

# 基于Transformer编码器的中文命名实体识别模型

司逸晨, 管有庆

(南京邮电大学 物联网学院, 南京 210003)

**摘要:** 命名实体识别是自然语言处理中的重要任务,且中文命名实体识别相比于英文命名实体识别任务更具难度。传统中文实体识别模型通常基于深度神经网络对文本中的所有字符打上标签,再根据标签序列识别命名实体,但此类基于字符的序列标注方式难以获取词语信息。提出一种基于Transformer编码器的中文命名实体识别模型,在字嵌入过程中使用结合词典的字向量编码方法使字向量包含词语信息,同时针对Transformer编码器在注意力运算时丢失字符相对位置信息的问题,改进Transformer编码器的注意力运算并引入相对位置编码方法,最终通过条件随机场模型获取最优标签序列。实验结果表明,该模型在Resume和Weibo中文命名实体识别数据集上的F1值分别达到94.7%和58.2%,相比于基于双向长短期记忆网络和ID-CNN的命名实体识别模型均有所提升,具有更优的识别效果和更快的收敛速度。

**关键词:** 自然语言处理;中文命名实体识别;Transformer编码器;条件随机场;相对位置编码

开放科学(资源服务)标志码(OSID):



中文引用格式:司逸晨,管有庆.基于Transformer编码器的中文命名实体识别模型[J].计算机工程,2022,48(7):66-72.

英文引用格式:SI Y C, GUAN Y Q. Chinese named entity recognition model based on Transformer encoder[J]. Computer Engineering, 2022, 48(7): 66-72.

## Chinese Named Entity Recognition Model Based on Transformer Encoder

SI Yichen, GUAN Youqing

(School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**[Abstract]** Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP), and compared with English NER, Chinese NER is often more difficult to achieve. Traditional Chinese entity recognition models are usually based on deep neural networks used to label all characters in the text. Although they identify named entities according to the label sequence, such character-based labeling methods have difficulty obtaining the word information. To address this problem, this paper proposes a Chinese NER model based on the Transformer encoder. In the word embedding layer of the model, the word vector coding method is used in combination with a dictionary, such that the char vector contains the word information. At the same time, to solve the problem in which the Transformer encoder loses the relative position information of the characters during an attention calculation, this paper modifies the attention calculation method of the Transformer encoder and introduces a relative position coding method. Finally, a Conditional Random Field (CRF) model is introduced to obtain the optimal tag sequence. The experimental results show that the F1 value of this model when applied to the Resume dataset reaches 94.7%, and on the Weibo dataset reaches 58.2%, which are improvements in comparison with traditional NER models based on a Bidirectional Long Short-Term Memory (BiLSTM) network and Iterated Dilated Convolution Neural Network (ID-CNN). In addition, it achieves a better recognition and faster convergence speed.

**[Key words]** Natural Language Processing (NLP); Chinese Named Entity Recognition (NER); Transformer encoder; Conditional Random Field (CRF); relative position encoding

DOI: 10.19678/j.issn.1000-3428.0061432

## 0 概述

自然语言处理 (Natural Language Processing, NLP) 是计算机科学、人工智能领域的重要研究方

向,旨在使计算机理解人类的语言并进行有效交互。命名实体识别 (Named Entity Recognition, NER) 是自然语言处理中的关键技术,主要用于识别语句中人名、地名、机构名、专有名词等包含特定意义的实

基金项目:江苏省高校自然科学基金项目(05KJD520146)。

作者简介:司逸晨(1996—),男,硕士研究生,主研方向为自然语言处理;管有庆,副教授、硕士。

收稿日期:2021-04-25 修回日期:2021-08-13 E-mail: 857554195@qq.com

体,广泛应用于文献关键词提取、电子病历疾病特征抽取等任务,可细分为通用领域的命名实体识别以及金融、医疗、军事等特定领域<sup>[1]</sup>的命名实体识别。早期研究多数基于词典和规则进行命名实体识别,之后机器学习技术被广泛应用于命名实体识别任务中。近几年,随着计算机性能的不不断提升,深度学习技术大幅提升了命名实体识别的准确率。

基于深度神经网络的命名实体识别模型一般将命名实体识别任务视作序列标注任务,对文本中的每一个字打上对应的标签,根据标签序列识别命名实体。目前,主流的基于深度学习的序列标注模型通常采用字嵌入层、编码层和解码层三层结构,文本中的字首先通过字嵌入层生成对应的字向量,然后在编码层进行上下文编码以学习语义,最后在解码层中生成对应的标签,不同的命名实体识别模型均是针对这三层进行改进<sup>[2-3]</sup>。在自然语言处理任务中,循环神经网络(Recurrent Neural Network, RNN)被广泛应用于各种任务的编码层,其中双向长短期记忆(Bidirectional Long Short-Term Memory, BiLSTM)网络是命名实体识别任务中常见的循环网络结构。文献[3]提出基于BiLSTM和条件随机场(Conditional Random Field, CRF)的命名实体识别模型,利用BiLSTM的双向编码能力进行前后文编码,通过CRF学习标签间的序列顺序,是目前主流的命名实体识别模型。文献[4]提出的Lattice-LSTM模型在BiLSTM模型的基础上进行改进,通过对编码层进行修改可在字向量中编码词语信息。文献[5-7]研究表明BiLSTM采用的门结构虽然能帮助解决梯度消失问题,但是三个门单元也导致了计算量的增加,延长了模型训练时间,而Lattice-LSTM对编码层的改进进一步增加了模型训练负担<sup>[8]</sup>。近几年,文献[9]提出的Transformer机器翻译模型被广泛应用于各自然语言处理任务,其基于注意力机制获取文本中字符间的长距离依赖,采用的并行结构也可以提升模型训练效率。但在命名实体识别任务中,使用Transformer作为编码器的性能表现并不理想。文献[10-12]指出Transformer机器翻译模型采用的绝对位置编码在经过模型自身注意力运算后会丢失字符中的相对位置信息,影响最终识别效果。

虽然BiLSTM模型在命名实体识别任务中表现较好,但是BiLSTM训练速度较慢。Lattice-LSTM模型通过对编码层的改进在字向量中添加了词信息,但进一步增加了模型的计算负担。Transformer编码器因为丢失了字符相对位置信息,无法充分发挥其性能优势。针对上述问题,本文提出一种基于Transformer编码器的中文命名实体识别模型。在字嵌入层中,使用结合词典的字向量编码方法将词语信息嵌入字向量。在Transformer编码器层中,改进自注意力计算方式,同时引入相对位置编码方法,而在模型中加入相对位置信息。

## 1 中文命名实体识别模型

基于Transformer编码器的命名实体识别模型的整体可以分为字嵌入层、Transformer编码器层和条件随机场层三层。在字嵌入层中,使用结合词典的字向量编码方法生成包含词语信息的字向量。在Transformer编码器层中,对字向量进一步编码以学习前后文特征,同时通过修改注意力运算方式和引入相对位置编码,取得字符的相对位置信息。最终通过条件随机场层获取最优标签序列,根据标签序列识别命名实体。基于Transformer编码器的命名实体识别模型如图1所示,其中,输出的“B”标签代表命名实体的开头,“I”标签代表命名实体的结尾,“O”标签代表这个词不是命名实体,在Transformer编码器层中包含多个Transformer编码器。

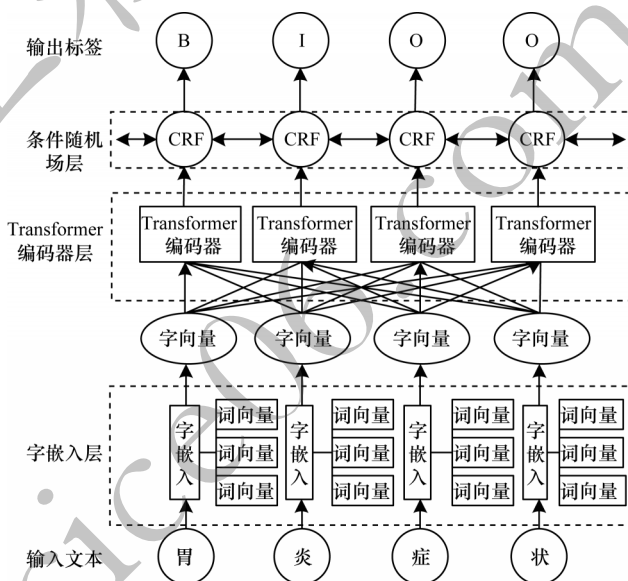


图1 基于Transformer编码器的中文命名实体识别模型

Fig.1 Chinese NER model based on Transformer encoder

### 1.1 结合词语信息的字嵌入层

在命名实体识别模型的字嵌入层中,需要将输入语句的每一个字映射为固定维度的字向量,以便后续的编码。在中文命名实体识别任务中,基于字符的编码方法难以利用词语的信息,因此本文提出一种结合词典的字向量编码方法,使生成的字向量可以包含词语的信息。

对于字向量的生成,首先需要进行字嵌入模型的选择。Word2Vec是一款经典的语言嵌入模型<sup>[13-15]</sup>,具体实现了Skip-Gram(跳字)和连续词袋(Continue Bag-of-Words, CBOW)两种模型,其中跳字模型的核心思想是使用中心字预测背景字,连续词袋模型的核心思想是使用背景字预测中心字。这两种模型都可以在不进行人工标注的前提下利用神经网络训练生成字向量,并且字向量中包含了上下文的信息<sup>[16]</sup>,然而在实际实验中,一般使用跳字模型生成字向量。

在选择好字嵌入模型后,将介绍融入词语信息的字向量编码方法。Lattice-LSTM模型<sup>[4]</sup>对LSTM的结构作了大幅修改,在字嵌入的同时引入词信息,并最终证明了在字向量中加入词语信息可以增强中文命名实体识别准确率<sup>[17]</sup>。但是,Lattice-LSTM模型<sup>[4]</sup>对LSTM的修改增加了训练时需要更新的参数量,增加了模型计算开销,同时这种修改难以应用于使用其他神经网络进行编码的命名实体识别模型。针对上述问题,本文提出一种相对简单的在字嵌入层引入词语信息的字向量编码方法。该方法只对命名实体识别模型的字嵌入层进行修改,从而保证了模型整体计算效率不受太大影响,同时该方法也具有较强的可移植性。

字向量编码方法的具体步骤如下:1)对于输入文本进行分句处理;2)使用Lattice-LSTM模型中开源的中文分词词典作为句中每个字对应的词典,其中约包括29万双字符词汇和28万三字符词汇;3)对于文本中的每一个字符 $c$ ,根据词典匹配句子中所有包含该字符的词,使用 $B(c)$ 、 $M(c)$ 、 $E(c)$ 3个集合编码这个字包含的词信息,其中, $B(c)$ 表示所有以字符 $c$ 开头且长度大于1的词, $M(c)$ 表示包含字符 $c$ 且字符 $c$ 不在开头和末尾的词, $E(c)$ 表示以字符 $c$ 结尾且长度大于1的词,如果集合为空,则添加一个特殊的空词None到集合中。如图2所示,字符 $c_5$ “胃”出现在词“肠胃炎”的中间、词“胃炎”的首部、词“肠胃”的底部,因此对应的词向量集合 $B(c_5)$ 为{“胃炎”}、 $E(c_5)$ 为{“肠胃”}、 $M(c_5)$ 为{“肠胃炎”},这样可将句中字符“胃”对应的3个词的信息“肠胃”、“胃炎”、“肠胃炎”通过字符的3个集合进行完整收录。

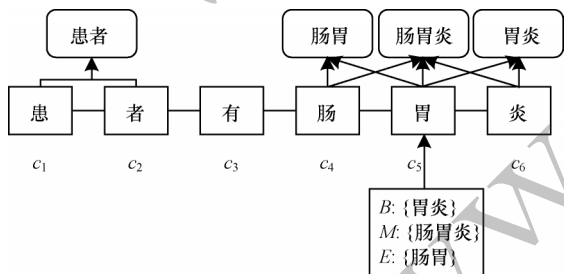


图2 融合词语信息的编码示意图

Fig.2 Schematic diagram of encoding fusing word information

在获得每个字符的 $B$ 、 $M$ 、 $E$ 3个词语集合后,根据创建的3个集合,将词语信息融入到字向量中,构造新的字向量,如式(1)所示:

$$\mathbf{x}^{\text{new}} = [\mathbf{x}^c; \mathbf{x}^{(B,M,E)}] \quad (1)$$

其中: $\mathbf{x}^{\text{new}}$ 表示最终生成的包含词语信息的字向量; $\mathbf{x}^c$ 表示根据跳字模型直接使用Word2Vec模型训练生成的字向量; $\mathbf{x}^{(B,M,E)}$ 表示根据 $B$ 、 $M$ 、 $E$ 3个词语集合生成的特征向量; $[\mathbf{x}^c; \mathbf{x}^{(B,M,E)}]$ 表示字向量和特征向量的拼接。 $\mathbf{x}^{(B,M,E)}$ 的具体生成方法如下:

$$\mathbf{x}^{(B,M,E)} = [\mathbf{v}^{(B)}, \mathbf{v}^{(M)}, \mathbf{v}^{(E)}] \quad (2)$$

其中: $[\mathbf{v}^{(B)}, \mathbf{v}^{(M)}, \mathbf{v}^{(E)}]$ 表示根据 $B$ 、 $M$ 、 $E$ 3个词语集合生成的特征向量的拼接。每个特征向量 $\mathbf{v}^{(s)}$ 的计算公式如下:

$$\mathbf{v}^{(s)} = \frac{1}{|s|} \sum_{w \in s} \mathbf{e}^w \quad (3)$$

其中: $s$ 表示 $B$ 、 $M$ 、 $E$ 中任意一个词语集合; $|s|$ 表示集合中词的总数; $\mathbf{v}^{(s)}$ 表示集合对应的特征向量; $w$ 表示词语集合中的词; $\mathbf{e}^w$ 表示词 $w$ 对应的词向量。通过式(3)实现了在字向量中加入词语信息,从而丰富了字向量的特征。

## 1.2 加入相对位置信息的Transformer编码器层

Transformer编码器的具体结构如图3所示,编码器的输入为之前生成的字向量,由于Transformer没有使用递归和卷积的方式编码字的位置信息,因此添加了一种额外的位置编码来表示序列中每个字的绝对位置信息。

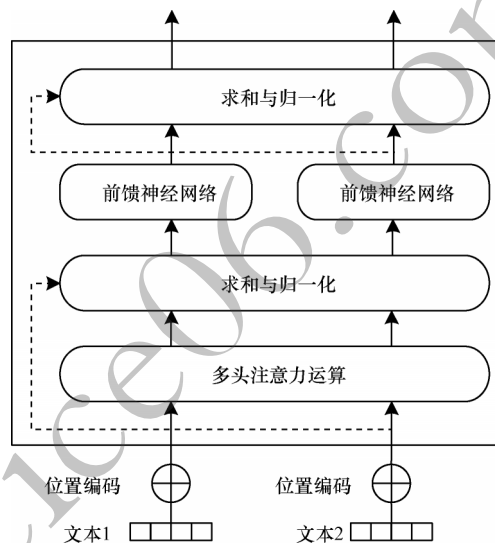


图3 Transformer编码器结构

Fig.3 Structure of Transformer encoder

位置编码的计算如式(4)和式(5)所示:

$$P_{PE_{(l,2i)}} = \sin\left(\frac{1}{10000^{2i/d}}\right) \quad (4)$$

$$P_{PE_{(l,2i+1)}} = \cos\left(\frac{1}{10000^{2i/d}}\right) \quad (5)$$

其中: $P_{PE}$ 为二维矩阵,矩阵的列数和之前生成的字向量维数相同, $P_{PE}$ 中的行表示文本中每一个字对应的位置向量,列表示位置向量的维度,位置向量的总维数等于字向量的总维数; $l$ 表示字在输入文本中的索引; $d$ 表示位置向量的总维数; $i$ 表示位置向量具体的维度,取值范围为 $\left[0, \frac{d}{2} - 1\right]$ ;  $P_{PE_{(l,2i)}}$ 表示索引为 $l$ 的字的位置向量在偶数维度的值,使用正弦函数计算; $P_{PE_{(l,2i+1)}}$ 表示索引为 $l$ 的字的位置向量在奇数维度的值,使用余弦函数计算;Transformer编码器中将 $\frac{1}{10000^{2i/d}}$ 作为三角函数的输入,使相对距离越大的



输入产生的相关性越弱,并将位置编码和字向量相加得到最终的字向量。

为便于计算,Transformer编码器使用绝对位置编码方法,但是这种编码方法在经过Transformer编码器内部的注意力运算后会丢失相对位置信息。假设输入序列为 $X$ ,根据Transformer编码器的注意力计算方法,序列中第 $i$ 个字和第 $j$ 个字的注意力计算分数如式(6)所示:

$$A_{Att_{ij}} = (W_q(V_i + P_i))^T(W_k(V_j + P_j)) \quad (6)$$

其中: $W_q$ 和 $W_k$ 是注意力计算中使用的生成查询向量的权重矩阵和生成键向量的权重矩阵; $V_i$ 和 $V_j$ 是第 $i$ 个字和第 $j$ 个字的字向量; $P_i$ 和 $P_j$ 是第 $i$ 个字和第 $j$ 个字的位置向量。对式(6)进行因式分解得到式(7):

$$A_{Att_{ij}} = V_i^T W_q^T W_k V_j + V_i^T W_q^T W_k P_j + P_i^T W_q^T W_k V_j + P_i^T W_q^T W_k P_j \quad (7)$$

其中: $V_i^T W_q^T W_k V_j$ 不包含位置编码; $V_i^T W_q^T W_k P_j$ 只包含序列中第 $j$ 个字的位置向量 $P_j$ ; $P_i^T W_q^T W_k V_j$ 只包含第 $i$ 个字的位置向量 $P_i$ ; $P_i^T W_q^T W_k P_j$ 中同时包含序列中第 $i$ 个字和第 $j$ 个字的位置向量 $P_i$ 和 $P_j$ 。事实上,根据Transformer编码器的编码方式, $P_i^T P_j$ 包含相对位置信息。对于文本中任意一个字符 $i$ ,将位置向量展开如式(8)所示:

$$P_i = \begin{bmatrix} \sin(\omega_0 i) \\ \cos(\omega_0 i) \\ \vdots \\ \sin\left(\omega_{\frac{d}{2}-1} i\right) \\ \cos\left(\omega_{\frac{d}{2}-1} i\right) \end{bmatrix} \quad (8)$$

其中: $\omega_n = \frac{1}{10000^{2n/d}}$ , $n$ 为序列长度。根据式(8)可以

得出 $P_i^T P_j$ 的运算结果,如式(9)所示:

$$P_i^T P_j = \sum_{n=0}^{\frac{d}{2}-1} [\sin(\omega_n i) \sin(\omega_n (i+k)) + \cos(\omega_n i) \cos(\omega_n (i+k))] \quad (9)$$

其中: $k$ 表示字符 $i$ 和字符 $j$ 的距离, $k=j-i$ 。由三角函数的性质可知, $\cos(a-b) = \sin(a)\sin(b) + \cos(a)\cos(b)$ ,因此将式(9)化简可得:

$$P_i^T P_j = \sum_{n=0}^{\frac{d}{2}-1} \cos(\omega_n i - (i+k)) = \sum_{n=0}^{\frac{d}{2}-1} \cos(\omega_n k) \quad (10)$$

由式(10)可知: $P_i^T P_j$ 的结果只与字符 $i$ 和字符 $j$ 的距离 $k$ 有关,即 $P_i^T P_j$ 表示字符 $i$ 和字符 $j$ 的相对位置关系,但是在 $P_i^T W_q^T W_k P_j$ 中加入了两个可训练的参数 $W_q$ 和 $W_k$ ,即进行一次未知的线性变换,这样会导致相对位置信息丢失。如图4所示,上方一条曲线表示 $P_i^T P_j$ 的运算结果,中间和下方两条曲线分

别表示 $P_i^T W_1 P_j$ 和 $P_i^T W_2 P_j$ ,其中 $W_1$ 和 $W_2$ 是两个随机的参数矩阵,可以看出随着字符 $i$ 和字符 $j$ 之间距离 $k$ 的变化, $P_i^T P_j$ 的曲线与 $k$ 有明显的关联并呈现对称性,这说明Transformer编码器使用的位置编码可以感知字符之间的相对距离变化,但是这种相对距离感知对方向不敏感。加入了随机参数矩阵的两条曲线失去了和距离 $k$ 的关联。这也证明了在经过注意力运算后Transformer的位置编码丢失了相对位置信息,从而验证了Transformer编码器中相对位置信息的丢失会影响其在命名实体识别任务中的性能表现<sup>[17]</sup>。

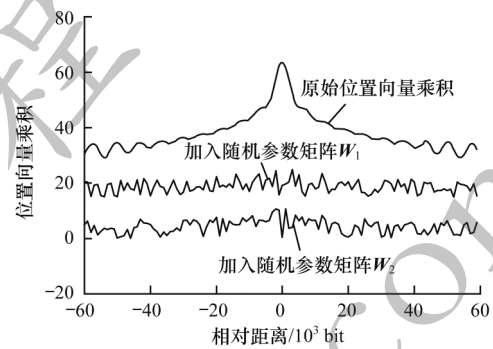


图4 Transformer位置向量乘积结果可视化

Fig.4 Visualization of product result of position vector

为加强Transformer编码器对相对位置的感知能力,在文献[17]研究的基础上,对式(7)中Transformer编码器的注意力计算公式进行修改。相比于文献[17],没有选择锐化Transformer的注意力矩阵,而是通过基于正弦函数的相对位置编码减少模型的注意力参数,同时保留字符间的距离信息和相对位置信息,提升模型在中文命名实体识别任务中的性能表现,计算公式如式(11)所示:

$$A_{Att_{ij}} = V_i^T W_q^T W_k V_j + V_i^T W_q^T W_k R_{i,j} + u^T W_k V_j + v^T W_k R_{i,j} \quad (11)$$

其中: $u$ 和 $v$ 表示可学习的参数向量; $R_{i,j}$ 是根据式(10)修改的相对位置编码。因为在引入相对位置编码后无需再使用注意力机制中的查询向量查询字符 $i$ 的绝对位置 $P_i$ ,所以使用参数向量 $u$ 和 $v$ 替换式(7)中的 $P_i^T W_q^T$ ,其中 $W_q$ 、 $u$ 和 $v$ 均是可学习的参数。 $R_{i,j}$ 表示字符 $i$ 和字符 $j$ 的相对位置编码,替换了式(7)中代表相对位置信息的 $P_i^T P_j$ ,相对位置编码的具体编码方式如式(12)所示:

$$R_{i,j} = \sum_{n=0}^{\frac{d}{2}-1} \sin(\omega_n k) \quad (12)$$

相对位置编码方法实质上是将式(10)中的 $\cos$ 函数替换成 $\sin$ 函数。在式(10)中因为三角函数 $\cos(-x) = \cos(x)$ 导致Transformer编码器使用的原始位置编码对相对距离的感知缺乏方向性,而 $\sin(-x) = -\sin(x)$ ,所以相对位置编码 $R_{i,j}$ 对方向敏感。通过上述修改,Transformer编码器在进行注意力运

算后不会再丢失相对位置信息,在感知字符距离变化的同时也具备了方向感知能力。

### 1.3 条件随机场层

在本文命名实体识别模型中,Transformer 编码器层只能获取包含进一步上下文信息的字向量,即使加入了词语信息和相对位置编码,也无法考虑最终预测标签之间的依赖关系,比如标签 I 必须在标签 B 后。因此,模型中采用条件随机场层考虑标签之间的相邻关系来获取全局最优的标签序列。条件随机场模型是一种经典的判别式概率无向图模型,该模型经常被应用于序列标注任务<sup>[18]</sup>,对于输入句子  $x=(x_1, x_2, \dots, x_n)$ , 句子标签序列  $y=(y_1, y_2, \dots, y_n)$  的打分如式(13)所示:

$$S(x, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (13)$$

其中:  $A$  为转移得分矩阵;  $A_{y_i, y_{i+1}}$  表示由标签  $y_i$  转移到标签  $y_{i+1}$  的转移得分;  $y_0$  和  $y_{n+1}$  表示句中起始和终止标签,这两个标签只在计算时临时添加;  $P_{i, y_i}$  表示第  $i$  个字被标记为  $y_i$  的概率。用 softmax 函数归一化得到  $y$  标签序列的最大概率,如式(14)所示:

$$P(y|x) = \frac{e^{S(x, y)}}{\sum_{\tilde{y} \in Y_x} e^{S(x, \tilde{y})}} \quad (14)$$

其中:  $\tilde{y}$  表示真实的标签序列;  $Y_x$  表示所有可能标签序列的集合。使用最大似然估计法求解模型的最小化损失函数值,如式(15)所示:

$$L_{\text{Loss}} = S(x, y) - \sum_{\tilde{y} \in Y_x} S(x, \tilde{y}) \quad (15)$$

其中:  $L_{\text{Loss}}$  表示损失函数。使用经过标注的文本迭代训练命名实体识别模型直至损失函数 Loss 小于阈值  $\varepsilon$ ,  $\varepsilon$  为事先设定好的常量。利用维特比算法求得全局最优序列,最优序列为最终命名实体识别模型的标注结果,如式(16)所示:

$$y^* = \underset{\tilde{y} \in Y_x}{\operatorname{argmax}} S(x, \tilde{y}) \quad (16)$$

其中:  $y^*$  为集合中使得分函数取得最大值的标签序列。

## 2 实验结果与分析

将基于 Transformer 编码器的命名实体识别模型与其他基于深度学习的命名实体识别模型进行性能对比,使用 Weibo 和 Resume 中文命名实体识别数据集进行实验,利用精确率、召回率以及 F1 值作为实验主要的评估指标,通过实验结果验证基于 Transformer 编码器的命名实体识别模型性能。

### 2.1 实验数据准备

Weibo 数据集来源于新浪微博上选取的标注信息,具体包括 2013 年 11 月至 2014 年 12 月约 1 900 条信息<sup>[8]</sup>。Resume 数据集来源于新浪金融上的中文简历信息,包含人名、种族、职称等 8 类实体,共涉及 4 731 条经过标注的中文简历信息<sup>[7]</sup>。2 个数据集的详细统计信息如表 1 所示。

表 1 数据集统计信息

名称	类型	训练集 样本量	验证集 样本量	测试集 样本量
Weibo	Sentence	1.4	0.27	0.27
	Char	73.8	14.50	14.80
Resume	Sentence	3.8	0.46	0.48
	Char	124.1	13.90	15.10

### 2.2 实验环境与参数设置

实验模型采用复旦大学提供的开源自然语言处理框架 FastNLP 搭建<sup>[19]</sup>,使用 Dropout 算法防止模型过拟合。实验环境设置如表 2 所示。实验中的超参数设置如表 3 所示。模型性能对于超参数学习率和 Batch Size 较为敏感。在实际操作中,Batch Size 选择 16,通过使用小批量的样本集增加模型迭代次数,更快达到拟合点,对应选择 0.001 的学习率以保持训练稳定性,同时将 Dropout 设为 0.3 以防止模型过拟合。

表 2 实验环境设置

实验环境	配置
操作系统	Windows10 家庭中文版
CPU	Intel Core i5-9300h
GPU	NVIDIA GeForce GTX 1660Ti
内存/GB	64
Python	3.7
FastNLP	0.6.0
Tensorflow	1.12

表 3 实验超参数设置

参数名	参数值
Word2Vec 字向量维度	300
词窗大小	10
学习率	0.001
最大序列长度	64
TransformerHiddenSize	200
注意力机制头数	6
Batch Size	16
Epoch	50
Dropout	0.3

2.3 与其他模型的对比结果与分析

引入基于ID-CNN+CRF的命名实体识别模型(简称为ID-CNN+CRF)<sup>[20]</sup>和经典的基于BiLSTM+CRF的命名实体识别模型(简称为BiLSTM+CRF)作为对比模型,在Weibo和Resume数据集上分别进行对比实验。由于基于Transformer编码器的命名实体识别模型中加入了相对位置信息,简称为Transformer+Relative Position+CRF。在Resume数据集上3种模型的实验结果如表4所示,F1值变化曲线如图5所示。从表4和图5可以看出,基于Transformer编码器的命名实体识别模型在Resume数据集上取得了最优结果,F1值达到了94.7%,略高于基于BiLSTM+CRF的命名实体识别模型和基于ID-CNN+CRF的命名实体识别模型。同时,基于Transformer编码器的命名实体识别模型在第20个Epoch时F1值开始增长缓慢,模型趋近于收敛,说明基于Transformer编码器的命名实体识别模型相比基于BiLSTM的命名实体识别模型和基于ID-CNN的命名实体识别模型具有更快的收敛速度。

表 4 Resume数据集上3种模型的实验结果

Table 4 Experimental results of three models on Resume dataset

on Resume dataset				%
模型	精确率	召回率	F1值	
Transformer+Relative Position+CRF	94.6	94.8	94.7	
BiLSTM+CRF	92.5	94.3	93.4	
ID-CNN+CRF	90.7	91.9	91.3	

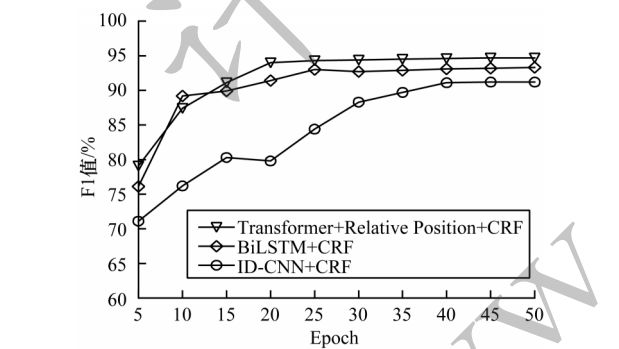


图5 3种模型在Resume数据集上的F1值变化曲线

Fig.5 F1 value change curves of three models on Resume dataset

在Weibo数据集上3种模型的实验结果如表5所示,F1值变化曲线如图6所示。从表5和图6可以看出,在Weibo数据集上3种模型的效果均不理想,基于Transformer编码器的命名实体识别模型的F1值仅达到58.2%,相比其他两个模型提升有限。根据对Weibo数据集的观察发现,3种模型识别效果均不佳的原因主要为:1)Weibo数据集的数据样本量较小,模型训练效果不佳;2)Weibo数据集中包含大量的人名类实体和地名类实体,基于深度学习的命名实体识别模型很难通过神经网络提取实体特征,从而影响了最终识别效果。

表 5 Weibo数据集上3种模型的实验结果

Table 5 Experimental results of three models on Weibo dataset

on Weibo dataset				%
模型	精确率	召回率	F1值	
Transformer+Relative Position+CRF	69.2	50.3	58.2	
BiLSTM+CRF	68.8	49.3	57.4	
ID-CNN+CRF	68.4	49.4	57.3	

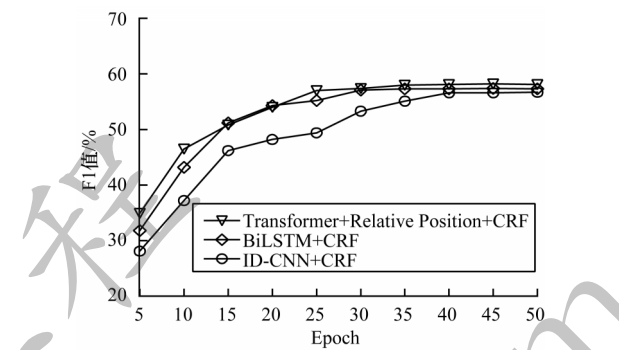


图6 3种模型在Weibo数据集上的F1值变化曲线

Fig.6 F1 value change curves of three models on Weibo dataset

在基于Transformer编码器的命名实体识别模型中,分别对字嵌入层和Transformer编码器层做了改进,其中字嵌入层使用融合词语信息的字向量编码方法,Transformer编码器层加入相对位置信息。为验证这些改动的有效性,引入原始基于Transformer+CRF的命名实体识别模型在Resume数据集上做进一步的对比实验,如图7所示。从图7可以看出,基于Transformer编码器的命名实体识别模型相比原始基于Transformer+CRF的命名实体识别模型,F1值约提升了2个百分点,证明了在字嵌入层中的词语信息及Transformer编码器层中的相对位置信息可有效提升命名实体识别模型的最终识别效果。

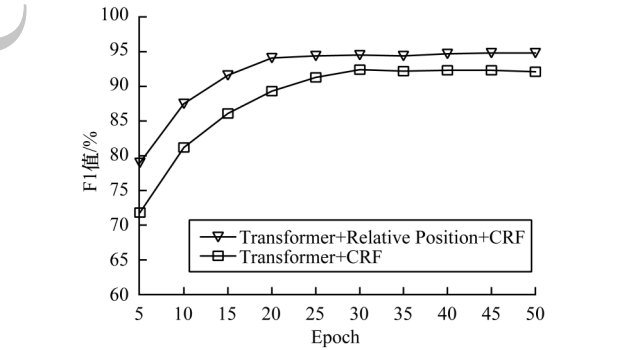


图7 2种模型在Resume数据集上的F1值变化曲线

Fig.7 F1 value change curves of two models on Resume dataset

3 结束语

本文针对中文命名实体识别过程中的词语信息丢失问题,提出一种基于Transformer编码器的中文命名实体识别模型。该模型使用结合词典的字向量编码方法使字向量中包含词语信息,通过改进Transformer编码器的注意力运算以及引入相对位置



编码方法增加字符的相对位置信息。在 Weibo 和 Resume 中文命名实体识别数据集上的实验结果表明,该模型相比于其他主流命名实体识别模型具有更好的识别效果。后续可在 MSRA 等数据集上,将该模型与其他基于深度学习的中文命名实体识别模型进行性能对比,进一步增强模型泛化能力。

### 参考文献

- [1] 殷章志,李欣子,黄德根,等.融合字词模型的中文命名实体识别研究[J].中文信息学报,2019,33(11):95-100,106.
- YIN Z Z,LI X Z,HUANG D G,et al. Chinese named entity recognition ensembled with character[J]. Journal of Chinese Information Processing, 2019, 33(11): 95-100, 106. (in Chinese)
- [2] 王红,史金钊,张志伟.基于注意力机制的LSTM的语义关系抽取[J].计算机应用研究,2018,35(5):1417-1420,1440.
- WANG H,SHI J C,ZHANG Z W. Text semantic relation extraction of LSTM based on attention mechanism[J]. Application Research of Computers, 2018, 35(5): 1417-1420, 1440. (in Chinese)
- [3] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2021-03-16]. <https://arxiv.org/abs/1508.01991v1>.
- [4] ZHANG Y,YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1554-1564.
- [5] 杜琳,曹东,林树元,等.基于BERT与Bi-LSTM融合注意力机制的中医病历文本的提取与自动分类[J].计算机科学,2020,47(S2):416-420.
- DU L,CAO D,LIN S Y,et al. Extraction and automatic classification of TCM medical records based on attention mechanism of BERT and Bi-LSTM[J]. Computer Science, 2020, 47(S2): 416-420. (in Chinese)
- [6] ZENG D H,SUN C J,LIN L,et al. LSTM-CRF for drug-named entity recognition[J]. Entropy, 2017, 19(6): 283.
- [7] YAN S,CHAI J P,WU L Y. Bidirectional GRU with multi-head attention for Chinese NER[C]//Proceedings of the 5th Information Technology and Mechatronics Engineering Conference. Washington D. C., USA: IEEE Press, 2020: 1160-1164.
- [8] DING R X,XIE P J,ZHANG X Y,et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 1462-1467.
- [9] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need?[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.
- [10] GUO S G,LIU Y P,LI H,et al. Transformer winding deformation detection based on BOTDR and ROTDR[J]. Sensors, 2020, 20(7): 2062.
- [11] DAI Z H,YANG Z L,YANG Y M,et al. Transformer-XL: attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 1-15.
- [12] SHAW P,USZKOREIT J,VASWANI A. Self-attention with relative position representations[C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1-25.
- [13] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. [2021-03-16]. <https://arxiv.org/abs/1301.3781>.
- [14] 张华伟.基于Word2Vec的神经网络协同推荐模型[J].网络空间安全,2019,10(6):25-28.
- ZHANG H W. Neural network cooperative recommendation model based on Word2Vec[J]. Cyberspace Security, 2019, 10(6): 25-28. (in Chinese)
- [15] 章跃琳.基于Word2Vec的在线商品特征提取与文本分类研究[D].温州:温州大学,2019.
- ZHANG Y L. Research on feature extraction and text classification of online commodity based on Word2Vec[D]. Wenzhou: Wenzhou University, 2019. (in Chinese)
- [16] LEI S. Research on the improved Word2Vec optimization strategy based on statistical language model[C]//Proceedings of International Conference on Information Science, Parallel and Distributed Systems. Washington D. C., USA: IEEE Press, 2020: 356-359.
- [17] YAN H,DENG B C,LI X N,et al. TENER: adapting Transformer encoder for named entity recognition[EB/OL]. [2021-03-16]. <https://arxiv.org/abs/1911.04474>.
- [18] 张应成,杨洋,蒋瑞,等.基于BiLSTM-CRF的商情实体识别模型[J].计算机工程,2019,45(5):308-314.
- ZHANG Y C,YANG Y,JIANG R,et al. Commercial intelligence entity recognition model based on BiLSTM-CRF[J]. Computer Engineering, 2019, 45(5): 308-314. (in Chinese)
- [19] DAI N,LIANG J Z,QU X P,et al. Style Transformer: unpaired text style transfer without disentangled latent representation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 5997-6007.
- [20] GAO M,XIAO Q F,WU S C,et al. An attention-based ID-CNNs-CRF model for named entity recognition on clinical electronic medical records[C]//Proceedings of International Conference on Artificial Neural Networks. Berlin, Germany: Springer, 2019: 231-242.

编辑 陆燕菲