融合 ERNIE 与改进 Transformer 的中文 NER 模型

罗 峦,夏骄雄
(上海理工大学 光电信息与计算机工程学院,上海 200093)

摘 要:命名实体识别是信息抽取和关系提取基础的关键任务。针对中文命名实体识别问题,提出了一种融合 ERNIE 和改进 Transformer 的中文命名实体识别深度学习模型——ERIT(combining ERNIE with Improved Transformer)。ERIT 使用 ERNIE 训练词向量作为嵌入层,摆脱了模型对于分词预处理过程的依赖,避免出现因分词错误以及信息缺失引起错误传播而导致准确率降低的情况,在兼顾输入文本识别精度的同时进一步优化输入语句的词向量,利用 Transformer 获取输入序列的上下文信息并进行特征提取,结合自注意力层对权重参数进行更新,并在此基础上,通过在自注意力层上增加约束正则项提高对参数约束性以提高每个生成标签的准确性,并加入计划采样机制以解决模型训练与测试过程中存在的不匹配问题。实验证明,ERNIE 作为嵌入层有效优化了词向量并提高了识别精度,且模型相较于其他实体识别模型取得了较好的效果。

关键词:自然语言处理;命名实体识别;深度学习;ERNIE;注意力机制
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)10-0120-06
doi:10.3969/j.issn.1673-629X.2022.10.020

Research on Chinese Named Entity Recognition Combining ERNIE with Improved Transformer

LUO Luan, XIA Jiao-xiong
(School of Option-electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Named entity recognition is the basic and key task of information extraction and relationship extraction. Aiming at the problem of Chinese named entity recognition, ERIT (combining ERNIE and Improved Transformer), a Chinese named entity recognition deep learning model combining ERNIE and improved Transformer, is proposed. ERIT uses Ernie training word vector as the embedding layer, which removes the dependence of the model on the word segmentation preprocessing process, avoids the situation that the accuracy is reduced due to the error of word segmentation and the lack of information. It further optimizes the word vector of the input sentence while taking into account the recognition accuracy of the input text. Transformer is used to obtain the context information of the input sequence and extract features, and the weight parameters are updated with the self-attention layer. On this basis, the attention regularization is added to the self attention layer in order to improve the accuracy of each generated label by improving the constraint of parameters, and the scheduled sampling mechanism is added to solve the mismatch problem in the process of model training and testing. Experiments show that as the input of embedding layer, ERNIE effectively optimizes the word vector and improves the recognition accuracy, and compared with other entity recognition models, the model achieves better results.

Key words: NLP; named entity recognition; deep learning; ERNIE; attention mechanism

0 引 言

随着互联网的普及和计算机技术的发展,人们获取信息的途径愈加丰富,同时网络上的文本信息呈指数级增长。这些海量文本中蕴含着科技创新的前沿趋势、热点话题、经济社会的舆情民意等重要信息,因此,如何从海量文本数据中挖掘出有价值的关键信息给研究者们带来了巨大挑战。在这一背景下,信息抽取[1]技术(Information Extraction, IE)应运而生。信息抽取主要是对文本中的非结构化信息进行处理,从文本中抽取出特定的实体或事件,帮助研究人员对海量文本

收稿日期:2021-04-07 **修回日期**:2021-08-10
基金项目:上海市自然科学基金项目(17ZR1428400)
作者简介:罗 峦(1996-),男,硕士,研究方向为自然语言处理、深度学习;通讯作者:夏骄雄,副研究员,硕导,研究方向为数据挖掘、智能决策支持系统、教育信息化等。

内容进行自动分类、提取和重构。命名实体识别(Named Entity Recognition,NER)作为信息抽取的重要子任务,受到了国内外研究者的广泛关注。

命名实体识别是自然语言处理领域的一项基础任务,主要任务是从文本数据中自动地发现信息实体以及识别它们对应的类别。自然语言处理研究领域的信息检索[2]、信息抽取、知识图谱[3]、问答系统[4]等多项任务均需要命名实体识别任务作为基础。随着自然语言处理技术的不断发展,对文本中包含的语义知识的挖掘变得愈发重要,丰富的语义知识会使得后续的任务取得更好的效果,而命名实体作为文本中的关键信息,包含了丰富的语义知识,因此正确识别文本中这些实体并进行分类,具有重要的研究意义。

1　相关工作

实体识别任务早期通常采用基于规则和词典的方法,这种方法考虑了数据的结构和特点,对于结构性强的实体(日期、时间、货币等)具有较好的识别效果,但对于结构性不强的实体识别效果差,且编写规则的过程复杂。基于统计模型的方法对特征选取要求更高,需要提取出对实体识别有影响的各种特征。基于统计的机器学习方法主要包括隐马尔可夫模型(Hidden Markov,HMM)、最大熵(Maximum Entropy,ME)、支持向量机(Support Vector Machine,SVM)等,这些方法需要人工提取特征,且特征的选取对识别效果有很大影响。

近几年,随着深度学习的不断发展,深度学习在自然语言处理的许多应用中都取得了长足进步,于是,NER 的研究热点逐渐从早期基于词典和规则的方法过渡到基于神经网络的方法。深度学习是一种从原始数据中自动学习特征的方法,具有较强的泛化能力,在很大程度上减弱了对繁琐的特征工程和专业知识的依赖。

文献[5]使用了多层叠加的双向长短期记忆网络(Long Short-Term Memory,LSTM)进行文本特征提取,在多种不同任务上进行实验,均取得了较好的识别效果,并证明了字符级的标注方法相较于词语级的标注方法效果更好;文献[6]通过多层叠加的 CNN 来扩大处理范围从而提高长距离信息捕获能力,并且使用空洞卷积在维持相同处理范围的同时有效减少了参数数量,在取得良好效果的同时降低了模型的运算复杂度;文献[7]使用大量的无标记语料来训练神经网络模型,将训练好的语言模型与序列标注模型进行拼接,将语言模型学到的知识作为额外特征以提高序列标注任务的性能;文献[8]提出了将带有局部注意力层卷积网络和带有全局注意力层的门控循环网络相结合的 CAN,以更好地捕获相邻字符和句子上下文的信息;文献[9]提出一种协同图神经网络(Collaborative Graph Network),通过词意信息自动构建词汇来解决命名实体识别中分词边界信息缺失问题;文献[10]提出一种词字-图卷积网络(WC-GCN),使用了一种交叉 GCN 模块同时处理两个方向的词字有向无环图,提高对长距离依赖的捕获能力;文献[11]将实体识别任务转变为问答任务以解决实体嵌套的问题。上述方法对实体识别的贡献大多体现在优化词向量或优化特征表达其中一方面,兼顾两者的方法大多结构复杂耗时较长。为此,该文提出的模型具有以下贡献:使用 ERNIE 训练词向量作为嵌入层,摆脱了模型对于分词预处理过程的依赖,避免出现因分词错误以及信息缺失引起错误传播而导致准确率降低的情况;通过自注意力层增加了约束正则项提高对参数约束性以提高每个生成标签的准确性;最后,加入计划采样机制以解决模型训练与测试过程中存在的不匹配问题。

2　模型实现

编解码器[12-13]解决问题的主要思路是通过神经网络模型在每个时间段不断地去学习输入文本序列的编码向量,然后根据存储的信息向量解析成目标序列。在每一步中,模型都是自循环,当生成下一个输出时,会将之前的输出作为附加的输入。

模型的结构如图 1 所示。

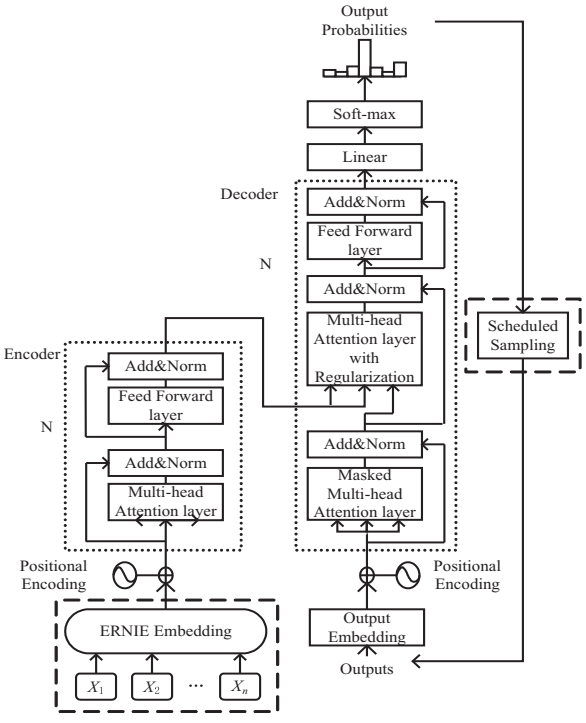

图 1　融合 ERNIE 和改进 Transformer 的模型示意图

该模型的整体运作流程是:首先输入文本序列,利用 ERNIE 预训练模型,获得包含序列总体信息的动态

词向量,接着将新的词向量输入到 multi-head 自注意力层进行特征提取,捕捉序列的特征信息,编码器通过学习将输入的文本序列编码成状态向量,状态向量会作为每个时刻特征输入,并结合之前时刻的输出进行学习并输出。编码器由 $N=6$ 个相同的基础层叠加组成,每个基础层由两个子层组成,分别是 multi-head 自注意力层和全连接层,其中每个子层都加了残差链接[14]和层标准化,子层的输出可表示为:

$$\mathrm{sub}_{\mathrm{layer}_{\mathrm{output}}} = \mathrm{LayerNorm}(x + (\mathrm{SubLayer}(x))) \quad (1)$$

其中,x 和 $\mathrm{SubLayer}(x)$ 分别对应 multi-head 自注意力层和全连接层的输入和输出。解码器由 $N=6$ 个相同的基础层叠加组成,但相比于编码器的基础层多了一个子层,该子层输入部分的 K、V 来自编码器,Q 来自上一位置解码器的输出,并且也包括残差链接和层标准化。

2.1 ERNIE 词嵌入

在自然语言处理中,文本序列处理由以下几个步骤组成。首先将原始文本序列嵌入至稠密的表征词嵌入,其次将词嵌入序列转化为定长的表征向量,最终输入后续任务中。对词嵌入表征的研究向来是自然语言处理的重中之重。然而,类似于 Word2vec 词向量模型存在无法解析一词多义以及上下文信息缺失等缺点,往往对性能的提升并不明显。ERNIE[15]作为一种预训练语言模型,是通过在海量语料中进行自监督学习而得到的一组适用性十分广泛同时还能在具体任务中动态优化词向量,所谓自监督学习是指在没有人工标注的数据上运行的监督学习。ERNIE Embedding 的结构如图 2 所示。

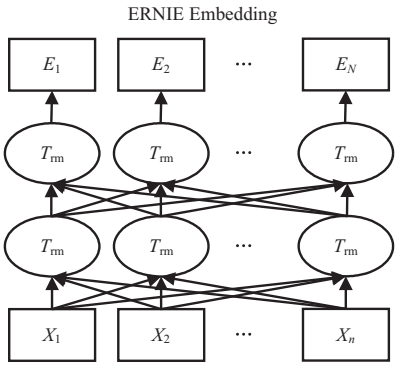

图 2 ERNIE Embedding 的结构

其训练过程与 BERT 类似,但 BERT 模型主要是聚焦在针对字或英文单词粒度的学习,没有充分注重到中文字级结构及语义单元之间的联系[16-17]。为此,ERNIE 训练过程通过对连续的语义单元掩码,对训练数据中的词法结构、语法结构、语义信息进行统一建模,使得模型学习完整概念的语义表示,进一步增强了在中文文本上的语义结构表示能力。

2.2 位置编码

由于多头自注意力层缺少对序列中字词顺序的表示,因此需要在序列中添加位置编码向量:Positional Encoding,该向量包含当前字词在序列中的位置信息,公式如下:

$$\mathrm{PE}(\mathrm{pos},2i) = \sin\left(\frac{\mathrm{pos}}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \quad (2)$$

$$\mathrm{PE}(\mathrm{pos},2i+1) = \cos\left(\frac{\mathrm{pos}}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \quad (3)$$

其中,pos 表示当前字词在句子中的位置,i 表示位置编码向量中的第几个维度,d_{model} 表示词向量的维度,位置编码的作用是将字词在句子中的位置 pos 映射成 d_{model} 维的位置向量,最后,将同维度的位置向量与词向量进行矩阵求和得到具有位置信息的词向量。

2.3 Multi-head 自注意力模块

Multi-head 自注意力模块在 Transformer 中首次被提出,其最大的特点是替代了传统的 RNN 和 CNN,整个结构完全由 attention 机制组成[18]。传统的 RNN(或者 LSTM,GRU 等)的计算限制为顺序的,也就是只能从左往右或者从右向左依次计算,这种机制带来了两个问题:t 时刻的计算必须依赖 $t-1$ 时刻的计算结果,这样限制了模型的并行能力;顺序计算的过程中信息会丢失,尽管在 RNN 上添加门控机制等方法一定程度上缓解了长期依赖的问题,但仍存在依赖递减的问题。而 Multi-head 自注意力模块有效解决了上述问题。

Multi-head 自注意力机制的计算过程相当于多个不同的单自注意力机制的集成,单自注意力机制输入由维数同为 d_k 的 query 和 key,以及维数为 d_v 的 value 组成,输出为:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (4)$$

其中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别代表 Query、Key、Value 三个参数矩阵,$\boldsymbol{Q} \in R^{n\times d_k}$,$\boldsymbol{K} \in R^{m\times d_k}$,$\boldsymbol{V} \in R^{m\times d_v}$,$\sigma$ 表示 softmax 函数,本质为三个维度 $n\times d_k$,$m\times d_k$,$m\times d_v$ 的矩阵相乘,最后输出的结果为一个维度为 $n\times d_v$ 的矩阵。公式中的 QK^{T} 为计算 Query 和 Key 的匹配程度,用点积计算匹配度类似于余弦相似性,这些匹配值相当于对源输入序列做权重处理,表示在生成一个标签时源序列中哪些词是需要被注意到的。$\sqrt{d_k}$ 进行尺度化的目的是避免因维度过高点积过大,使得经过 softmax 之后梯度贴近 0 或 1,不利于反向传播时的计算;最后,将经过 softmax 后得到的归一化的匹配值与对应的键值 $\boldsymbol{V}$ 相乘得到最终的注意力,以上计算过程又称为缩放点积注意力模块。

Multi-head 自注意力机制的计算过程相当于多个

不同的缩放点积注意力模块集成,如图 3 所示。

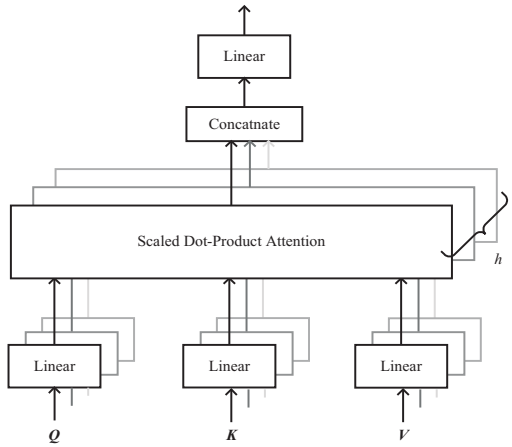

图 3　Multi-head 自注意力

首先将 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 经过一个线性变换,然后再做缩放点积注意力运算,整个过程重复 h 次。具体如公式(5)所示:

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V)\quad(i=1,2,\cdots,h) \tag{5}$$

其中,$\boldsymbol{W}_i^Q \in R^{d_{\text{model}}\times d_k}$,$\boldsymbol{W}_i^K \in R^{d_{\text{model}}\times d_k}$,$\boldsymbol{W}_i^V \in R^{d_{\text{model}}\times d_v}$,$\boldsymbol{W}_i^Q$,$\boldsymbol{W}_i^K$,$\boldsymbol{W}_i^V$ 为线性变换矩阵,是待训练的权值参数矩阵,作用是将 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 线性变换到 h 个不同的维度。每次对 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 进行的线性变换的参数 $\boldsymbol{W}$ 都是不一样的,这样做的目的是不同的 head 可以学习到不同的注意力,使得注意力更加多样化。得到多个 head 的计算结果后,将所有 head 的计算结果进行拼接,公式如下:

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_i,\cdots,\text{head}_h)\boldsymbol{W}^O \tag{6}$$

其中,$\boldsymbol{W}^O \in R^{hd_v\times d_{\text{model}}}$,表示再进行一次线性变换,最后得到所求注意力。

2.4　约束正则项

上面介绍了融合 ERNIE 和 Transformer 的模型,本节将引出约束正则项提高对参数约束性以提高每个生成标签的准确性,进一步对模型参数进行优化。因为需要为每个令牌都生成一个标签,因此在生成特定的标签时,应该更多注意到对应的令牌,而不应该让整个序列的标签过分地受到某一个令牌的影响,降低准确性。为解决此问题,即需要使令牌之间在整个解码过程中获得的注意力差不多,因此,在解码过程中,损失函数上加入注意力正则项:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{W}_i^Q \cdot \boldsymbol{K}\boldsymbol{W}_i^{K^{\mathrm{T}}} \tag{7}$$

$$\text{Regularization term} = \sum_n \left(\tau - \sum_{i=1}^h \sum_m A_{m,n}\right)^2 \tag{8}$$

其中,$\boldsymbol{A}$ 为一个 head 的注意力矩阵,m,n 表示注意力矩阵元素坐标,$\sum_m A_{m,n}$ 为一个 token 在当前 head 的注意力之和;$\sum_{i=1}^h \sum_m A_{m,n}$ 是一个令牌在解码过程中所有 head 的注意力之和,$\left(\tau - \sum_{i=1}^h \sum_m A_{m,n}\right)^2$ 的目的是让其注意力之和无限接近于一个常数超参 τ;$\sum_n$ 的目的是让每个令牌在整个解码过程中的注意力之和都接近于 τ,从而每个令牌在整个解码过程中获得的注意力都差不多。

2.5　计划采样机制

由于在训练时解码器每个时间点接入的都是真实序列标记,而在测试时每个时间点接入的都是上一时间点解码器的输出,这种训练和测试时编码器输入不匹配的差异会导致当在某一步做出错误预测后,后面会产生累积错误,也就是上一时间点生成了错误的标签,那么以它为输入生成的下一标签的准确性也会被影响。为解决此问题,引入一种计划采样策略:训练时不再完全采用真实序列标记作为下一时间点的输入,而是引入概率解决这个问题,以概率 p 选择真实标记,以 $1-p$ 选择模型自身的输出,p 采用反 sigmoid 函数,公式如下:

$$p = \frac{1}{1+\mathrm{e}^{k(2x-1)}} \tag{9}$$

其中,x 为已使用数据集比例,k 为超参;概率 p 的大小在训练过程中是动态变化的:开始时 p 尽量选择较大值,因为模型训练不充分,尽量使用真实标记。随着训练的进行,p 值逐渐减小,因为随着模型训练越来越充分,需要尽量选择模型自己的输出以避免训练测试不匹配问题。

3　实验研究

3.1　实验数据及评价指标

实验采用 1998 年 1 月的人民日报语料作为数据集,该数据集是北京大学计算语言学研究所和富士通研究中心共同制作的标注语料库,被作为原始数据应用于大量的研究和论文中。实验过程按 80%、10%、10% 将其随机划分为训练集、验证集和测试集。数据集使用 BIO 标注方式(Beginning,Inside,Outside)进行标注,其中包括 3 种实体类型,人名、地名和组织名(PER,LOC,ORG),共 7 种标签('B-PER','I-PER','B-LOC','I-LOC','B-ORG','I-ORG','O')。实验中采用精确率 P、召回率 R 和 F1 值作为模型的评价指标。计算公式如下:

$$P = \frac{\text{模型正确标注的实体个数}}{\text{模型标注的所有实体个数}} \times 100\% \tag{10}$$

$$R = \frac{\text{模型正确标注的实体个数}}{\text{样本中所有实体个数}} \times 100\% \tag{11}$$

$$\text{F1} = \frac{2\times P\times R}{P+R} \times 100\% \tag{12}$$

3.2 实验设置与结果分析

实验中,首先将文中模型与其他模型进行对比,验证模型的有效性;然后通过在不同比重数据集上进一步验证研究融合 ERNIE 的有效性;最后,通过消融实验进一步验证约束正则项及计划采样机制的有效性。

3.2.1 实验 1:文中模型与其他模型的比较

实验选取了几种效果较好且主流的模型包括 CRF、LSTM-CRF、GRU-CRF 以及 RD-CNN-CRF[19] 与文中模型进行对比,结果见表 1。实验结果显示,文中模型相比于其他模型在各个指标上都有所提升,通过在相同数据集上对比实验结果说明在实体识别任务中应用文中模型能够提高识别效果。

表 1 不同模型实验结果对比 %

模型	*P*	*R*	F1
CRF	72.41	63.53	67.68
LSTM-CRF	76.83	69.06	72.74
GRU-CRF	76.32	70.17	73.11
RD-CNN-CRF[14]	78.94	71.26	74.90
ERIT	83.74	75.52	79.41

3.2.2 实验 2:评估融合 ERNIE 后模型效果

按照不同比例从数据集中进行不放回随机取样生成不同规模的数据集,在 ERIT 上将词向量训练方式改为常规 word2vec 以此作为对比模型。实验结果见图 4。在不同规模的数据集上训练文中模型与对比模型,结果表明随着数据集规模增大,两者的结果均明显提高,表明了模型的有效性;同时在不同的规模数据集上文中模型表现均优于对比模型,说明融合 ERNIE 方式相较于 word2vec 词向量训练法可以提升实体识别模型的表现,这种情况在数据集较小时更加明显。

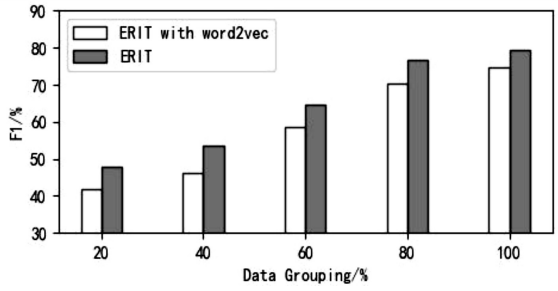

图 4 实验 2 结果

3.2.3 实验 3:消融研究

为了验证模型中约束正则项和计划采样机制的有效性,对文中模型进行消融研究。通过在训练过程中移除和添加这两部分然后在测试集上验证其对性能指标的影响。具体实验做法为分别对正则项和计划采样设置权重系数 λ_1、λ_2,当 λ_1 分别置为 0 或 1 时,即分别表示移除或添加了约束正则项;当 λ_2 置为 1 时即为在训练时解码器每个时间点接入的都是真实序列标记,当 λ_2 置为 p 时,即表示加入了计划采样机制,实验结果见表 2。

表 2 消融实验结果 %

评价指标	训练集 F1	测试集 F1
$\lambda_1=0, \lambda_2=1$	78.83±0.94	75.11
$\lambda_1=1, \lambda_2=1$	81.57±0.71	77.64
$\lambda_1=1, \lambda_2=p$	80.32±1.04	79.41

通过对比模型在没有采用约束正则项与计划采样机制($\lambda_1=0, \lambda_2=1$)时和只采用正则项($\lambda_1=1, \lambda_2=1$)时的实验结果表明,后者相较于前者取得了更好的效果,这验证了约束正则项的有效性,又与 $\lambda_1=0, \lambda_2=1$ 的结果对比发现将计划采样机制集成到模型中会进一步提高性能。值得注意的是,加入计划采样机制会使得模型在训练集上效果有所下降,但在测试集上有明显提升,因为加入计划采样机制,训练时不再完全采用真实标记,会策略性地选择上一时间点的输出接入解码器,这虽然降低了在训练集上的效果,但却使模型更具鲁棒性,所以实际测试效果更好。这些结果表明,提出的约束注意力正则项和计划采样机制是非常有效的。

4 结束语

提出的模型使用 ERNIE 训练词向量作为嵌入层对文本序列进行了特征提取,兼顾输入文本识别精度的同时进一步优化输入语句的词向量。利用 Transformer 获取输入序列的上下文信息并进行充分的特征提取,相比传统神经网络模型效果提升显著,在此基础上提出了约束正则项提高对参数约束性以提高每个生成标签的准确性,并加入计划采样机制以解决模型训练与测试过程中存在的不匹配问题,通过实验证明提出的约束注意力正则项和计划采样机制能进一步提升模型性能。

参考文献:

[1] LIN Y,JI H,HUANG F,et al. A joint neural model for information extraction with global features[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. [s. l.]:Association for Computational Linguistics,2020:7999-8009.

[2] 刘 萍,叶方倩,杨志伟. 认知建构视角下交互式信息检索模型研究[J]. 图书情报知识,2020(2):93-101.

[3] 谭 晓,张志强. 知识图谱研究进展及其前沿主题分析[J]. 图书与情报,2020(2):50-63.

[4] HUDSON D A,MANNING C D. Gqa:a new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:

6700-6709.

[5] KURU O, CAN O A, YURET D. Charner: character-level named entity recognition [C]//Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. Osaka: The COLING 2016 Organizing Committee, 2016: 911-921.

[6] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions [C]//Proceedings of the 2017 conference on empirical methods in natural language processing. Copenhagen: ACL, 2017: 2670-2680.

[7] PETERS M, AMMAR W, BHAGAVATULA C, et al. Semi-supervised sequence tagging with bidirectional language models[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). Vancouver: ACL, 2017: 1756-1765.

[8] ZHU Y, WANG G. 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). Minneapolis: ACL, 2019: 3384-3393.

[9] SUI D, CHEN Y, 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). Hong Kong, China: ACL, 2019: 3821-3831.

[10] TANG Z, WAN B, YANG L. Word-character graph convolution network for Chinese named entity recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1520-1532.

[11] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. [s. l.]: Association for Computational Linguistics, 2020: 5849-5859.

[12] 李晨斌,詹国华,李志华. 基于改进 Encoder-Decoder 模型的新闻摘要生成方法[J]. 计算机应用,2019,39(S2):20-23.

[13] CER D, YANG Y, KONG S, et al. Universal sentence encoder for English[C]//Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. Brussels: ACL, 2018: 169-174.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 770-778.

[15] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. Florence: ACL, 2019: 1441-1451.

[16] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. Florence: ACL, 2019: 4593-4601.

[17] HAO B, ZHU H, PASCHALIDIS I. Enhancing clinical BERT embedding using a biomedical knowledge base [C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: ICCL, 2020: 657-661.

[18] 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. Long Beach: Curran Associates Inc, 2017: 6000-6010.

[19] WANG Q, ZHOU Y, RUAN T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, 2019, 92: 103133.