计算机科学
COMPUTER SCIENCE

基于字词融合的低词汇信息损失中文命名实体识别方法

郭志强, 关东海, 袁伟伟

引用本文

郭志强, 关东海, 袁伟伟. 基于字词融合的低词汇信息损失中文命名实体识别方法[J]. 计算机科学, 2024, 51(8): 272-280.

GUO Zhiqiang, GUAN Donghai, YUAN Weiwei. Word-Character Model with Low Lexical Information Loss for Chinese NER [J]. Computer Science, 2024, 51(8): 272-280.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500047

基于字词融合的低词汇信息损失中文命名实体识别方法

郭志强　关东海　袁伟伟
南京航空航天大学计算机科学与技术学院　南京 211106
(529942688@qq.com)

摘　要　中文命名实体识别(CNER)任务是一种自然语言处理技术,旨在识别文本中具有特定类别的实体,如人名、地名、组织机构名等,它是问答系统、机器翻译、信息抽取等自然语言应用的基础底层任务。由于中文不具备类似英文这样的天然分词结构,基于词的 NER 模型在中文命名实体识别上的效果会因分词错误而显著降低,基于字符的 NER 模型又忽略了词汇信息的作用,因此,近年来许多研究开始尝试将词汇信息融入字符模型中。WC-LSTM 通过在词汇的开始字符和结束字符中注入词汇信息,使模型性能获得了显著的提升。然而,该模型依然没有充分利用词汇信息,因此在其基础上提出了基于字词融合的低词汇信息损失 NER 模型 LLL-WCM,对词汇的所有中间字符融入词汇信息,避免了词汇信息损失。同时,引入了两种编码策略平均(avg)和自注意力机制(self-attention)以提取所有词汇信息。在 4 个中文数据集上进行实验,结果表明,与 WC-LSTM 相比,该方法的 F1 值分别提升了 1.89%,0.29%,1.10%和 1.54%。

关键词:命名实体识别;自然语言处理;词汇信息损失;中间字符;编码策略
中图分类号　TP391

Word-Character Model with Low Lexical Information Loss for Chinese NER

GUO Zhiqiang,GUAN Donghai and YUAN Weiwei
School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China

Abstract　Chinese named entity recognition(CNER) task is a natural language processing technique that aims to recognize entities with specific categories in text,such as names of people,places,organizations. It is a fundamental underlying task of natural language applications such as question and answer systems,machine translation,and information extraction. Since Chinese does not have a natural word separation structure like English,the effectiveness of word-based NER models for Chinese named entity recognition is significantly reduced by word separation errors,and character-based NER models ignore the role of lexical information. In recent years,many studies have attempted to incorporate lexical information into character-based models,and WC-LSTM has achieved significant improvements in model performance by injecting lexical information into the start and end characters of a word. However,this model still does not fully utilize lexical information,so based on it,LLL-WCM(word-character model with low lexical information loss) is proposed to incorporate lexical information for all intermediate characters of the lexicon to avoid lexical information loss. Meanwhile,two encoding strategies average and self-attention mechanism are introduced to extract all lexical information. Experiments are conducted on four Chinese datasets,and the results show that the F1 values of this method are improved by 1.89%,0.29%,1.10% and 1.54%,respectively,compared with WC-LSTM.

Keywords　Named entity recognition,Natural language processing,Lexical information loss,Intermediate characters,Encoding strategy

1　引言

命名实体识别(Named Entity Recognition,NER),也称实体抽取[1],是自然语言处理领域的一项基础性工作。命名实体识别的主要任务是在文本中标识出命名实体并判断其对应的实体类型,如人名、地名、组织机构名、日期等[2]。NER 在许多自然语言处理的下游任务中发挥着重要作用,包括信息检索[3]、文本分类、问答系统[4]等。随着研究人员对其进行了大量的研究,许多方法被提出用于解决 NER 任务,包括隐马尔可夫模型(HMMs)[5]、最大熵(ME)[6]、支持向量机(SVM)[7]和条件随机场(CRF)[8]。随着深度学习的发展,神经网络被引入 NER 任务中。Lample 等使用 LSTM-CRF 模型[9],将字符信息集成到单词表示中,在英文领域的 NER 任务中获得了优异的效果。

到稿日期:2023-05-08　返修日期:2023-08-30
基金项目:航空基金(ASFC-20200055052005)
This work was supported by the Aviation Fundation(ASFC-20200055052005).
通信作者:关东海(dhguan@nuaa.edu.cn)

与具有天然分词特征的英文相比，中文由单个汉字构成，词语之间往往没有分割符号，且语言结构复杂，对实体边界的识别造成了很大的困难。因此，面对中文 NER 任务，一种直观的方式就是首先使用现有的中文分词方法对语句执行分词，然后将基于词级的序列标注模型应用于分词后的语句[10]。然而，现有的分词方法会不可避免地产生分词错误，从而导致实体边界的检测和实体类别的预测出现误差。为了避免分词错误的传播，大多数中文 NER 模型都是基于字符的。有研究表明，基于字符的模型在中文 NER 任务中优于基于词的模型[11]。

虽然基于字符的模型在中文 NER 上取得了不错的性能，但其依然没有利用中文语句中蕴含的词汇信息，而词汇的边界通常很有可能与实体的边界重合，因此引入词汇信息可有效提高实体边界预测的准确性。Zhang 等提出了 Lattice-LSTM[12]，它在 LSTM-CRF 的基础上，通过添加额外的路径，将词汇的开始字符和结束字符的存储单元相连接，将词汇信息融入字符中。但是，因其不同字符间连接的节点数不一致而无法进行批训练，以及词汇信息单一，所以仍然存在计算性能低下和信息损失的问题。针对上述问题，Liu 等提出了一种新的词-字符混合模型 WC-LSTM[13]，通过固定词汇向量大小、将词汇信息集成到输入向量中，以及更进一步地将词汇信息引入开始字符中，简化了 Lattice-LSTM 复杂的模型结构，加快了计算速度，并在一定程度上缓解了词汇信息损失的问题。

虽然 WC-LSTM 在 Lattice-LSTM 的基础上已经取得很大的进步，但其依然没有充分利用词汇信息。如图 1 所示，“天”字在第二层 LSTM 融入了“天安门”词汇信息，“门”字在第一层 LSTM 融入了“天安门”词汇信息，皆对该字符标签的正确预测产生了积极影响。由于 WC-LSTM 模型并不会对存在于词汇中间的字符注入词汇信息，因此“安”字丢失“天安门”词汇信息，这一定程度上降低了该字符被预测正确的可能性。针对该问题，本文提出在 BiLSTM 两层的输入上，对于每个字符，不仅融入以该字符为开头和结尾的词汇信息，还融入以该字符存在于其他词汇中间的词汇信息，分别采用两种编码策略来提取所有词汇信息，最终使用 CRF 得到全局最优标签序列。为验证本文模型的性能优势，在公开数据集 Weibo，Resume，MSRA，Taobao 上进行实验。

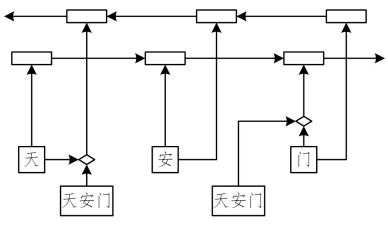


图 1 WC-LSTM 融入词汇信息的示例

Fig. 1 Example of WC-LSTM incorporating lexical information

本文的主要贡献如下：

1）提出了一种基于字词融合的低词汇信息损失的 NER 模型，在对字符融入以该字符为开始和结尾词汇信息的基础上，额外融入以该字符为中间字符的词汇，避免了词汇信息损失。

2）引入两种编码策略，能够固定模型输入和无遗失地采纳各词汇信息，充分利用词汇信息。

3）在 4 个中文数据集上进行实验，结果表明，本文方法在效果上均优于对比模型。

2 相关工作

在命名实体识别任务被提出后，经历了 3 个大的发展阶段。早期的命名实体识别技术发展不成熟，通常是通过人工构造规则和模板来实现的。但该方法需要领域专家自定义规则，不仅耗时耗力，而且对一个领域构造的规则模板不能应用于其他领域，通用性和可移植性较差。在将机器学习应用于命名实体识别任务之后，领域专家不再需要手动构建规则或模板，而是使用带注释的语料库来训练模型，代表性的模型有隐马尔可夫模型和条件随机场模型。这两种方法都是基于机器学习的，因此需要提取特征，在训练期间，可能存在误差传播的问题。鉴于此，学者们逐渐开始将研究重点转向深度学习。

本文从输入表示的角度出发，将这些方法分为基于词、基于字和基于字-词混合的 3 类模型，并分别对它们进行介绍。

2.1 基于词的 NER 模型

基于词的中文命名实体识别模型需要先对输入文本进行分词，然后将分词结果作为输入，输入命名实体识别模型中，再利用不同的算法进行实体标注和分类。文献[14]提出了第一个基于词的命名实体识别模型，其特征由人工构建的词法特征和词典组成，不仅耗时，而且手工特征通常基于规则或启发式方法设计，可能无法完全涵盖所有可能的情况，这导致某些稀有或特殊的实体类型可能无法被充分捕捉。文献[15]提出了 BiLSTM-CRF 模型。BiLSTM 可以同时从前向和后向对文本序列进行建模，有效地捕捉了词汇在整个句子中的语义和上下文相关性；CRF 层考虑了标签之间的依赖关系和转移概率，进一步提高了标注的准确性；BiLSTM-CRF 结合两者的优势，成为了 NER 领域的一种常用模型。文献[9]进一步利用 LSTM 提取了词汇的拼写特征，将拼写特征与词嵌入向量结合，可以为模型提供更丰富的特征表示，这有助于更好地捕捉词汇之间的语义和形态信息。文献[16]额外引入了汉字拼音特征，提出了 Visphone 模型。该模型使用两个相同的交叉转换器编码器，将输入字符的部首和语音特征与文本嵌入相融合。文献[17]通过加入汉字语音特征，解决了实体边界潜在词歧义问题。然而，当上述模型应用于中文命名实体识别时，作为模型输入，对中文进行分词是必须的，因此不可避免地都会遭受分词错误的影响，从而导致实体边界的检测和实体类别的预测出现误差。

2.2 基于字的 NER 模型

与基于词的模型不同，基于字的模型无需对文本分词，而是直接以单字为输入，这样可以减少分词错误带来的负面影响，并且通常能够提升模型性能。文献[18]首次将基于字符的 BiLSTM-CRF 运用于中文 NER 任务中，并在字符的基础上引入了额外的部首特征，部首特征能够捕捉中文字符的字形结构信息，还能够提供一定的字义相关性信息，从而提供更

丰富的特征表示，增强模型对中文命名实体的识别能力。文献[19]使用 BERT 进行特征提取，通过在 BERT 输出层添加一个 CRF 层来进行序列标注，基于大规模语料的预训练模型，BERT 能够学习到丰富的语义表示以及具有较强的泛化能力。文献[20]提出将实体抽取转换成字与字之间的关系抽取问题，通过构建字与字之间的二维网络，有效地捕获了近距离和远距离字对之间的相互作用。文献[21]通过在词之间插入边界符，巧妙地将词的边界信息融入模型，解决了词的 OOV 问题。文献[22]提出了一种简单有效的规律性启发的识别网络(RICON)，规则性感知模块捕捉每个实体的内部规则性，以便更好地预测实体类型，而规则性诊断模块则被用来确定实体的边界。虽然基于字符的命名实体识别通常优于基于词的方法，但是它忽略了词汇信息，而词汇信息对于确定实体边界非常重要。

2.3 基于字词混合的 NER 模型

由于缺乏足够的中文命名实体识别标注数据资源，因此如何在基于字符的模型中引入词汇信息成为近年来 NER 任务研究的重点。在一些中文命名实体识别任务上，使用词汇信息的方法可以媲美甚至优于大语料预训练模型 BERT。

文献[23]首次将中文命名实体识别任务和中文分词任务(CWS)进行联合训练，利用 CWS 任务向 CNER 任务提供词边界信息，在社交媒体数据的 NER 任务中其效果获得了显著提升。文献[24]提出将外部词典知识与 BERT 模型结合，通过 Lexicon Adapter 层直接将外部词典知识融合到 BERT 层中。文献[25]提出了一种基于 BERT 的序列标记任务的插件词典合并方法 DyLex，它采用了一种有效的监督词汇的方法来消除匹配噪声。文献[12]首先提出了一种基于格的长短期记忆网络(Lattice-LSTM)模型，它在 LSTM-CRF 的基础上，通过有向无环图来连接单词开始和结束字符之间的存储单元，从而利用词汇信息，在多个中文数据集上都达到了当时最优的效果。但 Lattice LSTM 中每个字符只能获取以它为结尾的词汇，且数量是不固定的，导致信息损失和运行效率低下。文献[26]借鉴了晶格结构，引入 transformer 作为编码器，从而能够进行批量运算，并引入多孔机制解决了 transformer 无法捕捉长距离依赖的问题。文献[27]在利用 Lattice 结构来引入词典信息的基础上，使用外部无标签数据对词频进行计数，为词典信息的每个部分动态分配权重。文献[13]对 Lattice-LSTM 进行了进一步的改进，通过固定模型输入大小，加快了模型的训练速率；将词汇信息引入词汇的开始或结尾字符中，弥补了词汇信息的损失问题。但这种方法仍然不能充分利用词汇信息。本文研究在其只利用以该字符为开始和结尾的词汇的基础上，为每个字符额外融入以该字符存在于其他词汇中间的词汇信息，进一步充分利用词汇信息。

3 LLL-WCM

3.1 模型架构

本文采用 LSTM-CRF 作为主要网络结构，与目前广泛应用于中文命名实体识别任务的模型相同。该模型的主要结构如图 2 所示。首先，将一段中文序列的每个字符映射到密集向量上；其次，对于每个字符，分别获取以它为开头、以它为结尾以及包含它(在词汇内部)的词汇，并且引入两种编码策略，提取固定大小的词汇信息；然后，将字符嵌入和为其分配的词汇嵌入进行连接，作为 LSTM 的输入来提取词和字符特征；最后，将 LSTM 输出结果通过 CRF 层，获得全局最优标签序列。接下来，将对词-字符嵌入层、词汇编码策略、LSTM 层和 CRF 层这 4 个部分分别进行阐释。

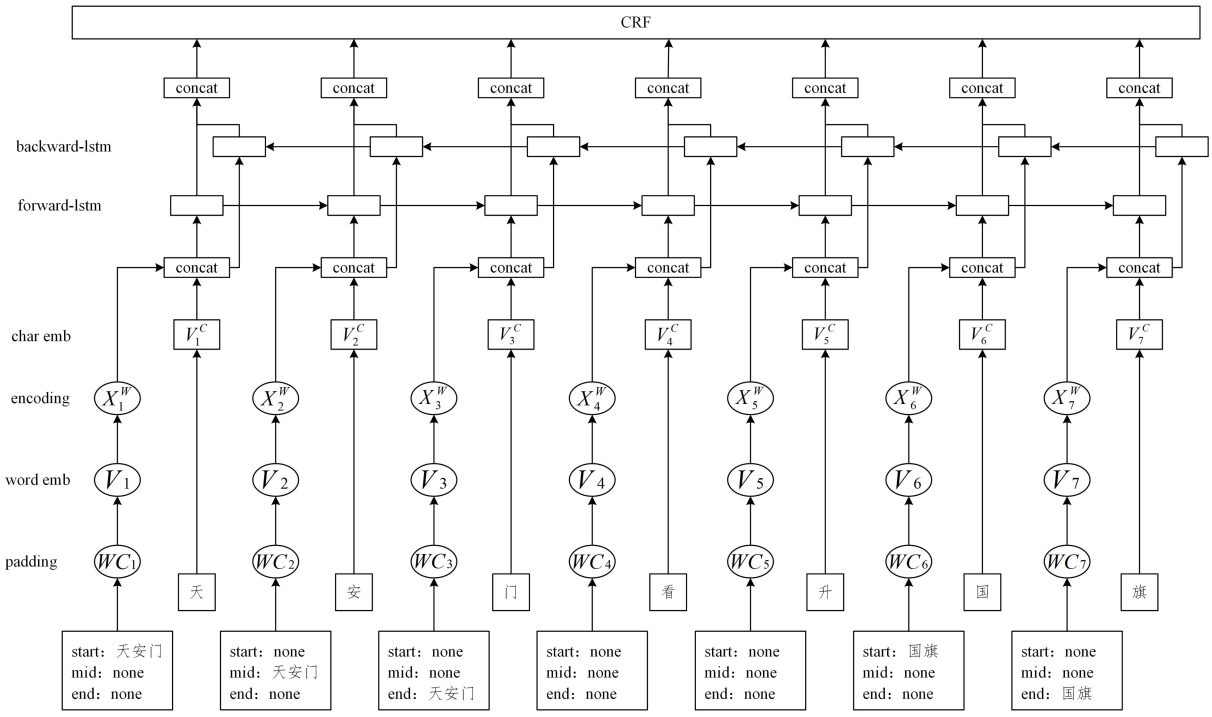


图 2 LLL-WCM 模型的总体架构

Fig. 2 Overall structure of LLL-WCM

3.2 符号表示定义

在正式介绍之前，需要对中文语句、词汇等的符号表示进行定义。将一条语句表示为 $S=\{c_1,c_2,\cdots,c_n\}$，S 表示整个语句，c_i 表示其中的第 i 个字符。$w_{i,j}$ 表示其中的一段字符序列$\{c_i,c_{i+1},\cdots,c_j\}$，$\overrightarrow{wc_i}$表示在前向 LSTM 中分配给 c_i 的所有词汇，且$\overrightarrow{wc_i}$是由不同$w_{i,j}$ 构成的集合，同理$\overleftarrow{wc_i}$表示在后向 LSTM 中分配给c_i的所有词汇。至此，我们为每个字符分配词汇，形成一系列字-词对，可以表示为$\overrightarrow{\boldsymbol{sp}}=\{(c_1,\overrightarrow{wc_1}),(c_2,\overrightarrow{wc_2}),\cdots,(c_n,\overrightarrow{wc_n})\}$。由于本文提出的模型在两层 LSTM 的输入相同，因此在后续介绍中，我们仅以前向 LSTM 为例。

3.3 嵌入层

对于每个字符c_i，其嵌入表示为：

$$\boldsymbol{v}_i^c=\boldsymbol{t}^c(c_i) \tag{1}$$

其中，$\boldsymbol{t}^c$表示一张字符嵌入查找表，它提供了大量单字的向量表示。

$\overrightarrow{wc_i}$表示分配给c_i的所有词汇，其一共由 3 种类型的词汇构成，以c_i为开头的词汇、以c_i为结尾的词汇以及c_i处在词汇内部的词汇，可表示为：

$w_{i,j}(i\leqslant j)$

$w_{h,i}(h\leqslant i)$

$w_{m,n}(m<i<n)$

不同字符分配到的词汇数量各不相同，c_i融入的词汇数量表示为n_i^r。为了统一模型输入大小，我们将一个批次中每个字符的n_i^r通过 padding 扩充至n_i^a，n_i^a的取值为：

$$n_i^a=\max(n_i^r) \tag{2}$$

因此，$\overrightarrow{wc_i}=\{\overrightarrow{w_{i1}},\overrightarrow{w_{i2}},\cdots,\overrightarrow{w_{in_i^a}}\}$，对于任一词汇$\overrightarrow{w_{ik}}$，其向量表示为：

$$\overrightarrow{\boldsymbol{v}_{ik}^w}=\boldsymbol{t}^w(\overrightarrow{w_{ik}}) \tag{3}$$

其中，$\boldsymbol{t}^w$表示一张词嵌入查找表，它提供了大量词的向量表示。$\overrightarrow{wc_i}$的向量表示为：

$$\boldsymbol{V}_i=(v_{i1}^{\overrightarrow{w}},\cdots,v_{in_i}^{\overrightarrow{w}}) \tag{4}$$

3.4 编码策略

为了实现批量化训练，需要确保模型的输入大小一致。同时，为了充分利用词汇信息，需要无遗失地采纳各词汇信息，因此本文提出了两种编码策略，并使用$\boldsymbol{x}_i^{\overrightarrow{w}}$表示$\boldsymbol{V}_i$经过编码后的最终向量表示。

1）平均策略。顾名思义，就是对$\boldsymbol{V}_i$包含的除 padding 以外的所有词汇嵌入求和并取平均值。若V_i仅包含 padding，则对所有 padding 取平均。$\boldsymbol{x}_i^{\overrightarrow{w}}$可表示为：

$$\boldsymbol{x}_i^{\overrightarrow{w}}=\begin{cases}\dfrac{1}{n_i^r}\sum\limits_{k=1}^{n_i^r}\boldsymbol{v}_{ik}^{\overrightarrow{w}}, & n_i^r>0\\ \dfrac{1}{n_i^a}\sum\limits_{k=1}^{n_i^a}\boldsymbol{v}_{ik}^{\overrightarrow{w}}, & n_i^r=0\end{cases} \tag{5}$$

对于不止包含 padding 的$\boldsymbol{V}_i$，我们均匀地对每个非 padding 词汇嵌入取其$\frac{1}{n_i^r}$，从而可以采纳所有词汇信息，$\boldsymbol{x}_i^{\overrightarrow{w}}$的大小和一个词汇嵌入的大小相同。

2）自注意力机制策略。自注意力机制可以有效捕捉文本序列的上下文信息，帮助识别命名实体。文献[28]以 LSTM 输出的隐藏状态为输入，使用注意力机制获取其线性组合。受此启发，将$\overrightarrow{wc_i}$的向量表示$\boldsymbol{V}_i$作为输入，计算式如下：

$$\boldsymbol{a}_i=\mathrm{softmax}(\boldsymbol{W}_{s2}\tanh(\boldsymbol{W}_{s1}\boldsymbol{V}_i^{\mathrm{T}})) \tag{6}$$

其中，$\boldsymbol{W}_{s2}$和$\boldsymbol{W}_{s1}$是两个可学习的参数矩阵。$\boldsymbol{W}_{s1}$的维度是 $m*n$，n 为词嵌入维度的大小；$\boldsymbol{W}_{s2}$的维度是 $m*1$，m 是一个可任意设置的参数。W_{s1}用于将$\boldsymbol{V}_i$映射到一个中间表示矩阵，然后再通过 tanh 激活函数和$\boldsymbol{W}_{s2}$得到初始注意力权重矩阵。我们的目标是将一个可变数量和长度的词汇信息编码成一个固定长度的嵌入，因此还需要使用 softmax 对权重矩阵进行归一化处理。依据该权重矩阵$\boldsymbol{a}_i$，最终可以得到词汇嵌入的线性组合：

$$\boldsymbol{x}_i^{\overrightarrow{w}}=\begin{cases}\sum\limits_{k=1}^{n_i^r}a_{ik}\boldsymbol{v}_{ik}^{\overrightarrow{w}}, & n_i^r>0\\ \sum\limits_{k=1}^{n_i^a}a_{ik}\boldsymbol{v}_{ik}^{\overrightarrow{w}}, & n_i^r=0\end{cases} \tag{7}$$

其中，a_{ik}表示第 k 个词汇嵌入的权重，我们依据权重系数对所有非 padding 词汇提取信息，从而可以采纳所有词汇信息，$\boldsymbol{x}_i^{\overrightarrow{w}}$最终大小和一个词汇嵌入的大小相同。

3.5 字词融合的 LSTM

命名实体识别任务需要考虑长期依赖问题，即当前词汇的实体类型可能受到前面或后面较远距离的单词的影响。LSTM 作为循环神经网络（RNN）的变种，可以通过其长短期记忆机制来解决传统 RNN 中普遍存在的长期依赖问题，从而有效地传递和表达长时间序列中的信息并且不会导致较长时间之前的有用信息被遗忘。因此，我们选择 LSTM 作为网络结构的一部分。使用上述编码策略对词汇嵌入编码后，LSTM 的输入为：

$$\boldsymbol{x}_i=\boldsymbol{v}_i^c\oplus\boldsymbol{x}_i^{\overrightarrow{w}} \tag{8}$$

式(8)表示将字符嵌入和词嵌入的连接作为 LSTM 的输入。LSTM 拥有 3 种类型的门结构，即遗忘门、输入门和输出门，其内部计算过程如下：

$$\begin{bmatrix}\boldsymbol{i}_j\\ \boldsymbol{o}_j\\ \boldsymbol{f}_j\\ \tilde{\boldsymbol{c}}_j\end{bmatrix}=\begin{bmatrix}\sigma\\ \sigma\\ \sigma\\ \tanh\end{bmatrix}(\boldsymbol{W}^{\mathrm{T}}\cdot\begin{bmatrix}\boldsymbol{x}_j\\ \boldsymbol{h}_{j-1}\end{bmatrix}+\boldsymbol{b}) \tag{9}$$

$$\boldsymbol{c}_j=\boldsymbol{c}_{j-1}\cdot\boldsymbol{f}_j+\boldsymbol{i}_j\cdot\tilde{\boldsymbol{c}}_j$$

$$\boldsymbol{h}_j=\boldsymbol{o}_j\cdot\tanh(\boldsymbol{c}_j)$$

其中，$\boldsymbol{i}_j$，$\boldsymbol{o}_j$，$\boldsymbol{f}_j$分别表示输入因子、输出因子和遗忘因子；$\boldsymbol{W}^{\mathrm{T}}$是权值；$\boldsymbol{b}$是偏置向量；$\sigma$表示 sigmoid 激活函数。

由于单向 LSTM 只能利用过去的信息，而双向 LSTM 可以同时利用过去和未来的信息来进行预测，这样可以提高预测的准确性和鲁棒性。因此，本文模型使用了双向 LSTM，其隐藏状态输出$\boldsymbol{h}_i$为两层 LSTM 隐藏状态输出的连接，表示为：

$$\boldsymbol{h}_i=\overrightarrow{\boldsymbol{h}_i}\oplus\overleftarrow{\boldsymbol{h}_i} \tag{10}$$

3.6 CRF 序列标记

对于 LSTM 的输出$\boldsymbol{h}_i$，我们将它投射到线性层进行维度映射，然后使用 softmax 进行分类，便能得到字符对应于不同标签的分数，令$\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$。

$$\begin{aligned} \boldsymbol{E}&=\boldsymbol{wH}+\boldsymbol{b}\\ \boldsymbol{E}&=\mathrm{softmax}(\boldsymbol{E}) \end{aligned} \tag{11}$$

其中，$\boldsymbol{w}$ 为权重向量，$\boldsymbol{b}$ 为偏置向量。

由于相邻字符标签之间可能会存在约束关系，而 CRF 层可以从训练数据中学习到标签约束，从而减少无效预测序列，提高标签预测准确性。CRF 损失函数由真实路径分数和所有路径总分数两部分组成。

$$Loss=-\log\frac{P_{\mathrm{Realpath}}}{P_1+P_2+\cdots+P_N} \tag{12}$$

在训练过程中，模型的参数值将随着训练过程的迭代不断更新，使得真实路径分数所占的比例越来越大。路径分数 p 由两部分组成，即发射分数和转移分数。设某一预测序列为 $y=\{x_1,x_2,\cdots,x_n\}$，则：

$$\begin{aligned} P_y&=E_y+T_y\\ E_y&=e_{x_1}+e_{x_2}+\cdots+e_{x_n}\\ T_y&=t_{\mathrm{start}\to x_1}+t_{x_1\to x_2}+\cdots+t_{x_n\to\mathrm{end}} \end{aligned} \tag{13}$$

其中，E_y 为发射分数，e_{x_1} 表示字符被预测为标签 x_i 的分数，该值可由式(11)计算得到。T_y 为转移分数，$t_{x_i\to x_j}$ 表示标签 x_i 的下一个标签为 x_j 的分数，概率越大则分数越高，在模型训练之前，可以随机初始化转移矩阵的分数，这些分数将随着训练的迭代过程被更新，从而学习到序列之间的约束关系。在解码时，我们使用维特比算法找到获取最高分数的标签序列。

4 实验

4.1 数据集

本文在 MSRA，Weibo，Resume 和 Taobao 这 4 个数据集上进行实验。MSRA 数据集包含了新闻、博客、论坛等多种类型的文本，共计 48442 个句子，其中包含了 3 种实体类型；Weibo 数据集包含了微博文本，共计 1890 个句子，其中包含了 7 种实体类型；Resume 数据集包含了简历文本，共计 4761 个句子，其中包含了 8 种实体类型；Taobao 数据集包含了商品名称、型号等多种类型的文本，其中包含了 9 种实体类型，共计 7998 个句子。数据集详情如表 1 所列。

表 1 数据集统计

Table 1 Statistics of datasets

Datasets	Type	Train	Dev	Test
Weibo	Sentence	1.4×10^3	0.27×10^3	0.27×10^3
	Char	73.8×10^3	14.5×10^3	14.8×10^3
Resume	Sentence	3.8×10^3	0.46×10^3	0.48×10^3
	Char	124.1×10^3	13.9×10^3	15.1×10^3
MSRA	Sentence	41.7×10^3	4.6×10^3	4.4×10^3
	Char	1955.9×10^3	214.1×10^3	172.6×10^3
Taobao	Sentence	6.0×10^3	0.998×10^3	1.0×10^3
	Char	180.5×10^3	30.1×10^3	30.0×10^3

本文使用的预训练字符嵌入和词嵌入与 Lattice-LSTM 使用的相同，该嵌入通过在自动分词的大规模中文预料上使用 word2vec 预训练得到。词嵌入中，包含 5.7×10^3 个单字、291.5×10^3 个双字词、278.1×10^3 个三字词和 129.1×10^3 个其他字词。

4.2 实验设置

在参数设置上，对于较小数据集 Weibo，Resume 和 Taobao，设置 LSTM 隐藏状态维度大小为 100，以防过拟合；对于大数据集 MSRA，设置 LSTM 隐藏状态维度大小为 200，使用更大的 LSTM 隐藏状态维度可以更好地捕捉数据中的复杂关系。设置 Weibo 的训练轮次为 50，MSRA，Resume 和 Taobao 的训练轮次为 75。字符嵌入大小和词汇嵌入大小分别设置为 100 和 50。输入序列最大长度为 250，dropout 置为 0.5，学习率最初设为 0.015，按 0.05 的速率衰减。使用 SGD 作为模型优化算法，使用 Precision，Recall，F1 作为模型的评价指标，计算方式如式(14)所示：

$$F_1=2\cdot\frac{precision\cdot recall}{precision+recall} \tag{14}$$

我们的实验环境为 python3.8、pytorch1.11.0，GPU RTX3090，win10。

4.3 对比模型

本文选用基于字符的 BiLSTM-CRF 作为基准模型，它是实际应用中的常用方法。

在 Weibo 数据集上，除基准模型、Lattice-LSTM、PLTE 和 WC-LSTM 外，还引入了 4 种对比模型，这 4 种方法均针对中文社交媒体数据而提出。

1)Peng and Dredze(a)[29]：使用了基于 LSTM 的神经网络模型，同时使用了字符级和词级的嵌入表示。通过联合训练识别任务和嵌入任务，使得模型在学习嵌入表示的同时能够更好地进行实体识别任务。

2)Peng and Dredze(b)[23]：提出了一种基于词分割表示学习的方法，该方法通过学习词汇之间的相关性来提高命名实体识别的准确性。

3)Sun and He(a)[30]：为解决社交媒体中大量的网络用语和短语不存在中文词典中的问题，提出了一种基于 F-score 的最大边际神经网络方法。

4)Sun and He(b)[31]：提出了一个统一的模型，该模型基于 Transformer 和 BiLSTM-CRF 模型，并使用了多任务学习和迁移学习技术。

在 MSRA 数据集上，引入了 6 种对比模型，以下方法在被提出时在该数据集上获得了优异的表现。

1)Chen 等[32]：提出了基于条件概率模型的方法，包括基于隐马尔可夫模型(HMM)和最大熵模型(MEM)，用于进行中文命名实体识别任务。

2)Zhang 等[33]：提出了一个基于概率特征的最大熵(ME)模型作为模型基本框架，联合多源知识，以提高分词和命名实体识别的性能。

3)Zhou 等[34]：提出了一种基于联合识别和分类的方法来进行中文命名实体识别。首先采用基于条件随机场(CRF)来识别句子中的实体，并使用特征模板来捕捉实体的上下文信息。然后，将识别出的实体和其上下文信息作为输入，利用支持向量机(SVM)进行实体类别的分类。

4)Lu 等[35]：提出了一种基于多原型的方法，将每个汉字表示成多个原型的组合，并引入了一种新的训练策略，通过最大化相似原型对之间的相似度，来优化多原型汉字表示的质量。

5)Dong 等[18]：提出了一个基于字符级别 LSTM-CRF 模型的中文命名实体识别方法，并引入了部首级别的特征来进一步提高模型的性能。

6）Cao 等[36]：提出了一种新颖的对抗性训练框架，通过在源域和目标域之间共享知识来提高目标域的性能。此外，还引入了自注意力机制，以提高模型对命名实体的识别能力。

应用于 Resume 和 Taobao 数据集的方法很少，因此在该数据集上本文方法与基准模型、Lattice-LSTM、PLTE 以及 WC-LSTM 进行了比较。

4.4 实验结果

各模型在 Weibo NER 任务上的结果如表 2 和图 3 所示。NE，NM，Overall 这 3 列表示各模型方法在命名实体（NE）、名义实体（NM）以及全部数据（NE+NM）上的 F1 值。

表 2 Weibo NER 任务对比结果

Table 2 Comparison results on Weibo NER

Models	NE	NM	Overall
Peng and Dredze(a)[29]	51.96	61.05	56.05
Peng and Dredze(b)[23]	55.28	62.97	58.99
Sun and He(a)[30]	54.50	62.17	58.23
Sun and He(b)[31]	50.60	59.32	54.82
Lattice-LSTM	53.04	62.25	58.79
PLTE	53.50	64.81	59.66
Baseline	47.86	57.79	52.80
WC-LSTM+avg	53.37	64.77	59.07
WC-LSTM+self-attention	53.44	64.48	59.39
LLL-WCM+avg	53.95	64.33	60.96
LLL-WCM+self-attention	54.64	65.32	**61.19**

注：最好的结果用粗体突出表示。

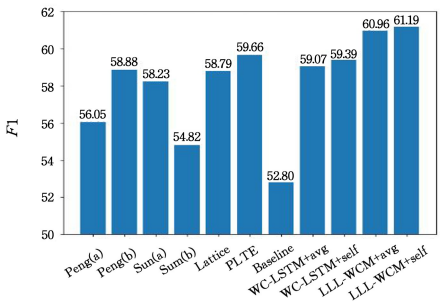


图 3 Weibo 数据集的结果

Fig. 3 Results on Weibo NER

文献[29]利用标记和未标记文本联合训练目标嵌入，文献[23]用 LSTM-CRF 模型联合训练 NER 和单词分割任务，进一步改善了 NER 效果。文献[30-31]利用跨域学习和半监督学习，缓解了稀有数据限制，改善了 NER 效果。然而，这些方法需要用到跨域数据或者半监督数据。Lattice-LSTM 在基于格的 LSTM 结构上融入了结尾词汇信息，与仅使用字符作为输入的基准模型相比，在全部数据上 F1 提升了 5.99%，说明了词汇信息在中文 NER 任务中具有重要的作用。PLTE 在格结构上利用位置关系表征增强了自注意力的能力，较 Lattice-LSTM 提升了 0.87%。WC-LSTM 从嵌入层进一步融入开始词汇信息，在两种策略上性能超越了 Lattice-LSTM，较之分别提升了 0.28%和 0.60%。本文模型在平均和自注意力两种策略上较 WC-LSTM 分别提升了 1.89%和 1.80%，这说明通过融入中间词汇进一步避免词汇信息损失，可一定程度上提升模型效果。在自注意力策略上，本文模型获得了最好的效果，这说明通过对输入序列中的各词汇嵌入进行自动加权，可以帮助模型更好地理解上下文，从而更准确地识别实体。

各模型在 Resume NER 任务上的结果如表 3 和图 4 所示。

表 3 Resume NER 任务对比结果

Table 3 Comparison results on Resume NER

Models	P	R	F1
Lattice-LSTM	94.81	94.11	94.46
PLTE	94.91	95.01	94.96
Baseline	93.13	93.24	93.18
WC-LSTM+avg	94.93	94.87	94.90
WC-LSTM+self-attention	94.77	94.56	94.66
LLL-WCM+avg	95.21	95.09	**95.15**
LLL-WCM+self-attention	95.13	94.78	94.95

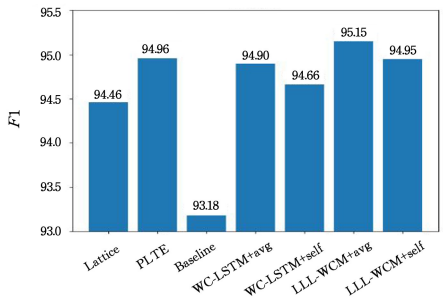


图 4 Resume 数据集结果

Fig. 4 Results on Resume NER

Lattice-LSTM 在词汇的结尾字符引入词汇信息，相比基于字符的基准模型，其 F1 值提升了 1.28%。PLTE 在基于格的 transformer 结构上融入了结尾词汇信息，较基准模型提升了 1.78%。WC-LSTM 进一步在词汇的开头字符融入词汇信息，在两种编码策略上与 Lattice-LSTM 相比分别提升了 0.44%和 0.20%。本文模型通过额外融入中间词汇信息，在两种编码策略上相比 WC-LSTM 分别提升了 0.25%和 0.29%。我们期望自注意力机制策略获取最佳效果，而表 3 和图 4 显示平均策略微优于自注意力机制策略，我们推测使用更复杂的编码策略在该小数据集上发生了过拟合。表 3 所列结果充分说明，丰富的词汇信息可以有效提升模型在中文 NER 任务上的性能。

各模型在 MSRA NER 任务上的结果如表 4 和图 5 所示。

表 4 MSRA NER 任务对比结果

Table 4 Comparison results on MSRA NER

Models	P	R	F1
Chen 等[32]	91.22	81.71	86.20
Zhang 等[33]	92.20	90.18	91.18
Zhou 等[34]	91.86	88.75	90.28
Lu 等[35]	—	—	87.94
Dong 等[18]	91.28	90.62	90.95
Cao 等[36]	91.73	89.58	90.64
Lattice-LSTM	92.58	91.77	92.17
PLTE	94.15	92.39	93.26
Baseline	88.91	87.28	88.09
WC-LSTM+avg	93.36	91.78	92.56
WC-LSTM+self-attention	93.11	91.37	92.23
LLL-WCM+avg	94.45	92.88	**93.66**
LLL-WCM+self-attention	94.23	92.33	93.27

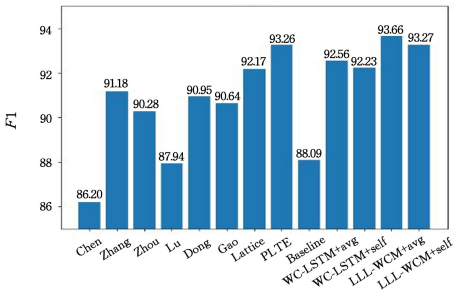


图 5 MSRA 数据集的结果

Fig. 5 Results on MSRA NER

文献[32]将条件随机场和最大熵模型应用于中文 NER 任务，取得了令人满意的效果。文献[33]以一个基于概率特征的最大熵(ME)模型作为基本框架，结合外部多源知识，提高了实体识别性能。文献[34]联合执行分词和实体识别任务，解决了传统的序列标记方法无法将一串连续的字符作为一个命名实体的候选者的问题，从而可以进行更准确的预测。文献[35]提出了一种对位置敏感的 skip-gram 模型来学习中文字符嵌入的多义性。文献[36]应用对抗性迁移学习框架将任务共享的词边界信息整合到汉语 NER 任务中。文献[18]在 Baseline 的基础上，首次利用了汉字的部首特征，性能提升了 2.86%；Lattice-LSTM 首次将词汇嵌入融入字符嵌入中，性能在 Baseline 的基础提升了 4.08%。PLTE 在格结构上引入注意力机制和多孔机制，较 Lattice-LSTM 提升了 1.09%。本文模型充分利用了词汇信息，在平均策略上获得了最好的效果，相比仅利用字符特征的 Baseline，F1 值提升了 5.57%。通过融入中间词汇信息，在 WC-LSTM 的基础上分别提升了 1.10%和 1.04%。自注意力机制策略效果不及平均策略，通过分析实验结果发现，MSRA 数据集中的新闻文本经常出现长文本的人员名单，且这些名字之间没有分割符，而平均策略能够为这些字符无偏倚地引入词汇信息，因此预测更为准确。

各模型在 Taobao NER 任务上的结果如表 5 和图 6 所示。

表 5 Taobao NER 任务对比结果

Table 5 Comparison results on Taobao NER

Models	*P*	*R*	*F*1
Lattice-LSTM	78.27	78.89	78.58
PLTE	82.23	80.44	81.32
Baseline	69.32	77.41	73.14
WC-LSTM+avg	80.41	79.96	80.19
WC-LSTM+self-attention	82.59	79.43	80.98
LLL-WCM+avg	81.20	82.27	81.73
LLL-WCM+self-attention	81.94	82.39	**82.16**

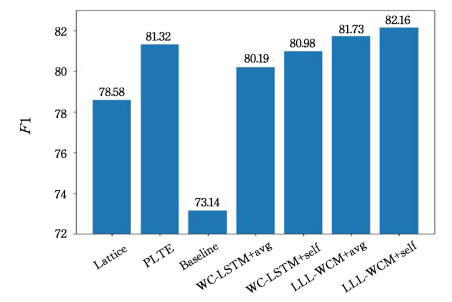


图 6 Taobao 数据集的结果

Fig. 6 Results on Taobao NER

通过在词汇的结尾字符引入词汇信息，Lattice-LSTM 精度获得了显著的提升，较 Baseline 提升了 5.44%。PLTE 继承了 Lattice-LSTM 的晶格结构，使用 transformer 作为编码器，并引入了多孔机制，解决了 transformer 无法捕捉长距离依赖的问题，F1 较 Lattice-LSTM 提升了 2.74%。WC-LSTM 在 Lattice 的基础上，通过对首字符融入词汇信息，在两种编码策略上较之分别提升了 1.61%和 2.40%。LLL-WCM 通过额外融入中间词汇信息来进一步减小词汇信息损失，在两种编码策略上相比 WC-LSTM 分别提升了 1.54%和 1.18%，其中自注意力编码方案获得了最佳效果。

4.5 消融实验

4.5.1 编码器选择

本文模型使用了 LSTM 作为编码器。为了验证在 NER 任务中使用 LSTM 作为编码器的优势，我们将它与 NLP 任务中常用的 CNN，Transformer 编码器进行对比。我们在 4 个数据集上进行实验，每个数据集选择最佳编码策略。

表 6 不同编码器的表现

Table 6 Performance of different encoders

	Weibo	Resume	MSRA	Taobao
Ours(LSTM)	**61.19**	**95.15**	**93.66**	**82.16**
Ours(CNN)	59.42	94.64	92.13	80.96
Ours(Transformer)	60.81	94.21	90.39	81.12

从表 6 中可以得知，基于 LSTM 的模型性能普遍优于基于 CNN 和 Transformer 的性能。经过分析，对于 NER 任务来说，距离和方向信息是十分重要的，离实体较近的字或词更有可能与实体相关，当前字词的类型也能够帮助模型预测左右两侧字符的标签。CNN 只能捕捉局部特征，无法捕捉长距离的依赖关系，Transformer 无法感知上下文信息的方向性。BiLSTM 不仅能够捕捉长距离的依赖关系，还能够区分上下文信息的方向性，因此在 NER 任务中使用 LSTM 作为编码器性能更佳。

4.5.2 预训练嵌入

为了验证预训练词汇嵌入对于命名实体识别的意义，我们参照预训练词汇嵌入表的格式和大小，保留词汇信息，使用均匀分布的方式为每个词汇生成同样维度大小的向量，取代先前的从大规模文本数据中通过无监督训练获得的嵌入。在 4 个数据集上进行实验，结果如表 7 所列。

表 7 有无预训练嵌入的差异

Table 7 Differences with or without pre-trained embedding

	Weibo	Resume	MSRA	Taobao
Baseline	52.80	93.18	88.09	73.14
LLL-WCM+uniform	58.25	94.06	90.45	80.45
LLL-WCM+pretain	**61.19**	**95.15**	**93.66**	**82.16**

相比完全不使用词汇嵌入，使用均匀分布的词汇嵌入在 4 个数据集上 F1 值分别提升了 5.45%，0.88%，2.45%和 7.31%，这说明均匀分布的词汇嵌入能够为模型提供一定程度的语义信息，且在数据量较小的任务上可能产生良好的效果。相比使用均匀分布的词汇嵌入，使用了预训练词汇嵌入的模型在 4 个数据集上的 F1 值分别提升了 2.94%，1.09%，3.21%和 1.71%。首先，这表明词汇信息能够有效地提高

NER 模型的效果；其次，通过大规模文本数据中无监督训练获得的词汇向量，更好地捕捉了词语之间的语义关联，还可以为模型提供词义、语法等更丰富的特征表示，从而更加有效地提高了实体识别的性能。

结束语 本文提出了一种利用词汇信息进行中文 NER 任务的新方法。本文模型可以有效地利用词汇信息并减轻分词错误的影响。为了解决 WC-LSTM 中存在的词汇信息损失问题，我们在其基础上引入了中间词汇信息，并使用两种编码策略来利用所有词汇信息。我们在 Weibo，Resume，MSRA 和 Taobao 这 4 个中文公开数据集上进行了实验，实验结果证明，通过融入中间词汇信息来减少词汇信息损失，可以进一步提高模型的识别效果。但该方法主要适用于扁平实体的识别，未考虑嵌套实体（一个实体中包含另一个实体）和非连续实体（构成实体的字符是非连续的）的情况。未来，我们将探索本文方法的变体，以满足对嵌套实体和非连续实体识别的需要。

参 考 文 献

[1] YU L，GUO Z，CHEN G，et al. Review of Knowledge Extraction Technology for Knowledge Graph Construction[J]. Journal of the University of Information Engineering，2020，21(2)：227-235.

[2] ZHONG S S，CHEN X，ZHAO M H，et al. Incorporating word-set attention into Chinese named entity recognition Method[J]. Journal of Jilin University(Engineering and Technology Edition)，2022，52(5)：1098-1105.

[3] CHEN Y，XU L，LIU K，et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing(Volume 1：Long Papers). 2015：167-176.

[4] DIEFENBACH D，LOPEZ V，SINGH K，et al. Core techniques of question answering systems over knowledge bases：a survey[J]. Knowledge and Information systems，2018，55：529-569.

[5] SAITO K，NAGATA M. Multi-language named-entity recognition system based on HMM[C]// Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition. 2003：41-48.

[6] CHIEU H L，NG H T. Named entity recognition with a maximum entropy approach[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. 2003：160-163.

[7] EKBAL A，BANDYOPADHYAY S. Named entity recognition using support vector machine：A language independent approach[J]. International Journal of Electrical and Computer Engineering，2010，4(3)：589-604.

[8] FENG Y，SUN L，LV Y. Chinese word segmentation and named entity recognition based on conditional random fields models[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006：181-184.

[9] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of NAACL-HLT. 2016：260-270.

[10] YANG J，TENG Z，ZHANG M，et al. Combining discrete and neural features for sequence labeling[C]// Computational Linguistics and Intelligent Text Processing：17th International Conference，CICLing 2016. Springer International Publishing，2018：140-154.

[11] HE J，WANG H. Chinese named entity recognition and word segmentation based on character[C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

[12] ZHANG Y，YANG J. Chinese NER Using Lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). 2018：1554-1564.

[13] LIU W，XU T，XU Q，et al. An encoding strategy based word-character LSTM for Chinese NER[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies(Long and Short Papers). 2019：2379-2389.

[14] COLLOBERT R，WESTON J. A unified architecture for natural language processing：Deep neural networks with multitask learning[C]// Proceedings of the 25th International Conference on Machine Learning. 2008：160-167.

[15] CHEN X，QIU X，ZHU C，et al. Long short-term memory neural networks for chinese word segmentation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：1197-1206.

[16] ZHANG B，CAI J，ZHANG H，et al. VisPhone：Chinese named entity recognition model enhanced by visual and phonetic features[J]. Information Processing & Management，2023，60(3)：103314.

[17] MAI C，LIU J，QIU M，et al. Pronounce differently，mean differently：A multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features[J]. Information Processing & Management，2022，59(5)：103041.

[18] DONG C，ZHANG J，ZONG C，et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]// Natural Language Understanding and Intelligent Applications：5th CCF Conference on Natural Language Processing and Chinese Computing，NLPCC 2016，and 24th International Conference on Computer Processing of Oriental Languages，ICCPOL 2016. Springer International Publishing，2016：239-250.

[19] KENTON J D M W C，TOUTANOVA L K. Bert：Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of naacL-HLT. 2019：4171-4186.

[20] LI J，FEI H，LIU J，et al. Unified named entity recognition as word-word relation classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022，36(10)：10965-10973.

[21] LI L，DAI Y，TANG D，et al. MarkBERT：Marking Word Boundaries Improves Chinese BERT[J]. arXiv：2203.06378，2022.

[22] GU Y，QU X，WANG Z，et al. Delving Deep into Regularity：A

Simple but Effective Method for Chinese Named Entity Recognition[C]// Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 1863-1873.

[23] PENG N, DREDZE M. Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016: 149-155.

[24] LIU W, FU X, ZHANG Y, et al. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5847-5858.

[25] WANG B, ZHANG Z, XU K, et al. DyLex: Incorporating Dynamic Lexicons into BERT for Sequence Labeling[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 2679-2693.

[26] MENGGE X, YU B, LIU T, et al. Porous lattice transformer encoder for Chinese NER[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 3831-3841.

[27] HUANG S, SHA Y, LI R. A chinese named entity recognition method for small-scale dataset based on lexicon and unlabeled data[J]. Multimedia Tools and Applications, 2023, 82(2): 2185-2206.

[28] LIN Z, FENG M, DOS SANTOS C, et al. A structured self-attentive sentence embedding[C]// International Conference on Learning Representations (ICLR). 2017.

[29] PENG N, DREDZE M. Named entity recognition for chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 548-554.

[30] HE H, SUN X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[31] HE H, SUN X. F-Score Driven Max Margin Neural Network for Named Entity Recognition in Chinese Social Media[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 713-718.

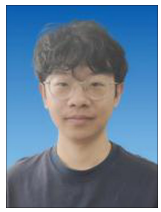
[32] CHEN A, PENG F, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 173-176.

[33] ZHANG S, QIN Y, HOU W J, et al. Word segmentation and named entity recognition for sighan bakeoff3[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 158-161.

[34] ZHOU J, QU W, ZHANG F. Chinese named entity recognition via joint identification and categorization[J]. Chinese Journal of Electronics, 2013, 22(2): 225-230.

[35] LU Y, ZHANG Y, JI D. Multi-prototype Chinese character embedding[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016: 855-859.

[36] CAO P, CHEN Y, LIU K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 182-192.



GUO Zhiqiang, born in 1998, postgraduate. His main research interests is knowledge graph.

GUAN Donghai, born in 1981, Ph.D, associate professor, graduate supervisor. His main research interests include data mining, knowledge inference, etc.

(责任编辑:喻藜)