

基于改进的 Transformer 编码器的中文命名实体识别

郑洪浩, 于洪涛, 李邵梅

(信息工程大学, 河南 郑州 450002)

摘要: 为了提高中文命名实体识别的效果, 提出了基于 XLNET-Transformer_P-CRF 模型的方法, 该方法使用了 Transformer_P 编码器, 改进了传统 Transformer 编码器不能获取相对位置信息的缺点。实验结果表明, XLNET-Transformer_P-CRF 模型在 MSRA、OntoNotes4.0、Resume、微博数据集 4 类数据集上分别达到 95.11%、80.54%、96.70%、71.46% 的 $F1$ 值, 均高于中文命名实体识别的主流模型。

关键词: 中文命名实体识别; Transformer 编码器; 相对位置信息

中图分类号: TP391

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2021041

Chinese NER based on improved Transformer encoder

ZHENG Honghao, YU Hongtao, LI Shaomei

Information Engineering University, Zhengzhou 450002, China

Abstract: In order to improve the effect of chinese named entity recognition, a method based on the XLNET-Transformer_P-CRF model was proposed, which used the Transformer_P encoder, improved the shortcomings of the traditional Transformer encoder that couldn't obtain relative position information. Experiments show that the XLNET-Transformer_P-CRF model achieves 95.11%, 80.54%, 96.70%, and 71.46% $F1$ values on the four types of data sets: MSRA, OntoNotes4.0, Resume, and Weibo, which are all higher than other mainstream chinese NER model.

Keywords: Chinese named entity recognition, Transformer encoder, relative position information

1 引言

命名实体识别 (NER, named entity recogni-

tion) 最早是在 1995 年 11 月的 MUC-6 会议中提出的信息抽取子任务, 主要是识别文本中的 Entity Name (人名、地名、机构名), Temporal Expressions

收稿日期: 2020-08-13; 修回日期: 2020-12-25

通信作者: 郑洪浩, 1140820290@qq.com

基金项目: 国家自然科学基金青年基金(62002384), 国家重点研发计划(2016QY03D0502), 郑州市协同创新重大专项(162/32410218)

Foundation Items: The National Natural Science Foundation of China (62002384), The National Key R&D Program of China (2016QY03D0502), Major Collaborative Innovation Projects of Zhengzhou (162/32410218)

论文引用格式: 郑洪浩, 于洪涛, 李邵梅. 基于改进的 Transformer 编码器的中文命名实体识别[J]. 网络与信息安全学报, 2021, 7(5): 105-112.

ZHENG H H, YU H T, LI S M. Chinese NER based on improved Transformer encoder[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 105-112.

(日期、时间、持续时间)和 Number Expressions (货币、度量衡、百分比表达式)^[1]。当前,命名实体识别成为自然语言处理任务的重要组成部分,在智能问答^[2]、机器翻译^[3]、信息检索^[4]等复杂的自然语言处理领域发挥着重要的作用。命名实体识别的基本原理是通过序列标注方法,在对每个字进行标注的基础上,预测实体的边界和类型。早期的命名实体识别主要使用基于规则和字典的方法,该方法效率较低、灵活性差,且往往需要大量的领域知识。

随着深度学习技术的发展及其在自然语言处理领域的广泛应用,基于深度学习的命名实体识别逐渐成为主流。深度学习技术可以直接将原始文本通过多步的特征抽取、转换和组合得到一种特征表示,并进一步输入预测函数得到实体识别结果^[5]。深度学习不需要手工特征提取这一耗时费力的工作,且端到端的学习模式避免了错误传播。

基于深度学习的命名实体识别模型通常可以分为嵌入层、编码层和输出层三部分,其中,嵌入层旨在将字词级别的特征转化为特征向量,编码层旨在获取文本上下文特征,输出层旨在获取序列之间的规则特征并对编码层输出的特征向量进行分类^[6]。目前,基于深度学习的主流研究是围绕这三层的功能实现展开的,各个模型的异同也主要体现在这三层的结构上。

对于编码层,基于 RNN 结构的编码器因其在处理序列类时间流数据上的优势被广泛使用^[7-10],但其在结构上存在串行计算、梯度消失(爆炸)^[11]和单向建模的问题,这些问题限制了基于 RNN 结构的编码器在命名实体识别任务上的效果。Transformer 编码器^[12]可以有效地解决基于 RNN 结构的编码器存在的 3 个问题,具体解决方案如下。①Transformer 编码器不同于基于 RNN 结构的编码器的串行计算结构,采用了并行计算结构以充分利用计算机的并行计算资源。②Transformer 编码器采用自注意力机制,在结构上消除了梯度消失和梯度爆炸的问题,可以获取长文本的依赖信息。③Transformer 编码器不同于基于 RNN 结构的编码器的双向拼接,可以实现双向参数的统一更新,不会割裂上下文

关系。然而,Guo 等^[13]的研究表明 Transformer 编码器直接应用到命名实体识别领域并不会得到有效提升。

本文对 Transformer 编码器在命名实体识别任务上效果较差的现象进行了分析,并提出了具体的改进。Transformer 编码器的功能实现主要源于自注意力机制,自注意力机制是无法从结构上获取相对位置信息的^[14-15],其中相对位置信息主要指字之间的距离和方向信息。然而,相对位置信息在命名实体识别任务中起到了重要的作用,如在句子“华纳兄弟创立了华纳兄弟公司”中,地点实体通常在“创立了”之后,人名实体通常在“创立了”之前。可见,字之间的相对位置信息在命名实体识别任务中极为重要。针对该问题,本文对 Transformer 编码器嵌入相对位置信息,即通过对注意力机制的扩展,将输入结构建模成一个含有方向性信息的结构。实验结果显示,嵌入相对位置信息的改进有效提升了 Transformer 模型在命名实体识别任务上的效果,改进后的 Transformer 编码器命名为 Transformer_P。

此外,本文对嵌入层模型进行了研究与实验。传统的深度学习模型采用 Word2vec 等模型^[16]生成静态字向量,静态字向量虽然可以携带字的简单语义信息,但是无法表征字的多义性,如在句子“陆军上校重新校对了作战方案”中,静态字向量对于“校”一字都以固定的向量来表示,然而,前后两个“校”字的语义信息截然不同,前者代指一种军衔,后者代指一种动作,这类一字多义的问题影响了命名实体识别的效果。针对该问题,杨飘等^[17]提出了一种用 BERT 预训练模型^[18]生成的动态字向量替换静态量的方法,实验表明基于动态字向量的方法有更好的表现。然而,以下两个问题会限制 BERT 的使用效果。一是忽视了训练时被掩盖(Mask)掉的符号(token)之间的相关关系;二是训练时使用掩盖策略,而实际应用中并没有,造成训练与实际应用不一致的问题。2019 年, Yang 等^[19]提出 XLNET 模型,该预训练语言模型利用自回归语言模型天然优势,避免了 BERT 模型训练应用不一致和无法获取 token 之间相关关系的问题。鉴于此,本文将 XLNET 模型引入嵌入层以表征字的多义性。综上所述,

为了提升中文命名实体识别的效果, 本文改进了传统 Transformer 编码器不能获取相对位置的缺陷, 并提出了基于 XLNET-Transformer_P-CRF 模型的中文命名实体识别方法。

2 相关工作

Transformer 是 Vaswani 等^[12]提出的一种利用自注意力机制的编码器, 其编码层由两个子层构成, 自注意力层和全连接层。与基于 RNN 结构的编码器相比, Transformer 在结构上并不能直接获取绝对位置信息和相对位置信息^[12,18]。为避免该问题, Vaswani 等^[12]设计了由不同频率的正弦编码组成的位置向量, 并采用位置向量与字向量相加的方式嵌入位置信息。本节对传统 Transformer 编码器的核心模块自注意力层和位置向量分别进行了介绍。

2.1 自注意力层

Transformer 编码能力的实现主要依赖于多头自注意力机制。对于每头注意力, 输入为序列 $H = \{h_1, \dots, h_n\}$, $h_i \in \mathbb{R}^{d_H}$, 输出为序列 $Z = \{z_1, \dots, z_n\}$, $z_i \in \mathbb{R}^{d_Z}$, 其中, n 为输入序列的长度, d_H 为输入序列的维度, d_Z 为输出向量的维度。自注意力层计算过程如下。

首先, 将每个输入向量 h_i 映射到 3 个不同的空间, 得到查询向量 $q_i \in \mathbb{R}^{d_Z}$ 、键向量 $k_i \in \mathbb{R}^{d_Z}$ 和值向量 $v_i \in \mathbb{R}^{d_Z}$ 。

$$\begin{cases} q_i = h_i W^Q \\ k_i = h_i W^K \\ v_i = h_i W^V \end{cases} \quad (1)$$

其中, $W^Q, W^K, W^V \in \mathbb{R}^{d_H \times d_Z}$ 为可训练的参数矩阵。

其次, 对于输入向量 h_j 和 h_i , 通过式(2), 可以得到输入向量之间的注意力得分 A_{ji} 。

$$A_{ji} = q_j k_i^T \quad (2)$$

然后, A_{ji} 通过缩放因子 $A_{i,j} = Q_i K_j^T$ 缩放后, 归一化为权重 a_{ji} 。

$$a_{ji} = \frac{\exp\left(\frac{A_{ji}}{\sqrt{d_Z}}\right)}{\sum_{m=1}^n \exp\left(\frac{A_{jm}}{\sqrt{d_Z}}\right)} \quad (3)$$

最后, 利用式(4)加权求和, 得到输出向量 z_j 。

$$z_j = \sum_{l=1}^n a_{jl} v_l \quad (4)$$

其中, $j, l = [1, n]$ 为输入序列中向量的位置。

2.2 位置向量

对于既不使用卷积也不使用递归结构的 Transformer 来说, 在结构上是无法直接获取位置信息的^[20]。Vaswani 等^[12]的解决方案是使用频率不同的正弦编码构建位置向量。

$$\begin{cases} U_{i,2m} = \sin(i / 10\,000^{2m/d}) \\ U_{i,2m+1} = \cos(i / 10\,000^{2m/d}) \end{cases} \quad (5)$$

其中, U 表示位置向量, i 表示位置, $m \in \left[0, \frac{d}{2}\right]$, d 代表输入的维度。

3 XLNET-Transformer_P-CRF 模型

针对已有方法的不足, 本文提出了基于 XLNET-Transformer_P-CRF 模型的中文命名实体识别方法。如图 1 所示, 首先文本输入嵌入层, 利用 XLNET 预训练模型得到动态字向量; 然后在编码层, Transformer_P 编码器对 XLNET 输出的向量进行编码, 通过嵌入相对位置信息, 最大化获取上下文语义信息; 最后, 通过输出层的 CRF 模型获取标签之间的规则特征, 并输出概率最大的标签序列。

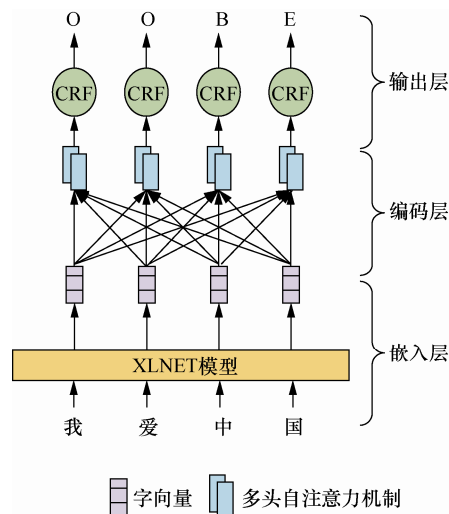


图 1 XLNET-Transformer_P-CRF 模型
Figure 1 Model of XLNET-Transformer_P-CRF

与传统的命名实体识别方法相比, 基于 XLNET-Transformer_P-CRF 模型的方法主要改进在于: 将改进后的 Transformer_P 编码器替换基于 RNN 结构的编码器, 解决了基于 RNN 结构的编码器存在的串行计算、梯度消失(爆炸)和单向建模的问题。此外, 引入了 XLNET 预训练模型, 该模型是在大规模无监督语料上训练所得, 可以通过上下文计算动态字向量以表征字的多义性。

3.1 嵌入层——XLNET

XLNET 是一种可以获得双向上下文信息的自回归语言模型。传统的自回归模型采用单向预测, 无法获取双向信息。XLNET 使用排列语言模型来获取双向的上下文信息, 同时维持自回归模型原有的单向形式。对于长度为 L 的文本, 只有 $L!$ 种不同的排列方式, 如果能考虑到所有排列顺序的文本, 就变相地获取了双向上下文信息, 其损失函数的具体公式如下。

$$\max_{\theta} E_{Z \sim Z_T} \left[\sum_{t=1}^T \log p_{\theta}(x_{Z_t} | x_{Z_{<t}}) \right] \quad (6)$$

其中, Z_T 代表长度为 T 的序列的所有排列组成的集合, Z 是其中一种排列组合方式。 x_{Z_t} 是预测的第 t 个元素, $x_{Z_{<t}}$ 代表 t 元素之前的 $1-t$ 个元素。

此外, 该模型引入了双流自注意力机制和循环机制。前者为排列语言模型分离了位置信息与内容信息。后者整合了 Transformer-XL^[14] 到预训练模型中, 并将 Transformer-XL 中的两项重要技术(相对位置编码范式和分割循环机制)融合进 XLNET, 使 XLNET 在处理长文本时具有更强大的优势。

XLNET 预训练模型与其他的预训练模型相比, 可以充分利用字两边的信息, 更好地表征字的多义性。

3.2 编码层——Transformer_P

本节对传统 Transformer 编码器的位置嵌入方式进行分析, 并针对其不能获取相对位置信息的缺陷做出了具体改进。

对于每个输入向量 h_i , 传统 Transformer 编码器使用字向量 e_i 与位置向量 u_i 相加的方式嵌入位置信息, 如式(7)所示。

$$h_i = e_i + u_i \quad (7)$$

其中, $e_i, u_i \in \mathbb{R}^{d_H}$ 。

对于输入向量 h_i 和 h_j , 将式(2)展开可得式(8)中所示的(a)、(b)、(c)、(d)4项。

$$A_{jl} = \underbrace{e_j W^Q (W^K)^T e_l}_{(a)} + \underbrace{e_j W^Q (W^K)^T u_l}_{(b)} + \underbrace{e_j W^Q (W^K)^T u_l}_{(c)} + \underbrace{u_j W^Q (W^K)^T u_l}_{(d)} \quad (8)$$

显然, 包含位置向量 u_l 的为(b)、(c)和(d)3项。

其中, (b)和(c)项分别只包含第 l 个元素和第 j 个元素的位置向量, 只能反映绝对位置信息。同时, Yan 等^[21]的研究证明了(d)项不能表征相对位置信息。综上可得, 传统的 Transformer 提取器并不能反映相对距离信息的结论。

为了改进现有的 Transformer 编码器不能获取相对位置信息的不足, 本文引入 Shaw 等^[22]的工作。通过对注意力机制的扩展, 将输入建模成一个含有方向性信息的结构。具体改动如式(10)和式(11)所示: 一是在计算注意力分数式(2)处添加相对位置编码 $a_{jl}^K \in \mathbb{R}^{d_z}$; 二是在计算输出序列式(4)处添加相对位置编码 $a_{jl}^V \in \mathbb{R}^{d_z}$, a_{jl}^K 和 a_{jl}^V 表示输入 h_i 和 h_j 之间的相对位置信息。

$$\widetilde{A}_{jl} = q_j (k_l + a_{jl}^K)^T \quad (9)$$

$$\widetilde{z}_j = \sum_{l=1}^n q_{jl} (v_j + a_{jl}^V) \quad (10)$$

其中, $\widetilde{A}_{jl}$ 表示改进后的 A_{jl} , $\widetilde{z}_j$ 表示改进后的 z_j 。

本文假设距离过长的相对位置信息对于命名实体的识别来说并不会起到信息增益的效果, 反而会带来噪声。本文将相对位置的最大距离设置为 k 。

$$a_{jl}^K \in w_{\text{clip}(l-j, k)}^K \quad (11)$$

$$a_{jl}^V \in w_{\text{clip}(l-j, k)}^V \quad (12)$$

$$\text{clip}(x, k) = \max(-k, \min(k, x)) \quad (13)$$

最后, 对于两处相对位置编码 a_{jl}^K 和 a_{jl}^V , 分别通过训练学习到 $2k+1$ 个向量, $w^K = (w_{-k}^K, \dots, w_k^K)$ 和 $w^V = (w_{-k}^V, \dots, w_k^V)$, $w_i^K, w_i^V \in \mathbb{R}^{d_z}$ 。

3.3 输出层——CRF

本文模型中的输出层采用主流的 CRF, 通过标签之间的依赖关系获得全局最优标签序列^[23]。

CRF 的步骤如下。

对于给定的输入序列 $s = \{s_1, \dots, s_T\}$ ，输出的标签序列为 $y = \{y_1, \dots, y_T\}$ ， $Y(s)$ 为所有标签的集合，计算 y 概率的过程如下。

$$P(y|s) = \frac{\sum_{t=1}^T e^{f(y_{t-1}, y_t, s)}}{\sum_{y' \in Y(s)} \sum_{t=1}^T e^{f(y'_{t-1}, y'_t, s)}} \quad (14)$$

其中， $f(y_{t-1}, y_t, s)$ 为计算从 y_{t-1} 到 y_t 的转移分数的函数，优化目标为最大化 $P(y|s)$ 。

4 实验和结果分析

4.1 训练方法

在训练中，实验采用反向传播算法更新全部参数（包括对 XLNET 模型的微调）。同时，实验采用了随机梯度下降和动量联合的方法优化损失函数，其中，学习率更新采用三角法^[24]。

对于嵌入层，实验选取了哈尔滨工业大学讯飞联合实验室在 5.4 B 词数的百科、新闻、问答类数据集上训练而成的 XLNET 模型，该模型共包含 24 层、768 个隐层、12 头注意力， 2.09×10^8 个参数。

对于编码层，Transformer_P 编码器除了在 MSRA 数据集上采用 6 头注意力，在其余数据集上均采用 4 头注意力，相对位置的最大距离均遍历 1 到 10，选取最好的效果。此外，本实验借鉴 Yan 等^[21]的训练方法，将式（3）中的 $\sqrt{d_z}$ 取消，目的在于解决中文命名实体识别数据集中实体较为稀疏的问题。

4.2 数据

（1）数据集

实验共选取了 4 个来源多样的中文数据集，其中，MSRA、OntoNotes4.0 主要来源于新闻行

业，Resume 数据集来自简历摘要，微博数据集主要来自社交媒体。

① MSRA 数据集^[25]是微软公开的数据集，包含人名、机构名、地名 3 类实体。

② OntoNotes4.0^[26]是一个多语言的大型数据集，本文实验只选取了其中的中文数据部分。

③ Resume 数据集^[27]是中文简历数据集，包含国家、教育机构、地点、人名、组织名、职业、民族、职务等 8 类实体。

④ 微博数据集^[28]是社交媒体类数据集，共包含地缘政治、人名、地名和组织名 4 类实体。

数据集详细信息如表 1 所示。

（2）标注规则

实验采用 BMES 标注规则，其中“B”表示一个实体的开始位置，“M”表示一个实体的中间位置，“E”表示一个实体的末尾位置，“S”表示一个单独的实体。

4.3 对比实验

实验一 Transformer_P 编码器有效性验证

为了验证 Transformer_P 编码器在中文命名实体识别任务上的有效性，本部分在 4 类数据集上进行了 Transformer_P 编码器与其他编码器的对比实验，采用的指标为 $F1$ 值。为控制实验变量，嵌入层统一使用 Word2vec 模型，输出层统一使用 CRF 模型。

如表 2 所示，在 MSRA、OntoNotes4.0、Resume 和微博 4 类数据集上，基于 Transformer 编码器的模型取得的 $F1$ 值均不如基于 RNN 结构编码器的模型。在加入了相对位置信息之后，基于 Transformer_P 编码器的模型在 4 类数据集上的 $F1$ 值有明显提升，分别高于基于传统 Transformer 编码器的模型的 $F1$ 值 2.68%、5.81%、1.57%、5.82%。基于 Transformer 编码器的模型除在微博

表 1 中文命名实体识别数据集详细信息
Table 1 Detailed statistics about the Chinese named entity recognition datasets

数据集	训练集		验证集		测试集	
	句子	字符	句子	字符	句子	字符
MSRA	464 00	2 169 900	—	—	4 400	172 600
OntoNotes4.0	15 700	491 900	4 300	200 500	4 300	208 100
Resume	3 800	124 100	460	13 900	480	15 100
微博	1 400	73 500	270	14 500	270	14 800

数据集上达到次好的 $F1$ 值以外,在 MSRA、OntoNotes4、Resume 三类数据集上达到了最高的 $F1$ 值。Transformer_P 编码器在微博数据集上未达到最高 $F1$ 值的原因主要是:微博数据集中数据较少,致使 Transformer_P 编码器的参数无法全部收敛,限制了效果。

表 2 Transformer_P 编码器与其他编码器的 $F1$ 值对比结果
Table 2 $F1$ score comparison results of Transformer_P encoder and other encoders

模型	MSRA	OntoNotes4.0	Resume	微博
BiGRU	88.03%	67.38%	93.78%	54.42%
BiLSTM	88.35%	67.00%	93.25%	56.23%
Transformer	87.67%	65.33%	92.64%	50.21%
Transformer_P	90.35%	71.14%	94.21%	56.03%

综上所述,Transformer_P 编码器被证明是有效的,对传统 Transformer 编码器进行嵌入相对位置信息的改进后,可以更有效地获取上下文信息。

实验二 XLNET-Transformer_P-CRF 模型的有效性验证

为验证本文所提模型 XLNET-Transformer_P-CRF 的有效性,实验在 4 类数据集上进行了本文所提模型与主流模型的对比实验,采用的指标为召回率、精确率和 $F1$ 值,其中,Max 指获取最佳效果时,相对位置的最大距离。

如表 3 所示,在 MSRA 数据集上,Chen 等^[29]、Zhang 等^[30]和 Zhou 等^[31]做了大量的特征工程,是该数据集上表现较好的统计模型。Dong 等^[32]使用 BiLSTM-CRF 模型和字符特征进行实体识别,相对于词级特征的实体识别,效果显著提升。

表 3 MSRA 数据集上的各指标对比结果
Table 3 Index comparison results of MSRA datasets

模型	精确率	召回率	$F1$ 值
文献[29]	91.22%	81.71%	86.20%
文献[30]	92.20%	90.18%	91.18%
文献[31]	91.86%	88.75%	90.28%
文献[32]	91.28%	90.62%	90.95%
文献[27]	93.57%	92.79%	93.18%
文献[33]	94.01%	92.93%	93.47%
文献[34]	—	—	94.35%
本文模型 Max=4	93.64%	96.63%	95.11%

Zhang 等^[27]、Sui 等^[33]和 Li 等^[34]使用了字词融合的方式提升中文命名实体识别的效果。通过比较,本文模型在 MSRA 数据集上达到了最好的效果, $F1$ 值高于次好的模型 0.76%。

如表 4 所示,在 OntoNotes4.0 数据集上,实验将本文模型和基于该数据集效果最佳的中文命名实体识别模型进行比较。Wang 等^[35]采用一种有效利用双语数据半监督学习的方法,在该数据集上获得了 74.32% $F1$ 值的效果。Che 等^[36]采用将不同语言的约束信息提高中文命名实体识别的效果,相对于基线效果, $F1$ 值提升了近 5%。Yang 等^[37]在文献[36]的基础上,丰富了输入特征,提升了命名实体识别的效果。Zhang 等^[27]、Sui 等^[33]和 Li 等^[34]使用了字词融合的方式提升中文命名实体识别的效果。通过比较,本文模型在 OntoNotes4.0 数据集上达到了最好的效果, $F1$ 值高于次好的模型 4.84%。

表 4 OntoNotes4.0 数据集上的各指标对比结果
Table 4 Index comparison results of OntoNotes4.0 datasets

模型	精确率	召回率	$F1$ 值
文献[35]	76.43%	72.32%	74.32%
文献[36]	77.71%	72.51%	75.02%
文献[37]	65.59%	71.84%	68.57%
文献 [27]	76.35%	71.56%	73.88%
文献[33]	75.06%	74.52%	74.79%
文献[34]	—	—	75.70%
本文模型 Max=4	76.68%	84.80%	80.54%

如表 5 所示,在 Resume 数据集上,Zhang 等^[27]将词典信息与字向量通过 LSTM 网络相融合,在训练中统一更新权重,有效解决了字词融合的问题。在此基础上,Liu 等^[38]利用了 4 种不同的策略来将单词信息编码为固定大小的矢量,使其可以分批训练并适应各种应用场景。Gui 等^[39]提出对图节点进行分类,从而实现序列标注。Li 等^[34]使用了 Transformer 的自注意机制使字符能够直接与潜在的单词交互,实现更好的字词融合效果。通过比较,本文模型在 Resume 数据集上达到了最好的效果, $F1$ 值高于次好的模型 1.77%。

表5 Resume数据集上的各指标对比结果
Table 5 Index comparison results of Resume datasets

模型	精确率	召回率	F1 值
文献[27]	94.81%	94.11%	94.46%
文献[38]	95.14%	94.79%	94.96%
文献[39]	95.28%	95.46%	95.37%
文献[34]	—	—	94.93%
本文模型 Max=4	96.23%	97.18%	96.70%

如表6所示,在微博数据集中,Peng等^[28]、He等^[40]、Zhang等^[27]3种效果较佳的命名实体识别模型分别利用了丰富输入特征、多领域特征和半监督方法获得数据、字词融合特征的方法。Sui等^[33]提出一种字符级的协作图神经网络,全方位获得词语信息。Li等^[34]将所有字符与自匹配词直接交互,利用了潜在的单词信息。通过比较,本文模型在微博数据集上达到了最好的效果,F1值高于次好的模型8.04%。

表6 微博数据集上的各指标对比结果
Table 6 Index comparison results of Weibo datasets

模型	精确率	召回率	F1 值
文献[28]	—	—	56.05%
文献[40]	—	—	58.23%
文献[27]	—	—	58.79%
文献[33]	67.31%	48.61%	56.45%
文献[34]	—	—	63.42%
本文模型 Max=5	68.34%	74.88%	71.46%

从表3~表6可以看出,与现有的方法相比,本文提出的基于XLNET-Transformer_P-CRF模型的方法更具有竞争力,在4类数据集上都达到了最好的效果。因此,该方法被证明是十分有效的。Transformer_P编码器采用自注意力机制,与同样基于自注意力机制的预训练语言模型XLNET相结合,可以更为显著地提高中文命名实体识别的效果。

5 结束语

本文提出了基于XLNET-Transformer_P-CRF深度学习模型方法。与传统的方法相比,本文创新点体现在对Transformer编码器进行改进,克服

了Transformer编码器不能获得相对位置关系的缺陷。实验表明,本文方法具有有效性,在4类数据集上都达到了最好的效果。

根据实验发现,Transformer_P编码器参数量大、对数据的依赖性强,在小数据集中效果一般。因此,后续需要研究如何在保证模型效果的前提下,减少模型的参数。

参考文献:

- [1] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: a brief history[C]//International Conference on Computational Linguistics, 1996: 466-471.
- [2] PIZZATO L A, MOLLA D, PARIS C. Pseudo relevance feedback using named entities for question answering[C]//Proceedings of the 2006 Australian Language Technology Workshop (ALTW-2006). 2006: 89-90.
- [3] BABYCH B, HARTLEY A. Improving machine translation quality with automatic named entity recognition[C]//Proceedings of the 7th International EAMT Workshop on MT and other Language Technology Tools, Improving MT Through other Language Technology Tools: Resources and Tools for Building MT, Association for Computational Linguistics. 2003: 1-8.
- [4] MANDL T, WOMSER-HACKER C. The effect of named entities on effectiveness in cross-language information retrieval evaluation[C]//Proceedings of the 2005 ACM Symposium on Applied Computing. 2005: 1059-1064.
- [5] 邱锡鹏. 神经网络与深度学习[M]. 北京: 机械工业出版社, 2020. QIU X P, Neural networks and deep learning[M]. Beijing: China Machine Press, 2020.
- [6] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020: 1.
- [7] HAMMERTON J. Named entity recognition with long short-term memory[C]//North American Chapter of the Association for Computational Linguistics. 2003: 172-175.
- [8] HUANG Z, XU W, YU K, et al. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: Computation and Language, 2015.
- [9] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[J]. arXiv: Learning, 2016.
- [10] CHIU J P, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4(1): 357-370.
- [11] BENGIO Y, SIMARD P Y, FRASCONI P, et al. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.
- [12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Neural Information Processing Systems. 2017: 5998-6008.
- [13] GUO Q P, QIU X P, LIU P F, et al. Star transformer[C]//NAACL. 2019: 1315-1325.
- [14] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: attentive language models beyond a fixed-length context[J]. arXiv: Learning, 2019.

- [15] HUANG A, VASWANI A, USZKOREIT J, et al. Music transformer: generating music with long-term structure[C]//International Conference on Learning Representations, 2019.
- [16] MIKOLOV T, CHEN K, CORRADO G S, et al. Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations. 2013.
- [17] 杨飘, 董文永. 基于 BERT 嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(04): 40-45, 52.
YANG P, DONG W Y. Chinese NER based on BERT embedding[J]. Computer Engineering, 2020, 46(4): 40-45, 52.
- [18] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv: Computation and Language, 2018.
- [19] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[J]. arXiv: Computation and Language, 2019.
- [20] PARIKH A P, TACKSTROM O, DAS D, et al. A decomposable attention model for natural language inference[C]//Empirical Methods in Natural Language Processing. 2016: 2249-2255.
- [21] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv: Computation and Language, 2019.
- [22] SHAW P, USZKOREIT J, VASWANI A, et al. SELF-attention with relative position representations[C]//North American Chapter of the Association for Computational Linguistics. 2018: 464-468.
- [23] LAFFERTY J, MCCALLUM A, PEREIRA F, et al. Conditional random fields: probabilistic models for segmenting and Labeling Sequence Data[C]//International Conference on Machine Learning. 2001: 282-289.
- [24] SMITH L N. Cyclical learning rates for training neural networks[C]//Workshop on Applications of Computer Vision, 2017: 464-472.
- [25] LEVOW G. The third international chinese language processing Bakeoff: word segmentation and named entity recognition[C]//Meeting of the Association for Computational Linguistics. 2006: 108-117.
- [26] Ralph Weischedel. Ontonotes release 4.0 LDC2011T03[S]. 2011.
- [27] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Meeting of the Association for Computational Linguistics. 2018: 1554-1564.
- [28] PENG N, DREDZE M. Named entity recognition for chinese social media with jointly trained embeddings[C]//Empirical Methods in Natural Language Processing. 2015: 548-554.
- [29] CHEN A, PENG F, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models[C]//Meeting of the Association for Computational Linguistics. 2006: 173-176.
- [30] ZHANG S, QIN Y, WEN J, et al. Word segmentation and named entity recognition for SIGHAN Bakeoff3[C]//Meeting of the Association for Computational Linguistics. 2006: 158-161.
- [31] ZHOU J S, QU W G, ZHANG F. Chinese named entity recognition via joint identification and categorization[J]. Chinese Journal of Electronics, 2013, 22(2): 225-230.
- [32] DONG C H, ZHANG J J, ZONG C Q. Character based LSTM-CRF with radical-level features for Chinese named entity recognition[C]//International Conference on Computer Processing of Oriental Languages. 2016: 239-250.
- [33] SUI D, CHEN Y, LIU K, et al. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network[C]//International Joint Conference on Natural Language Processing. 2019: 3828-3838.
- [34] QIU X P, LI X N, YAN H. Flat chinese ner using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6836-6842.
- [35] WANG M, CHE W, MANNING C D, et al. Effective bilingual constraints for semi-supervised learning of named entity recognizers[C]//National Conference on Artificial Intelligence. 2013: 919-925.
- [36] CHE W, WANG M, MANNING C D, et al. Named entity recognition with bilingual constraints[C]//North American Chapter of the Association for Computational Linguistics. 2013: 52-62.
- [37] YANG J, ZHANG Y, DONG F, et al. Neural word segmentation with rich pretraining[C]//Meeting of the Association for Computational Linguistics. 2017: 839-849.
- [38] LIU W, XU T, XU Q, et al. An encoding strategy based word-character LSTM for chinese NER[C]//North American Chapter of the Association for Computational Linguistics. 2019: 2379-2389.
- [39] GUI T, ZOU Y, ZHANG Q, et al. A lexicon-based graph neural network for chinese NER[C]//International Joint Conference on Natural Language Processing. 2019: 1040-1050.
- [40] HE H, SUN X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]//National Conference on Artificial Intelligence. 2017: 3216-3222.

[作者简介]



郑洪浩 (1992-), 男, 山东济宁人, 信息工程大学硕士生, 主要研究方向为命名实体识别、关系抽取。



于洪涛 (1970-), 男, 辽宁丹东人, 博士, 信息工程大学研究员, 主要研究方向为大数据与人工智能。



李邵梅 (1982-), 女, 湖北钟祥人, 博士, 信息工程大学副研究员, 主要研究方向为计算机视觉。