ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 4,Jul. 2024,pp. 1020－1032
DOI:10. 16337/j. 1004-9037. 2024. 04. 020

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

提示学习框架下融合多层级特征信息的中文命名实体识别

王　昕[1]，魏楚元[2]，张　蕾[2]，万珊珊[2]

(1. 北京建筑大学机电与车辆工程学院,北京 102616; 2. 北京建筑大学电气与信息工程学院,北京 102616)

摘　要：目前基于预训练-微调模式下的命名实体识别任务预训练与微调之间会出现差距,难以有效地对实体与上下文之间的关系进行建模,并且当前中文命名实体识别方法不能获取足够的字形或词义。针对上述问题,本文提出一种基于提示学习且融合多层级特征信息的命名实体识别方法。首先根据提示学习机制构建提示文本,再将输入文本的字符、词和实体级别特征信息与之拼接作为预训练模型的输入,以有效捕捉上下文之间的语义信息,缩小预训练模型与下游任务之间的差距,提高模型对命名实体识别的感知能力。本文提出的方法充分利用先验知识,提升模型的学习质量,提高在中文复杂多变语义环境下命名实体识别的效果。在人民日报、MSRA、Weibo、Resume和CMeEE数据集上的F_1值分别达到了97.09%、96.68%、83.44%、97.48%和76.05%。实验结果表明,本文提出方法总体优于目前主流的中文命名实体识别方法。

关键词：命名实体识别;语义特征;提示学习;多层级特征信息

中图分类号：TP391.1; TP183　　**文献标志码**：A

Chinese Named Entity Recognition Based on Prompt Learning and Multi-level Feature Fusion

WANG Xin[1], WEI Chuyuan[2], ZHANG Lei[2], WAN Shanshan[2]

(1. School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; 2. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China)

Abstract: The current named entity recognition task based on the pre-training-fine-tuning model has a gap between pre-training and fine-tuning, which makes it difficult to effectively model the relationship between entities and contexts, and the current Chinese named entity recognition methods cannot obtain sufficient character or word meanings. To address above problems, this paper proposes a named entity recognition method based on cue learning and incorporating multi-level feature information. Firstly, the cue text is constructed based on the cue learning mechanism, and then the character, word and entity-level feature information of the input text is spliced with it, which is taken as the input of the pre-trained model to effectively capture the semantic information between the contexts, narrow the gap between the pre-trained model and the downstream task, and improve the perceptive ability of the model for named entity recognition. The proposed method makes full use of prior knowledge to increase the learning ability of the model and improve the effectiveness of named entity recognition in the complex and variable semantic

基金项目:教育部人文社会科学研究一般项目(22YJAZH110);北京市教育科学"十四五"规划项目(CHAA22061)。
收稿日期:2023-03-19;**修订日期**:2023-08-10

environment of Chinese. The F_1 values reach 97.09%, 96.68%, 83.44%, 97.48% and 76.05% on the People's Daily, MSRA, Weibo, Resume and CMeEE datasets, respectively. Experimental results show that the proposed method is generally better than the current mainstream Chinese named entity recognition methods.

Key words: named entity recognition; semantic feature; prompt learning; multi-level feature information

引　　言

作为自然语言处理领域中的基础任务之一,命名实体识别(Named entity recognition,NER)能够对文本中具有特殊意义的实体进行检测,比如人名、地名以及组织名称等。近些年,NER技术在自然语言处理领域引起了广泛的兴趣和关注。一些研究者将该技术用于多种自然语言处理的下游任务中,比如文本摘要[1]、情感分析[2]和机器翻译[3]等。对于中文命名实体识别来说,文本中词语含义丰富,而单个字符并不能很好地表征词汇信息,并且汉字在不同语境中往往具有不同语义,很容易忽视字与句之间的隐含逻辑信息。

针对上述问题,本文采用了自然语言处理(Natural Language process, NLP)领域的第四范式,即提示学习(Prompt learning),并且融合多层级特征用于命名实体识别任务上,将预训练模型并不了解的命名实体识别任务转换为输出空间有限的、模型熟悉的掩码语言模型(Masked language model, MLM)任务。这样不仅可以规避参数过多导致的异步问题,还有助于减少预训练任务与微调之间的差异,让模型更好地适应中文文本。在提示学习框架基础上,利用不同的掩码机制分别对字、词和实体进行掩码,进而将上述得到的字、词和实体级别的特征信息融合后送入双向长短期记忆网络(Bi-directional long short-term memory, BiLSTM)网络中,以此更加全面地表示句子的语义信息。最后,在应用提示学习的约束条件下,完成中文实体命名识别任务。

近些年来,基于深度学习的命名实体识别方式逐渐占据主导地位。Dong等[4]受到象形文字的启发,从结构角度拆解汉字,将其分解为不同的偏旁与部首,然后将这些偏旁部首作为字级别特征输入到BiLSTM。然而,这种方法仅仅关注字形结构,没有考虑其在上下文中的相关信息,从而难以准确地捕捉完整的实体信息。针对这一情形,Jia等[5]在此基础上融入卷积神经网络(Convolutional neural networks, CNN)来提取字形结构与语义等特征信息。但是中文汉字在不同的场景下有着不同的含义,即仅基于字级别特征的方法无法解决中文命名实体识别中的多样性问题。例如,"轧钢车间的工人很团结,没有相互倾轧的现象。"这一句话中第一个"轧"表示挤压,第二个"轧"表示排挤,这两个"轧"即为截然不同的含义。

为合理地解决一词多义的问题,Matthew等[6]利用数据对模型分别进行正反两个方向的训练,同时将句子分成单独的字符送入循环神经网络(Recurrent neural networks, RNN)中获得其隐层状态表示,最后与之前的网络隐层进行融合从而进行序列标注。盛剑等[7]以BiLSTM为基础,引入卷积神经网络以深化词语的局部特征提取,从而提出一种精细化的命名实体识别方法。他们借助网络词典对部分数据进行标注,获得相对粗粒度的文本数据。为减少噪声和冗余信息对实体识别的干扰,首先确定实体的大类标签,然后进一步细化命名实体的标签。然而,这两种方法忽略了句子中词语之间的潜在逻辑联系。Zhu等[8]提出了一种基于注意力机制(Attention)的卷积神经网络(Convolutional attention network,CAN)。该方法合理地运用全局与局部的注意力机制,更好地捕捉相邻字符和句子上下文的信息。此外,CAN与其他方法的不同之处在于它不依赖于任何外部资源,其字符嵌入方式在实际应用场景中更加实用。

伴随着提示学习的兴起，NLP任务的范式从以往的“预训练-微调”模式发展到“预训练-提示-预测”的新模式。Schick等[9]提出了PET(Pattern exploiting training)，其主要思路在于借助自然语言构建模板，将下游任务转化为完型填空任务，从而利用BERT(Bidirectional encoder representation from transformers)的MLM模型进行预测。这样做不仅使模型更充分地利用了先验知识，而且解决了中文场景下语义灵活、可学到知识不足的问题。

仅依赖字符或词级别特征的实体识别方法，无法有效地兼顾这两个层级特征的优势，从而难以准确地捕捉实体特征信息。本文提出的中文命名实体识别方法基于提示学习并融合多层级特征信息，更有效地解决这一挑战。首先借助提示学习思想，根据输入文本构建模板内容，将实体识别任务转为完型填空任务，更合理地利用先验知识，解决中文场景下语义灵活、可学到知识不足的问题。其次，为了同时获取词语内部的相关性信息，构建一种多层级掩码神经网络，将字符、词和实体特征信息拼接。最后将多层级的特征信息与提示学习融合，以全面地表示词的语义信息，进而提升模型性能。

本文的主要贡献可总结为以下3方面：(1)将提示学习机制用于命名实体识别任务，通过构建提示模板信息，拉近上下游任务距离，成功地捕捉句子和实体之间的隐含逻辑信息，从而增强了模型在命名实体识别任务中的感知能力；(2)获取字、词和实体级别的特征信息，利用BiLSTM网络捕获深度语义表征，赋予实体更大的权重，最后将其与提示学习框架相融合；(3)通过MSRA、Resume、Weibo、CMeEE和人民日报数据集的广泛实验，验证了所提出的方法在效果上优于当前主流方法，证明了本文方法的有效性。

1 模　　型

本文提出基于提示学习且综合考虑多层级特征信息的中文命名实体识别模型，框架如图1所示。首先将原始句子分别输入到向量表示层与模板构造层中，分别获得融合了多层级特征的向量与含有提示信息的向量；然后将上述向量整合为预训练模型的输入，随后通过标签映射层进行转化，得到最终的实体识别结果。

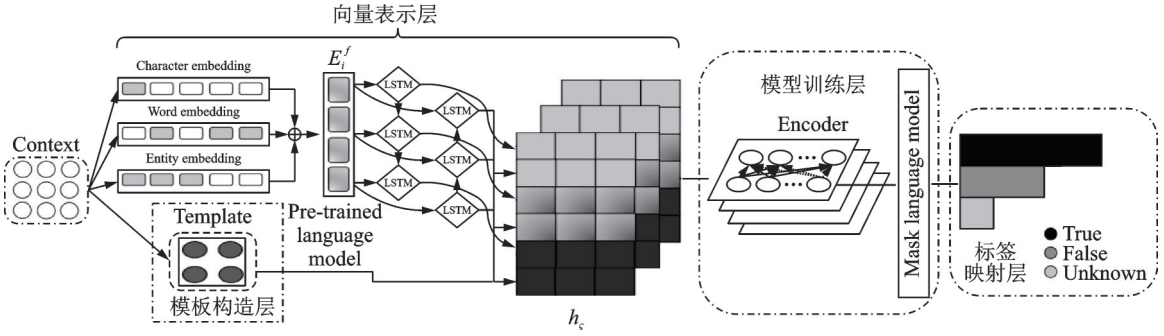

图1　本文模型框架图

Fig.1　Frame diagram of the proposed model

1.1 向量表示层

假设文本输入为$E_i(i=1,2,\cdots,n)$，其中i表示文本中的第i个字符。在获得字符级特征表示时采用WordPiece分词方法。将输入的文本按照字母和字符组成的子词进行切割。具体而言，首先将输入文本进行分词，将文本切分成一系列的子词。例如，将单词“playing”切分成“play”与“##ing”两个子词。这样不仅能够处理未登录词(Out of vocabulary, OOV)和词形变化等问题，还可以利用更细粒度的字级别信息。分词后，将每个子词映射到一个固定的字向量，然后将这些子词的字向量输入到模型进行处

理。通过多层的Transformer编码器,模型能够学习到文本E_i的字级别特征表示为$\boldsymbol{E}_i^{\mathrm{c}}$。由于仅考虑字级别特征不能很好地捕获词间所隐含的信息,本文使用SentencePiece进行中文词表扩展,在中文语料库上训练一个中文分词模型,然后将中文Tokenizer与模型的Tokenizer进行合并,通过组合它们的词汇表,最终获得一个合并后的分词模型,再进行词级别的分词,模型学习后得到词级别特征$\boldsymbol{E}_i^{\mathrm{w}}$。同时,采用Spacy对句中实体进行粗略的切分,并获得带有标记的分词结果:"[ent_start][ent][ent_end]",依据此标签模型学习后获得实体级别特征$\boldsymbol{E}_i^{\mathrm{ent}}$。最后将词级别特征与实体级别特征一并融入,所表示的融合向量为

$$\boldsymbol{E}_i^{\mathrm{f}}=\boldsymbol{E}_i^{\mathrm{c}}\oplus\boldsymbol{E}_i^{\mathrm{w}}\oplus\boldsymbol{E}_i^{\mathrm{ent}} \tag{1}$$

单向LSTM仅关注了输入文本序列的前向信息。为了更全面地考虑信息,本文采用双向LSTM,同时兼顾前、后向序列信息,使得文本中字符可以获得全局信息。将式(1)中融合了字符、词和实体特征信息的向量作为双向LSTM的输入,首先输入到遗忘门中,保留重要信息,即

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{W}_{\mathrm{f}}\left[\boldsymbol{h}_{t-1},\boldsymbol{E}_t^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{f}}\right) \tag{2}$$

式中:σ为sigmoid激活函数;$\boldsymbol{W}_{\mathrm{f}}$为遗忘门的权重矩阵;$\boldsymbol{h}_{t-1}$为$t-1$时刻的输出;$\boldsymbol{E}_t^{\mathrm{f}}$表示$t$时刻的输入;$\boldsymbol{b}_{\mathrm{f}}$为遗忘门的偏置向量。进一步地,更新输入门信息,以及记忆单元状态,即

$$\boldsymbol{i}_t=\sigma\left(\boldsymbol{W}_{\mathrm{i}}\times\left[\boldsymbol{h}_{t-1},\boldsymbol{E}_t^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{inp}}\right) \tag{3}$$

$$\boldsymbol{C}_t=\tanh\left(\boldsymbol{W}_{\mathrm{c}}\times\left[\boldsymbol{h}_{t-1},\boldsymbol{E}_t^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{c}}\right) \tag{4}$$

式中:$\boldsymbol{i}_t$为输入门;$\boldsymbol{W}_{\mathrm{i}}$为输入门的权重矩阵;$\boldsymbol{b}_{\mathrm{inp}}$为输入门的偏置向量;$\boldsymbol{C}_t$为$t$时刻的状态;$\boldsymbol{W}_{\mathrm{c}}$为记忆单元的权重矩阵;$\boldsymbol{b}_{\mathrm{c}}$为记忆单元的偏置向量;tanh为双曲正切函数。接着捕获输出门的信息,获取最终隐状态向量,即

$$\boldsymbol{O}_t=\sigma\left(\boldsymbol{W}_{\mathrm{o}}\times\left[\boldsymbol{h}_{t-1},\boldsymbol{E}_t^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{oup}}\right) \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{O}_t*\tanh\left(\boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\tanh\left(\boldsymbol{W}_{\mathrm{c}}\times\left[\boldsymbol{h}_{t-1},\boldsymbol{E}_t^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{c}}\right)\right) \tag{6}$$

式中:$\boldsymbol{W}_{\mathrm{o}}$为输出门的权重矩阵;$\boldsymbol{b}_{\mathrm{oup}}$为输出门的偏置向量;本文将经过LSTM网络获取的前、后向隐向特征分别表示为$\boldsymbol{h}_{\mathrm{f}}$、$\boldsymbol{h}_{\mathrm{b}}$,从而得到最终的隐层表示,即

$$\boldsymbol{h}_{\mathrm{f+b}}=[\boldsymbol{h}_{\mathrm{f}},\boldsymbol{h}_{\mathrm{b}}] \tag{7}$$

1.2　模板构造层

模板构造层的主要职责在于应用提示学习机制,创建带有提示信息的模板。引入提示信息的目的不仅在于帮助模型获取上下文与实体之间的隐含逻辑信息,还可以拉近下游任务与预训练模型之间的距离,充分发挥MLM的能力。本文采用混合模板构建的方式,提示可以十分灵活,可以将其插入到上下文中,具体设计流程如图2所示。假设原始输入文本为X,首先指定模板初始化的值表示为S,S_i指代初始模板S中第i个提示字符。本文初始化时所定义模板表示为

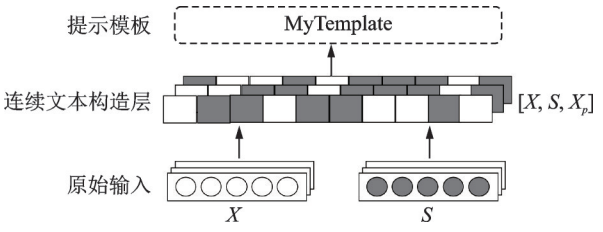

图2　提示文本构建流程图

Fig.2　Prompt text building flowchart

$$\begin{cases}\mathrm{MyTemplate}=\text{文本}X_\alpha\text{的实体},X_\beta\text{是否包含?}\\ \text{回答(Ans):[MASK]}\end{cases} \tag{8}$$

式中:汉字部分为指定初始值的连续模板信息;X_α为输入文本切分为字符形式的信息;X_β为原输入文本

信息。

传统的使用离散prompt搜索方法是直接将模板中每个token映射为对应的embedding，然后为整个模板生成一个得分，本文将模板中的S_i映射为一个可训练的参数$\boldsymbol{h}_i$，$\boldsymbol{h}_i(0\leqslant i<m)$为可训练的嵌入张量。这部分token称为soft-prompt token，在后续优化过程中，因soft-prompt token存在序列关系，所以使用双向LSTM对模板中的序列进行表征，最后采用梯度下降法更新连续的参数，获得与下游任务关联程度较高的提示模板X_p，模板自动更新。上述过程如式(9)所示。

$$\{\boldsymbol{h}_0,\boldsymbol{h}_1,\cdots,\boldsymbol{h}_i,X,\boldsymbol{h}_{i+1},\cdots,\boldsymbol{h}_m,\text{Ans}\} \tag{9}$$

1.3 模型训练层

本文采用RoBERTa模型作为模板层的学习器。模型训练层主要由堆叠在一起的Transformer编码器组成，每个编码器包含多头自注意力层和前馈神经网络层两个子层。多头自注意力机制相当于多个不同的自注意力模型组合在一起，不同的自注意力头可以抽取不同的特征，提升句子中关键信息的权重。图3展示了本文模型训练层中编码层的网络结构。

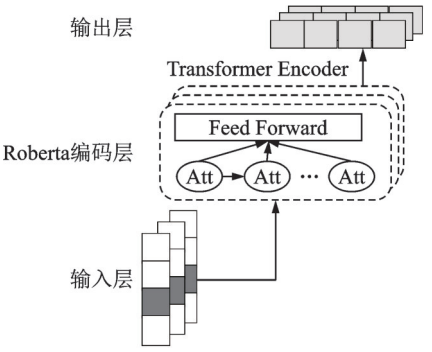

图3 编码层网络结构

Fig.3 Encode layer network structure

本文将问题的定义设定为在连续空间向量中寻找最佳向量，以最小化标注样本上的损失，具体如下。

定义语言模型为$\boldsymbol{\Upsilon}$，原始输入文本表示为$E=\{E_0,E_1,\cdots,E_n\}$，标签对应的词为K，模板为$T=\{T_0,T_1,\cdots,T_u\}$，定义$\boldsymbol{\Phi}(E)$表示E的词嵌入，$\boldsymbol{\Phi}(K)$为K的词嵌入，因此输入方式为

$$\text{Input}=\{\boldsymbol{\Phi}(E),\boldsymbol{\Phi}(T),\boldsymbol{\Phi}(K)\} \tag{10}$$

式中$\boldsymbol{\Phi}(T)$部分代表着可计算的模板向量，其相对于输入内容中原始文本和标签词的位置可以自由调整。首先初始化一个$(u,768)$维的向量，然后通过训练和随机梯度下降优化，以不断改进与模板部分相关的参数，最后利用下游损失函数Loss随机梯度下降，有

$$\boldsymbol{\Phi}(T_{0:u})=\arg\min_{T}\text{Loss}\left(\boldsymbol{\Upsilon}\left(\boldsymbol{\Phi}(E),\boldsymbol{\Phi}(K)\right)\right) \tag{11}$$

通过获取与提示模板相关联的向量后，将两种特征向量进行融合，有

$$\boldsymbol{h}_{\varsigma}=\text{concat}\left(\boldsymbol{h}_{r+1},\boldsymbol{\Phi}(T_{0:u})\right) \tag{12}$$

然后将其拼接到RoBERTa的词嵌入层中转化为词向量，通过预训练模型中的MLM任务对[MASK]标记进行概率预测。

1.4 MLM层

本文使用RoBERTa的MLM层来执行完型填空任务，预测句中缺失部分的概率对应于候选词集合，并评估插入的内容在文本中的语义合理性。具体过程如下所示。

定义候选词集$G=\{G_0,G_1,\cdots,G_z\}$，对于每个句子C，$\boldsymbol{\Upsilon}(C|G)$表示语言模型在空缺位置为$G$时的得分。进一步，每个候选词对应的得分有

$$\text{Logits}\left(G_j\right)=\boldsymbol{\Upsilon}\left(C|G_j\right) \tag{13}$$

最后，通过应用softmax函数，将这些得分转化为概率分布，以确定哪个词具有最高的概率，从而填入被掩码的部分。

1.5　标签映射层

在得到掩码词之后,将结果传送至标签映射层。标签映射层的主要任务是确定MLM任务中标签与候选词之间的关联关系。例如根据不同的下游任务将[positive,negative]、[体育,军事,…,金融]等集合作为完型填空的候选词,构建标签映射关系。

当预训练语言模型预测[MASK]部分在标签词集上的概率分布时,将标签词集K中的单词概率映射为原始标签集合的概率有

$$P\left(K|E_\rho\right)=\delta\left(P_\tau\left([\mathrm{MASK}]=K|E_\rho\right)|K\in K_y\right) \tag{14}$$

式中:δ表示将标签词集中每个单词在[MASK]位置的预测概率转换为原始标签预测概率的函数;K_y表示原始标签集合中某个特定类别标签y所对应的词集;$P_\tau\left([\mathrm{MASK}]=K|E_\rho\right)$表示模型预测标签词集$K$中每个单词在[MASK]位置的概率;$E_\rho$表示基于Prompt机制处理后的文本。

在模型训练过程中,对于不符合语义的标签词赋予较小的权重,以最小化它们对最终预测结果的影响。随后通过归一化操作,得到每个字符的权重,具体步骤为

$$\lambda_K=\frac{\exp(W_K)}{\sum\limits_{a\in K_y}(\boldsymbol{W}_a)} \tag{15}$$

式中:W_K代表每个标签词赋予的可学习权值,初始化值为0。进一步地,利用argmax函数将得分最高的标签作为最终的预测结果,预测标签为

$$\hat{y}=\arg\max_{y\in Y}\frac{\exp\left(s(y|E_\rho)\right)}{\sum\limits_{y'}\exp\left(s(y'|E_\rho)\right)} \tag{16}$$

式中$s(y|E_\rho)$表示定义在输入E_ρ和输出y上的分数函数。将式(15)得到的λ_K作为平均权重,把标签词得分的加权平均值作为预测得分,有

$$s\left(y|E_\rho\right)=\sum_{K\in K_y}\lambda_K\ln P_\tau\left([\mathrm{MASK}]=K|E_\rho\right) \tag{17}$$

2　实　　验

2.1　实验数据集

为了验证本文提出方法的有效性,在公开数据集1998版《人民日报》(以下简称"PD_1998")以及2004版《人民日报》(以下简称"PD_2004")、MSRA、Resume、Weibo和CMeEE上分别进行实验。具体数据集划分如表1所示。

(1)人民日报数据集。取材于人民日报的正式新闻体语料,主要由中文文章组成,包括人物、地点和机构3种实体信息。2004版本的数据集除上述3类实体,额外增加时间实体。

(2)MSRA数据集。由微软亚洲研究院提供的一个数据集,主要是简体中文新闻,包括人物、地点和组织3种实体信息。

(3)Weibo数据集。由微博提供,含有社交媒体相关内容,表达方式相对灵活。由7类实体构成,如:组织名特

表1　数据集划分

Table 1　Dataset partition

数据集	训练集	验证集	测试集
PD_1998	20 864	2 318	4 636
PD_2004	200 000	50 000	36 268
MSRA	42 451	2 363	2 391
Weibo	1 350	270	270
Resume	3 821	463	477
CMeEE	15 000	5 000	3 000

指(ORG.NAM)、组织名泛指(ORG.NOM)等。

(4) Resume数据集。主要包括上市公司内管理人员的姓名、学历、职位以及工作经历等相关简历实体信息。

(5) CMeEE数据集。包含504种常见的儿科疾病、7 085种身体部位、12 907种临床表现、4 354种医疗程序等九大类医学实体。

2.2 参数设置及评价指标

本文采用Linux操作系统,配备Tesla V100 32 GB显卡,使用Python3.8.13开发环境和Pytorch1.12.1开发框架。实验超参数设置如表2所示。

表2 参数设置

Table 2 Parameter setting

参数	设定值
预训练模型	roberta-wwm-ext-large
最大文本长度	100
学习率	1×10^{-5}
droupout	0.1
batch_size	16
epoch	30
优化器	AdamW
词向量维度	768

为验证本文方法的命名实体识别效果,本文使用精确率P(Precision)、召回率R(Recall)和F_1值作为模型的评价指标。

2.3 实验结果分析

2.3.1 对比实验

为验证模型的有效性,将本文模型与近年提出的基线模型进行实验对比。

(1) 在人民日报数据集上进行对比实验,其对比实验结果如表3、4所示。从表3与表4可以看出,本文模型相比于Lattice-LSTM-CRF、BERT-BiLSTM-CRF和ALBERT BiLSTM-Self-Attention-CRF三个模型的F_1值分别提高了6.32%、3.44%和3.12%,证明了本文方法的有效性。并且,对比近年来的$\text{LERT}_{\text{base}}$、$\text{LERT}_{\text{large}}$、$\text{MacBERT}_{\text{large}}$及$\text{PERT}_{\text{large}}$四个模型的$F_1$值分别提高了1.49%、0.79%、1.29%和0.99%,可知本文模型融入提示信息后,缩短了预训练任务与微调之间的差距,并且融合多层级信息增加了模板的可解释性,在此数据集上得到了最好的表现。

表3 人民日报(1998版)数据集对比实验结果

Table 3 Comparative experimental results on People's Daily(1998) dataset %

模型	P	R	F_1
Lattice-LSTM-CRF[10]	90.65	90.89	90.77
BERT BiLSTM-CRF[11]	93.08	93.88	93.65
ALBERT BiLSTM-Self-Attention-CRF[12]	94.33	93.90	93.97
$\text{LERT}_{\text{base}}$[13]	—	—	95.60
$\text{MacBERT}_{\text{large}}$[14]	—	—	95.80
$\text{PERT}_{\text{large}}$[15]	—	—	96.10
$\text{LERT}_{\text{large}}$[13]	—	—	96.30
Ours	**96.33**	**94.05**	**97.09**

表4 人民日报(2004版)数据集对比实验结果

Table 4 Comparative experimental results on People's Daily(2004) dataset %

模型	P	R	F_1
Lattice-LSTM-CRF[10]	93.57	92.79	93.18
BERT BiLSTM-CRF[11]	94.43	93.86	94.14
ALBERT BiLSTM-Self-Attention-CRF[12]	93.13	94.21	93.67
$\text{LERT}_{\text{base}}$[13]	—	—	96.74
$\text{MacBERT}_{\text{large}}$[14]	—	—	96.31
$\text{PERT}_{\text{large}}$[15]	—	—	96.52
$\text{LERT}_{\text{large}}$[13]	—	—	96.91
Ours	**97.4**	**97.33**	**97.36**

(2) 在MSRA数据集上进行对比实验,对比实验结果如表5所示。从表5可知,本文模型相比于Lattice-LSTM-CRF、LGN和LR-CNN三个模型的F_1值分别提高了3.5%、3.22%、2.97%。综合表3~5可知,Lattice-LSTM-CRF识别效果低于其他模型,这表明虽然Lattice-LSTM-CRF融合了字、词级

别的特征,但是神经网络处理自然语言的能力弱于预训练模型。对比近年来的$LERT_{base}$、$LERT_{large}$、$MacBERT_{large}$及$PERT_{large}$四个模型,本文模型F_1值分别提高了0.98%、0.38%、0.48%和0.48%,比AAMwWIE、MWAM两个模型的F_1值分别提高了3.44%、2.54%。虽然MWAM模型在召回率上具有优势,但其F_1值较本文模型低,差距较为明显,证明了本文模型在中文命名实体识别任务上的有效性。

(3)在Weibo数据集上进行对比实验,其实验结果如表6所示。从表6可以看出,Weibo数据集多为社交媒体平台内容,文字使用较为灵活,表达更加丰富,本文模型比Star-GAT、Soft-Lexicon、NNBA、W2Ner与结合实体边界线索的模型F_1值分别提高了13.3%、12.94%、13.32%、11.02%和9.8%,证明了本文方法在复杂语义环境下性能依旧有较大提升。

(4)在Resume数据集上进行对比实验,对比实验结果如表7所示。从表7可以观察到,本文模型分别比Lattice-LSTM、Soft-Lexicon(LSTM)、SVR-BiGRU-CRF、PLTE和AELT模型的F_1值分别提升了3.16%、2.17%、2.28%、2.37%、2.08%和2.03%。在所对比的4个数据集中,本文模型的F_1值均优于对比基线模型,本文模型在获取字、词和实体信息时不仅考虑了不同层级的特征,还考虑了上下文语义信息,并且在兼顾三者优点的同时使用提示学习方式,刺激模型获取文本先验知识,从而提升中文命名实体识别的效果。

(5)在CMeEE数据集上进行对比实验。与最近发表的医疗实体识别模型进行比较。在相同参数和数据集的条件下,实验结果如表8所示。由表8可得,本文所提出模型比其他方法具有更强的泛化能力,能够更加充分地挖掘文本中的语义信息。对于文本中有错别字的情况,本文模型仍可以很好地识别。该数据集由多轮中文医疗对话构成,其中包括一些不包含医疗实体的对话文本,在数据构成复杂的情况下,能够识别出更多的医疗实体,并且达到最高的F_1值。

表5　MSRA数据集对比实验结果

Table 5　Comparative experimental results on MSRA dataset　%

模型	P	R	F_1
Lattice-LSTM-CRF[10]	93.57	92.79	93.18
LGN[16]	94.19	92.73	93.46
LR-CNN[17]	94.50	92.93	93.71
AAMwWIE[18]	93.52	92.97	93.24
MWAM[19]	94.02	**94.25**	94.14
$MacBERT_{large}$[14]	—	—	96.20
$PERT_{large}$[15]	—	—	96.20
$LERT_{base}$[13]	—	—	95.70
$LERT_{large}$[13]	—	—	96.30
Ours	**96.94**	93.55	**96.68**

表6　Weibo数据集对比实验结果

Table 6　Comparative experimental results on Weibo dataset　%

模型	P	R	F_1
Star-GAT[20]	70.85	67.12	70.14
Soft-Lexicon[21]	70.94	67.02	70.50
NNBA[22]	70.11	68.12	70.12
W2Ner[23]	70.84	73.87	72.32
结合实体边界线索[24]	69.93	77.53	73.54
Ours	**81.91**	**82.04**	**83.44**

表7　Resume数据集对比实验结果

Table 7　Comparative experimental results on Resume dataset　%

模型	P	R	F_1
Lattice-LSTM-CRF[10]	94.75	93.89	94.32
Soft-Lexicon(LSTM)[25]	95.10	95.53	95.31
SVR-BiGRU-CRF[26]	95.13	95.25	95.20
LR-CNN[17]	95.37	94.84	95.11
PLTE[27]	95.34	95.46	95.40
AELT[28]	95.80	96.06	95.93
Ours	**96.85**	**96.53**	**97.48**

2.3.2 消融实验

本文在2.1节中所提到的数据集中随机挑选4个数据集分别进行了消融实验，以F_1值作为评价标准。词嵌入是指将单词从高维空间嵌入到低维连续向量空间中，将每个单词或词组映射为实数域上的向量。在模型的底层输入中，词嵌入方法的选择在一定程度上决定了模型的最终表现。为了研究本文进行多种词嵌入表示的有效性，本节针对字、词及实体级特征进行消融实验。此外，为了验证提示学习在模型中的作用，本文还将完整的模型网络结构拆分为不使用提示学习的框架，结果如表9所示。

表8 CMeEE数据集对比实验结果

Table 8 Comparative experimental results on CMeEE dataset %

模型	P	R	F_1
GP(ALBERT)[29]	57.80	50.76	54.05
GP(BERT)[29]	52.61	61.42	56.67
GP(RoBerta)[29]	53.81	64.47	58.66
FLAT[30]	60.56	65.17	62.78
BERT-IDCNN-GAT[31]	71.15	70.28	70.71
Ours	**75.47**	**76.67**	**76.05**

表9 消融实验结果

Table 9 Ablation experiment results %

模型	PD_2004	Resume	CMeEE	MSRA
Ours	**97.36**	**97.48**	**76.05**	**96.68**
Without prompt	93.28	94.65	68.13	92.86
Only token	96.12	95.73	70.61	94.02
Only word	96.94	96.08	73.61	95.65
Only entity	97.04	97.01	73.97	95.13

从词嵌入的角度来看，不论是哪种结构要素，本文模型优于其他嵌入方式，不仅考虑了字、词特征信息，还考虑了实体特征信息，使模型以更加全面的形式表示句子的语义信息。在加入提示学习后可以有效缓解因模型解释性较差所导致分类效果不佳的问题，并且在各种灵活的语义环境下取得了较高F_1值，由此可见模型每一部分都起着重要的作用。

2.3.3 提示模板分析

(1) 提示模板性能分析

手工设计的提示模板会对模型的效果产生一定的波动，本文将评估手工设计模板对模型性能产生的影响，实验结果如表10所示。从结果可以看出，模型的性能会受到提示模板较大的影响。具体地，在数据集上对模板采用了前缀式与后缀式的形式进行评测。相比之下，在中文数据集上最大与最小值相差1.42%，这表明提示模板对模型的准确率影响有一定的关系。通过优化模板的形式可以较大程度地提升模型的性能。

表10 不同提示模板的准确率

Table 10 Accuracy with different prompt templates %

模板	准确率
文本＜text＞中的实体，＜text＞是否包含？＜MASK＞	**97.36**
＜MASK＞是文本＜text＞中的实体	96.27
在＜text＞中，请确认是否存在实体：＜MASK＞	95.94
文本＜text＞中的实体属于哪个？＜MASK＞	97.14

(2) 推理词形式性能分析

本文方法将序列标注任务转换为基于自然语言推理形式的完型填空任务，同时受到P-tuning方法的启发，推理词不仅可以是自然语言形式，也可以是非自然语言形式，因此本文综合了上述两种方式的优点，使用混合模板形式。本文对上述形式的推理词进行性能评估，实验结果如表11所示。从结果可以看出，非自然语言形式的推理词较为稳定，模型的性能对比自然语言推理词有所提升。具体地，对于形式简单、数据区分度高的数据集，如Resume和PD_1998，自然语言形式的推理词表现较为出众。对于相对复杂的数据集，如PD_2004和MSRA，非自然语言形式的推理词具备更好的性能，说明非自然语言形式的推理词可以从众多的上下文信息中学习到推理词的连续化表达形式。而对于类别数多且复杂、语言环境灵活的任务，如Weibo、CMeEE，本文模型性能较好。这是由于它可以从具体任务中自主

学习到更适合当前模板的推理词形式,而不受自然语言形式的限制。在此基础上,为了防止连续模板在学习过程中偏离具体任务,本文指定了初始化模板,有效避免了推理词的影响,极大地提升了提示模板在下游任务的鲁棒性。

表11　数据集推理词形式性能比较

Table 11　Performance comparison of dataset inference words　%

方式	PD_1998	PD_2004	Weibo	Resume	CMeEE	MSRA
本文推理词	**97.09**	**97.36**	**83.44**	**97.48**	**76.05**	**96.68**
自然语言推理词	96.71	96.47	73.11	96.65	68.13	95.86
非自然语言推理词	96.19	96.89	72.83	96.02	69.31	96.02

(3)提示模板长度分析

为分析不同长度的提示模板对模型分类效果的影响,参照实验数据集的样本长度分布,本文选取[5,50]之间10组不同长度的提示模板在改进后的P-tuning方法上进行实验,实验结果如图4所示。

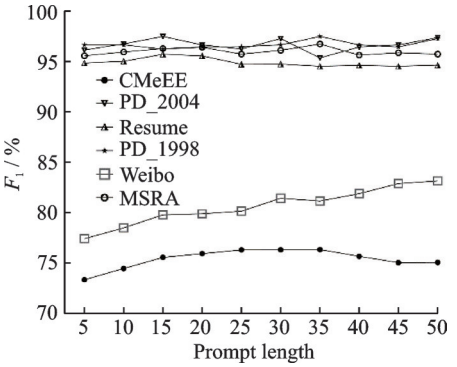

图4　提示模板长度分析

Fig.4　Length analysis of prompt templates

提示学习方法的应用要对模型参数进行冻结。然而,当选择较短的提示模板时,由于优化过程中可学习参数不足,提示模板所传递的信息主要是来自于预训练模型的知识,而非下游任务的相关知识。此外,将提示模板与输入文本的词嵌入拼接时,并不能有效提升拼接后的词向量对原始输入文本的表示效果,反而可能影响模型对下游任务的性能。当适当长度的提示模板被选择时,模板能够保留适量的下游任务知识,提升模型对下游任务的理解能力。相对地,如果选择较长的提示模板,则可能会过度保留训练集的知识,导致模型过拟合,导致准确率不佳。因此,在选择提示模板长度时需要权衡模型复杂性与任务需求,以获得最佳性能。

2.3.4　*参数分析*

为了研究本文提出的方法中不同参数对实验结果的影响,选择了实验轮数、批大小以及学习率作为研究因素,在MSRA数据集上进行了实验,结果如图5~7所示。经过实验,观察到随着迭代轮次数量的增加,模型的准确性总体上呈现上升趋势,这表明适度增加训练轮数可以提高模型性能。然而,随着轮数的继续增加,模型的性能逐渐趋于稳定,但时间复杂度也随之增加。因此,在主要实验中,选择了30轮的迭代轮次。批大小的选择对实验结果有一定程度的影响,发现其在16时得到较好的效果。随着模型学习率的变化,实验结果也在一定范围内波动,但是当学习率过大时模型会无法收敛,在最优值附近徘徊。

2.3.5　*模型速率分析*

为了评估本文提出的方法在时间性能方面的表现,在相同的实验环境下将本文方法与部分基线模型进行了比较,图8为不同模型在人民日报、MSRA和Weibo数据集上完成10轮迭代次数的平均训练时间。从图8可以明显看出,与BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF模型相比,本文提出的方法在每个数据集上的训练时间最短。这是因为BERT-BiLSTM模型在计算上相对较为复杂,导致了计算时间的增加;ERNIE-BilSTM采用三阶段补充特征的方式无形中增加了计算成本。而本文模型所构建的提示文本较为精简,只更新模板对应的参数,从而极大地减小了模型学习所用的参数量,并且解

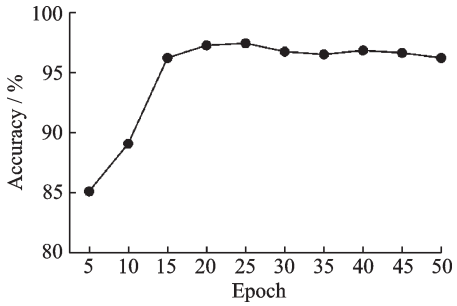

图5 不同迭代轮次下准确率

Fig.5 Accuracy with different iteration rounds

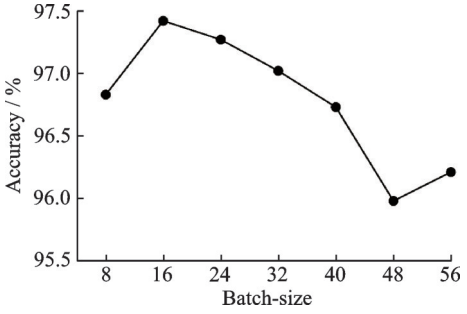

图6 不同批次大小下的准确率

Fig.6 Accuracy with different batch sizes

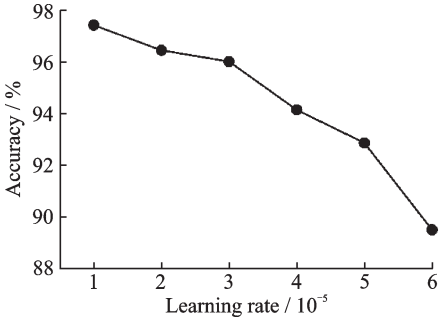

图7 不同学习率下的准确率

Fig.7 Accuracy with different learning rates

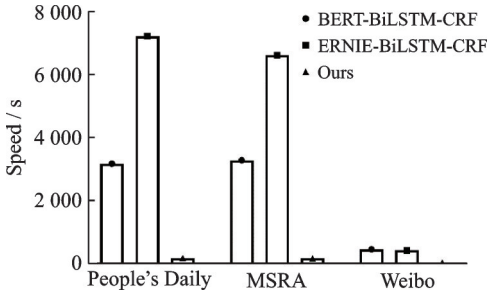

图8 不同模型完成10轮迭代的平均时间

Fig.8 Average runtime of ten iterations for different models

决了BERT模型上下游任务不一致的问题,对时间性能影响较小。此外,标签映射词表可以根据输入的提示文本自动构建标签词预测范围内的词表,限制模型的搜索空间,以此提升时间利用率。

3 结束语

本文提出了一种基于提示学习的中文命名实体识别方法。该方法结合了多层级特征信息,增强了模型对句子语义的表征能力。此外,通过自动构建混合模板与映射词表,能够有效提取标签词表中的语义信息,进一步提高了识别性能。实验结果表明,与现有模型相比,本文方法在中文命名实体识别方面取得了更好的效果并且缩减了执行时间。然而,本文方法仍然有改进的空间,尽管模板可以自动构建,但它们的解释性较差,并且缺乏广泛覆盖的映射词表,这些因素都可能对最终的预测结果产生不利影响。未来的研究可以继续优化提示模板的设计,以及如何更好地进行提示工程,进一步提升提示学习在命名实体识别任务上的应用效果。此外,还可以探索将该方法应用于其他自然语言处理任务,如在生成式任务或多轮对话任务中,如何提供更精准的实体识别结果。

参考文献:

[1] FARMAKIOTOU D, KARKALETSIS V, KOUTSIAS J, et al. Rule-based named entity recognition for Greek financial texts[C]//Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000). Patras, Greece: WCL, 2000: 75-78.

[2] LUO Gang, HUANG Xiaojiang, NIE Zaiqing, et al. Joint entity recognition and disambiguation[C]//Proceedings of the 2015

Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: ACL, 2015: 879-888.

[3] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2001, 2(11): 45-66.

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

[5] JIA Yaozong, XU Xiaobin. Chinese named entity recognition based on CNN-BiLSTM-CRF[C]//Proceedings of the 9th International Conference on Software Engineering and Service Science (ICSESS). Piscataway, NJ, USA: IEEE, 2018: 1-4.

[6] MATTHEW E P, WALEED A, CHANDRA B, et al. Semi-supervised sequence tagging with bidirectional language models [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017: 1756-1765.

[7] 盛剑, 向政鹏, 秦兵, 等. 多场景文本的细粒度命名实体识别[J]. 中文信息学报, 2019, 33(6): 80-87.
SHENG Jian, XIANG Zhengpeng, QIN Bing, et al. Fine-grained named entity recognition for multi-scenario[J]. Journal of Chinese Information Processing, 2019, 33(6): 80-87.

[8] ZHU Yuying, WANG Guoxin, KARLSSON B F. CAN-NER: Convolutional attention network for Chinese named entity recognition[EB/OL]. (2020-07-15). http://arxiv.org/pdf/1904.02141.

[9] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference[C]//Proceedings of Conference of the European Chapter of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2021.

[10] ZHANG Yue, ZHANG Jie. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 1554-1564.

[11] DAI Zhenjin, WANG Xutao, NI Pin, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]//Proceedings of the 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics. Piscataway, NJ, USA: IEEE, 2019: 1-5.

[12] 游乐圻, 裴忠民, 罗章凯. 融合自注意力的ALBERT中文命名实体识别方法[J]. 计算机工程与设计, 2023, 44(2): 605-611.
YOU Leqi, PEI Zhongmin, LUO Zhangkai. Chinese named entity recognition based on ALBERT fused with self-attention[J]. Computer Engineering and Design, 2023, 44(2): 605-611.

[13] CUI Yiming, CHE Wanxiang, WANG Shijin, et al. LERT: A linguistically-motivated pre-trained language model[EB/OL]. (2022-03-22). http://arxiv.org/abs/2011.05344.

[14] CUI Yiming, CHE Wanxiang, LIU Ting, et al. Revisiting pre-trained models for chinese natural language processing[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP. [S.l.]:ACL, 2020: 657-668.

[15] CUI Yiming, YANG Ziqing, LIU Ting. PERT: Pre-training BERT with permuted language model[EB/OL]. (2022-04-05). https://arxiv.org/abs/2203.06906.

[16] GUI Tao, ZOU Yicheng, ZHANG Qi, 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 (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019: 1040-1050.

[17] GUI Tao, MA Ruotian, ZHANG Qi, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019.

[18] 赵萍, 窦全胜, 唐焕玲, 等. 融合词信息嵌入的注意力自适应命名实体识别[J]. 计算机工程与应用, 2023, 59(8): 167-174.
ZHAO Ping, DOU Quansheng, TANG Huanling, et al. Attention adaptive model with world information embedding for named entity recognition[J]. Computer Engineering and Applications, 2023, 59(8): 167-174.

[19] 占文韬, 吴晓鸽, 凌捷. 基于多窗口注意力机制的中文命名实体识别[J]. 小型微型计算机系统, 2024, 45(6): 1325-1330.
ZHAN Wentao, WU Xiaoling, LING Jie. Chinese named entity recognition based on multi-window attention mechanism[J]. Journal of Chinese Computer Systems, 2024, 45(6): 1325-1330.

[20] CHEN Chun, KONG Fang. Enhancing entity boundary detection for better Chinese named entity recognition[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.]: ACL, 2021: 20-25.

[21] MA Ruotian, PENG Minlong, ZHANG Qi, 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.]:ACL, 2020: 59515960.

[22] CHEN Yanping, WU Yuefei, QIN Yongbin, et al. Recognizing nested named entity based on the neural network boundary assembling model[J]. Intelligent Systems, 2019, 35(1): 74-81.

[23] LI Jingye, FEI Hao, LIU Jiang, et al. Unified named entity recognition as word-word relation classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 10965-10973.

[24] 黄蓉, 陈艳平, 扈应, 等. 结合实体边界线索的中文命名实体识别方法[J]. 计算机工程与应用, 2024, 60(6): 199-206.
HUANG Rong, CHEN Yanping, HU Ying, et al. Chinese named entity recognition methods combined with entity boundary cues[J]. Computer Engineering and Applications, 2024, 60(6): 199-206.

[25] MA Ruotian, PENG Minlong, ZHANG Qi, et al.Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.Dublin, Ireland: ACL, 2022: 5951-5960.

[26] 张召武, 徐彬, 高克宁, 等. 面向教育领域的基于SVR-BiGRU-CRF中文命名实体识别方法[J]. 中文信息学报, 2022, 36(7): 114-122.
ZHANG Zhaowu, XU Bin, GAO Kening, et al. SVR-BiGRU-CRF based Chinese named entity recognition for education domain[J]. Journal of Chinese Information Processing, 2022, 36(7): 114-122.

[27] XUE Mengge, YU Bowen, LIU Tingwen, et al. Porous lattice transformer encoder for Chinese NER[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: ICCL, 2020: 3831-3841.

[28] 韩晓凯, 岳顾, 褚晶, 等. 基于注意力增强的点阵Transformer的中文命名实体识别方法[J]. 厦门大学学报(自然科学版), 2022, 61(6): 1062-1071.
HAN Xiaokai, YUE Qi, CHU Jing, et al. Chinese named entity recognition based on attention-enhanced lattice Transformer [J]. Journal of Xiamen University Nature Science, 2022, 61(6): 1062-1071.

[29] SU J, MURTADHA A, PAN S, et al. Global pointer: Novel efficient span-based approach for named entity recognition[EB/OL]. (2022-06-06). http://arxiv.org/abs/2208.03054.

[30] LI Xiaonan, HANG Yan, QIU Xipeng, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2020: 6836-6842.

[31] 梁文桐, 朱艳辉, 詹飞, 等. 基于深度学习多模型融合的医疗命名实体识别[J]. 计算机应用与软件, 2022, 39(10): 162-168,229.
LIANG Wentong, ZHU Yanhui, ZHAN Fei, et al. Medical named entity recognition based on deep learning multi-model fusion[J]. Computer Appliactions and Software, 2022, 39(10): 162-168,229.

作者简介:

王昕(1999-),男,硕士研究生,研究方向:自然语言处理、数据挖掘。

魏楚元(1977-),**通信作者**,男,博士,教授,硕士生导师,研究方向:自然语言处理、数据挖掘、机器学习等,E-mail:weichuyuan@bucea.edu.cn。

张蕾(1981-),女,博士,副教授,研究方向:城市时空数据的机器学习与挖掘算法。

万珊珊(1980-),女,博士,副教授,研究方向:数据挖掘、智能教学系统、自然语言处理、图像识别和智慧城市。

(编辑:刘彦东)