计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):300−308 [doi: 10.15888/j.cnki.csa.008986]

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

数字人文环境下融入多特征的词命名实体识别①

张　朦[1], 刘忠宝[1,2]

[1](中北大学 软件学院, 太原 030051)
[2](北京语言大学 语言智能研究院, 北京 100083)
通信作者: 刘忠宝, E-mail: liuzb@nuc.edu.cn

摘　要: 近年来, 数字人文受到广泛关注, 数字人文环境下的词命名实体识别研究日渐兴起, 但鲜有研究从字特征的特征表示能力、分词的准确性、领域知识的有效性等方面进行探究. 鉴于此, 针对汉字的象形文字特点和词文本的特殊性, 在字特征的基础上, 引入部首特征、格律特征和声韵特征, 提出特征增强单元和特征抽取单元, 并将词牌知识三元组通过 ANALOGY 得到的知识向量表示为词牌知识向量, 通过双向长短时记忆网络、注意力机制等模型将部首向量、字向量、格律向量、声韵向量、词牌知识向量进行深度融合, 最终构建出融入多特征的词命名实体识别方法. 在《花间集全译》自制语料上的对比实验和消融实验的结果表明, 本文所提方法能够有效利用多特征提升词命名实体识别性能. 其 *F*1 值达到了 85.63%, 完成了词命名实体识别任务.

关键词: 命名实体识别; 多特征; 格律; 数字人文; 诗词

引用格式: 张朦,刘忠宝.数字人文环境下融入多特征的词命名实体识别.计算机系统应用,2023,32(3):300–308. http://www.c-s-a.org.cn/1003-3254/8986.html

Named Entity Recognition of Poetry by Integrating Multi-features in Digital Humanities

ZHANG Meng[1], LIU Zhong-Bao[1,2]

[1](School of Software, North University of China, Taiyuan 030051, China)
[2](Institute of Language Intelligence, Beijing Language and Culture University, Beijing 100083, China)

Abstract: In recent years, research on the named entity recognition of poetry in digital humanities is emerging, but few studies have been conducted with regard to the feature expressiveness of character features, word segmentation accuracy, and the effectiveness of domain-specific knowledge in poetry texts. According to the characteristics of Chinese pictographs and the particularity of poetry texts, a recognition method of named poetry entities with a feature enhancement unit and a feature extraction unit is proposed, which integrates multiple features such as characters, radicals, sounds, and metrical rules. The method presents the knowledge vectors obtained from the knowledge triples of tune pattern titles through the ANALOGY model as the knowledge vectors of tune pattern titles. Then, the radical vector, character vector, metrical rule vector, sound vector, and knowledge vector of tune pattern titles are deeply fused through the bidirectional long short-term memory network and attention mechanism models. In this way, the recognition method of named poetry entities fusing multi-features is constructed. The results of comparative experiments and ablation experiments on the self-made corpus of Translation of Among Flowers (Hua Jian Ji) (《花间集全译》) show that the proposed method can effectively use multi-features to improve the recognition performance of named entities, and its F1 score reaches 85.63%, which means it completes the recognition task of named poetry entities.

Key words: named entity recognition; multi-features; metrical rule; digital humanities; poetry

① 基金项目: 教育部哲学社会科学研究后期项目 (21JHQ081)
收稿时间: 2022-08-17; 修改时间: 2022-09-15; 采用时间: 2022-09-27; csa 在线出版时间: 2022-12-02
CNKI 网络首发时间: 2022-12-05

1 相关研究

伴随自然语言处理技术的进步, 为如今词的数字化、语义化以及知识挖掘等数字人文研究的日渐兴盛打下了坚实的基础. 而命名实体识别作为语义知识提取的关键环节, 无论是对于自然语言处理研究, 还是对于数字人文环境下的知识组织都具有重要的学术价值和现实意义.

命名实体识别的相关研究从早期基于规则匹配的研究发展到基于统计学习的研究, 再到基于预训练模型、注意力机制等深度学习模型的研究, 逐步取得了一系列研究成果.

基于规则匹配、统计学习的命名实体识别中条件随机场 (conditional random fields, CRF) 与规则匹配的方法相结合是较为常用的. 因为其既可以发挥条件随机场模型利用文本序列信息的优势, 又可以融合准确率高、领域专业性强的规则匹配方法. 这种方法致力于收集、总结实体的使用规律, 将其构造为规则进行命名实体识别[1–3].

基于深度学习模型的命名实体识别研究是近几年本领域的前沿和热点所在. Tang 等[4] 收集唐诗和相关典故作为实验语料, 使用 BERT 模型对唐诗和典故进行向量化, 通过计算唐诗的例句、候选句以及候选句引用的典故的相似度, 进行例句的典故实体识别. Yan 等[5] 以中文核心典籍 CCT 作为语料, 采用卷积神经网络抽取字的部首特征、结构特征, 并将部首特征、结构特征和字特征拼接起来进行命名实体识别. 谢靖等[6] 以《素问》为实验语料, 采用 SikuBERT 获取字向量, 以 flat-lattice Transformer (FLAT) 为主要结构进行命名实体识别. Zhou 等[7] 利用 ALBERT 作为 Embedding 层, 双向长短时记忆网络模型 (bi-directional long short term memory, Bi-LSTM) 作为特征提取层并引入多头自注意力机制, CRF 模型作为输出标注层构建 ALBERT-Bi-LSTM-MHA-CRF 命名实体识别框架. 在诗词语料的实验中证明 ALBERT-Bi-LSTM-MHA-CRF 模型提高了古文语料实体识别效果.

此外, 随着 Strubell 等[8] 将膨胀卷积神经网络 (iterated dilated convolutional neural network, IDCNN) 应用于命名实体识别中, 由于 IDCNN 较传统 CNN 具有更大的接受域, 因此在提取序列特征时能够很好地兼顾到局部特征, 故在命名实体识别中 IDCNN 得到应用. Yu 等[9] 在 IDCNN 后加入 CRF 模型, 使 IDCNN 关注到标签的邻近关系, 使得命名实体识别效果得到提升.

另一方面, 研究人员认识到了多特征在命名实体识别中的重要性. 如 Tan 等[10] 将部首特征嵌入 ZEN2 预训练模型获取的字特征中, 提升字特征的语义表达能力, 实现命名实体识别. Wu 等[11] 利用 CNN 抽取部首特征, 并将部首特征融入字特征中实现命名实体识别. 崔丹丹等[12] 提出基于字、词特征, 利用 Lattice LSTM 模型的命名实体识别方法, 在《四库全书》语料上的实验表明, Lattice LSTM 模型相比传统的 Bi-LSTM-CRF 模型提高了实体识别效果.

对相关研究进行梳理可以看出, 学者们在命名实体识别研究中取得了一系列研究成果. 但针对古文的命名实体识别仍是具有挑战性的课题[13]. 首先, 基于规则匹配、统计学习的古文实体识别研究相比现代文更加依赖相关领域专家依据语料人工标注数据并制定相关识别规则, 这将消耗大量人力、物力成本; 其次, 古文命名实体识别中采用基于规则匹配、统计学习的方式实验性能很难进一步提升[14]; 再次, 基于深度学习模型的古文命名实体识别研究认识到字、词特征的重要性, 并将字、词特征进行融合, 以助力古文命名实体识别, 但分词效果不尽如人意的缺陷尚未探讨[15]; 最后, 除字、词特征外, 能否引入外部知识和更多有效特征实现更高性能的古文命名实体识别尚未深入研究. 这些问题是本文在古文中词的命名实体识别中尝试探索的方向.

2 研究方法

词命名实体识别有 3 个显著特点: 一是汉字与部首. 词是由汉字组成的, 其作为最为古老的象形文字之一蕴含着丰富的语义信息. 而部首作为汉字特有的组成部分有其特定的含义. 如: “城”“塔”和“坛”的部首为“土”, 其表示建筑物、房屋的一部分等含义, 即与地点相关; “草”“艾”“芝”, 的部首为“艹”, 其表示草本植物的总称, 即与事物相关. 这样的例子有很多, 在此不再赘述. 故联合部首特征和字特征可以在命名实体识别中发挥作用. 二是词牌. 词牌决定着词的字数、句式、节奏等规则. 相比于传统命名实体识别中消耗大量人力、物力成本的人工构造规则, 引入词牌特征减少了工作量, 可以一定程度提升词的实体识别性能. 三是格律和声韵. 一首词有其特有的格律规则, 除某些特有名词外, 一首词可按格律划分为单字词、双字词[16]. 即在一首词中, 单字词、双字词有其特定的格律特征. 而声韵刻画了格律的字音, 可以增强对格律语义的理解. 故在模型中同时引入声韵特征和格律特征可以为实体词

的识别提供线索, 代替分词差强人意的词特征在命名实体识别上发挥作用.

根据以上分析, 本文提出融入多特征的词实体识别方法 (attention-based multi-features named entity recognition model, AM-NER), 本文的研究思路是: 首先, 输入层在对字、部首、格律、声韵向量化表示的基础上, 通过 ANALOGY 模型将词牌名知识图谱表示为词牌知识向量; 其次, 特征提取层分别将部首向量、字向量与词牌知识向量, 格律向量、声韵向量与词牌知识向量三三融合, 获得字特征向量和格律特征向量; 最后, 将字、格律特征向量通过特征增强层进行向量增强, 输出层获取实体识别结果实现命名实体识别. 本文所提模型 AM-NER 的模型结构如图 1 所示, 其由输入层、特征提取层、特征增强层和输出层 4 部分组成.

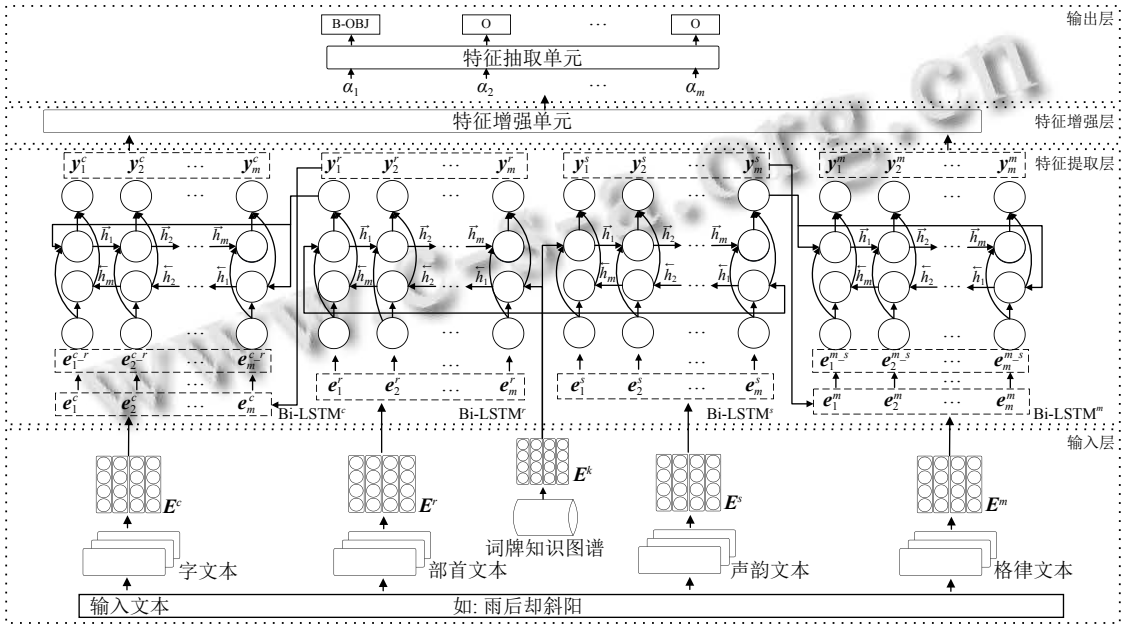

图 1　AM-NER 模型结构

2.1　输入层

输入层负责对词句文本、词牌文本进行预处理、生成输入数据; 将输入数据表示为相应向量 (向量化词牌、格律、部首的数据集可见第 3.1.3 节).

2.1.1　输入数据获取

图 2 以“雨后却斜阳”词句文本与其对应的“菩萨蛮”词牌文本为例, 给出了输入数据的获取过程.

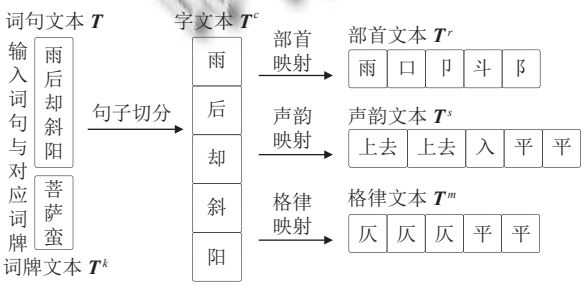

图 2　文本转换过程实例

图 2 中, 首先输入词句 $\boldsymbol{T}$ 和对应词牌 $\boldsymbol{T}^k$. 对于词句 $\boldsymbol{T}$, 它有 m 个字组成, 即字文本 $\boldsymbol{T}^c=\{c_1, c_2, \cdots, c_m\}$, 其中 c_i (i=1, 2, ⋯, m) 表示中的每个字; 部首文本 $\boldsymbol{T}^r$ 是根据新华字典的部首映射关系, 得到部首文本 $\boldsymbol{T}^r=\{r_1, r_2, \cdots, r_m\}$, 其中 r_i (i=1, 2, ⋯, m) 表示 $\boldsymbol{T}^c$ 中每个字的部首; 声韵文本 $\boldsymbol{T}^s=\{s_1, s_2, \cdots, s_m\}$ 是以搜韵网 (https://sou-yun.cn) 的标注四声功能查询 $\boldsymbol{T}^c$ 的声韵并获取, 其中 s_i (i=1, 2, ⋯, m) 表示 $\boldsymbol{T}^c$ 中每个字的所有声韵; 格律文本 $\boldsymbol{T}^m=\{m_1, m_2, \cdots, m_m\}$ 是以诗词吾爱网 (https://www.52shici.com) 的格律匹配功能查询 $\boldsymbol{T}^c$ 的格律并获取, 其中 m_i (i=1, 2, ⋯, m) 表示 $\boldsymbol{T}^c$ 中每个字的格律. 由上述分析可知, $|\boldsymbol{T}^c|=|\boldsymbol{T}^r|=|\boldsymbol{T}^s|=|\boldsymbol{T}^m|$, $|\cdot|$表示文本规模.

2.1.2　输入数据向量表示

自然语言处理领域中, BERT 等预训练模型可以充分利用到字符上下文信息, 生成字符向量具有表达多义性, 并为下游任务节省了大量的时间和计算资源. GuwenBERT (https://github.com/Ethan-yt/guwenbert) 是用殆知阁文献作为语料, 基于中文 BERT-wwm 模型继续训练的预训练模型. 在第 3.1.4 节实验中, GuwenBERT

模型较其他模型在生成字向量方面具有优异的效果. 故本文使用该模型对字文本 $\boldsymbol{T}^c$ 进行向量化表示. 如图 1 所示, $\boldsymbol{E}^c=\{e_1^c,e_2^c,\cdots,e_m^c\}$ 表示字向量集合, 其中 e_i^c (i=1, 2, ⋯, m) 表示字向量. 如图 3 给出了 BERT 模型的网络结构, 其中使用 Transformer 作为编码器计算文本中每个字之间的相关程度, 使输出字向量都包含文本的上下文语义信息.

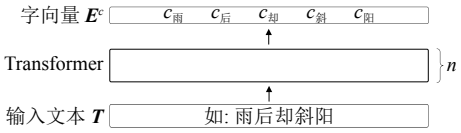

图 3 BERT 模型结构

目前还未有基于部首、格律的预训练语言模型, 与传统的 one-hot 等方式表示文本相比, 使用 Word2Vec 模型[17] 表示文本, 一是解决了数据离散的问题, 二是起到了扩充特征的作用, 故本文使用该模型得到输入数据 $\boldsymbol{T}^r$ 和 $\boldsymbol{T}^m$ 的向量集合. 如图 1 所示, $\boldsymbol{E}^r=\{e_1^r,e_2^r,\cdots,e_m^r\}$ 表示部首向量集合, 其中 e_i^r(i=1, 2, ⋯, m) 表示部首向量; $\boldsymbol{E}^m=\{e_1^m,e_2^m,\cdots,e_m^m\}$ 表示格律向量集合, 其中 e_i^m(i=1, 2, ⋯, m) 表示格律向量.

本文在词牌知识图谱作为数据的 ANALOGY 模型[18] 与其他常用知识图谱嵌入模型的实验中, ANALOGY 模型表现出优异的性能. 故本文通过 ANALOGY 模型将词牌知识图谱中的三元组进行分布式向量表示. 如图 1 所示, $\boldsymbol{E}^k=\{e_1^k,e_2^k,\cdots,e_m^k\}$ 表示词牌知识向量集合, 其中 e_i^k(i=1, 2, ⋯, m) 表示词牌知识向量.

2.2 特征提取层

鉴于中文文本具有显著的序列特征, 故本文采用 Bi-LSTM 模型作为特征提取层的基础模型.

本文以字向量、部首向量、词牌知识向量融合获得字特征向量为例, 介绍该层的工作原理及工作流程, 其结构如图 4 所示.

图 4 中, 首先将 $\boldsymbol{E}^r$ 对应的词牌知识向量集合 $\boldsymbol{E}^k$ 进行线性变换为 $\boldsymbol{k}$, 然后将 Bi-LSTM^r 模型的初始状态和隐藏状态均置为 $\boldsymbol{k}$, 并将对应的部首向量集合 $\boldsymbol{E}^r$ 输入 Bi-LSTM^r 模型, 得到部首特征向量集合 $\boldsymbol{Y}^r=\{y_1^r,y_2^r,\cdots,y_m^r\}$. 其计算如式 (1)、式 (2) 所示:

$$\boldsymbol{k}=\text{Linear}(\boldsymbol{E}^k) \tag{1}$$

$$\boldsymbol{Y}^r=\text{Bi-LSTM}^r(\boldsymbol{E}^r,\boldsymbol{k},\boldsymbol{k}) \tag{2}$$

多头注意力 (multi-head attention, MultiHead) 机制通过将多个注意力机制进行横向拼接来增强模型的关注能力, 进而可以表征不同位置、不同方面的语义信息. 故本文通过多头注意力机制将 $\boldsymbol{Y}^r$ 与 $\boldsymbol{E}^c$ 进行特征融合, 得到融合后的向量集合 $\boldsymbol{E}^{c\_r}$. 其计算如式 (3) 所示:

$$\boldsymbol{E}^{c\_r}=MultiHead(\boldsymbol{E}^c,\boldsymbol{Y}^r,\boldsymbol{Y}^r) \tag{3}$$

最后, 将 $\boldsymbol{E}^{c\_r}$ 作为 Bi-LSTM^c 的输入向量集合并将 Bi-LSTM^r 模型最后时刻的隐层状态传递给 Bi-LSTM^c 模型作为初始状态, 进而得到字特征向量集合 $\boldsymbol{Y}^c=\{y_1^c,y_2^c,\cdots,y_m^c\}$, 其中 y_i^c(i=1, 2, ⋯, m) 表示字特征向量, 其计算如式 (4) 所示:

$$\boldsymbol{Y}^c=\text{Bi-LSTM}^c(\boldsymbol{E}^{c\_r}) \tag{4}$$

格律特征向量集合 $\boldsymbol{Y}^m=\{y_1^m,y_2^m,\cdots,y_m^m\}$ 与字特征向量集合的获取过程类似, 在此不再赘述.

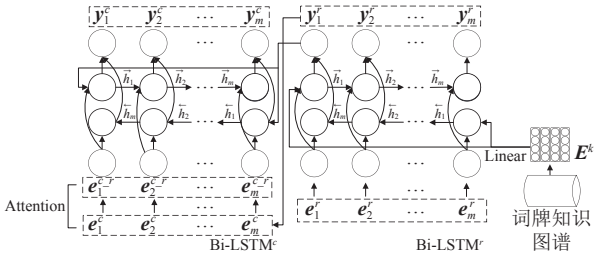

图 4 字特征向量生成示意图

2.3 特征增强层

特征提取层中, 仅依靠 Bi-LSTM 无法解决序列长短不一的问题. 故本文受 Raffel 等[19]、袁健等[20] 的启发, 在该层中提出特征增强单元, 以此从多角度、多层次获取文本的相关特征. 其结构如图 5 所示.

图 5 中, ⊙表示点积运算, ⊕表示加和运算.

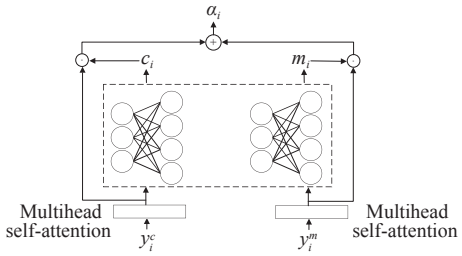

图 5 特征增强单元结构

特征增强单元首先通过多头自注意力机制将格律特征向量集合 $\boldsymbol{Y}^m=\{y_1^m,y_2^m,\cdots,y_m^m\}$ 作为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 进行特征融合, 进而得到格律特征增强向量集合 $\boldsymbol{Y}^{m\_h}=\{y_1^{m\_h},y_2^{m\_h},\cdots,y_m^{m\_h}\}$, 其中 $y_i^{m\_h}$(i=1, 2, ⋯, m) 为经过多头自注意力机制的格律特征增强向量, 其计算过程如

式 (5) 所示:

$$\boldsymbol{Y}^{m\_h} = MultiHead(\boldsymbol{Y}^{m}, \boldsymbol{Y}^{m}, \boldsymbol{Y}^{m}) \tag{5}$$

字特征增强向量集合 $\boldsymbol{Y}^{c_h}=\{y_1^{c_h}, y_2^{c_h}, \cdots, y_m^{c_h}\}$的计算过程与格律特征增强向量集合的计算过程类似, 在此不再赘述.

其次, 将字特征增强向量集合中的向量$y_i^{c_h}$(i=1, 2, ⋯, m) 和格律特征增强向量集合中的$y_i^{m_h}$(i=1, 2, ⋯, m) 输入到权重公式中, 计算$y_i^{c_h}$的权重 c_i 和$y_i^{m_h}$的权重 m_i, 其计算过程如式 (6)–式 (8) 所示:

$$c_i = \frac{\exp(s(y_i^{c\_h}))}{\exp(s(y_i^{c\_h})) + \exp(s(y_i^{m\_h}))} \tag{6}$$

$$m_i = \frac{\exp(s(y_i^{m\_h}))}{\exp(s(y_i^{c\_h})) + \exp(s(y_i^{m\_h}))} \tag{7}$$

$$s(\eta) = \sigma(\boldsymbol{W}_2 \tanh(\boldsymbol{W}_1 \times \eta))) \tag{8}$$

其中, exp 表示以自然常数 e 为底的指数函数, σ表示 logistic 函数, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ 表示权重矩阵, ×表示矩阵乘法.

最后, 将$y_i^{c_h}$和$y_i^{m_h}$按各自权重融合得到特征增强向量 $\boldsymbol{a}_i$ (i=1, 2, ⋯, m), 最终得到特征增强向量集合 $\boldsymbol{A}=\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m\}$, 其计算过程如式 (9) 所示:

$$\alpha_i = c_i \cdot y_i^{c\_h} + m_i \cdot y_i^{m\_h} \tag{9}$$

利用加入多头自注意力机制的特征增强单元, 首先, 其较传统的向量拼接不会引起维度过高的问题; 其次, 增强了 $\boldsymbol{Y}^c$ 和 $\boldsymbol{Y}^m$ 的自身主要信息的关注能力, 弱化了无关信息; 最后, 可以让模型自行训练出 $\boldsymbol{Y}^c$ 和 $\boldsymbol{Y}^m$ 的结合比例, 以得到更全面、丰富的嵌入式表示.

2.4 输出层

输出层负责让特征增强层输出的特征增强向量过滤噪声, 进行主要特征抽取并将主要特征映射为标签序列以实现命名实体识别. 本文受 Xuan 等[21] 的启发, 在该层应用特征抽取单元, 其结构如图 6 所示. 其中, *out* 表示每层输出维度大小, Conv2d 表示卷积操作, *pool* 表示最大池化操作, *reshape* 表示改变形状, *o* 表示过滤器个数, *k* 表示过滤器大小, *s* 表示步长, *d* 表示扩张率参数.

与使用传统的线性层收束特征相比, 卷积神经网络中的卷积层能够很好地描述数据的局部特征, 通过池化层可以进一步抽取出局部特征中最具有代表性的部分. 而相比于卷积网络, 迭代膨胀卷积[22] 由多层不同膨胀宽度的膨胀卷积块组成, 其通过去掉池化层、增大感受野来覆盖更多的特征范围, 因而主要信息抽取效果更佳. 故特征抽取单元首先将特征增强向量 $\boldsymbol{a}_i$ (i=1, 2, ⋯, m) 渲染为 32×32 单通道输入, 利用多个膨胀卷积块替换卷积网络抽取特征.

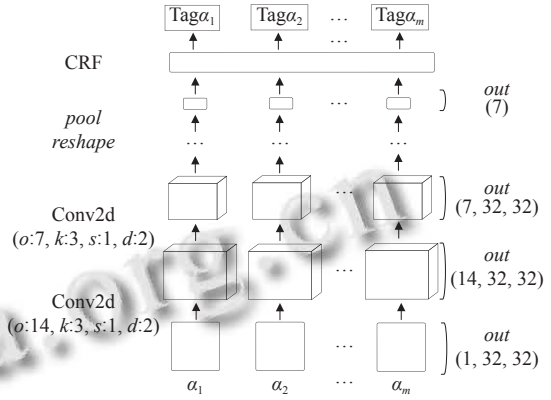

图 6 特征抽取单元结构

其次, 为了过滤噪声, 将膨胀卷积网络抽取到的特征压平, 并采用一维最大池化来处理, 压缩并保留输入特征中的 7 维主要特征.

在获取主要特征后, 理论上只需要输出其中最大的分数值即可作为输入文本的标签类别. 但是这样做只能依据所抽取的特征对当前文本进行预测, 忽略了输出类别标签的依赖关系, 而 CRF 模型能够通过考虑标签的邻近关系获得全局最优标签序列. 故特征抽取单元最后加入 CRF 模型作为分类器得到最终结果.

3 实验

3.1 实验数据集

3.1.1 《花间集全译》词句采集

本文通过人工标注《花间集全译》[23] 中去除序、附录等无关内容, 获得共 500 首词作为实验数据集. 将实验数据集划分为训练集、测试集和验证集, 其比例为 6:2:2.

3.1.2 实验数据集标注

晚唐五代时期, 花间词的作者绝大多数在比较安定的西蜀为官或为蜀地人. 在这样一个比较安定的环境里, 统治者沉溺声色, 穷奢极欲. 当时词作迎合了这样的社会风气, 写出了不少主题为咏物、恋情、边塞等的词作. 在恋情词中, 作者特别注意身体部位的描写. 咏物词中, 总观全集, 所咏之物有杨柳、鸳鸯、宝马、明月等. 作品中作者将思想寄托于山水之间的作品也不乏少数.

经过分析, 本文采用 IBO 格式对实验语料集进行

命名实体标注, “B-XXX”表示命名实体的开始, “I-XXX”表示命名实体的内部, “O”表示非实体字符. 本文将人物以及描写身体部位的词语标注为人物实体即 PER, 将物体标记为物体实体即 OBJ, 将地点标记为地点实体即 LOC. 故 3 种实体类型总共有 6 种标签, 加上不是实体的 O 标签总共有 7 种标签.

经过对实验数据集的统计, 500 首词中有 381 首词至少涉及 2 类命名实体, 108 首词涉及全部 3 类命名实体. 词内容中, 人物实体有 467 种, 共出现 997 次, 其中, 人, 君, 头, 眉等有较高频次; 物体实体有 1 595 种, 共出现 3 785 次, 其中花, 风, 月等有较高频次; 地点实体有 484 种, 共出现 768 次, 其中画堂, 玉楼等有较高频次. 500 首词共有 21 261 个字, 3 类实体的字数有 8 570 个, 占含有实体的词字数的 40.3%, 从实体的整体规模上来看, 实验数据集具有一定的代表性.

3.1.3 词牌、格律、部首数据集采集

为了使输入层获取质量更优的词牌、格律、部首向量, 本文做了以下工作.

本文采集赵崇祚等编撰的《花间集全译》[18] 中所有词的词牌, 并通过网络上收集词牌的别称, 经过繁简转化、去重等预处理后, 汇总为词牌数据集. 如表 1 所示, 由于篇幅所限仅列出 5 个词牌的 3 个别称.

表 1 词牌信息

词牌	别称1	别称2	别称3
菩萨蛮	子夜歌	重叠金	花间意
南歌子	南柯子	怕春归	春宵曲
渔歌子	渔父	渔父乐	渔父词
虞美人	一江春水	玉壶水	巫山十二峰
杨柳枝	柳枝	杨柳	—

本文通过筛选清华大学“文脉”知识图谱[24] 和思知知识图谱, 截取实体中含有词牌数据集中词牌相关字符的三元组, 经预处理后, 构建的局部知识图谱作为词牌领域知识图谱. 利用 ANALOGY 模型将词牌领域知识图谱处理为第 2.1.2 节中的词牌知识向量.

本文通过爬虫获取词牌数据集中相关词牌的词句, 经过预处理后, 利用第 2.1.1 节的标记方法标记相关词句的部首、声韵、格律, 汇总为部首数据集、声韵数据集、格律数据集. 利用 Word2Vec 算法将部首数据集、声韵数据集、格律数据集处理为第 2.1.2 节中的部首向量、声韵向量、格律向量.

词牌数据集的词牌数量、相关词句的句子个数等统计信息如表 2 所示.

表 2 词牌、词句等信息

项目	数量
词牌	127
词句	10481
知识图谱实体	8096
知识图谱关系	1217

3.1.4 字嵌入实验

为了验证 GuwenBERT 模型生成字向量的有效性, 本文对古文预训练模型以及 Word2Vec 在实验数据集上进行命名实体识别效果对比, 这些方法如下.

(1) GuwenBERT-CRF: GuwenBERT 是用殆知阁古代文献作为语料, 基于中文 BERT-wwm 模型继续训练的预训练模型. 本文使用 GuwenBERT 模型构造字向量, 送入 CRF 模型实现命名实体识别.

(2) SikuRoBERTa-CRF: 该模型通过在《四库全书》语料上对 RoBERTa 继续训练得来. 本文将实验数据集的简体字符转换为繁体字符, 使用 SikuRoBERTa 模型构造字向量, 送入 CRF 模型实现命名实体识别.

(3) Word2Vec-CRF: 本文使用 Word2Vec 将实验数据集构造为字向量送入 CRF 模型, 实现命名实体识别.

(4) BERT-CCPoem-CRF: BERT-CCPoem 是通过在中国古典诗词的语料库 CCPC-Full v1.0 对 BERT 继续训练得来. 本文使用 BERT-CCPoem 模型构造字向量, 送入 CRF 模型实现命名实体识别.

表 3 给出了以上各个模型在实验数据集上的实验结果.

表 3 字嵌入实验结果 (%)

模型	实体	*P*	*R*	*F*1	平均*F*1
GuwenBERT-CRF	LOC	**73.7**	**75.7**	**74.7**	**81.2**
	PER	**76.8**	**78.6**	**77.7**	
	OBJ	**82.8**	**84.3**	**83.5**	
SikuRoBERTa-CRF	LOC	55.8	60.9	58.2	66.1
	PER	58.7	62.7	60.7	
	OBJ	66.1	72.5	69.2	
Word2Vec-CRF	LOC	38.7	30.7	34.3	49.3
	PER	43.0	34.7	38.4	
	OBJ	54.2	55.1	54.6	
BERT-CCPoem-CRF	LOC	58.8	66.1	62.2	70.0
	PER	57.2	66.9	61.7	
	OBJ	71.9	76.1	73.9	

从表 3 的实验结果可看出, GuwenBERT-CRF 模型的 3 类实体识别效果均高于其他 3 种模型, 尤其是在 OBJ 实体的识别中, GuwenBERT-CRF 模型的调和平均值均超过了 80%; Word2Vec-CRF 模型仅在句子的表面对上下文信息进行提取表示, 没有融入更多内

部特征, 故在 3 类实体识别实验中的表现均最差, 远低于其他实验的识别效果; BERT-CCPoem-CRF 模型的实验结果较为中庸, 没有展示特别突出的识别性能; SikuRoBERTa-CRF 模型的实验结果仅优于 Word2Vec-CRF 模型, 其原因在于 SikuRoBERTa 采用《四库全书》古文语料继续训练, 但未能获取到词领域的知识. 故本文使用 GuwenBERT 模型生成字向量.

3.2 参数设置

本文利用网格搜索法来确定 AM-NER 模型的参数. max_epoch 最大为 128 并设置连续 20 轮模型性能没有提升就停止训练, batch_size 在网格 [4, 8, 16] 中搜索选取, lr 在网格 [0.0002, 0.0006, 0.001] 中搜索选取, lstm_dropout 在网格 [0.1, 0.2, 0.6] 中搜索选取, num_heads 在网格 [4, 8, 16] 中搜索选取. 经过多次实验调参, 得出的识别效果较好的模型参数设置如表 4 所示.

表 4 参数设置

参数	含义	取值
max_epoch	最大迭代次数	128
batch_size	批量大小	16
lr	初始学习率	0.0002
lstm_dropout	Bi-LSTM模型丢失率	0.2
linear_dropout	线性层丢失率	0.1
num_heads	注意力机制头数	8
hidden_dim	隐藏层的神经元数目	256

3.3 评价指标

实验采用准确率 (precision, P)、召回率 (recall, R) 以及调和平均值 ($F1$-score, $F1$) 来衡量实体识别性能. 其计算公式如式 (10)–式 (12) 所示:

$$P = \frac{TP}{TP+FP} \times 100\% \tag{10}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

$$F1 = \frac{2 \times P \times R}{P+R} \times 100\% \tag{12}$$

其中, 真正例 (true positive, TP) 表示被正确分类的正例样本, 假正例 (false positive, FP) 表示被错误分类的正例样本, 假负例 (false negative, FN) 表示被错误分类的负例样本. 而 P 表示模型预测正确的正例样本占预测为正例的样本的比例, R 表示模型预测正确的正例样本中占实际为正例的样本的比例.

3.4 对比实验与消融实验

3.4.1 对比实验

为了验证 AM-NER 模型的有效性, 本文在实验数据集上对 AM-NER 模型和其他模型的识别结果进行对比, 证明所提模型的优越性. 这些方法如下.

(1) Bi-LSTM-CRF: 该模型以 Word2Vec 生成字向量作为数据, 采用 Bi-LSTM+CRF 为基础模型进行命名实体识别.

(2) ALBERT-Bi-LSTM-MHA-CRF[7]. 该模型使用 ALBERT 模型构造字向量, 然后引入自注意力机制并利用 Bi-LSTM 和 CRF 进行特征提取, 进行命名实体识别.

(3) RCBC[11]: 该模型通过融合字、部首向量进行特征提取, 实现命名实体识别.

(4) Lattice LSTM[12]: 该模型通过融合字、词进行特征提取, 实现命名实体识别.

表 5 给出了以上各个模型在实验数据集上的实验结果.

表 5 对比实验结果 (%)

模型	P	R	$F1$
Bi-LSTM-CRF	81.62	81.57	81.64
ALBERT-Bi-LSTM-MHA-CRF	83.65	84.17	83.92
RCBC	84.49	85.26	84.43
Lattice LSTM	82.65	82.72	82.86
AM-NER	**85.47**	**85.62**	**85.63**

通过将表 5 对比实验中的模型实验结果进行综合分析, 可以得出以下结论.

通过对比 Bi-LSTM-CRF、ALBERT-Bi-LSTM-MHA-CRF 可以发现, ALBERT-Bi-LSTM-MHA-CRF 较 Bi-LSTM-CRF 的 $F1$ 提高了 2.28%, 这主要因为预训练模型可以动态地表示文本向量, 且能根据命名实体识别任务为语义表征能力进行微调, 帮助模型学习领域知识, 进而产生更为丰富的语义特征, 因此带来了性能的提升.

其中 ALBERT-Bi-LSTM-MHA-CRF 为只使用字特征的模型, 其 $F1$ 值为 83.92%, 而 RCBC 为利用部首特征和字特征的双通道模型. RCBC 的 $F1$ 值相较于 ALBERT-Bi-LSTM-MHA-CRF 提升了 0.51%, 这表明利用部首特征能提升字特征的特征表达能力助力命名实体识别.

通过对比 ALBERT-Bi-LSTM-MHA-CRF、Lattice LSTM 可以发现, 前者的 $F1$ 值比后者提高了 1.06%, 分析发现, 前者使用字特征作为输入, 后者使用字特征和词特征作为输入, 因此造成 Lattice LSTM 的 $F1$ 值较 ALBERT-Bi-LSTM-MHA-CRF 略低的原因在于, 如: “楚女欲归南浦, 朝雨. ”中“楚女”切分为了“楚”“女”; “肠断塞门消息, 雁来稀. ”中“雁来”未正确切分为“雁”“来”等的分词错误导致.

本文所提 AM-NER 模型的 $F1$ 值达到了 85.63%, 超过了所有的对比模型, 这证明了 AM-NER 的有效性与优越性. 该模型相较于 RCBC 模型, 其 $F1$ 值提升了 1.2%. 分析得知 RCBC 模型是没有引入词牌知识向量、格律以及声韵特征的, 而 AM-NER 模型通过加入额外的有效特征助力命名实体识别, 进而使得 AM-NER 模型的性能超过了 RCBC.

3.4.2 消融实验

为了验证 AM-NER 模型每个模块的有效性, 本文在实验数据集上调整模型的结构进行对比, 这些方法如下.

(1) Two Bi-LSTM-CRF: 该模型使用两个 Bi-LSTM 模型分别对字向量和部首向量建模, 利用注意力机制融合部首向量和字向量, 将融合后的向量送入 CRF 模型进行命名实体识别.

(2) Four Bi-LSTM-CRF: 该模型在 Two Bi-LSTM 的基础上, 加入两个 Bi-LSTM 模型对格律向量和声韵建模, 将两个通道的输出简单拼接后送入 CRF 模型进行命名实体识别.

(3) A-NER: 该模型舍弃本文所提 AM-NER 模型中的特征抽取单元, 将特征增强层的输出向量送入 CRF 模型进行命名实体识别.

(4) M-NER: 该模型舍弃本文所提 AM-NER 模型中的词牌知识向量进行命名实体识别.

表 6 给出了以上各个模型在实验数据集上的实验结果.

表 6 消融实验结果 (%)

模型	P	R	$F1$
Two Bi-LSTM-CRF	83.92	83.76	83.84
Four Bi-LSTM-CRF	84.44	84.23	84.32
A-NER	84.86	85.24	85.05
M-NER	85.41	85.56	85.52
AM-NER	**85.47**	**85.62**	**85.63**

针对本文所提的实验数据集, 通过将表 6 消融实验中的模型实验结果进行综合分析, 可以得出以下结论.

A-NER 模型的 $F1$ 值为 85.05%, 与 Four Bi-LSTM-CRF 模型相差为 0.73%, 对比两类模型可以发现, Four Bi-LSTM-CRF 模型只是将不同特征通过简单拼接进行特征融合, 在特征提取过程中, 字、格律特征间均没有进行任何信息交互, 而 A-NER 模型利用特征增强单元将字特征与格律特征进行信息交互, 进而提升了命名实体识别性能.

通过对比 AM-NER 模型与 A-NER 模型可以发现, AM-NER 模型 $F1$ 值较 A-NER 模型上升了 0.58%. 分析发现, AM-NER 采用特征抽取单元, 进行特征抽取. 特征抽取单元中膨胀卷积过滤了噪声, CRF 使得模型考虑标签的邻近关系, 进而提升了命名实体识别效果.

为了进一步验证对比实验中 AM-NER 与 RCBC 模型的对比结果. 本文分析了 Two Bi-LSTMs-CRF、Four Bi-LSTM-CRF 模型以及 AM-NER、M-NER 模型的实验结果, 结论如下所示.

通过对比 Two Bi-LSTMs-CRF、Four Bi-LSTM-CRF 模型可以发现, Two Bi-LSTMs-CRF 为利用字特征、部首特征的双通道模型, 其 $F1$ 值为 83.84%. Four Bi-LSTM-CRF 模型为综合利用字特征、部首特征、格律特征与声韵特征的四通道模型. Four Bi-LSTM-CRF 模型的 $F1$ 值比 Two Bi-LSTMs-CRF 提高了 0.48%, 这表明加入格律、声韵等有效特征能为词命名实体识别带来增量.

本文所提模型 AM-NER 的 $F1$ 值达到了 85.63%, 超过了所有的消融模型, 这证明了 AM-NER 模型的有效性与优越性. 该模型相较于 M-NER 模型, 其 $F1$ 值提升了 0.11%. 分析得知 M-NE 模型没有引入外部知识, 而 AM-NER 模型将词牌知识向量引入到 AM-NER 模型中, 进而使得 AM-NER 模型的性能超过了 M-NER 模型. 这表明在命名实体识别中, 引入外部有效知识能够指导深度模型进行更为精准的命名实体识别分析, 进而带来性能的提升.

4 结束语

数字人文的出现为自然语言处理研究带来了新的活力, 为词的深度利用带来了新的视角. 本研究针对词文本的特殊性, 在字特征的基础上, 引入部首特征、声韵特征、格律特征以及词牌特征, 提出融入多特征的词命名实体识别方法. 《花间集全译》实验语料集上的对比实验和消融实验的结果表明, 本文方法的 $F1$ 值较其他模型有一定提升. 需要注意到本研究的不足, 如语料规模需要进一步扩大, 应对词的实体类别进行更加精细的划分等. 在后续研究中应着重探讨上述问题, 但希望能抛砖引玉, 充分挖掘词的价值.

参考文献

1 Zhang Y, Li YK, Zhang J, *et al*. A method for place name recognition in Tang poetry based on feature templates and conditional random field. Proceedings of the 4th International Joint Conference on Web and Big Data. Tianjin: Springer, 2020. 627–635. [doi: 10.1007/978-3-030-60

259-8_46]
2 李章超, 李忠凯, 何琳. 《左传》战争事件抽取技术研究. 图书情报工作, 2020, 64(7): 20–29. [doi: 10.13266/j.issn.0252-3116.2020.07.003]
3 Long YF, Xiong D, Lu Q, *et al*. Named entity recognition for Chinese novels in the Ming-Qing dynasties. Proceedings of the 17th Chinese Lexical Semantics Workshop (CLSW 2016). Singapore: Springer, 2016. 362–375.
4 Tang XM, Liang SC, Zheng JY, *et al*. Automatic recognition of allusions in Tang poetry based on BERT. Proceedings of the 2019 International Conference on Asian Language Processing. Shanghai: IEEE, 2019. 255–260. [doi: 10.1109/IALP48816.2019.9037679]
5 Yan CX, Wang J. Exploiting hybrid subword information for Chinese historical named entity recognition. Proceedings of the 2020 IEEE International Conference on Big Data. Atlanta: IEEE, 2020. 4795–4801. [doi: 10.1109/BigData50022.2020.9378009]
6 谢靖, 刘江峰, 王东波. 古代中国医学文献的命名实体识别——以 Flat-lattice 增强的 SikuBERT 预训练模型为例. 图书馆论坛, 2022, 42(10): 51–60.
7 Zhou FG, Wang C, Wang JP. Named entity recognition of ancient poems based on Albert-BiLSTM-MHA-CRF model. Wireless Communications and Mobile Computing, 2022, 2022: 6507719. [doi: 10.1155/2022/6507719]
8 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2670–2680. [doi: 10.18653/v1/D17-1283]
9 Yu BH, Wei JX. IDCNN-CRF-based domain named entity recognition method. Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology. Weihai: IEEE, 2020. 542–546. [doi: 10.1109/ICCASIT50869.2020.9368795]
10 Tan HX, Yang ZH, Ning JZ, *et al*. Chinese medical named entity recognition based on Chinese character radical features and pre-trained language models. Proceedings of the 2021 International Conference on Asian Language Processing (IALP). Singapore: IEEE, 2021. 121–124. [doi: 10.1109/IALP54817.2021.9675274]
11 Wu YF, Wei X, Qin YB, *et al*. A radical-based method for Chinese named entity recognition. Proceedings of the 2nd International Conference on Big Data Technologies. Jinan: Association for Computing Machinery, 2019. 125–130. [doi: 10.1145/3358528.3358562]
12 崔丹丹, 刘秀磊, 陈若愚, 等. 基于 Lattice LSTM 的古汉语命名实体识别. 计算机科学, 2020, 47(S2): 18–22. [doi: 10.11896/jsjkx.200500090]
13 黄水清, 王东波. 古文信息处理研究的现状及趋势. 图书情报工作, 2017, 61(12): 43–49. [doi: 10.13266/j.issn.0252-3116.2017.12.005]
14 苏祺, 胡韧奋, 诸雨辰, 等. 古籍数字化关键技术评述. 数字人文研究, 2021, 1(3): 83–88.
15 Li XY, Meng YX, Sun XF, *et al*. Is word segmentation necessary for deep learning of Chinese representations? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 3242–3252. [doi: 10.18653/v1/P19-1314]
16 罗凤珠. 诗词语言切分与语意分类标记之系统设计及应用. 第四届数位典藏技术研讨会, 2005. 1–25.
17 Mikolov T, Sutskever I, Cheng K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.
18 Liu HX, Wu YX, Yang YM. Analogical inference for multi-relational embeddings. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR.org, 2017. 2168–2178.
19 Raffel C, Ellis DPW. Feed-forward networks with attention can solve some long-term memory problems. arXiv:1512.08756, 2015.
20 袁健, 章海波. 多粒度融合嵌入的中文实体识别模型. 小型微型计算机系统, 2022, 43(4): 741–746. [doi: 10.20009/j.cnki.21-1106/TP.2020-0972]
21 Xuan ZY, Bao R, Jiang SY. FGN: Fusion glyph network for Chinese named entity recognition. Proceedings of the 5th China Conference on Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence. Nanchang: Springer, 2020. 28–40. [doi: 10.1007/978-981-16-1964-9_3]
22 Yang ZC, Hu ZT, Salakhutdinov R, *et al*. Improved variational autoencoders for text modeling using dilated convolutions. Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017. 3881–3890.
23 赵崇祚. 花间集全译. 崔黎明, 译. 贵阳: 贵州人民出版社, 2008.
24 Lin YK, Liu ZY, Luan HB, *et al*. Modeling relation paths for representation learning of knowledge bases. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 705–714.

(校对责编: 孙君艳)