计算机科学
COMPUTER SCIENCE

基于标签信息融合与多任务学习的中文命名实体识别

廖梦, 贾真, 李天瑞

引用本文

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 Meng, JIA Zhen, LI Tianrui. Chinese Named Entity Recognition Based on Label Information Fusion and Multi-task Learning [J]. Computer Science, 2024, 51(3): 198-204.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200114

基于标签信息融合与多任务学习的中文命名实体识别

廖 梦[1] 贾 真[1] 李天瑞[1,2,3]
1 西南交通大学计算机与人工智能学院 成都 611756
2 四川省制造业产业链协同与信息化支撑技术重点实验室 成都 611756
3 综合交通大数据应用技术国家工程实验室 成都 611756
(liaomeng28@163.com)

摘 要 随着中文命名实体识别研究的不断深入,大多数模型关注融入词汇或字形信息来丰富特征表示,但是却忽略了标签信息。因此文中提出了一种融合标签信息的中文命名实体识别模型。首先,通过预训练模型 BERT-wwm 得到字符的嵌入表示,并将标签向量化,使用 Transformer 解码器结构将字符表示与标签表示进行交互学习,捕捉字符与标签的相互依赖关系,丰富字符的特征表示。为了促进标签信息的学习,构建了基于文本句的监督信号,增加了多标签文本分类任务,采用多任务学习的方式进行训练。其中,命名实体识别任务采用条件随机场进行解码预测,多标签文本分类任务采用双仿射机制进行解码预测,两任务共享除解码层以外的所有参数,保证了不同的监督信息反馈到每个子任务。在公开数据集 MSRA,Weibo 和 Resume 上进行了多组对比实验,分别获得了 95.75%,72.17%,96.23%的 F1 值。与多个基准模型相比,所提模型的实验效果有一定的提升,证明了该模型的有效性与可行性。

关键词:命名实体识别;标签信息;注意力机制;双仿射机制;预训练模型
中图分类号 TP391

Chinese Named Entity Recognition Based on Label Information Fusion and Multi-task Learning

LIAO Meng[1],JIA Zhen[1] and LI Tianrui[1,2,3]
1 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province,Chengdu 611756,China
3 National Engineering Laboratory of Integrated Transportation Big Data Application Technology,Chengdu 611756,China

Abstract With the development of Chinese named entity recognition research,most models focus on enriching feature representation by integrating vocabulary or glyph information but ignore label information. Therefore,a Chinese named entity recognition model integrating label information is proposed in this paper. Firstly,the embedding representation of characters is obtained by pre-trained model BERT-wwm,and labels are represented as vectors. The character representation and label representation are interactively learned by using the Transformer decoder structure to capture the interdependence between characters and labels and enrich the feature representation of characters. To promote the learning of label information,a supervision signal based on text sentences is constructed,multi-label text classification tasks are added,and multi-task learning is used for training. Among them,the named entity recognition task uses a conditional random field for decoding and prediction,and the multi-label text classification task uses a biaffine mechanism for decoding and prediction. The two tasks share all parameters except the decoding layer,which ensures that different supervision information is fed back to each subtask. Several groups of comparative experiments are carried out on the public data sets MSRA,Weibo,and Resume,and the F1 values of 95.75%,72.17%,and 96.23% are obtained respectively. Compared with several benchmark models,experimental result of the proposed model is improved to some extent,which validates its effectiveness and feasibility.

Keywords Named entity recognition,Label information,Attention mechanism,Biaffine mechanism,Pre-trained model

到稿日期:2023-02-17 返修日期:2023-06-06
基金项目:国家自然科学基金面上项目(62176221)
This work was supported by the National Natural Science Foundation of China(62176221).
通信作者:李天瑞(trli@swjtu.edu.cn)

1　引言

随着自然语言处理技术的不断发展与应用，命名实体识别（Named Entity Recognition，NER）作为自然语言处理的关键基础任务之一，也备受关注。该任务旨在在给定文本中识别出一些具有特殊含义的实体文本片段，其关键在于如何识别出该实体的前后边界，并对实体进行准确的分类。命名实体识别已经广泛应用于问答系统、关系抽取、机器翻译等下游任务中，实体能否被正确识别起着至关重要的作用。因此，研究命名实体识别是十分有必要的。

中文任务与英文任务最大的区别在于输入的粒度。英文文本格式简单，单词间有分隔符，因此将单词作为最小粒度是显然的。但是，中文文本汉字字符之间紧密相连，若以词汇作为最小粒度，可能存在分词错误，进而影响模型效果；若以字符作为最小粒度，可能会缺失部分语义特征。现有模型大多是基于字符的，为了弥补字符信息单一的缺陷，很多模型提出加入词汇、字形、义原等特征信息来提升模型效果。这些额外的特征信息有利于中文命名实体识别任务，但却很少考虑标签信息。一般地，命名实体识别数据集往往只包含字符对应的实体标签，以此作为监督信号，从而对字符向量表示进行解码预测。这种方式仅能通过字符相关的通道进行信息传递，若考虑引入标签信息，虽然标签相关的参数能得到一些反馈调整，但效果可能不佳。若能提供额外的监督信号，并对标签向量表示进行解码预测，不仅对标签表示有益，而且有助于字符表示。

综上，受细粒度实体分类论文[1-2]启发，本文提出一种用于中文命名实体识别的模型 BIFT（BERT and Interactive Fusion Transformer），将字符信息与标签信息进行交互融合，增强了字符的特征表示。为防止监督信号单一导致融合不充分的问题，本文额外构建监督信号，增加多标签文本分类（Multi-Label Text Classification，MTC）任务，采用多任务学习方式训练两个子任务参数，促进标签表征的更新，从而更利于字符与标签的信息融合。将本文模型在 3 个中文命名实体识别数据集上进行了验证，其效果优于对比的基准模型，证明了所提模型能够有效提升命名实体识别效果。

2　相关工作

中文命名实体识别的模型结构一般分为嵌入层、建模层和解码层。根据模型特点，本文从基于字符的模型、融合字形信息的模型、融合词汇信息的模型和其他模型这 4 方面进行介绍。

中文命名实体识别模型存在以词作为输入的模型，如 Li 等[3]以词汇作为最小粒度，并使用 softmax 函数进行解码。但是，中文往往伴有分词错误问题，因此现有的模型大多都是基于字符的模型。Jia 等[4]使用双向长短期记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）获取上下文特征表示，再通过条件随机场（Conditional Random Field，CRF）进行解码。该类模型是解决命名实体识别的经典模型。Peng 等[5]使用双向门控循环单元（Bi-directional Gated Recurrent Unit，BiGRU）替代 BiLSTM 来提取上下文特征，并结合语言预测任务来增强字符表示。除循环神经网络（Recurrent Neural Network，RNN）类模型外，卷积神经网络（Convolutional Neural Network，CNN）和 Transformer[6]等结构也常常被使用。Wang 等[7]提出了一种门控卷积神经网络来获取文本字符的局部特征表示，与 RNN 类模型相比，该模型训练效率获得极大提升。Yan 等[8]采用 Transformer 编码器结构代替了传统的 RNN 或 CNN，其中的多头自注意力机制有利于深入地学习全局上下文表示。由于各类模型建模的特征具有显著区别，Jin 等[9]提出了一种常规卷积与空洞卷积结合的混合卷积神经网络来获取局部特征，再使用 BiLSTM 来提取上下文特征，并设计了一种带门控机制的自注意力网络来自适应聚合特征信息。Chang 等[10]通过预训练模型 BERT 来得到字符的嵌入式表示，并同时使用 BiLSTM 和多层空洞卷积神经网络来提取全局和局部的特征。

汉字是属于表意体系的一种文字，其中包括偏旁部首、笔画等字形信息，这类信息中可能包含一定的语义信息，因此一些模型提出引入字形信息用于增强命名实体识别效果。Dong 等[11]使用 BiLSTM 对字符所包含的字形信息进行建模，有效获取前向和后向的依赖关系，并拼接字符表示作为最终的嵌入表示。Liu 等[12]通过 CNN 结构对五笔字形的嵌入式表示进行特征提取，针对 RNN 只能计算短距离依赖的问题，提出了一种迭代学习的训练策略，有效提升了模型的全局建模能力。Zhang 等[13]提出了两种融合五笔字形信息的方法，一种是在嵌入层引入五笔字形信息，从而丰富字符向量表示；另一种是对五笔序列进行编码建模。上述模型是直接对字形进行嵌入式表示，也有一些模型是从字符图像获取字形信息。Meng 等[14]提出了一种基于字符图像的方法，采用简体中文、繁体中文、隶书等字体的字符图像，通过多层的 CNN 结构提取字形信息，额外增加图像分类任务来避免过拟合。Song 等[15]使用更深层的 CNN 结构来聚合字形信息，在 CNN 层之间加入批归一化、最大池化和暂退法等操作来改善过拟合和复杂度过高等问题，最终将字形表示与 BERT 输出的嵌入表示进行拼接。Xuan 等[16]不再以简单的拼接方式来融合字形和字符信息，而是使用滑动窗口机制得到多个字符与字形信息融合的切片表示，然后通过注意力机制自适应聚合切片表示并作为字符的最终表示。

中文词汇中包含字符信息和边界信息，因此将词汇信息融入到字符表示显然有助于命名实体识别任务。Zhang 等[17]提出一种晶格状 LSTM，该结构将词汇与词汇的开始和结束字符连接，除原生 LSTM 的信息流外，额外增加从词汇的开始字符到结束字符的信息流，中间经过词汇，从而将词汇信息融入词汇的结束字符中。Gui 等[18]使用图神经网络融入词汇信息，以字符作为节点，以词汇作为边，实现词汇信息的聚合。Sui 等[19]提出一种基于图注意力网络的方法，其包含多种图结构，除了词汇的开始和结束字符，还将词汇的中间字符和前后字符与词汇连接，将词汇信息融入更多字符之中。Li 等[20]提出一种基于 Transformer 编码器的模型，将文本与匹配词汇都作为输入；为了建模字符和词汇之间的关系，设计了一种巧妙的位置编码，实现了字符与词汇间的信息交互。上述模型在建模层引入词汇信息，但这些模型不具有可迁移

性，解决方法是在嵌入层融合词汇信息。Liu 等[21]以固定编码形式表示字符匹配的词汇信息，将词汇编码与字符编码拼接后送入建模层，提高了模型效率。Ma 等[22]根据词汇中匹配字符的位置将词汇分为 4 类，分别代表该字符位于词汇的开始位置、中间位置、结束位置和唯一位置，对 4 类词汇信息分别进行聚合后，与字符信息进行拼接。Liu 等[23]对 BERT 结构进行改进，在每一层 Transformer 之间加入词典适配器，通过字典树得到字符匹配的词汇集合，使用双线性注意力和残差机制融合词汇和字符信息。

以上工作均采用序列标注的方式进行命名实体识别，除此之外，还有一些与众不同的创新方法。Li 等[24]使用实体的标签生成对应问题，以机器阅读理解的方式，通过各类标签的问题去匹配文本中的实体答案。Yan 等[25]采用序列到序列(Sequence to Sequence，Seq2seq)模型来解决命名实体识别问题，利用 Brat 编码器编码文本字符信息，再通过 Brat 解码器生成实体位置和标签类别。Jimenez 等[26]将 GPT-3 模型用于识别实体。GPT-3 作为一种大型预训练模型，包含 96 层 Transformer 解码器结构，能够理解文本的深层语义信息。Li 等[27]提出一种新的标注方案，解决了嵌套实体问题，并采用跨度分类的方式进行解码。

3 问题定义与概念说明

3.1 中文命名实体识别

中文命名实体识别是找到文本中的实体并分类，根据实体的表现形式，可分为连续型/非连续型、嵌套型/非嵌套型。其中，连续型表示实体的所有字符是连续相邻的，嵌套型表示两个实体间存在位置重合。本文解决的是连续非嵌套型中文命名实体识别的问题，采用序列标注方法进行识别。输入长度为 q 的文本 $c=(c_1,c_2,\cdots,c_q)$，预测标签序列为 $y=(y_1,y_2,\cdots,y_q)$，将真实标签序列 y^* 作为监督信号。

3.2 多标签文本分类

多标签文本分类是识别文本中含有的标签集合，本文采用多个二分类的方法进行处理。输入长度为 q 的文本 $c=(c_1,c_2,\cdots,c_q)$，预测所有标签类别的得分序列为 $z'\in\mathbb{R}^m$，将 01 序列 $z\in\mathbb{R}^m$ 作为监督信号，其中 m 表示标签类型数量，标签类型数量与具体样本无关。

显然，多标签文本分类在测试时是无效的，该任务的主要作用是在训练时帮助标签信息的表示学习。

3.3 标签说明

中文命名实体识别数据集使用了特别的标注体系，如 BIO，BIOS 和 BIOES 等。以 BIOS 标注体系为例，“B”代表实体的开始位置，“I”代表实体的非开始位置，“O”代表非实体位置，“S”代表单字符实体的位置。

标签由字符在实体中的位置和实体类型组成，如“B-LOC”表示地名实体的开始位置。标签类型数量由数据集和标注体系决定，假设某数据集采用 BIOS 标注体系，实体类型共 s 种，那么该数据集的标签类型数量为 $3s+1$。

3.4 标签信息说明

标签信息的具体形式是将所有类型的标签进行向量化，也就是一个向量矩阵，向量的数量等于标签类型数量，并不等于输入文本的长度。在每一步模型训练中不断调整优化所有类型标签的向量表示，从而在训练集上得到标签信息。对于每一个样本，标签信息都是同一个向量矩阵，包含所有类型的标签，并不是只含有该样本的真实标签类型。例如，对于 Resume 数据集中的样本“高勇：男，中国国籍。”，该样本中的真实标签序列为：{B-NAME，I-NAME，O，O，O，B-CONT，I-CONT，I-CONT，I-CONT，O}。真实标签序列只含有 5 种标签，但该数据集共有 25 种标签，所以标签信息是由 25 个标签向量组成的向量矩阵。该数据集的其他样本也是同一个标签向量矩阵。

由于标签信息的内容全部来源于训练集，并且代表着所有类型的标 签，与具体样本的真实标签序列无关，因此即使是测试的时候，标签信息作为输入也不会造成真实标注结果的泄露。

4 模型

本文的命名实体识别模型结构如图 1 所示，分为嵌入层、交互融合层、解码层。在嵌入层，使用预训练模型得到文本的字符表示。在交互融合层，提出一种改进的 Transformer 结构，通过多头互注意力机制捕获双向信息依赖，实现字符信息与标签信息的交互融合。在解码层，通过 CRF 解码实体标签，并基于双仿射(Biaffine)机制进行多标签文本分类，两个任务以多任务学习方式同时训练。

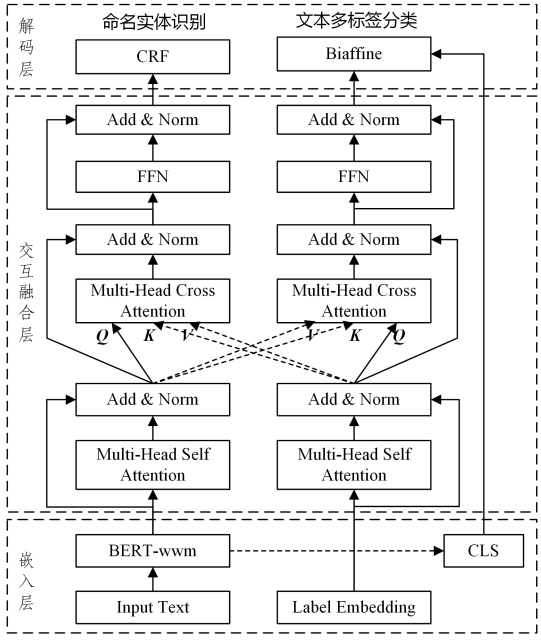

图 1 BIFT 模型

Fig. 1 BIFT model

4.1 嵌入层

4.1.1 字符嵌入表示

本文使用预训练模型 BERT-wwm[28]对文本字符进行上下文语义表征。该模型采用全词掩盖的方法进行遮掩语言模型(Masked Language Model，MLM)预训练，包含一定的词汇信息。给定一个长度为 q 的文本 $c=(c_1,c_2,\cdots,c_q)$，那么文本的字符嵌入表示为：

$$\boldsymbol{x}^c=BERT(c) \tag{1}$$

其中，$\boldsymbol{x}^c \in \mathbb{R}^{n\times d_1}$，$n$ 表示 BERT-wwm 编码的句子长度，d_1 为字符嵌入表示维度；$\boldsymbol{x}_1^c \in \mathbb{R}^{d_1}$ 代表“CLS”向量表示，本文将其作为整个文本的嵌入表示。

4.1.2　标签嵌入表示

给定一组数量为 m 的标签集合 $l=(l_1,l_2,\cdots,l_m)$，标签嵌入表示通过查询标签向量表得到。对于标签 l_k，其嵌入表示为：

$$\boldsymbol{x}_k^l = e^l(l_k) \tag{2}$$

其中，e^l 表示标签向量表，标签向量表采用随机初始化的方法产生；$\boldsymbol{x}^l \in \mathbb{R}^{m\times d_1}$ 表示所有类型标签的嵌入表示，也代表着标签信息，对应图 1 中的“Label Embedding”。

4.2　交互融合层

如何有效融合标签信息和字符信息，是本模型的关键之处。受 Transformer 模型结构以及序列标注论文[29]的启发，本文提出了 IFT 结构，采用两个 Transformers 解码器结构分别对字符信息和标签信息增强特征表示，取消了原型 Transformer 解码器中使用的遮掩多头自注意力机制，具体过程如下所述。

在嵌入层得到字符嵌入表示 $\boldsymbol{x}^c$ 和标签嵌入表示 $\boldsymbol{x}^l$ 后，将 $\boldsymbol{x}^c$ 和 $\boldsymbol{x}^l$ 输入到多头自注意力层，通过多头自注意力机制的作用，捕获了字符间的全局依赖，同时也发现了标签之间的依赖关系。多头自注意力层将产生两个输出 $\boldsymbol{H}^{sc}$ 和 $\boldsymbol{H}^{sl}$，计算方法如式(3)和式(4)所示。

$$\boldsymbol{H}^{sc} = MultiHead(\boldsymbol{x}^c,\boldsymbol{x}^c,\boldsymbol{x}^c) \tag{3}$$

$$\boldsymbol{H}^{sl} = MultiHead(\boldsymbol{x}^l,\boldsymbol{x}^l,\boldsymbol{x}^l) \tag{4}$$

其中，$\boldsymbol{H}^{sc} \in \mathbb{R}^{n\times d_1}$ 表示字符向量表示；$\boldsymbol{H}^{sl} \in \mathbb{R}^{m\times d_1}$ 表示标签向量表示；$MultiHead$ 表示多头注意力的计算公式，具体计算过程如式(5)－式(7)所示。

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}^O \tag{5}$$

$$\boldsymbol{head}_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{6}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

其中，$Attention$ 表示单头注意力计算公式；$\boldsymbol{Q}$，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示查询向量、键向量和值向量；$\boldsymbol{head}_i$ 表示第 i 个注意力头；h 表示注意力头数量；$\boldsymbol{W}^O \in \mathbb{R}^{hd_2\times d_1}$ 表示映射矩阵，$\boldsymbol{W}_i^Q \in \mathbb{R}^{d_1\times d_2}$，$\boldsymbol{W}_i^K \in \mathbb{R}^{d_1\times d_2}$ 和 $\boldsymbol{W}_i^V \in \mathbb{R}^{d_1\times d_2}$ 分别表示第 i 个注意力头中对 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 的转换矩阵，并且 $d_2=d_1/h$。

为了防止梯度消失，并且加速模型收敛，使用了残差机制和层归一化函数，即如图 1 所示的“Add&Norm”结构。具体操作过程如下：对于每一层，假设输入为 $\boldsymbol{I}$，输出为 $\boldsymbol{H}$，令 $\boldsymbol{M}=\boldsymbol{I}+\boldsymbol{H}$，最后对 $\boldsymbol{M}$ 进行层归一化，得到最终输出 $\boldsymbol{S}$。为了减少符号表示，$\boldsymbol{H}^{sc}$ 和 $\boldsymbol{H}^{sl}$ 在进行上述操作后，仍采用原符号 $\boldsymbol{H}^{sc}$ 和 $\boldsymbol{H}^{sl}$ 表示输出。值得注意的是，IFT 结构中的多头互注意力层和全连接层也将使用同样的方法进行残差连接和层归一化，后文不再赘述。

接下来，在多头互注意力层进行标签信息与字符信息的融合。多头互注意力机制能够捕获字符和标签之间的相互依赖关系，从而获得对方信息的聚合表示。对于输入 $\boldsymbol{H}^{sc}$ 和 $\boldsymbol{H}^{sl}$，多头互注意力层输出为 $\boldsymbol{H}^{cc}$ 和 $\boldsymbol{H}^{cl}$，计算方法如式(8)和式(9)所示。

$$\boldsymbol{H}^{cc} = MultiHead(\boldsymbol{H}^{sc},\boldsymbol{H}^{sl},\boldsymbol{H}^{sl}) \tag{8}$$

$$\boldsymbol{H}^{cl} = MultiHead(\boldsymbol{H}^{sl},\boldsymbol{H}^{sc},\boldsymbol{H}^{sc}) \tag{9}$$

其中，$\boldsymbol{H}^{cc} \in \mathbb{R}^{n\times d_1}$ 表示字符依赖的标签信息，$\boldsymbol{H}^{cl} \in \mathbb{R}^{m\times d_1}$ 表示标签依赖的字符信息。

考虑到仅使用注意力机制对复杂过程的拟合程度不够，并不能提取到理想特征，通过前向反馈网络层(Feed Forward Network，FFN)来增强网络能力。FFN 将向量空间先扩大再恢复，中间穿插非线性激活函数。FFN 虽然只是两层的全连接层，但 Dong 等[30]证明了取消 FFN 结构将导致归纳偏差问题。FNN 的具体计算过程如式(10)－式(12)所示。

$$\boldsymbol{H}^{fc} = FFN(\boldsymbol{H}^{cc}) \tag{10}$$

$$\boldsymbol{H}^{fl} = FFN(\boldsymbol{H}^{cl}) \tag{11}$$

$$FFN(\boldsymbol{x}) = \max(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)_2+\boldsymbol{b}_2 \tag{12}$$

经过 IFT 结构的建模之后，得到了增强的字符向量表示和标签向量表示。多层 IFT 结构可以提取到更深层次的特征，有利于字符信息与标签信息的充分融合，因此将 IFT 层数设置为 N。

4.3　解码层

对于中文命名实体识别任务，以 BIOS 标注体系为例，真实的标签序列具有 3 种特点：1)由于不涉及嵌套实体，那么某种实体类型的连续标签中不能出现其他实体类型的标签；2)“S”型标签的上一个标签不能是“B”型标签；3)“I”型标签上一个标签只能是“B”或“I”型标签。因为 CRF 使用转移矩阵来表示标签之间的依赖关系，并能获得全局最优的预测序列，所以本文采用 CRF 对字符向量表示进行解码预测，具体过程如下所述。

对于输入 $\boldsymbol{x}=\boldsymbol{H}^{fc}$，序列 y 作为输出序列的概率为 $P(y|\boldsymbol{x})$，计算方法如式(13)－式(15)所示。

$$P(y|x) = \frac{e^{score(\boldsymbol{x},y)}}{\sum\limits_{y'\in Y} e^{score(\boldsymbol{x},y')}} \tag{13}$$

$$Score(\boldsymbol{x},y) = \sum_{i=1}^{n}\boldsymbol{T}_{i,y_i} + \sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_{i+1}} \tag{14}$$

$$\boldsymbol{T} = \sigma(\boldsymbol{x}\boldsymbol{w}_t+\boldsymbol{b}_t) \tag{15}$$

其中，Y 表示所有可能的输出序列集合，$\boldsymbol{T}$ 表示标签概率矩阵，$\boldsymbol{A}$ 表示转移概率矩阵，$\boldsymbol{w}_t$ 和 $\boldsymbol{b}_t$ 表示全连接层参数。

命名实体识别任务的损失函数为 $loss_{\mathrm{ner}}$，计算式如下：

$$loss_{\mathrm{ner}} = -\log P(y^*|\boldsymbol{x}) \tag{16}$$

其中，y^* 表示真实标记序列。模型经过训练后，预测时可通过维特比算法得到全局最优标签序列。

对于多标签文本分类任务，本文采用多个二分类的方法进行处理，监督信号表示为 $z\in\mathbb{R}^m$。采用双仿射机制来进行分类任务，具体计算过程如式(17)和式(18)所示。

$$loss_{\mathrm{mtc}} = -\frac{1}{m}\sum_{i=1}^{m}(z_i\log p_t(i)+(1-z_i)\log(1-p_t(i))) \tag{17}$$

$$p_t(i) = \sigma((\boldsymbol{x}^{\mathrm{cls}})^{\mathrm{T}}\boldsymbol{U}_t\boldsymbol{H}_i^{fl}+(\boldsymbol{x}^{\mathrm{cls}}\oplus\boldsymbol{H}_i^{fl})^{\mathrm{T}}\boldsymbol{W}_t+b_t) \tag{18}$$

其中，$x^{\mathrm{cls}}=x_1^c$，$\boldsymbol{U}_t\in\mathbb{R}^{d_1\times d_1}$，$\boldsymbol{W}_t\in\mathbb{R}^{2d_1}$，$b_t$ 为双仿射函数参数。当 $p_t(i)\geqslant 0.5$ 时，表示文本中含有第 i 种标签。

为了同时训练命名实体识别任务和多标签文本分类任务，本文模型的最终损失函数表示为：

$$loss = loss_{\mathrm{ner}}+loss_{\mathrm{mtc}} \tag{19}$$

5 实验

5.1 实验数据

实验使用了3个中文公开数据集，包括MSRA[31]，Weibo[32]和Resume[17]。数据集采用了BIOS标注体系，本文模型同样也适用于其他标注体系的数据，不同之处在于标签数量的差异。

MSRA数据集是由微软亚洲研究院标注新闻语料产生的，共包含50729条样本，涉及机构(ORG)、地点(LOC)、人物(PER)3种实体类型，标签类型数量为10。

Weibo数据集通过新浪微博数据过滤标注而成，共包含1890条样本，涉及政治实体(GPE)、地址(LOC)、机构组织(ORG)、人物(PER)4种实体类型，并且实体还可细分为泛指和特指，标签类型数量为25。

Resume数据集使用上市公司工作人员的简历作为语料，共包含4761条样本，涉及种族(RACE)、国籍(CONT)、职称(TITLE)、专业(PRO)、组织机构(ORG)、籍贯(LOC)、学历(EDU)、人名(NAME)8种实体类型，标签类型数量为25。

本文对数据集进行句子级的数据规模统计，具体如表1所列。

表1 数据集统计结果

Table 1 Dataset statistics

数据集	训练集	验证集	测试集
MSRA	41728	4636	4365
Weibo	1350	270	270
Resume	3821	463	477

5.2 评价指标

模型评估采用正确率(P)、召回率(R)和Micro-F1值作为评价指标，P表示预测正确实体数占预测实体总数的比例，R表示预测正确实体数占真实实体数的比例，F1为P和R的调和平均值，计算方法如式(20)－式(22)所示。

$$P=\frac{TP}{TP+FP}\times 100\% \tag{20}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{21}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{22}$$

其中，TP表示预测正确的实体数，FP表示预测错误的实体数，FN表示没被识别出的实体数。

5.3 实验设置

本文将BERT-wwm的处理长度n设为150，标签嵌入维度与字符嵌入维度d_1设为768，注意力头数量h设为6，dropout设为0.5，IFT层数N设为2，batch-size设为32，epoch设为100，BERT-wwm参数学习率设为0.00001，其他参数学习率设为0.001，使用Adam优化器进行训练。

5.4 实验结果

5.4.1 基准模型

为了验证本文模型的实验效果，将近年来效果较好的模型作为对比模型。

Lattice-LSTM[17]：在建模层引入词汇，通过改进的LSTM融合词汇信息。

CAN-NER[33]：在建模层引入词汇，通过改进的CNN融合词汇信息。

Glyce[14]：在嵌入层引入字形，通过改进的CNN进行字形信息提取，并使用拼接的简单方式进行融合。

LGN[18]：在建模层引入词汇，通过改进的图神经网络融合词汇信息。

FGN[16]：在建模层引入字形，通过改进的CNN进行字形信息提取，并利用滑动窗口和注意力机制融合字形信息。

SoftLexicon[22]：在嵌入层引入词汇，通过BIES标记分层表示并融合词汇信息。

LEBERT[23]：在嵌入层引入词汇，通过改进的Transformer结构融合词汇信息。

5.4.2 实验结果

BIFT模型在MSRA，Weibo和Resume数据集上的实验结果如表2－表4所列，其分别取得了95.75%，72.17%和96.23%的F1值，在MSRA和Weibo数据集上均优于其他模型。在文本格式不规范且语料数据量少的Weibo数据集上，BIFT模型效果显著，说明BIFT模型能够充分提取标签信息。在Resume数据集上，BIFT模型比融合词汇信息的模型效果更佳，但融合字形信息的模型表现最好，可能是字形信息比较适用于用词规范的数据。

表2 MSRA实验结果

Table 2 Experimental results in MSRA

(%)

Model	P	R	F1
Lattice-LSTM	93.57	92.79	93.18
CAN-NER	93.53	92.42	92.97
Glyce	95.57	95.51	95.54
LGN	94.19	92.73	93.46
FGN	95.45	**95.81**	95.64
SoftLexicon	95.75	95.10	95.42
LEBERT	—	—	95.70
BIFT	**96.34**	95.17	**95.75**

表3 Weibo实验结果

Table 3 Experimental results in Weibo

(%)

Model	P	R	F1
Lattice-LSTM	—	—	58.79
CAN-NER	—	—	59.31
Glyce	67.68	67.71	67.60
LGN	55.34	64.98	60.21
FGN	69.02	**73.65**	71.25
SoftLexicon	—	—	70.50
LEBERT	—	—	70.75
BIFT	**73.62**	70.77	**72.17**

表4 Resume实验结果

Table 4 Experimental results in Resume

(%)

Model	P	R	F1
Lattice-LSTM	94.81	94.11	94.46
CAN-NER	95.05	94.82	94.94
Glyce	**96.62**	96.48	96.54
LGN	95.28	95.46	95.37
FGN	96.49	**97.08**	**96.79**
SoftLexicon	96.08	96.13	96.11
LEBERT	—	—	96.08
BIFT	95.97	96.50	96.23

从实验结果可以发现本文模型十分有效，证明引入标签

信息是可行的，同时也说明 BIFT 模型确实能捕捉字符与标签之间的依赖，丰富了两者的特征表示。BIFT 模型与融入字形或词汇信息等特征的模型相比，也具有竞争力。

5.5 IFT 层数影响

为了观察 IFT 层数对 BIFT 模型实验效果的影响，设计了 IFT 层数分别为 1，2，3，4 的 4 组实验，分别在 Weibo 和 Resume 数据集上进行实验，结果如表 5 所列。

表 5　BIFT 在不同 IFT 层数下的实验结果

Table 5　Experimental results of BIFTwith different number of layers of IFT

(%)

IFT 层数	Weibo			Resume		
	P	*R*	*F*1	*P*	*R*	*F*1
1	70.70	63.53	66.92	95.85	96.24	96.04
2	**73.62**	**70.77**	**72.17**	**95.97**	**96.50**	**96.23**
3	71.07	67.63	69.31	95.43	96.13	95.78
4	68.97	67.63	68.29	95.76	95.64	95.70

由实验结果可知，模型在 IFT 层数为 2 时，实验效果最佳。当只有 1 层 IFT 时，模型可能无法有效建模字符与标签之间的关系。而当 IFT 层数超过 2 时，模型复杂度逐渐增高，导致模型过拟合，实验效果越来越差。

5.6 消融实验

为了验证 BIFT 模型各部分的有效性，进行了一系列的消融实验，结果如表 6 所列。“-Biaffine”代表“CLS”向量表示与标签向量表示以拼接的简单方式进行结合；“-TMC”表示去除多标签文本分类任务，仅保留中文命名实体识别任务；“-IFT”表示去除 IFT 结构，也就是 BERT-CRF 模型。“-TMC”比“-IFT”实验效果更佳，证明了 IFT 结构的有效性。BIFT 模型与“-Biaffine”的 F1 值均高于“-TMC”，验证了增加多标签文本分类任务能够提升中文命名实体识别效果。BIFT 模型比“-Biaffine”实验效果更佳，说明双仿射机制相对于拼接方式，能更好地建模文本表示与标签信息之间的关系。综上，BIFT 模型各结构都是有效且可行的。

表 6　消融实验结果

Table 6　Ablation experiment results

(%)

Model	MSRA			Weibo			Resume		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
BIFT	**96.34**	95.17	**95.75**	**73.62**	70.77	**72.17**	**95.97**	**96.50**	**96.23**
-Biaffine	95.61	**95.26**	95.43	70.33	**71.01**	70.67	95.73	96.44	96.08
-TMC	95.52	95.17	95.34	69.05	70.05	69.54	95.32	96.44	95.88
-IFT	95.31	95.04	95.18	70.25	67.87	69.04	95.42	95.88	95.65

结束语　本文提出了一种融合标签信息的中文命名实体识别模型，通过 Transformer 结构来进行字符信息与标签信息的交互融合，并将中文命名实体识别和多标签文本分类以共享参数的多任务学习方式进行训练。实验证明，本文模型有效可行，在多个数据集上获得了不同程度的提升，与融合字形、词汇信息的多个模型相比具有竞争力。本文模型适用于标签类型较少的数据，如果标签类型过多，可能导致难以捕捉字符与标签之间的关系，模型复杂度过高。在后续工作中，将考虑改变标签信息的表现形式，降低其数据规模，从而使其广泛适用于标签类型过多的中文命名实体识别任务。

参 考 文 献

[1] LI J Q，CHEN X J，WANG D K，et al. Enhancing Label Representations with Relational Inductive Bias Constraint for Fine-Grained Entity Typing[C]// International Joint Conferences on Artificial Intelligence. 2021：3843-3849.

[2] LIN Y，JI H. An attentive fine-grained entity typing model with latent type representation[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). 2019：6197-6202.

[3] LI J Q，ZHAO S H，YANG J J，et al. WCP-RNN：a novel RNN-based approach for Bio-NER in Chinese EMRs[J]. The journal of supercomputing，2020，76(3)：1450-1467.

[4] JIA Y Z，MA X P. Attention in character-Based BiLSTM-CRF for Chinese named entity recognition[C]// Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence. 2019：1-4.

[5] PENG D L，WANG Y R，LIU C，et al. TL-NER：A transfer learning model for Chinese named entity recognition[J]. Information Systems Frontiers，2020，22(6)：1291-1304.

[6] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017：5998-6008.

[7] WANG C Q，CHEN W，XU B. Named entity recognition with gated convolutional neural networks[C]// Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. 2017：110-121.

[8] YAN H，DENG B C，LI X N，et al. TENER：adapting transformer encoder for named entity recognition[J]. arXiv：1911.04474，2019.

[9] JIN Y L，XIE J F，GUO W S，et al. LSTM-CRF neural network with gated self attention for Chinese NER[J]. IEEE Access，2019，7：136694-136703.

[10] CHANG Y，KONG L，JIA K J，et al. Chinese named entity recognition method based on BERT[C]// 2021 IEEE International Conference on Data Science and Computer Application(ICDSCA). 2021：294-299.

[11] DONG C H，ZHANG J J，ZONG C Q，et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]// 5th CCF Conference on Natural Language Processing and Chinese Computing. 2016：239-250.

[12] LIU Y H，LIU C J，XU R F，et al. Utilizing glyph feature and iterative learning for named entity recognition in finance text[J]. Journal of Chinese Information Processing，2020，34(11)：74-83.

[13] ZHANG D, WANG M T, CHEN W L. Named entity recognition combining wubi glyphs with contextualized character embeddings[J]. Computer Engineering, 2021, 47(3): 94-101.
[14] MENG Y X, WU W, WANG F, et al. Glyce: Glyph-vectors for chinese character representations[J]. Advances in Neural Information Processing Systems, 2019, 32: 2746-2757.
[15] SONG C H, SEHANOBISH A. Using chinese glyphs for named entity recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13921-13922.
[16] XUAN Z Y, BAO R, JIANG S Y. FGN: Fusion glyph network for Chinese named entity recognition[C]// China Conference on Knowledge Graph and Semantic Computing. 2020: 28-40.
[17] 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.
[18] 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 (EMNLP-IJCNLP). 2019: 1039-1049.
[19] SUI D B, CHEN Y B, LIU K, et al. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network[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). 2019: 3821-3831.
[20] 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. 2020: 6836-6842.
[21] LIU W, XU T G, XU Q H, 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, Volume 1 (Long and Short Papers). 2019: 2379-2389.
[22] MA R T, PENG M L, 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. 2020: 5951-5960.
[23] 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 (Volume 1: Long Papers). 2021: 5847-5858.
[24] LI X Y, FENG J R, MENG Y X, et al. A Unified MRC Framework for Named Entity Recognition[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5849-5859.
[25] YAN H, GUI T, DAI J Q, et al. A Unified Generative Framework for Various NER Subtasks[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: 5808-5822.
[26] JIMENEZ G B, MCNEAL N, WASHINGTON C, et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again[C]// Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 4497-4512.
[27] LI J Y, 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: 10965-10973.
[28] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Au-dio, Speech, and Language Processing, 2021, 29: 3504-3514.
[29] CUI L Y, ZHANG Y. Hierarchically-Refined Label Attention Network for Sequence Labeling[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). 2019: 4115-4128.
[30] DONG Y, CORDONNIER J B, LOUKAS A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth[C]// International Conference on Machine Learning. 2021: 2793-2803.
[31] LEVOW G A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.
[32] 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.
[33] ZHU Y Y, WANG G X. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 3384-3393.

LIAO Meng, born in 1997, postgraduate. His main research interests include information extraction and natural language processing.

LI Tianrui, born in 1969, Ph.D, professor, Ph.D supervisor, is a distinguished member of CCF (No. 05237D). His main research interests include big data intelligence, urban computing, rough sets and granular computing.

(责任编辑:李亚辉)