基于 Wobert 与对抗学习的中文命名实体识别

倪渊[1,2],廖世豪[3*],张健[1,2]

(1.北京信息科技大学经济管理学院,北京 100192;2.绿色发展大数据决策北京市重点实验室,北京 100192;
3.北京信息科技大学计算机学院,北京 100192)

摘 要: 由于自然语言处理(NLP)将中文命名实体识别(NER)任务建模为序列标注任务,将文本中每个字符映射至一个标签,每个字符相对独立且信息有限,因此在 NER 领域词汇信息的加入能够解决字符间缺乏联系的问题。针对现有中文 NER 模型多需要额外构建词汇表、提取词汇信息方式繁琐、词级嵌入与字级嵌入因来源不同导致信息难以融合的问题,提出一种基于 Wobert 与对抗学习的中文 NER 模型 ALWAE-BiLSTM-CRF。与传统预训练模型相比,Wobert 预训练模型在预训练阶段就已将文本分词,充分学习了词与字两个层次的信息,因此 ALWAE-BiLSTM-CRF 通过 Wobert 预训练模型获取字符词向量,再使用 Wobert 分词器获取预训练模型中已存在的词汇向量,接着使用 BiLSTM 模型获取两者的时序信息,随后通过多头注意力机制将词汇级别的信息要素融入字符词向量,同时通过对抗学习攻击生成对抗样本以增强模型泛化性,最后使用条件随机场(CRF)层对结果进行约束,获得最佳的预测序列。在 Resume 数据集与瓷器领域的自建数据集 Porcelain 上进行对比实验与消融实验,结果表明,ALWAE-BiLSTM-CRF 模型的 F1 值分别达到 97.21%与 85.7%,证明了其在中文 NER 任务中的有效性。

关键词: 深度学习;命名实体识别;注意力机制;特征融合;条件随机场

中图分类号: TP391.1 **文献标志码:** A **DOI:** 10.19678/j.issn.1000-3428.0068258

Chinese Named Entity Recognition Based on Wobert and Adversarial Learning

NI Yuan[1,2], LIAO Shihao[3*], ZHANG Jian[1,2]

(1. College of Economics and Management, Beijing Information Science and Technology University, Beijing 100192, China;
2. Beijing Key Laboratory of Green Development Big Data Decision, Beijing 100192, China;
3. College of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China)

【Abstract】 Natural Language Processing (NLP) models the Chinese Named Entity Recognition (NER) task as a sequence annotation task and maps each character in the text to a label. Each character is relatively independent and has limited information. Therefore, the addition of vocabulary information to the NER field can solve the problem of the lack of connections between characters. To address the challenges of existing Chinese NER models that require additional vocabulary construction, employ a cumbersome extraction process of vocabulary information, and have difficulties integrating information due to different sources of word-level embedding, this study proposes a Chinese NER model based on Wobert and adversarial learning named ALWAE-BiLSTM-CRF. Unlike traditional pre-training models, the Wobert pre-training model segments the text in advance (i.e., during the pre-training stage), thereby fully learning information at both the word and character levels. Accordingly, the proposed model obtains character word vectors through the Wobert pre-training model and then uses the Wobert word splitter to obtain the existing vocabulary vector in the pre-training model. The proposed model next uses the BiLSTM model to obtain the temporal information of the two. The model then utilizes a multi-head attention mechanism to integrate vocabulary-level information elements into the character word vector while simultaneously generating adversarial samples through adversarial learning attacks to enhance model generalization. Finally, the proposed model utilizes a Conditional Random Field (CRF) layer to constrain the results and obtain the best prediction sequence. The study conducted comparative and ablation experiments on the Resume and self-built Porcelain datasets in the field of porcelains, the results show that the ALWAE-BiLSTM-CRF model achieves 97.21% and 85.7% F1 values on the two datasets, proving its effectiveness in the Chinese NER task.

【Key words】 deep learning; Named Entity Recognition(NER); attention mechanism; feature fusion; Conditional Random Field(CRF)

收稿日期: 2023-08-17 **修回日期:** 2023-12-26

基金项目: 国家重点研发计划青年科学家项目(2021YFF0900200)。

通信作者 E-mail: * lsh123@bistu.edu.cn

0 引言

命名实体识别(NER)[1]任务的目标是在非结构化文本中尝试提取隐含于其中的实体,并将识别出来的实体分类,NER 是自然语言处理(NLP)的基础任务之一。实体类别包括时间、地理位置、人名、组织机构名、专有名词等,其作为知识图谱(KG)构建流程中的一环,在自然语言处理领域受到了广泛的关注。以往的 NER 任务主要针对英文文本,随着中文命名实体识别在中文知识图谱构建等任务中的广泛应用,针对中文文本的命名实体识别研究越来越受到工业界以及学术界的关注。

与英文命名实体识别不同,中文命名实体识别更为困难,因为中文短语中表达意思的单字或词语没有自然分割。一般的中文命名实体识别问题被形式化为序列标注问题,首先基于不同的规则将文本划分为多个 token,然后将其初始化为向量并输入到用于提炼信息的神经网络模型中,该模型最终输出与每个 token 相对应的预测标签。常用的 token 有两种形式,即字符和词语。然而,由于中文字符本身相对独立,导致预训练模型输出的向量中缺乏中文词语所具有的复杂语义信息,因此纯粹基于字符的模型可能面临实体类别识别错误和实体识别不完整等多种问题。

为了解决上述问题,研究人员主要考虑如何在字符嵌入的基础上引入与原文相关的单词信息,利用外部知识改进中文信息表示,而对于单纯基于字符的中文命名实体识别模型来说,增强词汇信息被证明是一个有效手段。现有方法的普遍做法是将整个句子作为原始输入,之后通过各种模型组合提取文本信息进行序列标注,在此基础上引入句子中的词语信息[2]或是词汇的边界信息[3]作为模型训练的辅助信息,但是上述两种新增的信息都与句子原本语义存在鸿沟,辅助信息与原始输入的融合有限,导致模型效果难以提升。

本文提出一种基于对抗学习和 Wobert(Word-based BERT)自适应嵌入的 BiLSTM-CRF 中文命名实体识别模型,即 ALWAE-BiLSTM-CRF 模型。与使用卷积神经网络(CNN)、注意力机制、BiLSTM 等模型的组合来提取特征不同[4-5],ALWAE-BiLSTM-CRF 模型使用 Wobert 对输入的文本序列分别进行字符级别与词语级别的两次编码,除了获得单个字符的词向量外,模型额外融入词语细粒度的词嵌入信息,由于字符词向量与词语向量出自同样的预训练模型映射的多维空间,因此两者更容易进行信息融合。此外,该模型将对抗学习添加至 Wobert 词嵌入层,通过在训练过程中计算梯度以获取扰动,将扰动加至训练样本编码中生成对抗样本,同时影响两种不同细粒度的词向量,以此来提升抗干扰性与泛化性。此方法不仅避免了繁琐的词字典构造,还提供了更为合理和有效的词嵌入信息来源。最后通过在 Resume 数据集与自建 Porcelain 数据集上进行不同层次的对比消融实验,以验证所提方法在中文命名实体识别领域的识别效果。

1 相关工作

命名实体识别任务以序列标注方法为主,将输入中的每一个元素都标记为一个标签,以此区分不同类型的实体与非实体。以往研究者们主要使用机器学习的方法研究命名实体识别问题,包括隐马尔可夫[6](HMM)、条件随机场[7](CRF)等机器学习算法。近年来,随着计算机算力的不断提升,命名实体识别领域的研究逐渐由基于机器学习转变为基于深度学习。由于 Word2vec 等静态语言模型无法感知语境对字词的影响,因此文献[8]在词向量方向进行改进,引入 BERT 模型对文本进行词向量表征,在包括命名实体识别在内的不同自然语言处理任务中都取得了出色的表现。

预训练语言模型[9]可以基于海量文本的预训练任务,自动化学习通用语言的信息表示,解决了一词多义、生僻字等多种自然语言中存在的工程难点,给深度学习模型提供了质量更高的初始化词向量嵌入,其具有较好的泛化性,缓解了下游任务的调优难度。随着预训练模型的进一步研究与发展,陆续出现了多种在 BERT 模型基础上进行各种优化与改进的预训练模型,并应用于中文命名实体识别任务中,如文献[10]提出基于 ALBERT 预训练模型的中文命名实体识别算法,通过 ALBERT 预训练模型提取词嵌入的特征,并利用空洞卷积网络捕捉文本全局语义。文献[11]则是提出了结合 ERNIE2.0 的实体识别模型,通过 ERNIE2.0 模型得到词的动态表征,同时使用 BiSRU 和软注意力分别提取全局高维度的序列特征和进行权重计算。2019 年,文献[12]提出了 RoBERTa 模型,RoBERTa 在 BERT 的基础上扩大模型规模,改进训练流程,同时增加大量训练数据,进一步提高了词汇表征的信息维度与准确度。

为了更好地提取文本词向量的信息,文献[13]首次通过 BiLSTM 将上下文特征信息结合起来,将 BiLSTM-CRF 模型应用于序列标注任务,进行命名

实体识别的训练，对比多种模型结果，验证了BiLSTM-CRF模型在序列标注任务中的有效性与鲁棒性。文献[14]构建BERT-BiLSTM-CRF模型，在非结构化历史文本中抽取实体，在历史文化领域中也取得了良好效果。文献[15]提出了一种多标签CNN方法，将实体识别任务转换成分类任务，改进多标签机制并添加至原本的输入层中，以此获取文本中相邻标签之间的交互信息，此方法在化合物和疾病名的识别任务中取得了相对较好的识别效果。文献[16]提出了一种多任务的深度神经网络，为了解决文本数据中存在的噪声干扰，其并行结合CNN与BiLSTM，从文本序列、词典信息以及语法信息中获取更高阶的特征。

除此之外，大量的研究者专注于将词汇信息添加至模型中进行训练。文献[17]提出HAN模型，通过两个不同层次的注意力机制感知词汇与句子，赋予模型基于词汇和句子的不同重要性来动态调整权重的能力，以提升模型性能。文献[18]提出了一种基于词典增强的中文序列标注模型，促进了BERT底层的词汇知识融合。文献[19]提出Lattice-LSTM模型，利用外部词典将词向量融入到字向量中，通过词汇信息增强实体信息，从而提升实体识别能力，但是Lattice-LSTM模型训练与推理速度过低，且模型结构难以应用到其他神经网络结构中。为此，文献[20]基于Lattice-LSTM的思想提出了一种更简便的实现方法，在基于字符的模型上整合每个字符所有可能匹配的单词，进而将词典信息编码至字符中，模型结构容易实现，也提升了适应其他模型结构的能力。文献[21]提出了Flat-Lattice-Transformer模型，该模型利用多头注意力机制捕获中文单词与词语的长距离依赖关系。

生成对抗网络(GAN)最早由文献[22]提出，最初应用于计算机视觉领域。GAN由鉴别器和生成器两个部分组成。鉴别器用于确定样本是原始真实样本还是生成器生成的假样本。生成器的目的是尝试生成难以由鉴别器正确确定的样本。随后的研究人员将对抗学习应用于自然语言处理领域，文献[23]结合对抗学习与注意力机制，在命名实体识别任务中整合了共享词的边界信息。文献[24]在词性标注任务中使用了对抗训练，不仅提升了整体词性标注的准确率，同时强化了模型的鲁棒性。

综上所述，中文命名实体识别的优化手段主要包括使用更强大的预训练模型丰富词嵌入向量特征信息，以及融合字符与词汇信息。以上方法因为词向量与字向量并非源于同一个预训练模型，或是词向量并非兼具词语中字符的信息与词语本身的整体独立信息，导致两者维度偏移且融合程度不够进而影响模型泛化能力。因此，本文提出基于Wobert与对抗学习的中文命名实体识别算法，为词汇信息的加入与融合提供更便捷的解决方案，无需额外构建词汇词表，同时使用对抗学习加速融合过程并增强模型泛化能力。

2 本文模型设计与实现

2.1 ALWAE-BiLSTM-CRF模型结构

ALWAE-BiLSTM-CRF模型整体框架如图1所示，模型的层次结构主要包括4个部分：第一部分是词向量嵌入层，使用Wobert对文本序列分别进行字符与词语级别的两次嵌入，得到两种包含不同信息要素的词向量；第二部分是BiLSTM层，BiLSTM将处理Wobert输出的向量，以捕获文本序列的长距离依赖关系；第三部分是Attention层，注意力机制进一步提取词语细粒度词向量中的最匹配信息，融入字符细粒度的词向量中；第四部分是CRF层，CRF层可以学习到标签之间的约束关系，对上层输出的预测概率进一步更新后解码得到最优的预测序列。

将文本按词语级别的细粒度切分是Wobert默认的分词方式，但是只需要将文本中的每一个字都以空格字符分隔，那么Wobert仍会将每个字都视为单独的token。本文将每条文本预处理为上述两种形式，通过Wobert嵌入之后会得到两种不同的词向量，这种方法以非常简单的途径就能获取文本的词汇知识，不需要在模型层接入额外的结构。

BiLSTM可以同时从正、反两个方向对文本序列的向量进行二次建模，可以有效地捕获每个字与上下文的相关性，而CRF层则可以弥补BiLSTM没有考虑到标签之间依赖关系与转移概率的缺陷，两者结合的方式已然成为命名实体识别领域的一种常见模型组合。本文以此为基础进一步在BiLSTM层与CRF层之间增加自注意力层，主要原因是NER任务的特殊性，需要将每一个token都预测为一个单独的标签，故而以单字级别的向量为基准来考虑如何将词汇信息融入到每个字的向量中是非常合理的思路。自注意力机制正是在向量层面寻找最优配比的算法，自注意力层可以自适应地选择合适的词汇信息并将其加入到相适应的字向量中，在向量进入CRF层之前，自注意力层为每个字都融入了符合当前语境的词汇信息，充分利用了向量中蕴含的词汇信息。

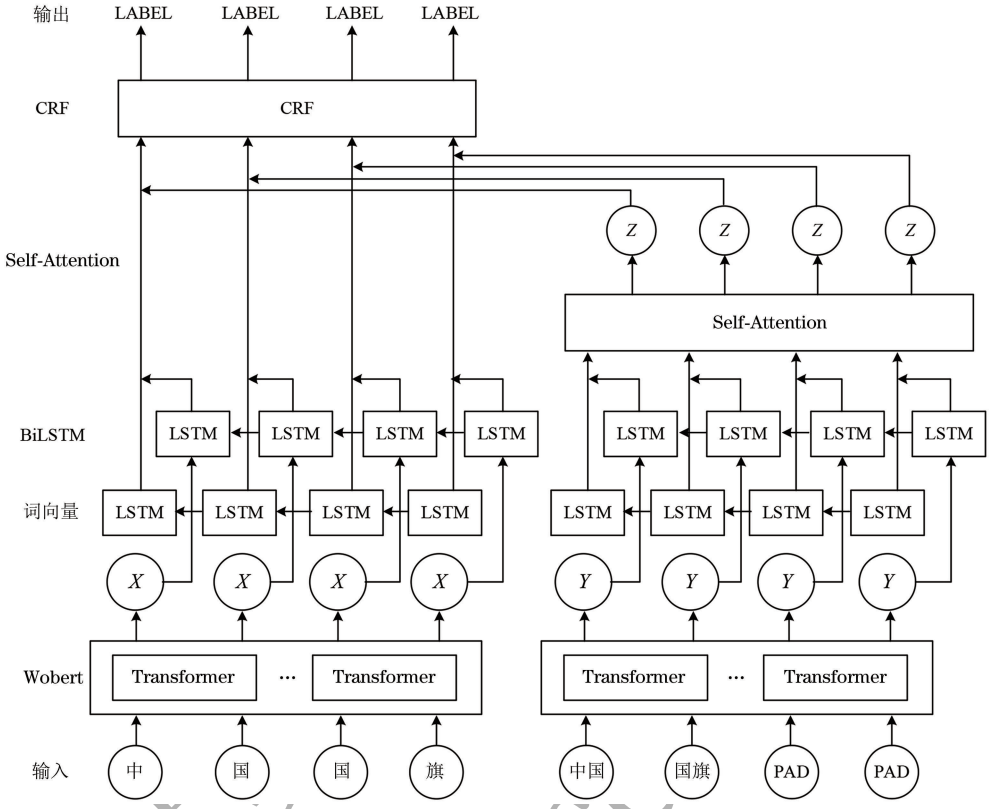

图 1 ALWAE-BiLSTM-CRF 模型结构

Fig. 1 Structure of ALWAE-BiLSTM-CRF model

来源于同一文本的两组向量之间必然存在部分重复的冗余信息，使用对抗学习在向量层添加扰动可以增加信息的多元性，在模型每次计算完原始梯度之后，通过计算对抗样本在反向传播后得到的对抗梯度，累加原始梯度与对抗梯度一起对模型的参数进行更新，不仅补充了文本语义广度，也加强了两组不同细粒度的文本信息的融合深度，提高了模型的泛化能力与稳定性。

2.2 Wobert 预训练模型

众多研究表明，使用分布式词向量可以显著提升模型性能。随着预训练模型的逐步发展，中文命名实体识别任务普遍使用 BERT 模型来提取文本的全局特征，BERT 模型是由多个 Transformer 结构堆叠的深度网络模型，能够敏锐地捕捉上下文的语义关系，动态地根据具体的文本语境准确地表达语义特征，因此可以很好地解决中文的一词多义问题。

RoBERTa-wwm-ext[25] 是一种对 BERT 改进的预训练语言模型。与 BERT 的主要区别在于，RoBERTa-wwm-ext 在预训练阶段将模型的训练策略转变为全词掩码(wwm)，即使用 MASK 标签替换完整的词，而不是单独的一个字，这个特点让该模型更适用于多种中文领域自然语言处理任务，也包括中文命名实体识别任务。

Wobert[26] 是在 RoBERTa-wwm-ext 的基础上以词为单位继续进行预训练的大型预训练模型，Wobert 在词典中加入了中文词语，同时修改了分词器的分词规则，将中文词语的向量初始化后，在 RoBERTa-wwm-ext 基础上以 MLM 任务继续训练，进一步提升了中文文本的表示效果。考虑到分词算法的准确性可能对模型产生影响，Wobert 选择只保留最常见的一部分词汇，那么所有分词算法的分词结果都相差不大，对模型可能产生的影响也可以忽略。同时 Wobert 在初始化阶段没有采用常用的随机初始化，而是使用字 embedding 的平均作为词 embedding 的初始化，利用了一定量的先验知识，使模型具有快速提取词汇信息的能力。由于是在文本嵌入之前将文本进行分词，将词语看作一个整体参与模型训练，因此，相较于 RoBERTa-wwm-ext 模型，Wobert 拥有高一个维度的词汇信息。Wobert 输入实例如图 2 所示。

通过外部词典将词汇知识与 BERT 模型相结合的方法需要先将输入文本分词，然后将分词的结果输入模型，再利用不同的算法融合词汇知识。这种方法不仅结构复杂，同时需要调整表示层来适应 BERT 模型，损失了一定的词汇信息。而 Wobert

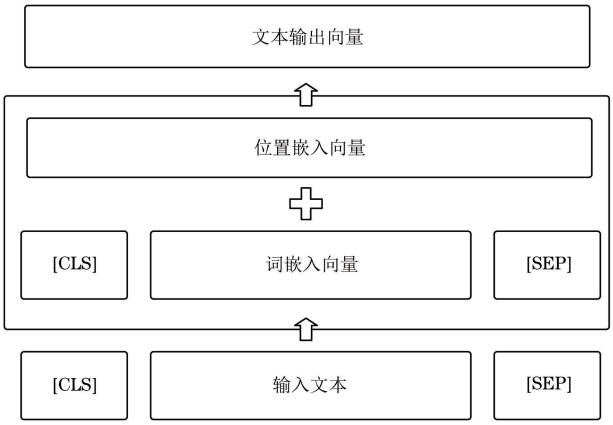

图 2 Wobert 输入实例

Fig. 2 Example of Wobert input

本身就具备词汇层级的分词能力，Wobert 默认按词级别将文本拆分为 token，同时只需要在原文本中相邻字之间插入空格即可获取单个文字的嵌入向量，基于这个特性，同一个文本便可以通过 Wobert 进行两次不同的嵌入，得到截然不同的两种词嵌入表示。不论是字级别还是词级别的信息都来自于同一个预训练模型，整合了更完整的词汇信息，同时避免了调整词汇信息适应预训练模型过程中产生的信息损失。使用 Wobert 同时获取字与词两个等级嵌入的方法易于实现，也避免了复杂的外部词典嵌入的模型结构，并且仅通过调整特征提取层就可以快速适应不同领域。Wobert 在词汇信息的提取方面因为简单同时有效而具备优越性。

2.3 BiLSTM

循环神经网络（RNN）是一种用于处理序列数据的神经网络，其在处理序列数据时获取句子的语序信息，但是当过往相关信息和当前位置的间隔变得非常大时，循环神经网络就会丧失学习远距离信息的能力。除此之外，循环神经网络在训练时还可能产生梯度爆炸、梯度消失等问题。LSTM 对这些问题进行改进，同时增强了模型对长距离信息的获取能力。与一般循环神经网络的主要区别在于，LSTM 中的“门”机制可以判断特定的信息是否有用。

LSTM 单元结构如图 3 所示。LSTM 能够记忆长期依赖关系，具备该能力的关键在于 LSTM 中除了记忆单元外还有 3 个“门”单元，包括输入门、输出门与遗忘门。输入门决定哪些信息会输入到记忆单元中，输出门决定信息是否需要从记忆单元中输出，遗忘门则在学习过程中判断哪些信息需要丢弃。

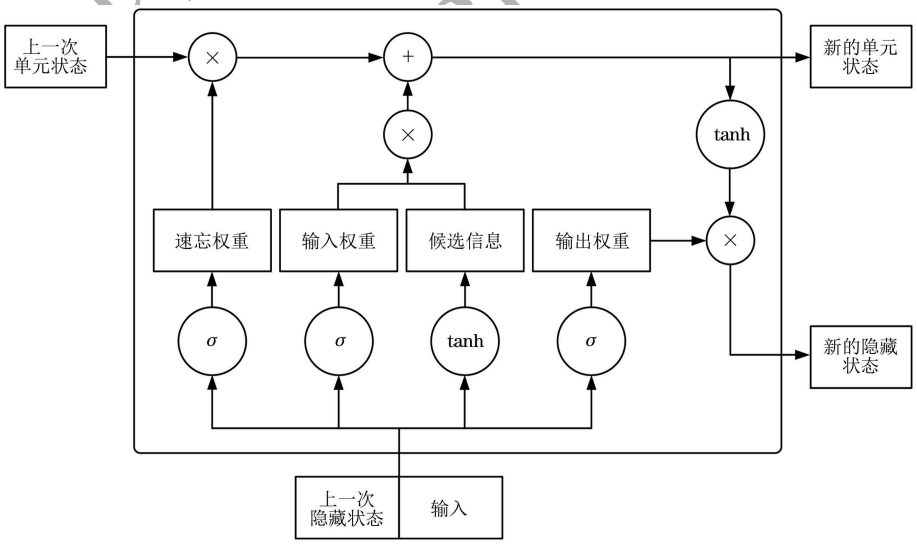

图 3 LSTM 单元结构

Fig. 3 LSTM unit structure

由于 LSTM 只能从前往后单方向地逐步处理文本，而在序列标注任务中，与每个 token 相联系的下文信息也有很重要的参考价值，实体的标签会同时受到历史与未来两个方向信息的影响。为了同时获取每个 token 的上下文信息，本文在这个阶段选择使用两个 LSTM 层组成的 BiLSTM 模型，两个方向的 LSTM 层分别用来提取前向和后向的相关性，将文本的词向量输入 BiLSTM 层中即可得到双向隐藏层的语义特征，将两者相拼接即可得到全局语义特征，既包括过去的显式信息也包括未来的隐藏信息。

BiLSTM 实现了对上下文信息的有效提炼与整合，保证了后续模型接收到的向量信息准确且充分。除此之外，BiLSTM 不仅在 Wobert 提供的全局上下文信息中提取了局部特征，同时还进一步融入了相对位置信息，强调每个字符的局部上下文信息，降低了长距离信息的干扰。NER 任务中的距离和方向的区分也非常重要，距离实体较近的字与当前实

体有关的可能性更高，当前字的信息也可以帮助模型预测左右两侧字符的类型，因此，选择 BiLSTM 作为编码器也在一定程度上提升了文本中零散实体标签序列的预测准确率。

2.4 多头自注意力机制

本文方法需要融合字词两种不同的信息要素，由于字符细粒度与词汇细粒度的两种文本信息通过 BiLSTM 层编码之后得到的向量具有相同的分配权重，如果将两者的信息视为同样重要，必然会降低向量信息的有效利用率，因此，进一步区分不同条件下的字词特征重要性非常有必要。通过实验分析，本文选择使用注意力机制对词汇细粒度的向量进行进一步提炼，再将其与字符细粒度的向量进行融合，可以使得两个不同细粒度级别的向量之间产生更为丰富的交互空间建模。

注意力机制源自于人类的视觉神经系统，当观察某个事物时，人们的注意力会主动聚焦于观测主体重点关注的目标区域，同时降低其他区域信息的接收能力，以此筛选出最具价值的信息。注意力机制能够在接收到上层模型输入向量后，进一步筛选训练当前任务重点关注的部分信息，赋予该部分更高的权重。本文模型利用多头自注意力机制构建一个融合词汇与字符不同级别信息的神经网络，可以同时读取整个文本信息，从每个级别中都能获得足够的上下文信息。在训练过程中，多头自注意力可以基于数据集来细化和更新其参数，从而更好地学习训练集的语义信息。通过注意力机制对词语级别的嵌入向量进行提纯，找到最适合融入字符级别词向量的信息，以此增强嵌入向量的语义信息体量与质量。注意力机制的计算公式如下：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为 query 向量、key 向量、value 向量；d 是 $\boldsymbol{Q}$、$\boldsymbol{K}$ 的维度；$\sqrt{d}$ 是惩罚因子，用于限制 $\boldsymbol{Q}$、$\boldsymbol{K}$ 的内积过大，防止梯度消失。

当嵌入文本序列的向量维度较高时，单次注意力无法有效地从过高的维度中提取有效信息。多头注意力机制允许多个头独立计算注意力矩阵，这相当于词向量分割后独立地并行计算注意力矩阵。最后，将每个头部的输出结果进行链接，从不同维度中获取更有效的信息：

$$\boldsymbol{h}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_{q_i},\boldsymbol{K}\boldsymbol{W}_{k_i},\boldsymbol{V}\boldsymbol{W}_{v_i}) \tag{2}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)\boldsymbol{W}_{\mathrm{o}} \tag{3}$$

式中：$\boldsymbol{W}_{q_i}$、$\boldsymbol{W}_{k_i}$、$\boldsymbol{W}_{v_i}$ 与 $\boldsymbol{W}_{\mathrm{o}}$ 都为可训练的参数矩阵。

为了有效地预测从模型中学习到的权重，减少输出序列中的冗余和噪声数据，提高命名实体识别的准确性，本文将多头注意力机制添加至经过 BiLSTM 层后的输出向量中，并根据向量的不同重要程度分配权重，通过调整 BiLSTM 输出概率矩阵的权重，使模型更关注关键信息，并动态平衡输入序列的局部特征，进一步从字符与词汇两个级别的向量中提取对模型预测更有效的特征信息，构成新的文本序列特征表示。相较于单纯地拼接词汇与字符两个不同的向量，多头注意力机制充分利用了数据之间的字符词汇相关性，提高了文本中潜在的关键词对实体识别结果的影响，同时通过剔除冗余信息提高了模型的训练效率，也进一步深化了词汇级与字符级的多粒度嵌入之间的信息交互。多头自注意力模型如图 4 所示。

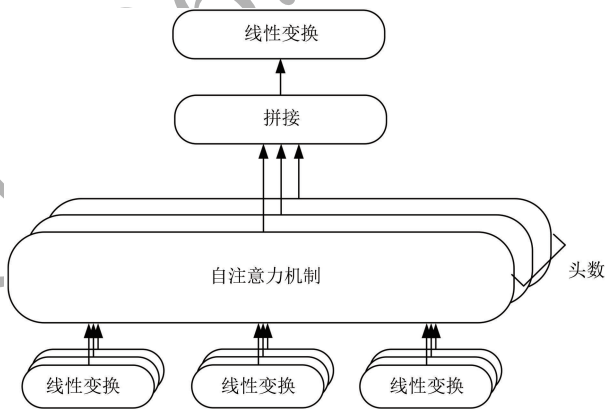

图 4 多头自注意力模型

Fig. 4 Multi-head self-attention model

2.5 CRF 层

CRF 是在最大熵模型和隐马尔可夫模型的基础上提出的一种判别概率、无定向图学习模型，一般用于序列数据分析。由于 BiLSTM 层只能获得上下文信息表示之间的关系，Attention 层也不会考虑序列文本预测的连续标签之间的依赖关系，因此本文选择将 Attention 层输出的向量输入至 CRF 层进行标签序列预测，CRF 能学习到相邻标签之间的约束关系，得到更优的预测序列，避免违背常理的标签序列出现，提高模型预测的准确性与合理性，例如“I”标签只可能出现在“B”标签之后，不能作为标签序列的开头。

对于给定的文本序列 $X=(x_1,x_2,\cdots,x_n)$ 与其对应的预测标签序列 $Y=(y_1,y_2,\cdots,y_n)$，定义 CRF 的评估分数为：

$$S(X,Y)=\sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_{i+1}}+\sum_{i=0}^{n}\boldsymbol{P}_{i,y_i} \tag{4}$$

式中：$\boldsymbol{A}$ 和 $\boldsymbol{P}$ 分别是 CRF 内的转移得分矩阵和上层输出的得分矩阵。使用 Softmax 函数将所有可能存在的预测序列归一化，即可得到输出概率：

$$P(Y \mid X)=\frac{e^{S(X,Y)}}{\sum_{\tilde{Y}\in Y_x} e^{S(X,\tilde{Y})}} \tag{5}$$

式中：Y_x 表示文本序列 X 对应的所有可能存在的预测序列。至此，则可以通过对数似然得到最后的损失分数：

$$L_{\text{loss}}=-\sum_{t\in T}\ln(P(Y_t \mid X_t)) \tag{6}$$

式中：T 表示训练数据中所有句子的集合；X_t 和 Y_t 分别代表句子 t 所对应的输入与预测序列。在解码阶段使用维特比(Viterbi)算法得到的分数最高的标签序列即是预测结果。

CRF 层可以从训练集中获取约束性规则，降低非法标签序列出现在预测结果中的概率，以保证模型最后输出的预测标签具有合理性。

2.6 对抗学习

由于字符细粒度与词汇细粒度来源于同一文本，不可避免地存在一定比例的冗余重复信息，这部分信息在模型训练过程中会导致特定参数更新速度过快，进而降低模型的鲁棒性。因此，本文考虑在词嵌入层添加扰动因子以生成对抗样本，引入噪声用于平衡局部重复信息带来的负面影响，增强模型的抗干扰能力，进而提升模型的预测能力。

对抗训练是一种在训练过程中引入噪声的方法，通过对模型的可学习参数进行正则化来提升模型的泛化能力。对抗训练的假设是：在给输入加上扰动后，输出分布和原来的分布一致。本文使用的对抗学习策略为 FGM(Fast Gradient Method)，FGM 在 Wobert 词向量上添加的扰动为：

$$r_{\text{adv}}=\frac{\varepsilon\times \boldsymbol{g}}{\|\boldsymbol{g}\|_2} \tag{7}$$

式中：$\boldsymbol{g}$ 为梯度；本文将 ε 设置为固定的值 1.0。

在模型训练时根据词向量矩阵的梯度计算得到 r_{adv}，与当前已有的词向量相加，使用融合向量再一次通过模型，将得到的扰动梯度累加到原梯度上后，再将词向量恢复到扰动之前的值，最后以累加后的梯度对模型参数进行更新。融合向量表示会与对抗样本一起发送到 BiLSTM 中进行训练。对抗样本将模拟数据集标签中的自然错误，使模型更能容忍由错误标签引起的参数波动，提高了模型鲁棒性。对抗学习对字符与词语两个级别的 Wobert 词向量进行扰动，能明显提升泛化性，使得模型在测试集上的预测能力得到显著提升。

3 实验结果与分析

3.1 数据集与评价指标

3.1.1 数据集

本文将在公开使用的中文数据集 Resume 和瓷器领域的自建数据集 Porcelain 上进行一系列的对比实验与消融实验，以验证本文模型的有效性。Resume 数据集是根据新浪财经网上的上市公司中高级经理人的简历摘要而制作的数据集，通过筛选过滤与人工标注而构建。Resume 数据集包含 1 027 份简历摘要，实体标注分为人名、国籍、籍贯、种族、学位、专业、机构、职称等 8 个类别。瓷玉器数据集是根据雅昌艺术网中的瓷器拍品的描述数据而制作的数据集，通过筛选清洗并经由专业数据标注机构的标注与验证构建而成。瓷器数据集包含 6 000 条训练集、1 029 条验证集和 1 029 条测试集数据，实体标注共有 3 种，包括瓷器釉种、瓷器纹饰与象征寓意。上述数据集均采用“BIO”标注法，文本中的所有非实体字符会被统一标记为“O”，而对于需要特殊标注的实体，将实体的第一个字标注为“B-实体名称”，其余部分均标注为“I-实体名称”。两个数据集的详细信息如表 1 所示，Porcelain 数据集部分展示如图 5 所示，Resume 数据集部分展示如图 6 所示。

表 1 数据集详细信息

Table 1 Dataset details 单位：个

数据集	训练集数	验证集数	测试集数
Resume	3 821	463	477
Porcelain	6 000	1 029	1 029

示例文本	盘 O 内 O 壁 O 一 O 周 O 以 O 粉 O 彩 O 绘 O 折 O 枝 O 四 O 季 O 花 O 卉 O 及 O 蟠 B-WEN 桃 I-WEN 纹 I-WEN，O 寓 O 意 O 四 B-HAN 季 I-HAN 富 I-HAN 贵 I-HAN 长 I-HAN 寿 I-HAN，O 盘 O 心 O 椭 O 圆 O 开 O 光 O 内 O 以 O 矾 O 红 O 描 O 绘 O 双 B-WEN 龙 I-WEN 戏 I-WEN 珠 I-WEN 图 I-WEN

图 5 Porcelain 数据集部分展示

Fig.5 Partial display of Porcelain dataset

示例文本	2 O 0 O 0 O 7 O 年 O 7 O 月 O 至 O 2 O 0 O 1 O 2 O 年 O 6 O 月 O 任 O 原 O 平 B-ORG 安 M-ORG 银 M-ORG 行 E-ORG 副 B-TITLE 行 M-TITLE 长 E-TITLE，O 并 O 自 O 2 O 0 O 0 O 7 O 年 O 6 O 月 O 至 O 2 O 0 O 1 O 2 O 年 O 6 O 月 O 任 O 原 O 平 B-ORG 安 M-ORG 银 M-ORG 行 E-ORG 执 B-TITLE 行 M-TITLE 董 M-TITLE 事 E-TITLE 。O

图 6 Resume 数据集部分展示

Fig.6 Partial display of Resume dataset

3.1.2 评价指标

本文模型使用的评价指标包括准确率(P)、召回率(R)和 F1 值(F_1)。准确率表示识别正确的实体占识别出的全部实体的比例,召回率表示识别正确的实体占样本标注实体总数的比例,F_1 是结合准确率和召回率的综合评价指标。各评价指标的计算公式如下:

$$P=\frac{T_{TP}}{T_{TP}+F_{FP}} \tag{8}$$

$$R=\frac{T_{TP}}{T_{TP}+F_{FN}} \tag{9}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{10}$$

式中:T_{TP} 为真正例的数量;F_{FP} 为假正例的数量;F_{FN} 为假反例的数量。

3.2 实验环境

本文的实验均在 Python 3.8、PyTorch 1.11.0 下运行,采用的 CPU 为 Intel Xeon Platinum 8260C,GPU 为 NVIDIA GeForce RTX 3090,CUDA 版本为 11.3。

3.3 实验参数设置

本文模型使用 Wobert 预训练模型生成词向量,表 2 列出了实验模型的不同模块所设置的超参数。

表 2 模型参数设置

Table 2 Model parameter settings

参数	设置
训练轮数	40
Batch size	16
Wobert 词嵌入维度	768
最大句子长度	128
LSTM 隐藏层维度	192
LSTM 层数	2
Self-Attention 头数	32
Dropout 率	0.1
Wobert 学习率	2×10^{-5}
其他模块学习率	1×10^{-3}
优化器	Adam(betas 为(0.9,0.9))
权重衰减率	0.01
学习率衰减率	0.01

3.4 对比实验

3.4.1 与其他模型的对比实验

为验证模型的有效性,在 Resume 数据集上将本文模型与 Lattice-LSTM、LGN、LR-CNN、SoftLexicon、NFLAT 模型进行实验对比。对比模型具体如下:

1)Lattice-LSTM[19]。一种网格状(Lattice)的 LSTM,用来获取句子中的字典词(Lexicon word)的表征,将潜在的词汇信息插入基于字符的 LSTM-CRF 中。

2)LGN[27]。在图神经网络的框架上添加词汇信息,借助图神经网络独有的结构,将实体识别任务重新建模为节点分类任务。

3)LR-CNN[28]。使用 Rethinking 机制在 Feedback 层中添加更高层次的特征,同时通过调整字典词汇之间的权重,解决字典词汇存在的词汇冲突问题。

4)SoftLexicon[20]。使用向量形式来表征字典信息,提高了字典模块与其他神经网络模块的兼容性。

5)NFLAT[29]。具有多头 Inter-attention 的网格结构,能够同时对不同长度的词汇与字符建模。

6)BSNER[30]。一种边界平滑的方法,将预测实体的概率从实体边界分配至边界周围,使得模型产生了更平滑的预测结果。

从表 3 可以看出:Lattice-LSTM 在词汇的结尾字符引入词汇信息进行融合,提升了模型的精度,证明了文字字符词汇信息的融合可以提升模型识别效果,但是在词汇信息完整度与利用率方面仍有提升空间;LR-CNN 与 LGN 分别基于卷积神经网络与图神经网络对局部特征进行提取,与 BiLSTM 和注意力机制的结合相比,它们在局部信息的提取与融合方面能力仍然有限;SoftLexicon 虽然对字符表示层进行细微调整并引入了词典信息,但其仍然没有充分利用词汇信息;NFLAT 与 BSNER 也从不同的角度去增强词汇与字符的信息融合程度,提升了模型预测的准确率;本文模型相较之下可以更有效地利用词汇信息,减轻了分词算法误差的影响,缓解了词汇信息在融合过程中的损失以及利用不充分的缺陷,从而进一步地提高了模型识别效果。在 Resume数据集上,本文模型的准确率、召回率、F1

表 3 模型对比结果

Table 3 Comparison results of models %

对比模型	P	R	F_1
Lattice-LSTM(2018 年)	94.81	94.10	94.46
LR-CNN(2019 年)	95.37	94.84	95.11
LGN(2019 年)	95.28	95.46	95.37
SoftLexicon(2020 年)	96.08	96.13	96.11
NFLAT(2022 年)	95.63	95.52	95.58
BSNER(2022 年)	96.63	96.69	96.66
本文模型	97.51	96.90	97.21

值分别为 97.51%、96.90%、97.21%，相较于目前实体识别效果较好的 LGN、SoftLexicon、NFLAT、BSNER 模型，本文模型的 F1 值分别提升了 1.84、1.10、1.63、0.55 个百分点，表明本文模型的中文命名实体识别效果更好。

此外，为了进一步验证本文方法的先进性，将本文模型与最近在中文命名实体识别领域的相关研究进行对比分析，训练语料均采用 Resume 数据集。由表 4 可知，4 种对比的中文命名实体识别模型都是以预训练模型获得全局特征，然后通过不同的神经网络结构提取局部特征，以及融合词汇、字符、词汇边界等信息，以提升模型在中文命名实体识别领域的识别效果，但是性能均逊色于本文模型，表明本文方法能够更有针对性地融合词汇与字符两个层级的信息特征。

表 4 本文模型与最近研究成果的对比

Table 4 Comparison between the model proposed in this paper and recent research findings %

对比模型	P	R	F_1
CMH[31]（2023 年）	96.79	96.86	96.83
R-DBBC[32]（2023 年）	96.91	97.49	97.20
BIFT[33]（2023 年）	95.97	96.50	96.23
MGA_CV[34]（2023 年）	95.23	96.06	95.64
本文模型	97.51	96.90	97.21

3.4.2 消融实验

为了验证模型的有效性，在 Resume 数据集与 Porcelain 数据集上针对各模块进行消融实验，结果如表 5 与表 6 所示。

表 5 Resume 数据集上的消融实验结果

Table 5 Ablation experiment results on Resume dataset %

对比模型	P	R	F_1
完整模型	97.51	96.90	97.21
去掉对抗学习模块的模型	97.20	96.84	97.02
去掉词汇信息模块的模型	97.01	96.53	96.77
去掉上述 2 个模块的模型	96.76	96.04	96.40
去掉 BiLSTM 模块的模型	95.68	96.52	96.10
去掉自注意力模块的模型	96.69	95.85	96.27

表 6 Porcelain 数据集上的消融实验结果

Table 6 Ablation experiment results on Porcelain dataset %

对比模型	P	R	F_1
完整模型	85.70	88.36	87.01
去掉对抗学习模块的模型	85.10	87.91	86.48
去掉词汇信息模块的模型	86.35	87.30	86.82
去掉上述 2 个模块的模型	84.48	87.49	85.96
去掉 BiLSTM 模块的模型	83.30	87.34	85.27
去掉自注意力模块的模型	85.00	85.95	85.47

由表 5 与表 6 看出：对抗学习与词汇预训练信息的引入，在 2 个实验数据集上均对准确率、召回率有大幅提升，证明了通过 Wobert 预训练模型引入词汇预训练词向量对实体识别的有效性，同时对抗学习进一步增强了 2 种不同层级的词向量的融合，两者相辅相成对模型性能提升较大；BiLSTM 与多头注意力机制也在很大程度上影响了模型对于局部特征的提取与整合能力，如果去掉会大幅降低模型的实体识别精准度。综上所述，本文提出的基于 Wobert 与对抗训练的中文实体识别模型通过多个模块的深度交互融合，实现了相对较好的实体识别能力。

4 结束语

针对目前中文命名实体识别领域词汇信息获取方法复杂、词汇信息利用率不高且融合程度低的问题，本文提出一种基于 Wobert 与对抗学习的中文命名实体识别模型。首先通过 Wobert 获取字符与词汇两种不同细粒度的词向量，随后使用 BiLSTM 模块提取文本的上下文语义特征，通过多头自注意力机制进一步更新词汇信息与字符信息的权重比，两者深度融合后由 CRF 进行最后的约束，同时在训练过程中配合 FGM 对抗学习增强 Wobert 词向量的泛化性。通过对比实验与消融实验验证了本文模型的有效性以及相比其他模型的优势。在未来的工作中，尝试将模型扩展至更大规模的命名实体识别数据集，在保持模型性能的情况下减少模型训练时间。

参考文献

[1] TJONG KIM SANG E F, DE MEULDER F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. [S. l.]: Association for Computational Linguistics, 2003: 142-147.

[2] HUANG S B, SHA Y P, LI R S. 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.

[3] LI L Y, DAI Y, TANG D Y, et al. MarkBERT: marking word boundaries improves Chinese BERT[EB/OL]. [2023-07-05]. https://arxiv.org/abs/2203.06378.

[4] ZHU Y, WANG G. CAN-NER: convolutional attention network for Chinese named entity recognition[C]// Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2019: 3384-3393.

[5] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the

Association for Computational Linguistics, 2016, 4: 357-370.
[6] SAITO K, NAGATA M. Multi-language named-entity recognition system based on HMM[C]//Proceedings of ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition. [S. l.]: Association for Computational Linguistics, 2003: 41-48.
[7] FENG Y, SUN L, LV Y. Chinese word segmentation and named entity recognition based on conditional random fields models[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. [S. l.]: Association for Computational Linguistics, 2006: 181-184.
[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2019: 4171-4186.
[9] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing [C]// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. [S. l.]: Association for Computational Linguistics, 2020: 657-668.
[10] LI D, LONG J, QU J, et al. Chinese clinical named entity recognition with ALBERT and MHA mechanism [J]. Evidence-Based Complementary and Alternative Medicine, 2022, 2022: 2056039.
[11] 张付领. 结合 ERNIE2.0 的医疗中文命名实体识别模型[J]. 电子设计工程, 2023, 31(4): 38-42.
ZHANG F L. Medical Chinese named entity recognition model combined with ERNIE2.0 [J]. Electronic Design Engineering, 2023, 31(4): 38-42. (in Chinese)
[12] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2023-07-05]. http://arxiv.org/abs/1907.11692v1.
[13] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2023-07-05]. https://arxiv.org/abs/1508.01991.
[14] LIU S, YANG H, LI J Y, et al. Chinese named entity recognition method in history and culture field based on BERT [J]. International Journal of Computational Intelligence Systems, 2021, 14(1): 163.
[15] ZHAO Z, YANG Z, LUO L, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network[J]. BMC Medical Genomics, 2017, 10(5): 73.
[16] AGUILAR G, MAHARJAN S, LÓPEZ MONROY A P, et al. A multi-task approach for named entity recognition in social media data[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text. [S. l.]: Association for Computational Linguistics, 2017: 148-153.
[17] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2016: 1480-1489.
[18] LIU W, FU X Y, 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. [S. l.]: Association for Computational Linguistics, 2021: 5847-5858.
[19] LARSEN B, AONE C. Fast and effective text mining using linear-time document clustering[C]//Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 1999: 16-22.
[20] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 5951-5960.
[21] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 6836-6842.
[22] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. [2023-07-05]. http://arxiv.org/abs/1406.2661v1.
[23] CAO P F, CHEN Y B, 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. [S. l.]: Association for Computational Linguistics, 2018: 182-192.
[24] YASUNAGA M, KASAI J, RADEV D. Robust multilingual part-of-speech tagging via adversarial training [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2018: 976-986.
[25] XU Z. RoBERTa-wwm-ext fine-tuning for Chinese text classification[EB/OL]. [2023-07-05]. http://arxiv.org/abs/2103.00492v1.
[26] 苏剑林. 提速不掉点: 基于词颗粒度的中文 WoBERT[EB/OL]. [2023-07-05]. http://kexue.fm/archives/7758.
SU J L. Speed up without dropping points: Chinese WoBERT based on word granularity[EB/OL]. [2023-07-05]. http://kexue.fm/archives/7758. (in Chinese)
[27] GUI T, ZOU Y C, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2019: 1040-1050.
[28] GUI T, MA R T, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. [S. l.]: Association for Computational Linguistics, 2019: 4982-4988.
[29] WU S, SONG X N, FENG Z H, et al. NFLAT: non-flat-lattice transformer for Chinese named entity recognition [EB/OL]. [2023-07-05]. http://arxiv.org/abs/2205.05832v3.
[30] ZHU E W, LI J P. Boundary smoothing for named entity recognition[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2022: 7096-7108.
[31] 于右任, 张仰森, 蒋玉茹, 等. 融合多粒度语言知识与层级信息的中文命名实体识别模型[J]. 计算机应用, 2024, 44(6): 1706-1712.
YU Y R, ZHANG Y S, JIANG Y R, et al. Chinese named entity recognition model incorporating multi-granularity linguistic knowledge and hierarchical information[J]. Journal of Computer

Applications,2024,44(6):1706-1712. (in Chinese)
[32] 杨长沛,廖列法. 基于门控空洞卷积特征融合的中文命名实体识别[J]. 计算机工程, 2023, 49(8): 85-95.
YANG C P, LIAO L F. Chinese named entity recognition based on dilated gated convolution feature fusion [J]. Computer Engineering, 2023, 49(8): 85-95. (in Chinese)
[33] 廖梦,贾真,李天瑞. 基于标签信息融合与多任务学习的中文命名实体识别[J]. 计算机科学,2024,51(3):198-204.
LIAO M, JIA Z, LI T R. Chinese named entity recognition based on label information fusion and multi-task learning[J]. Computer Science,2024,51(3):198-204. (in Chinese)
[34] 王庆人,王银子,仲红,等. 面向中文的字词组合序列实体识别方法[J]. 清华大学学报(自然科学版),2023,63(9):1326-1338.
WANG Q R, WANG Y Z, ZHONG H, et al. Chinese-oriented entity recognition method of character vocabulary combination sequence [J]. Journal of Tsinghua University (Science and Technology),2023,63(9):1326-1338. (in Chinese)

编辑 吴云芳