智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向科技学术会议的命名实体识别研究

于润羽，杜军平，薛哲，徐欣，奚军庆

引用本文:
于润羽, 杜军平, 薛哲, 等. 面向科技学术会议的命名实体识别研究[J]. 智能系统学报, 2022, 17(1): 50–58.
YU Runyu, DU Junping, XUE Zhe, et al. Research on named entity recognition for scientific and technological conferences[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(1): 50–58.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107010

DOI: 10.11992/tis.202107010
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20211221.1704.002.html.

面向科技学术会议的命名实体识别研究

于润羽[1], 杜军平[1], 薛哲[1], 徐欣[1], 奚军庆[2]
(1. 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; 2. 司法部信息中心, 北京 100020)

摘 要: 针对通用领域的命名实体识别算法难以充分挖掘到科技学术会议论文数据中语义信息的问题, 提出一种结合关键词–字符长短期记忆网络和注意力机制的科技学术会议命名实体识别算法。首先对论文数据集中的关键词特征进行预训练, 获得词汇层面的潜在语义信息, 将其与字符级别的语义信息融合, 解决错误的词汇边界影响识别准确率的问题。然后, 将双向长短期记忆网络和注意力机制输出的向量进行融合, 同时考虑上下文和全局信息。最后利用条件随机场进行实体的识别。实验表明, 所提出的算法在不同数据集上都取得了较好的识别效果, 和对比算法相比, 准确率、召回率、F_1 指数均有一定程度的提升。

关键词: 命名实体识别; 长短期记忆网络; 注意力机制; 字词融合; 精准画像; 自然语言处理; 信息抽取; 预训练模型
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2022)01−0050−09

中文引用格式: 于润羽, 杜军平, 薛哲, 等. 面向科技学术会议的命名实体识别研究 [J]. 智能系统学报, 2022, 17(1): 50–58.
英文引用格式: YU Runyu, DU Junping, XUE Zhe, et al. Research on named entity recognition for scientific and technological conferences[J]. CAAI transactions on intelligent systems, 2022, 17(1): 50–58.

Research on named entity recognition for scientific and technological conferences

YU Runyu[1], DU Junping[1], XUE Zhe[1], XU Xin[1], XI Junqing[2]
(1. Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. Judicial Information Centre, Beijing 100020, China)

Abstract: Aiming at the problem that the named entity recognition algorithm in the general field cannot fully mine the semantic information in the scientific and technological academic conference paper data, a scientific and technological conference named entity recognition algorithm based on the combination of keyword-character long-short term memory (LSTM) and attention mechanism is proposed. First, pretraining of keyword features in the data set is conducted to obtain the latent semantic information at the vocabulary level, and merge it with the semantic information at the character level to solve the problem that the wrong vocabulary boundary affects recognition accuracy. Then, the bi-directional long-short term memory (BiLSTM) and the vector outputs of the attention mechanism are fused, and the contextual and global information is considered. Finally, conditional random field (CRF) is used to identify entities. Experimental results show that the proposed algorithm has achieved better recognition results on different data sets. Compared with the comparison algorithms, the accuracy, recall, and F1 index of the proposed algorithm have been improved.
Keywords: named entity recognition; long-short term memory network; attention mechanism; character-word fusion; accurate portrait; natural language processing; information extraction; pre-trained models

科技大数据[1-2]可以定义为与科研相关的活动产生的海量数据, 其以论文数据为主体, 具有数据规模大、内容专业化、特征属性繁多的特点。科技学术会议数据包含某个领域内的论文集合。以学术会议为单位进行画像的构建, 可以帮助科研人员快速获得有价值的科研信息, 而构建画像的核心工作即为命名实体识别。

命名实体识别是自然语言处理中知识抽取领

收稿日期: 2021−07−09. **网络出版日期**: 2021−12−21.
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(61772083,61802028); 广西科技重大专项(桂科 AA18118054).
通信作者: 杜军平. E-mail: junpingdu@126.com.

域中的重要研究方向, 其目的是将给定文本中的实体按照预定义好的类别进行分类[3-4]。学术会议论文数据的命名实体识别与通用领域的识别有一定区别, 主要原因在于通用领域的数据集的文本有较为严格的组成规范。但由于科研领域技术更新迭代快, 导致论文数据集中有大量的专业术语[5]。同时实体之间的关系也相对复杂, 增加了实体识别的难度。

中文命名实体识别的准确率和中文分词结果直接相关, 如果在分词阶段发生错误, 会严重影响识别效果[6]。目前在中文命名实体识别过程中, 大部分方法是基于字符模型编码, 这种方式在通用领域的命名识别中取得了较好的效果, 但它无法挖掘到一串字符信息中的词级别的信息。为了解决这个问题, 可以把字符级模型和词级别的模型相结合, 降低歧义发生的概率[7]。然而在学术论文数据中, 由于专业词汇较多, 采用这种方式很有可能产生错误的词语边界。因此本文引入论文关键词特征,提出关键词−字符编码方式, 在编码阶段同时考虑到关键词级别和字符级别的语义信息。此外, 在长短期记忆网络 (long-short term memory, LSTM) 和条件随机场 (conditional random field, CRF) 为主体框架的基础上, 在 LSTM 层引入自注意力机制 (self-attention mechanism, SA), 弥补长短期记忆网络无法考虑到全局信息的缺陷, 最后将 LSTM 和注意力机制输出的结果进行融合再通过 CRF 进行标注, 兼顾了字符之间的依赖关系, 在论文数据集中取得了更好的识别效果。

本文的主要贡献:

1) 提出了一种结合基于关键词−字符 LSTM 和注意力机制 (keyword-character long-short term memory and attention mechanism, KCLA) 的命名实体识别方法, 利用论文数据集进行训练并进行命名实体的识别;

2) 使用预训练模型对关键词特征进行训练, 获得对应的词向量, 在神经网络中将其与字符级别特征进行融合, 获取文本中潜在的语义信息;

3) 为科技学术会议论文数据中的实体进行定义, 在网络层同时使用长短期记忆网络和注意力机制, 充分考虑文本中的上下文信息以及全局信息, 优化实体识别的效果。

1　命名实体识别的研究现状

近年来命名实体识别的研究方法快速发展, 包括基于统计机器学习的方法和基于深度学习的方法。机器学习方法一般是通过标注好的文本进行训练, 利用训练好的模型进行识别[8]。常用模型有隐马尔可夫模型[9]、最大熵模型、决策树、支持向量机等。

基于深度学习的方法近年来发展迅速, 可以通过不同的神经网络完成该任务, 首先是卷积神经网络 (convolutional neural network, CNN), Yao 等[10]提出提出了一种基于 CNN 的适合医学文本内容的训练的命名实体识别方法, 无需构建词典同时保证较高的准确率。Strubell 等[11] 提出了迭代扩张卷积神经网络 (iterated dilated convolutional neural networks, IDCNN) 命名实体识别的方法, 与下文提到的目前最具有表现力的 LSTM 模型相比, 该模型只需要 $O(N)$ 的时间复杂度, 在保持与 LSTM 相当的精度的条件下, 可以实现 8 倍的速度提升。Yang 等[12] 分别采用字符级 CNN 和词级别 CNN 的的方式进行命名实体识别, 在字符级 CNN 中使用单层 CNN, 词级别采用多层 CNN, 最后利用 Softmax 或者 CRF 的方式实现实体的标注。Kong 等[13] 提出了一种完全基于 CNN 的模型, 充分利用 GPU 并行性来提高模型效率, 模型中构造多级 CNN 来捕获短期和长期上下文信息, 在保证较高识别准确率的情况下大幅提高了效率。

循环神经网络循环神经网络 (recurrent neural network, RNN) 也可以用于命名实体识别, RNN 的变体 LSTM 在命名实体识别方面取得了显著的成就。Huang 等[14] 融合双向长短期记忆网络和条件随机场 (BiLSTM-CRF) 应用于自然语言处理基准序列标记数据集。Zhang 等[15] 提出了针对中文 NER 的 Lattice LSTM 模型。与基于字符的方法相比, 显式地利用了词序列信息, 达到了最佳结果。Han 等[16] 针对专业领域内命名实体识别通常面临领域内标注数据缺乏的问题, 将生成对抗网络与长短期记忆网络模型相结合, 在各项指标上显著优于其他模型。

近年来, 基于深度学习的命名实体识别研究除了基于卷积神经网络和循环神经网络的方法外, 还出现了一些更新的技术。首先, Transformer 模型[17-18] 不再使用传统的神经网络思想, 使用到的只有注意力机制[19]。BERT 模型于 2018 年被提出, 在自然语言处理的各个领域都取得了令人瞩目的效果[20], 在命名实体识别领域, Dai 等[21] 在中文电子病历表识别的应用上使用了 BERT+BiLISM+CRF 的网络结构, 取得了很好的效果, Li 等[22] 使用了多层变种网络结构进行中文临床命名实体识别, 同样取得了很好的识别效果。文献 [23] 中利用预训练的 BERT 模型结合 BiLSTM, 提高了

在 Weibo 中文数据集上命名实体识别的准确率。Li 等[24] 针对现有的 Lattice LSTM 结构复杂的问题, 提出了 FLAT, 在性能和效率上均有提升。Yoon 等[25] 提出一个新型的命名实体识别 (named entity recognition, NER) 模型, 由多个双向 LSTM 网络构成, 每个网络作为一个单独的任务识别某一种制定的实体类型, 多个任务将各自学习到的知识进行转移, 获得更准确的预测。

2　KCLA 命名实体识别算法

在本节中, 主要介绍结合关键词–字符 LSTM 和注意力机制的科技学术会议论文命名实体识别算法。

2.1　算法整体结构

本文提出一种结合关键词–字符 LSTM 和注意力机制的科技学术会议论文命名实体识别算法。如图 1 所示, 模型的分为向量表示层 (Embedding)、融合双向长短期记忆网络和自注意力机制层 (BiLSTM-SA), 以及条件随机场层 (CRF)。具体而言, 向量表示层抽取了字符级别的特征以及关键词特征, 挖掘了数据中潜在的语义信息, 生成向量作为后续网络的输入。BiLSTM-SA 层通过神经网络提取局部和全局的文本的特征, 最后通过 CRF 层获得最大概率的命名实体分类。

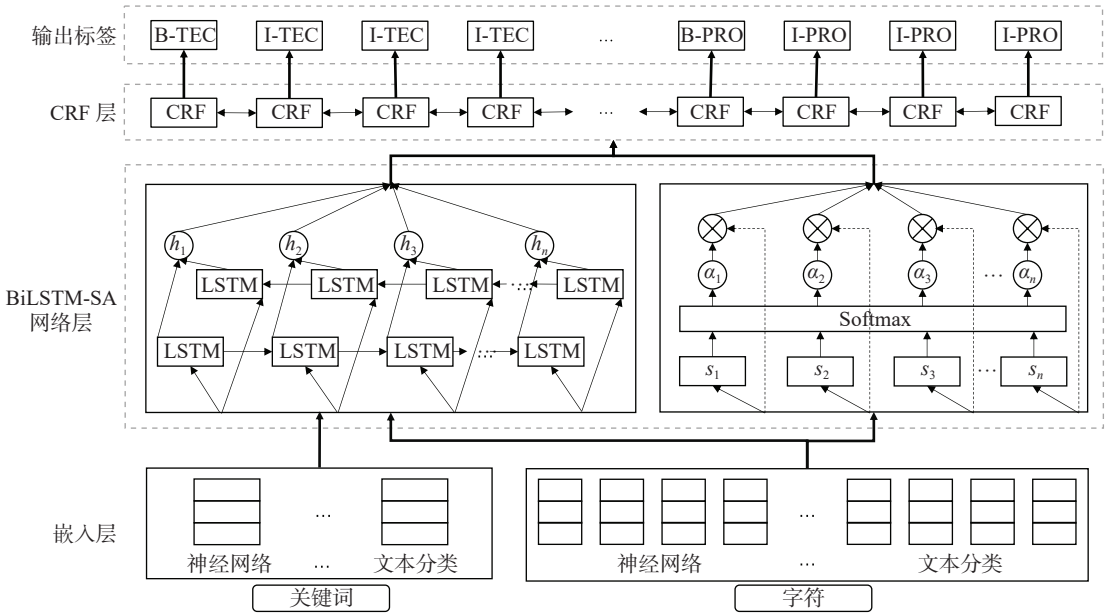

图 1　KCLA 算法整体框架

Fig. 1　Framework of KCLA algorithm

2.2　向量表示层

向量表示层主要将科技学术会议论文中的自然语言文本映射成后续层次能够识别计算的形式。向量表示层可以基于不同的模型实现, 下面分别介绍字符级别编码模型, 词级别编码模型以及本文提出的关键词–字符编码模型。

2.2.1　基于字符级别编码

基于字符级别编码模型是将按照每一个中文字符进行编码, 给定一个论文标题文本序列: 基于神经网络的文本分类, 可以将其表示为 $\boldsymbol{s}=[c_1\ c_2\ \cdots\ c_n]$, 其中 c_i 表示句子中的第 i 个字符, 每个字符经过式 (1) 的变换, 获得对应的输入向量。

$$\boldsymbol{x}_i^c=\boldsymbol{E}^c(c_i) \tag{1}$$

式中 $\boldsymbol{E}^c$ 代表字符级别的向量表示。最终的输入向量可以表示为 $\boldsymbol{x}^c=[x_1^c\ x_2^c\ \cdots\ x_n^c]$, 输入到 LSTM 网络中。

2.2.2　基于词级别编码

基于词级别编码模型是将按照词中文词汇进行编码, 同样给定文本序列: 基于神经网络的文本分类, 按照常规的中文分词方式对其进行切分, 然后按照词级别进行编码, 可以将其表示为 $\boldsymbol{s}=[w_1\ w_2\ \cdots\ w_n]$, 通过式 (2) 的变换, 获得对应的输入向量。

$$\boldsymbol{x}_i^w=\boldsymbol{E}^w(w_i) \tag{2}$$

式中 $\boldsymbol{E}^w$ 代表词级别的向量表示。最终的输入向量可以表示为 $\boldsymbol{x}^w=[x_1^w\ x_2^w\ \cdots\ x_n^w]$, 输入到 LSTM 网络中。

2.2.3　关键词–字符编码模型

关键词–字符编码模型主要考虑到了科技学术会议中论文数据本身的特点。由于论文数据专业性强, 因此常规的分词方式并不适用于论文数据集, 如果采用基本的字词融合, 可能会产生很

多错误的边界, 影响识别准确率。考虑到论文数据集中有关键词这一特征, 例如对于文本序列: 基于神经网络的文本分类模型, 在关键词字段中包含了神经网络、文本分类等词汇, 如果不考虑关键词信息, 该句会被切分为

$$s = 基于|神经|网络|的|文本|分类$$

对于本文想要识别的实体, 显然产生了错误的词汇边界, 因此要引入关键词特征, 构建词典, 对于例子中的文本序列, 需要将其正确切分为

$$s = 基于|神经网络|的|文本分类$$

获得了正确的词汇边界后, 在上述的文本序列中, 字符层面依然通过$\boldsymbol{x}_i^c = \boldsymbol{E}^c(c_i)$对输入的字符进行变换, 获得对应的向量。除此之外要考虑关键词层面的信息, 这里运用$w_{b,e}^k$来表示一个关键词信息, 例如: $w_{1,4}^k$表示关键词“神经网络”, $w_{6,9}^k$表示关键词“文本分类”, 通过式$\boldsymbol{x}_{b,e}^w = \boldsymbol{E}^w(w_{b,e})$进行变换。在实现的过程中, 首先利用 Word2Vec 对文本中的关键词进行预训练, 获得关键词的词向量模型, 在模型中提取出词向量矩阵, 然后和字符级别的向量共同输入到 LSTM 网络层中, 在 LSTM 中对二者进行融合, 整体结构如图 2 所示。

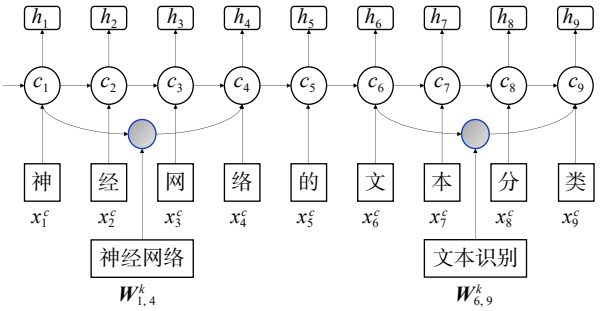

图 2　Keyword-Character 编码结构

Fig. 2　Structure of Keyword-Character

2.3　BiLSTM-SA 层

LSTM 是一种特殊的 RNN, 与传统的 RNN 相比, LSTM 同样是基于x_t和h_{t-1}来计算h_t, 但加入了输入门i_t、遗忘门f_t以及输出门o_t 3 个门和 1 个内部记忆单元c_t。

第 t 层的更新计算公式为

$$\begin{cases} i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \\ c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \\ h_t = o_t \otimes \tanh(c_t) \end{cases} \tag{3}$$

LSTM 模型按照文本序列的输入处理上文的信息, 而下文的信息对于科技学术会议论文数据的处理也有重要意义, 因此本模型采用 BiLSTM, 它由两层 LSTM 组成, 向量表示层得到的向量按照正序作为正向 LSTM 的输入, 即可以得到输出序列:

$$\boldsymbol{h}_L = [h_{L1}\ h_{L2}\ \cdots\ h_{Ln}]$$

再通过反向输入的方式, 得到逆向 LSTM 输出序列:

$$\boldsymbol{h}_R = [h_{R1}\ h_{R2}\ \cdots\ h_{Rn}]$$

将两层的输出进行融合, 得到包含上下文的特征$\boldsymbol{h}_n = [h_L\ h_R]$。

在本文提出的关键词–字符编码模型中, LSTM 的输入需要包含字符级关键词级信息。在 2.2.3 节中, 我们获得了字符级向量$\boldsymbol{x}_i^c$以及关键词级向量$\boldsymbol{x}_{b,e}^w$, 对于关键词级向量$\boldsymbol{x}_{b,e}^w$, 同样通过式 (3) 进行变换获得 LSTM 的单元$c_{b,e}^w$。但不需要输出门, 因为最终的预测是以字符为单位, 因此在词级别不需要进行输出。

对于字符级别的输入$\boldsymbol{x}_i^c$, 可以直接利用式(3)来获得输出。图 2 中对于单元c_4的输入包含了两个信息, 首先是包括x_4^c(络) 本身的信息, 此外需要输入$w_{1,4}^k$(神经网络)的信息, 对这二者, 利用式(4)进行链接。

$$c_j^c = \sum_b \alpha_j^c \otimes \tilde{c_j^c} + \alpha_{b,j}^w \otimes c_{b,j}^w \tag{4}$$

式中α_j^c和$\alpha_{b,j}^w$为归一化系数。

BiLSTM 在可以考虑到上下文的信息, 但对于全局信息无法充分的表达, 因此本模型将自注意力机制作为 BiLSTM 模块的补充, 提高命名实体识别的准确率。

Attention 的计算如式(5)所示。$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$三个矩阵均来自同一输入, 首先计算 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 之间的点乘, 然后除以一个尺度标度d_k, 然后将其结果归一化, 再乘以矩阵 $\boldsymbol{V}$ 就得到权重求和的表示。由于 Attention 本身就考虑到了全局的输入, 因此直接利用字符级别编码进行输入。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

获得了 BiLSTM 和 Attention 的输出之后, LSTM 的输出为$\boldsymbol{h} = [h_1\ h_2\ \cdots\ h_n]$。

Attention 层的输入为字符编码的向量, 输出通过式 (5) 的计算后, 输出为$\boldsymbol{a} = [a_1\ a_2\ \cdots\ a_n]$, 然后对这两个输出进行融合操作, 假设 BiLSTM-SA 层的输出为y_n, 在进行融合操作时采用归一化求和的形式, 即

$$y_i = \alpha_i^h \otimes h_i + \alpha_i^a \otimes a_i \tag{6}$$

其中α_i^h和α_i^a为归一化系数, 计算如式(7)所示:

$$\begin{cases} \alpha_i^h = \dfrac{\mathrm{e}^{h_i}}{\mathrm{e}^{h_i} + \mathrm{e}^{a_i}} \\ \alpha_i^a = \dfrac{\mathrm{e}^{a_i}}{\mathrm{e}^{h_i} + \mathrm{e}^{a_i}} \end{cases} \tag{7}$$

得到y_n后将其输入到 CRF 层中，获得命名实体识别最大概率的分类。

2.4　CRF 层

在预测当前标签时，CRF 通常可以产生更高的标记精度。由于论文数据相邻字符之间有较强的依赖关系，因此，在模型的最后一层，利用 CRF 来对前序层中得到的融合特征信息进行解码。

我们获得 LSTM-SA 层的序列输出为 $\boldsymbol{y}=[y_1\ y_2\ \cdots\ y_n]$，CRF 的标记过程为

$$o_i = W_s h_i + b_s \tag{8}$$

$$S(x,y) = \sum_{i-1}^{N}(O_{i,y_i} + T_{y_{i-1},y_i}) \tag{9}$$

式中：O_{i,y_i}表示第i个单词标记为y_i个标签的概率；$T_{i,j}$表示由标签转移到标签的概率。CRF 在语句 S 中标记序列的概率为

$$p(y|S) = \frac{\mathrm{e}^{S(x,y)}}{\sum_{y\in Y_x} S(x,y)} \tag{10}$$

最终的解码阶段通过 CRF 中的标准 Viterbi 算法，预测出最优的命名实体识别序列。

3　实验结果

本节进行实验并对结果进行分析。首先介绍算法的评价指标和实验参数，然后描述了在该评价指标和参数下 KCLA 算法的实验结果，并和其他网络结构进行了对比。

3.1　评价指标

本实验使用准确率 (precision)、召回率 (recall) 以及 F_1 值作为科技学术会议论文命名实体识别对比实验的评价指标。

准确率 P、召回率 R、F_1 值的公式分别为

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$F_1 = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$$

式中：TP 表示实际为真且预测为真的个数；FP 表示实际为假但预测为真的个数；FN 为实际为真但预测为假的个数。

3.2　实验采用数据集

本实验中，利用 scrapy 爬虫框架，对知网上的论文数据，按照不同的领域进行了爬取，利用按照领域爬取的数据进行训练及测试。实验中，获取了信息科学和机械工业分类下的论文数据各 20 000 条，按照 8:2 的比例构建训练集和测试集，将命名实体定义为研究技术 (TEC)、研究问题 (PRO)、研究形式 (MOD) 3 类实体，然后对数据集中的数据进行标注，数据集标注后数据分布情况如表 1、2 所示。

表 1　信息科技领域数据集分布情况

Table 1　Distribution of data sets in the field of information technology

数据集	训练集	测试集
论文数量	16 000	4 000
研究技术实体数量	9 389	1 363
研究问题实体数量	15 469	3 829
研究形式实体数量	9 810	2 633

表 2　机械工业领域数据集分布情况

Table 2　Distribution of data sets in the field of machinery industry

数据集	训练集	测试集
论文数量	16 000	4 000
研究技术实体数量	9 291	2 240
研究问题实体数量	15 399	3 855
研究形式实体数量	9 391	2 351

3.3　实验结果

在本文实验中，KCLA 算法的关键词的特征向量维度设置为 50，LSTM 的隐藏层维度为 128，batch size 设置为 32，学习率设置为 0.001，dropout 为 0.5，优化器使用 Adam。

本节使用 IDCNN、IDCNN+CRF、BiLSTM、BiLSTM+CRF 以及 Lattice-LSTM 这几种算法进行对比实验，实验结果在信息科学数据集下如表 3 所示，在机械工业数据集下如表 4 所示。

表 3　信息科学分类论文数据不同算法的对比实验

Table 3　Comparative experiment of different algorithms in information science papers

算法	准确率	召回率	F_1值
IDCNN	0.712 6	0.749 1	0.730 3
IDCNN+CRF	0.764 6	0.762 9	0.763 7
BiLSTM	0.721 3	0.809 3	0.762 8
BiLSTM+CRF	0.786 1	0.808 2	0.796 9
Lattice LSTM	0.785 8	0.809 2	0.797 3
KCLA	0.808 4	0.815 6	0.811 9

表 4　机械工业分类论文数据不同算法的对比实验

Table 4　Comparative experiment of different algorithms in mechanical industry papers

算法	准确率	召回率	F_1值
IDCNN	0.718 6	0.753 5	0.735 6
IDCNN+CRF	0.752 7	0.773 2	0.762 8
BiLSTM	0.715 9	0.784 7	0.748 7
BiLSTM+CRF	0.763 6	0.783 6	0.773 4
Lattice LSTM	0.737 8	0.786 2	0.761 2
KCLA	0.791 1	0.798 6	0.794 8

根据表 3 可以看出, 在信息科学数据集中, 本文提出的 KCLA 算法在性能方面要优于对比算法。首先, IDCNN 是 CNN 卷积神经网络的改进, 它通过引入空洞卷积的概念, 共享参数防止过拟合。IDCNN+CRF 则在 IDCNN 的基础上加入 CRF, 通过 Viterbi 算法, 预测出全局最优的标注序列。CNN 的优点在于时间复杂度相对较低, 但准确率不及以 BiLSTM 为主体的算法。对比算法中, BiLSTM+CRF 同时考虑到上下文的信息和字符之间的关联, 因此效果相对较好, 但其并没有挖掘到潜在词级别的语义信息, KCLA 算法通过融合关键词的特征, 对关键词信息进行预训练获得对应的词向量, 获取到了文本中潜在的语义信息、准确率、召回率、F_1 值均有一定程度的提高。

根据表 4 可以看出, 在机械工业数据集的对比算法中, KCLA 算法也取得了最好的识别效果。以 LSTM 为主体框架的算法的效果仍然优于 IDCNN 算法, 其中 Lattice LSTM 对比 BiLSTM+CRF 没有明显的提升, 主要原因在于其利用通用领域的词向量, 可能会产生错误的词汇边界, 因此影响了识别效果。

图 3 描述在信息科学数据训练过程中 loss 的变化趋势, 图 4 给出在第一个 epoch 中 loss 随 batch 的变化趋势。

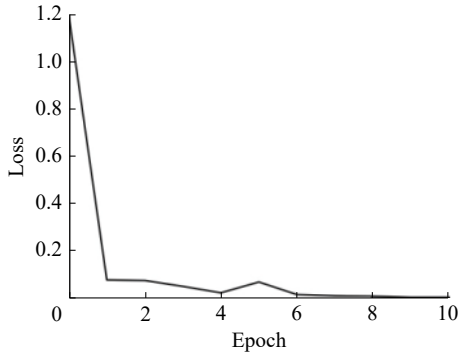

图 3　loss 随 epoch 的变化趋势

Fig. 3　Trend of loss with epoch

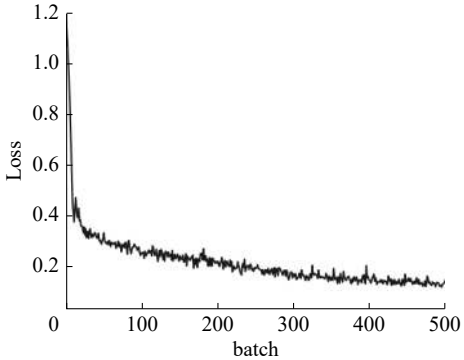

图 4　loss 随 batch 的变化趋势

Fig. 4　Trend of loss with batch

根据图 3 可以看出, epoch 到达 10 时基本收敛。本文实验中将训练 epoch 参数设置为 20, 但设置了提前终止条件: 如果两个周期内验证集准确率没有提升, 则提前停止训练。在实验中训练到第 10 个 epoch 时, 提前停止。

根据图 4 可以看出, loss 在第一个 epoch 中快速下降。在机械工业数据集中的 loss 变化与信息科学领域趋势相同。

3.4　网络参数对于模型性能的影响

3.4.1　LSTM 隐藏层参数对识别效果的影响

将 LSTM 的隐藏层维度设置不同数值进行实验, 确定其对论文数据命名实体识别效果的影响, 图 5 和图 6 分别给出隐藏层维度对信息科学和机械工业数据识别效果的影响。

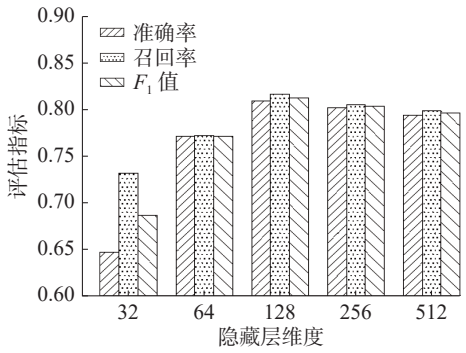

图 5　隐藏层维度对信息科学数据识别效果的影响

Fig. 5　Influence of hidden dimension in the information science data

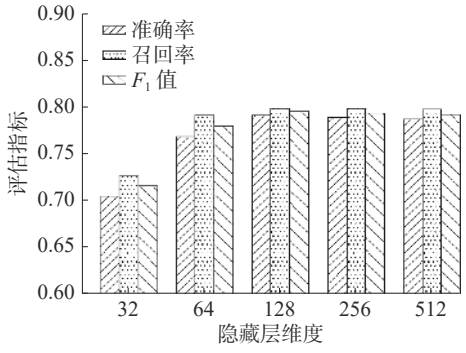

图 6　隐藏层维度对机械工业数据识别效果的影响

Fig. 6　Influence of hidden dimension in the machinery industry data

根据图 5 可以看出, 在信息科学数据集中, 隐藏层维度分别设置为 32、64、128、256、512。识别的各项评价指标开始随着隐藏层维度的增大而升高, 128 维时获得最好的识别效果, 对比 32 维的识别效果, 128 维的识别准确率、召回率、F_1 值分别提升了约 16%、8%、12%, 可见隐藏层维度是影响命名实体识别效果的重要参数。但随着维度的

继续增加，识别的效果并没有提升，甚至有轻微幅度的下降。

根据图 6 可以看出，在机械工业数据集中，隐藏层维度在 128 维和 256 维时都获得了很好的识别效果。对比 32 维时，128 维的准确率、召回率、F_1 值分别提高了约 9%、7%、8%。1 到达 512 维时有很微小的下降，结合图 5、6 可以得出结论：当隐藏层维度较低时，KCLA 不足以充分的学习到文本中的特征，影响了识别的效果。但如果维度设置的过高，可能导致过拟合现象，导致识别效果下降。

3.4.2　batch size 参数对识别效果的影响

将 batch size 设置不同数值进行实验，确定其对论文数据命名实体识别效果的影响，图 7、8 分别给出 batch size 对信息科学和机械工业数据识别效果的影响。

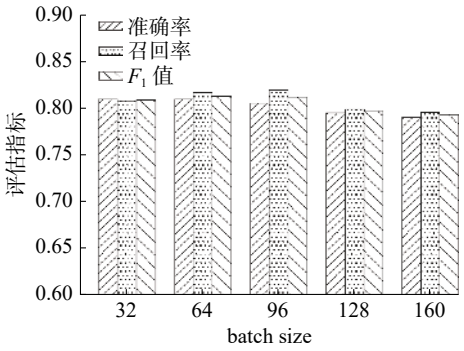

图 7　batch size 对信息科学数据识别效果的影响

Fig. 7　Influence of batch size in the information science data

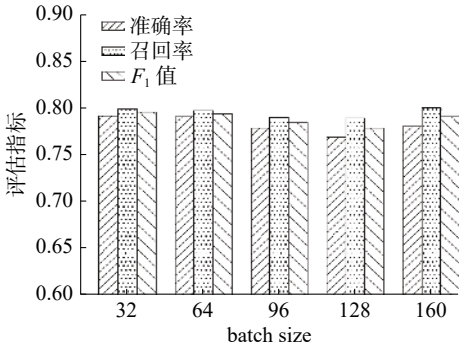

图 8　batch size 对机械工业数据识别效果的影响

Fig. 8　Influence of batch size in the machinery industry data

根据图 7 可以看出，在信息科学数据集下，从 F_1 值来看，batch size 为 64 时，识别效果最好，但和其他 size 相比，效果波动幅度很小，并没有明显的差异。

根据图 8 可以看出，在机械工业数据集中，F_1 值在 32 时获得了最好的识别效果，和在信息科学数据集中一样，在 batch size 从 32 增加到 160 的整个过程中，只有小幅度的变化。同时，准确率、召回率、F_1 值会有一定的波动，并没有在某一个 size 下共同取得最好的效果。结合图 7、8 可以得出结论，参数 batch size 对于 KCLA 算法影响较小。

4　结束语

本文针对科技学术会议论文数据，提出了结合关键词–字符 LSTM 和注意力机制的命名实体识别算法（KCLA），对学术会议中包含的论文信息进行实体定义，对数据集按照实体定义进行标注，然后利用 KCLA 算法对实体进行识别。其中 KCLA 算法模型由向量表示层、BiLSTM-SA 层和 CRF 层构成。实验数据表明，KCLA 算法可以对科技学术会议中论文数据的命名实体进行有效的识别。通过对比实验，将 KCLA 与 IDCNN, BiLSTM 等算法进行比较，KCLA 算法在科技学术会议论文数据集中有更好的表现。基于识别出的命名实体，结合论文数据中结构化的数据中获取到的关联关系，可以对学术会议数据构建知识图谱和精准画像，更加直观形象地展示出科技学术会议中潜在的语义信息，为科研人员进行科研信息的获取以及进行科研决策提供良好的数据支撑。

参考文献：

[1]　苏晓娟, 张英杰, 白晨, 等. 科技大数据背景下的中英双语语料库的构建及其特点研究 [J]. 中国科技资源导刊, 2019, 51(6): 87–92.
SU Xiaojuan, ZHANG Yingjie, BAI Chen, et al. Research of bilingual corpus construction and its characteristics in big data[J]. China science & technology resources review, 2019, 51(6): 87–92.

[2]　胡吉颖, 谢靖, 钱力, 等. 基于知识图谱的科技大数据知识发现平台建设 [J]. 数据分析与知识发现, 2019, 3(1): 55–62.
HU Jiying, XIE Jing, QIAN Li, et al. Constructing big data platform for sci-tech knowledge discovery with knowledge graph[J]. Data analysis and knowledge discovery, 2019, 3(1): 55–62.

[3]　何玉洁, 杜方, 史英杰, 等. 基于深度学习的命名实体识别研究综述 [J]. 计算机工程与应用, 2021, 57(11): 21–36.
HE Yujie, DU Fang, SHI Yingjie, et al. Survey of named

entity recognition based on deep learning[J]. Computer engineering and applications, 2021, 57(11): 21−36.

[4] 焦凯楠, 李欣, 朱容辰. 中文领域命名实体识别综述 [J]. 计算机工程与应用, 2021, 57(16): 1–15.
JIAO Kainan, LI Xin, ZHU Rongchen. Overview of Chinese domain named entity recognition[J]. Computer engineering and applications, 2021, 57(16): 1–15.

[5] 周园春, 王卫军, 乔子越, 等. 科技大数据知识图谱构建方法及应用研究综述 [J]. 中国科学:信息科学, 2020, 50(7): 957–987.
ZHOU Yuanchun, WANG Weijun, QIAO Ziyue, et al. A survey on the construction methods and applications of sci-tech big data knowledge graph[J]. Scientia sinica (informationis), 2020, 50(7): 957–987.

[6] LIU Wei, XU Tongge, XU Qinghua, et al. An encoding strategy based word-character LSTM for Chinese NER[C]// 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: ACL, 2019: 2379−2389.

[7] 张江英, 郝矿荣, 王直杰, 等. 基于 lattice LSTM-CRF 模型的中文紧急事件抽取 [C]//2020 中国自动化大会论文集. 上海: [s. n.], 2020: 770−775.

[8] 李嘉欣, 王平. 中文命名实体识别研究方法综述 [J]. 计算机时代, 2021(4): 18–21.
LI Jiaxin, WANG Ping. A review of research methods of Chinese named entity recognition[J]. Computer era, 2021(4): 18–21.

[9] PATIL N V, PATIL A S, PAWAR B V. HMM based Named Entity Recognition for inflectional language [C]//2017 International Conference on Computer, Communications and Electronics. New York, USA: IEEE, 2017: 565−572.

[10] YAO Lin, LIU Hong, LIU Yi, et al. Biomedical named entity recognition based on deep neutral network[J]. International journal of hybrid information technology, 2015, 8(8): 279–288.

[11] STRUBELL Emma, VERGA Patrick, BELANGER David, 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, Denmark. Stroudsburg, USA: Association for Computational Linguistics, 2017: 1−13.

[12] YANG Jie, LIANG Shuailong, ZHANG Yue. Design challenges and misconceptions in neural sequence labeling[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA. Association for Computational Linguistics 2018: 3879−3889.

[13] KONG Jun, ZHANG Leixin, JIANG Min, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of biomedical informatics, 2021, 116: 103737.

[14] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer science, 2015: 1–10.

[15] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2018: 1554−1564.

[16] ZHANG Han, GUO Yuanbo, LI Tao. Domain named entity recognition combining GAN and BiLSTM-attention-CRF[J]. Journal of computer research and development, 2019, 56(9): 1851.

[17] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, USA: ACL, 2020: 38−45.

[18] NICO E, VASILEIOS B, KLAUS D. Point transformer [J]. IEEE access, 2021, 9: 134826–134840.

[19] CHIARA Bartolozzi, GIACOMO Indiveri. A selective attention multi–chip system with dynamic synapses and spiking neurons[M]//Advances in Neural Information Processing Systems 19. Cambridge: MIT Press, 2007: 113−120.

[20] JACOB Devlin, CHANG Mingwei, LEE Kenton, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11) [2021-07-06].https://arxiv.org/abs/1810.04805.

[21] DAI Zhenjin, WANG Xutao, NI Pin, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. New York, USA: IEEE, 2019.

[22] LI Xiangyang, ZHANG Huan, ZHOU Xiaohua. Chinese clinical named entity recognition with variant neural

structures based on BERT methods[J]. Journal of biomedical informatics, 2020, 107: 103422.

[23] 毛明毅, 吴晨, 钟义信, 等. 加入自注意力机制的 BERT 命名实体识别模型 [J]. 智能系统学报, 2020, 15(4): 772–779.
MAO Mingyi, WU Chen, ZHONG Yixin, et al. BERT named entity recognition model with self-attention mechanism[J]. CAAI transactions on intelligent systems, 2020, 15(4): 772–779.

[24] LI Xiaonan, YAN Hang, QIU Xipeng, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020: 6836−6842.

[25] YOON W, SO C H, LEE J, et al. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition[J]. BMC bioinformatics, 2019, 20(Suppl 10): 249.

作者简介：

于润羽，硕士研究生，主要研究方向为深度学习、数据挖掘。

杜军平，教授，博士生导师，主要研究方向为人工智能、社交网络分析、数据挖掘、运动图像处理。主持国家重点研发计划项目 1 项、国家自然科学基金重点项目 1 项、发表论文 400 余篇，出版学术专著 6 部。

薛哲，副教授，主要研究方向为机器学习、人工智能、数据挖掘、图像处理。主持国家自然科学基金青年基金项目、参与国家重点研发计划项目 1 项。发表学术论文 30 余篇，出版专著 1 部。

2022 国际制导、导航与控制学术会议 (ICGNC2022)

由中国航空学会制导、导航与控制分会、飞行器控制一体化技术重点实验室共同主办，中国自动化学会导航制导与控制专业委员会、中国自动化学会无人飞行器自主控制专业委员会、中国自动化学会控制理论专业委员会协办，哈尔滨工程大学承办的 2022 国际制导、导航与控制学术会议 (ICGNC2022) 将于 2022 年 8 月 5—7 日在哈尔滨召开。

本届会议英文论文将由 Springer 出版社的 LectureNotes in Electrical Engineering 系列正式出版 (EI 收录)，所有录用的中文论文将被推荐到《中国科学》、《航空学报》、《宇航学报》、《智能系统学报》、《上海交通大学学报》、《北京航空航天大学学报》、《南京航空航天大学学报》等核心期刊发表，其中相关教学改革论文将被推荐到《实验室研究与探索》、《电气电子教学学报》等核心期刊发表。热忱欢迎国内外相关研究领域同行踊跃投稿并参会！详情请访问 ICGNC2022 会议官方网：http://icgnc.buaa.edu.cn/

重要日期：

投稿截止日期：2022 年 3 月 10 日

Special Session 申请截止日期：2022 年 2 月 25 日

终稿截止日期：2022 年 4 月 25 日

注册截止日期：2022 年 4 月 25 日