侧壁上并邻近所述耳机插孔，所述收容筒体的侧壁上设置有缠绕柱，所述缠绕柱用于缠绕所述耳机线缆。| H04R1/10|

图 3 专利文本数据格式

3.2 实验设计

本文采用准确率、召回率和 *F*1 测量值作为评估指标. *a* 表示正确分类的样本数, *b* 表示错误分类的样本数, *c* 表示属于该类别却被错误分类的样本数. 其中, 准确率 *p* 为正确分类的样本数和所有样本和之比. 准确率越高, 说明分类越准确, 如式 (14).

$$p = \frac{a}{a+b} \tag{14}$$

召回率 *r* 表示为样本中的正例被预测正确, 即正确分类的样本数与该类实际样本数的比值. 召回率越高, 说明在该类上预测时漏掉的样本越少, 如式 (15).

$$r = \frac{a}{a+c} \tag{15}$$

*F*1 值是将召回率和准确率综合考虑, 用于评价模型的总体性能, 计算公式如式 (16).

$$F1 = \frac{2pr}{p+r} \tag{16}$$

本文采用自主设计了一个文本分类系统, 基于 Druid 实现文本数据的分布式列存储, 通过 Zookeeper 对集群进行管理, 使用 Superset 对文本数据进行可视化展示. 首先通过系统从 Druid 中读取相关数据, 进行数据预处理, 在预处理阶段将专利文本数据进行分词, 去停用词, 使用 Word2Vec 方法进行词向量化. 之后调用本文提出的 BiLSTM_ATT_CNN 模型, 采用 mini-batch 的梯度下降方法进行训练学习, 避免了批量梯度下降收敛速度慢和随机梯度下降法容易收敛到局部最优解的现象. 实验中, 每次训练样本数为 64, 样本长度为 400, 能够使文本中包含的代表性的词语较丰富. 在 CNN 中, 设置卷积层使用的 4 种卷积核窗口尺寸分别为 3, 5, 7, 11, 个数均为 128. 将 BiLSTM_ATT_CNN 模型与经典 CNN、LSTM 方法相比, 通过训练测试后, 将结果分析上传保存到 Druid, 并通过 Superset 在界面展示.

为说明本文模型的分类效果, 本文使用 BiLSTM_ATT_CNN 模型的分类结果, 与经典的 CNN、LSTM 分类模型的分类结果进行对比.

3.3 实验结果和分析

经过多次实验, 结果如表 1 所示.

表 1 专利文本分类结果比较

模型	准确率	召回率	*F*1值
CNN	0.8812	0.8641	0.8734
LSTM	0.9351	0.9138	0.9203
BiLSTM_ATT_CNN	0.9742	0.9788	0.9786

由上表结果所示, 在相同的数据集上, 均使用结巴分词系统进行分词, 使用 Word2Vec 进行词向量化, 可以看出 BiLSTM_ATT_CNN 模型要优于传统的 CNN 和 LSTM 模型. 原因在于 BiLSTM_ATT 模型可以更好的捕捉双向的语义依赖, Attention 机制减弱数据稀疏性而造成重要信息的损失. CNN 模型可以弥补 BiLSTM_ATT 模型未能提取的局部特征信息的问题, 从而完善了专利文本特征提取, 使得分类效果有了一定的提升.

4 结束语

本文利用 BiLSTM_ATT_CNN 模型进行学习训练, 首先 BiLSTM_ATT 模型不仅考虑上文信息, 还兼顾下文信息, 另外, 在其隐层添加注意力机制, 能够计算历史节点对当前节点的影响力权重, 集注意力分配概率分布, 有效地防止信息丢失, 进一步优化了特征向量. CNN 模型能够捕获空间或时间结构的局部关系的能力, 让其在 BiLSTM_ ATT 模型获得的文本信息特征向量和原始文本的词向量X_i进行首尾连接的新文本特征向量上进行训练, 提升了 CNN 模型输入层数据的质量. 从实验结果可以看出, 与单一模型相比, 虽然将 BiLSTM 和 CNN 模型进行结合, 增加了训练时间增加, 提高了计算成本, 但是 BiLSTM_ATT_CNN 模型对专利文本分类效果要高于单一模型. 在下一阶段的研究中, 会集中考虑对新专利进行分类时, 可能会出现一些全新的词, 在分词时, 可能会因为分词词库中没有该词, 而进行了不恰当的分词, 导致有用信息丢失. 其次, 基于 IPC 专利分类体系为多级多标签分类, 考虑建立分层机制进行专利文本分类, 在部、大类、小类前 3 个级别采用同一个分类器, 在大组和小组级别上, 由于专利文本相似度的增加, 根据每个级别的类别数和专利文本数量设计不同的分类器进行训练学习.

参考文献

1 于红. 对《国际专利分类表》第七版一些变化的研究. 科技文献信息管理, 2001, (4): 22–27.

2 贾杉杉, 刘小安, 彭涛. 基于 IPC 的专利文本自动分类研究综述. 中国计算机用户协会网络应用分会 2017 年第二十一届网络新技术与应用年会论文集. 雄安, 中国. 2017. 26–28, 44.

3 刘红光, 马双刚, 刘桂锋. 基于机器学习的专利文本分类算法研究综述. 图书情报研究, 2016, 9(3): 79–86.

4 Kim Y. Convolutional neural networks for sentence classification. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1746–1751.

5 Rao A, Spasojevic N. Actionable and political text classification using word embeddings and LSTM. arXiv: 1607.02501, 2016.

6 方炯焜, 陈平华, 廖文雄. 结合 GloVe 和 GRU 的文本分类模型. 计算机工程与应用. https://www.cnki.net/KCMS/detail/11.2127.tp.20200331.1742.006.html. (2020-04-01)[2020-05-19].

7 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究 [硕士学位论文]. 南京: 南京大学, 2016.

8 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, TX, USA. 2015. 2267–2273.

9 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析. 计算机应用, 2018, 38(11): 3075–3080. [doi: 10.11772/j.issn.1001-9081.2018041289]

10 马建红, 王瑞杨, 姚爽, 等. 基于深度学习的专利分类方法. 计算机工程, 2108, 44(10): 209–214.

11 余本功, 张培行. 基于双通道特征融合的 WPOS-GRU 专利分类方法. 计算机应用研究, 2020, 37(3): 655–658.

12 胡杰, 李少波, 于丽娅, 等. 基于卷积神经网络与随机森林算法的专利文本分类模型. 科学技术与工程, 2018, 18(6): 268–272. [doi: 10.3969/j.issn.1671-1815.2018.06.042]

13 牛世雄. 中文专利的自动分类 [硕士学位论文]. 大连: 大连理工大学, 2017.

14 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings ofthe 1st International Conference on Learning Representations. Scottsdale, AZ, USA. 2013.

15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 5998–6008.

16 Doetsch P, Zeyer A, Ney H. Bidirectional decoder networks for attention-based end-to-end offline handwriting recognition. Proceedings of the 2016 15th International Conference on Frontiers in Handwriting Recognition. Shenzhen, China. 2016. 361–366.