

基于深度学习的中文命名实体识别研究

王雪梅, 陶宏才

(西南交通大学信息科学与技术学院, 四川 成都 611756)

摘要:针对经典 BiLSTM-CRF 命名实体识别模型训练时间长、无法解决一词多义及不能充分学习文本上下文语义信息的问题,提出一种基于 BERT-BiGRU-Attention-CRF 的中文命名实体识别模型。首先利用 BERT 语言模型预训练词向量,以弥补传统词向量模型无法解决一词多义的问题;其次,利用双向门控循环单元(BiGRU)神经网络层对文本深层次的信息进行特征提取,计算每个标签的预测分值,得到句子的隐藏状态序列;然后利用注意力机制(Attention)层对词加权表征,挖掘词间的关联关系,得到新预测分值,新状态序列;最后通过条件随机场(CRF)对新预测分值计算全局最优解,从而获得模型对实体标签的最终预测结果。通过在 MSRA 语料上的实验,结果表明文中模型的有效性。

关键词:中文命名实体识别;BERT;BiGRU;Attention;CRF

中图分类号:TP391

文献标志码:A

doi:10.16836/j.cnki.jcuit.2020.03.003

0 引言

命名实体识别是自然语言处理技术的一个重要组成部分,其功能是识别文本中人名、地名、组织机构名等命名实体,其识别效果在很多任务中有重要意义,如机器翻译、自动问答及关系抽取等任务。近年来,基于神经网络的命名实体识别方法被相继提出,其主要思路是先使用 CNN(convolutional neural network)、RNN(recurrent neural network)等网络结构提取序列隐含特征,然后利用条件随机场 CRF(conditional random field)求解最优序列,如经典模型 BiLSTM-CRF。然而,BiLSTM-CRF 模型存在训练时间长、无法解决一词多义、不能充分学习文本上下文信息的问题。因此,提出了 BERT-BiGRU-Attention-CRF 模型对中文语料进行命名实体识别。该模型使用经典 BiLSTM-CRF 实体识别模型作为基础模型,用 BiGRU(bi-directional gated recurrent Unit)替换 BiLSTM(bi-directional long short term memory)使参数更少,结构更简单,拥有更好的收敛性。采用预训练语言模型 BERT(bi-directional encoder representations from transformers)训练中文词向量,能够较完整地保存文本语义信息,可以较好地解决命名实体的一词多义问题。使用 Attention 机制挖掘文本序列之间的潜在特征,对语义信息的利用更充分,提升了模型对实体的识别率。将文中模型在 MSRA 语料上与其他相关模型进行对比实验,结果表明模型在中文命名实体识别方面的有效性。

1 相关工作

早期的命名实体识别方法需要人工设计规则模式,这种方法被称为基于词典和规则的方法。这种方法首先要求设计者足够专业,其次规则的设计依赖于数据集,数据集改变时规则必须重新设计,最后是该方法对于处理非结构化的数据效率低下。因此,为避免上述问题,基于统计学习的命名实体识别方法被提出。该方法被处理为一种序列标注问题。比较有代表性的基于统计学习的命名实体识别方法包括最大熵、隐马尔可夫模型、支持向量机和条件随机场模型等。虽然这种方法不需要人工设计规则模式,但是当提取特征时依然需要人工完成。人工提取特征成本代价高,缺乏领域的自适应性,模型的泛化能力和迁移能力欠缺。近年来,基于神经网络的命名实体识别方法被广泛研究。Collobert 等^[1]利用 CNN 取代人工进行特征提取,并且通过融合其他特征提出句子级别上的对数似然函数,取得了较好效果。而 RNN 的提出则解决了难以获得序列上下文长期依赖关系的问题,以及输入为可变长度文本的问题。处理时间序列数据时,在 RNN 基础上改进的许多模型可以很好地获取并保存序列上下文信息^[2-3]。此后,Huang 等^[4]提出 BiLSTM-CRF 命名实体识别模型,同时为提高模型性能还合并了其他语言功能。Google 的 Jacob 等^[5]推出基于 Transformer 的双向编码器 BERT 表示方法。王月^[6]使用 BERT 预训练词向量代替传统 Skip-gram、CBOW(continue bag of word)等方式来训练静态词向量,提升了词向量的表征

能力,解决了中文语料采用字向量训练时词语边界的划分问题,同时还使用注意力机制改进经典的命名实体识别模型架构 BiLSTM-CRF 以达到充分学习文本上下文信息的目的。古雪梅^[7]同样使用了 BERT 融合 BiLSTM-CRF 解决了推文中恶意软件名称识别任务存在的一词多义问题,同时还提出将多头自注意力机制 Self-attention 用于解决类别不均衡的问题。不过,王月与古雪梅的研究都是以 LSTM(long short term memory)模型为基础,而 LSTM 具有训练时间长、参数较多、内部计算复杂等问题。因此,使用参数更少、结构更简单、拥有更好收敛性的 BiGRU 来取代经典命名实体模型 BiLSTM-CRF 中的 BiLSTM,与 BERT 和 Attention 机制融合,提出基于 BERT-BiGRU-Attention-CRF 的中文命名实体识别模型。

2 相关模型

2.1 BERT 模型

语言模型是计算任意语言序列 w_1, w_2, \dots, w_n 出现概率 $P(w_1, w_2, \dots, w_n)$ 的方法,如式(1)所示。

$$P(s) = P(w_1, w_2, \dots, w_n) \quad (1)$$

比较典型的语言模型是从左到右计算下一个词的概率。传统的语言模型无法表示一个字的上下文、字的多义性、句子的句法特征等。针对这个问题,BERT 预训练语言模型被提出。BERT 预训练语言模型结构如图 1 所示。

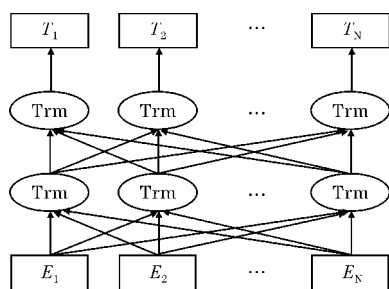


图1 BERT 预训练语言模型

BERT 预训练语言模型采用双向 Transformer 作为特征抽取器,可以获取字符左右两侧更长的上下文信息。BERT 模型拥有两个新的预测任务,分别是“Masked 语言模型”和“下一个句子预测”。它们分别捕捉词级别和句子级别的表示,通过联合训练达到目的。

“Masked 语言模型”目的是实现双向 Transformer 的预训练,该方法采用一个非常简单的方式,遮住句子某些单词,让 Transformer 通过上下文的词去预测被遮挡的单词。随机遮挡 15% 的单词作为训练样本。

这部分单词的遮挡方法如下:

- (1) 80% 被遮挡词用 masked token 替换。
- (2) 10% 被遮挡词用随机词替换。
- (3) 10% 被遮挡词不变。

“下一个句子预测”任务是在 BERT 预训练模型中使用一个二分类模型以学习句子之间的关系。具体做法是将文本中部分句子随机替换为其他句子,利用上一个句子对下一个句子做“是或否”的预测。

BERT 最重要的部分是双向 Transformer 编码结构。由自注意力机制(self-Attention)和前馈神经网络(feed forward network)构成每个单元,单元可以连续堆叠。其结构如图 2 所示。

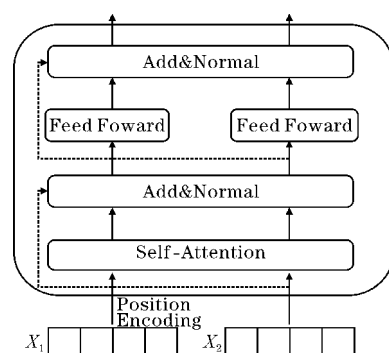


图2 Transformer 编码单元

Transformer 中的自注意力机制模块,其核心思想是先计算一句话中的每个词与这句话中所有词的相互关系,再利用相互关系调整每个词的权重以获得每个词新的表达。这个词新的表征和单纯的词向量相比是一个更加全局的表达,因为它不但蕴含该词本身,还蕴含其他词与这个词的关系^[8]。

BERT 通过联合调节内部的双 Transformer,利用 Self-Attention 在单词编码过程中学习上下文中其余词对当前词的贡献程度,从而增强上下文语义信息的提取。

2.2 BiGRU

LSTM 控制信息传递通过遗忘门、输入门和输出门实现,而 GRU 模型则是将 LSTM 的遗忘门和输入门合并为一个被称为更新门的单一门,是一种更加简单的网络模型。GRU 能取得与 LSTM 相当的结果,且参数更少,结构更简单,拥有更好的收敛性。

不过,因为 GRU 只能得到单向的文本信息,想要得到双向的文本信息,需要使用 BiGRU,它可以从正反方向上同时获得上下文信息。通过 BiGRU 对双向的文本信息进行提取,可以使特征提取的准确率更高。除此之外,BiGRU 还拥有响应时间快、复杂度低、更少依赖于词向量的优点^[9]。

2.3 Attention 机制

Attention 机制在图像领域、语音识别、自然语言处理等各种类型的深度学习任务中表现非凡。Attention 机制通过给予文本中部分词语更多关注来达到提高 Attention 层特征提取质量的目的,所采用的方式是概率权重分配的方式,即计算不同时刻词向量的概率权重^[10]。

2.4 CRF

条件随机场常用于命名实体识别、词性标注、句法分析等,是近几年自然语言处理领域常用的算法之一。在中文命名实体识别中,很强的依赖关系存在于输出标签之间,为保证最后输出标签的合法性,使用 CRF 为最后输出标签添加一些约束。例如, I-ORG 后面不能接 B-PER,可以看出 CRF 约束 I-ORG 之后可能出现的标签。具体地,若令 $X = \{x_1, x_2, \dots, x_n\}$ 为输入序列, $Y = \{y_1, y_2, \dots, y_n\}$ 为与输入序列对应的输出标签序列,则构建条件概率模型 $P(X|Y)$ 是条件随机场的目标。然后,通过此模型预测给定序列的全局最优标签序列。

3 一种新的中文命名实体识别模型

3.1 问题提出

LSTM 的出现是为了解决 RNN 无法很好地处理长距离依赖、梯度消失、梯度爆炸问题。LSTM 在自然语言处理的广泛应用中逐渐暴露很多缺点,比如 LSTM 参数较多、内部计算复杂、训练时间长。于是,2014 年一种更加简单的 GRU 模型被提出。然而目前看来,大多研究仍都是选用 LSTM 作为基础网络进行命名实体识别。

最近几年,在自然语言处理领域中,基于神经网络的深度学习方法取得了显著成功,成就得益于词向量技术的发展。在大部分神经网络命名实体识别模型中,训练词向量的工具最常见的是 Word2Vec。不过,尽管 Word2Vec 能够较好地获得文本序列上下文特征,但是因 Word2Vec 输入的上下文有限,故无法解决一词多义的问题。因此,提出基于 Transformer 的双向编码器 BERT 表示方法。不同于 Word2Vec, BERT 使用文本内容的左、右语境进行预训练得到文本的深度双向表征,有助于命名实体的识别^[11]。因此,如何将 BERT 与命名实体识别模型结合,成为近期研究的热点。

在从命名实体识别中提取特征的过程中,传统的

深度学习过于关注文本的全局特征,而忽略了局部特征对命名实体识别的重要影响。因此,迫切需要使用 Attention 机制为网络模型输出的信息分配不同的权重,以获得句子中最相关的局部信息,充分学习文本上下文信息^[12]。

针对以上提出的问题,利用 BERT、BiGRU、Attention 模型对经典模型 BiLSTM-CRF 改进,以提升命名实体识别的效果。

3.2 新模型结构

所提中文命名实体识别模型由 BERT 层、BiGRU 层、Attention 层和 CRF 层组成,模型结构如图 3 所示。

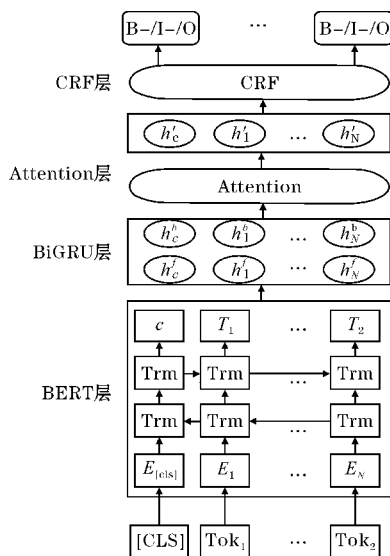


图3 新的中文命名实体识别模型结构

该结构的本质是将预处理好的文本送入 BERT 层得到词向量,然后将 BERT 的输出送入 BiGRU 层获取上下文特征表示,再将 BiGRU 层的输出送入 Attention 层以分配 BiGRU 网络结构中特定信息的权重。最后,利用 CRF 对 Attention 层输出序列进行标注,即第一步对标签转移概率进行建模,第二步利用建好的模型找到全局最优标签序列,以达到命名实体识别的目的^[13]。

3.3 新模型训练过程

3.3.1 BERT 层

在将分词后的句子输入 BERT 模型之前,首先在第一个句子的首尾分别嵌入 cls 和 sep 两个特殊的字符。两个句子之间用 sep 衔接。每个字符对应应有 3 个向量,分别是 token Embeddings(表示词向量)、Segment Embeddings(句向量)和 Position Embeddings(位置向量),如图 4 所示。token Embeddings 为词向量,第一个特殊字符 cls 也有对应的词向量,用于下游的分类

任务;Segment Embeddings 是句向量,用于区分不同句子,便于 BERT 预训练模型做下一个句子预测任务;Position Embeddings 会人为给定一个位置向量来表征序列的顺序信息。

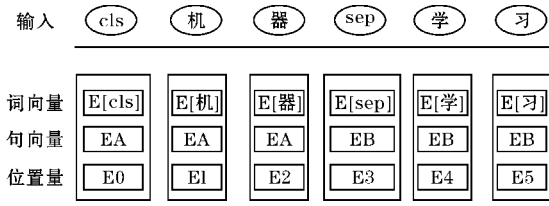


图4 BERT 预训练语言模型的向量构成

将词向量、句向量、位置向量乘以 3 个不同的权值矩阵 W_q, W_k, W_v 得到 Q, K, V , 它们分别是每个字符对应的 3 个不同的向量矩阵。将它们送入公式(2), 然后输出句子中所有词向量的带权和。

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

其中, $Q \in R^{n \times d_k}, K \in R^{m \times d_k}, V \in R^{m \times d_v}, d_k$ 为 Q, K 的其中一个维度。BERT 预训练模型使用由多个自注意力机制构成的多头注意力机制(Multihead-Attention), 首先可以用于获取句子级别的语义信息, 其次可以增大注意力单元的“表示子空间”, 以扩展模型专注于不同位置的能力^[14]。如式(3)和(4)所示, 其中 h 为多头数目, $i = 1, 2, \dots, h$; W_i^Q, W_i^K, W_i^V 为投影矩阵, $W_i^Q \in R^{d_{\text{model}} \times d_k}, W_i^K \in R^{d_{\text{model}} \times d_k}, W_i^V \in R^{d_{\text{model}} \times d_v}, W^O$ 为附加权重矩阵, $W^O \in R^{hd_v \times d_{\text{model}}}$ 。

$$MultiHead(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (3)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

此外, 通过在 Transformer 编码单元中加入残差网络和层归一化来解决深度学习中的退化问题, 如式(5)和(6)所示。

$$LN(x_i) = \alpha \times \frac{x_i - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}} + \beta \quad (5)$$

$$FFN = \max(0, xW_1 + b_1)W_2 + b_2 \quad (6)$$

通过式(2)~(6), 最终, 编码生成基于当前语境上下文的词嵌入向量。

3.3.2 BiGRU 层

BiGRU 层接收 BERT 层的词嵌入向量作为输入。通过式(7)~(10), 得到在 t 时刻的正向 GRU 和反向 GRU 的隐藏状态 \tilde{h}_t 和 \bar{h}_t 。最后, 使用最终的隐藏状态作为输出以达到获得上下文信息的目的^[15]。而最终的隐藏状态是通过连接正向和反向隐藏状态得到, 即 $h_t = [\tilde{h}_t, \bar{h}_t]$ 。具体计算公式如下:

$$z_t^c = \sigma(W_{zx}x_t^c + W_{zh}h_{t-1}^c + b^c) \quad (7)$$

$$r_t^c = \sigma(W_{rx}x_t^c + W_{rh}h_{t-1}^c + b^c) \quad (8)$$

$$\bar{h}_t^c = \tanh(W_{cx}x_t^c + W_{ch}(r_t^c \cdot h_{t-1}^c + b^c)) \quad (9)$$

$$h_t^c = (1 - z_t^c) \cdot h_{t-1}^c + z_t^c \cdot \bar{h}_t^c \quad (10)$$

其中, x_t^c 表示 t 时刻的输入向量, h_{t-1}^c 表示上一隐藏节点输出的激活值, \bar{h}_t^c 表示候选隐藏状态, h_t^c 是隐藏状态, 也是输出向量, 包含前面 t 时刻所有有效信息。 r_t^c 表示重置门, 其功能是决定前一时刻隐藏状态需要重置的信息有多少。当 r_t^c 接近于 0 时, 前一时刻的隐藏状态全部重置为当前时刻的输入。 z_t^c 表示更新门, 其功能是决定前一时刻的信息是否被丢弃。 z_t^c 越小, 表示前一时刻隐藏节点所包含的信息被丢弃得越多。此网络模型复杂度较低的原因就是忽略了某些没有用的信息。因此, 隐藏状态 h_t^c 是由 r_t^c 与 z_t^c 一起决定的。 W 和 b^c 都表示模型参数。 σ 表示 sigmoid 激活函数。 \cdot 表示 Hadamard 乘积。

3.3.3 Attention 层

Attention 层的输入为经过 BiGRU 神经网络层输出的隐藏状态 h_t , 通过 Attention 机制对隐藏状态加权表征, 以挖掘词间的关联关系, 最终得到新的状态序列作为输出。Attention 机制的计算公式为

$$e_i = v_i \tanh(w_i h_i + b_i) \quad (11)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{c=1}^n e_c} \quad (12)$$

$$s_t = \sum_{i=1}^k \alpha_i h_i \quad (13)$$

其中, v_i 和 w_i 表示第 t 时刻的权重系数矩阵; b_i 表示第 t 时刻相应的偏移量; e_i 表示隐层状态向量 h_i 在第 t 时刻得到的能量值。

3.3.3 CRF 层

CRF 层使用 CRF 模型对 Attention 层输出的新隐藏状态序列 $s = (s_1, s_2, \dots, s_n)$ 进行解码, 得到最终预测标签序列。CRF 模型的具体建模细节如下:

若给定输入句子 $h = (h_1, h_2, \dots, h_n)$, 输出标签序列 $y = \{y_1, y_2, \dots, y_n\}$, 将每对 h 和 y 的评估分数定义为 $f(h, y)$ 。 $f(h, y)$ 由两个部分组成, 第一部分: 为获得每个位置 t 处由 BiGRU 网络与 Attention 机制结合输出的分数矩阵, 将隐藏状态 s_t^c 与标注 y_t 对应的参数向量 $w_{y_t}^n$ 相乘。第二部分: 由于标签之间存在相关性, 因此通过建立一阶依赖关系来捕捉这种相关性。方法是在位置 t 处添加矩阵 A , 这种矩阵被称为转移得分矩阵, 也是一种参数矩阵, 其目的是定义不同标签对之间的相似性得分, $A_{i,j}$ 表示标签 i 转移到标签 j 的转移

得分。 $f(h, y)$ 公式为

$$f(h, y) = \sum_{t=0}^n A_{y_t, y_{t+1}} + \sum_{t=1}^n w_{y_t}^n s_t^c \quad (14)$$

通过归一化所有标注序列,得到有序列 y 的概率分布, $Y(h)$ 表示所有可能的标记序列,公式为

$$p(y | h) = \frac{e^{f(h, y)}}{\sum_{y' \in Y(h)} e^{f(h, y')}} \quad (15)$$

在训练过程中,将有关正确标签序列的对数概率最大化,公式为

$$\begin{aligned} \log(p(y | h)) &= f(h, y) - \log\left(\sum_{y' \in Y(h)} e^{f(h, y')}\right) \\ &= f(h, y) - \log y' \in Y(h) \text{ add } f(h, y') \quad (16) \end{aligned}$$

在预测阶段,采用动态规划算法 Viterbi 在新的隐藏状态序列 $s = (s_1, s_2, \dots, s_n)$ 上找到总得分最高的序列,求解最优路径,如式(17)。

$$y^* = \arg \max_{y' \in Y(s)} (s, y') \quad (17)$$

3.4 新模型的算法描述

模型使用 tensorflow 框架实现,算法描述如下:

输入: $C = \{c_1, c_2, \dots, c_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$ (其中 C 是原始字符, Y 为实体的最终标签)

输出:命名实体结果

BERT 层:

(1)使用 Google 训练好的 BERT 模型进行 embedding。

(2)获取对应 embedding 的词向量数据作为输入。

BiGRU 层:

(1)定义前向 GRU, $\text{forward} = \text{GRU}(c_i)$ 。

(2)定义后向 GRU, $\text{backward} = \text{GRU}(c_j)$ 。

(3)采用动态 RNN 建立 BiGRU 模型。通过动态 RNN 加载前向 GRU 和后向 GRU 得到一个元组 (output_fw, output_bw)。其元组中两个元素的维度一样。因此将元组中的两个元素直接进行拼接,即: $H = \text{contact}(\text{output_fw}, \text{output_bw})$ 。得到输出 H , 传入到下一层 Attention 机制中。

Attention 层:

(1)获得 GRU 的神经元数量 hiddenSize, 利用 hiddenSize 初始化一个权重向量 W , 这个权重向量是可训练的参数。

(2)利用公式(11), 首先对 BiGRU 的输出 H 用激活函数做非线性转换得到 M 。然后对 W 和 M 做维度转换, 最后让权重向量 W 和 M 做矩阵运算, 得到 new M 。

(3)再一次对 new M 做维度转换, 得到 restore M 。

(4)如公式(12), 使用 softmax 函数对 restore M 做

归一化处理得 alpha。

(5)利用公式(13), 将求得的 alpha 值通过矩阵运算操作与 H 进行加权求和得 r 。

(6)将三维 r 压缩成二维 squeeze R , 再用激活函数对 squeeze R 做非线性转换得到 sentenceRepre, 即 Attention 的输出。

CRF 层:

(1)利用公式(14)计算某个词被标记为某标签的得分向量。这个词的标签不仅受其他词的标签影响, 而且这个词的标签受相邻词的标签的影响。

(2)利用公式(15)并应用动态规划的思想计算文本所对应标签的概率分布。

4 实验与分析

4.1 实验数据及评价指标

在验证新模型有效性的实验过程中, 实验数据采用 SIGHAN 中文命名实体识别评测的 MSRA 语料。MSRA 语料属于网络上公开的数据, 各种预处理操作都已完成, 可以直接用于实验中。在实验时, 训练集用 46364 条句子, 测试集 4365 句子。

使用准确率 P 、损失率以及迭代时间作为评价标准来评估模型性能。其中 P 如公式(18): T_p 为模型识别正确的实体个数, F_p 为模型识别到的不相关实体个数。

$$P = \frac{T_p}{T_p + F_p} \times 100\% \quad (18)$$

实验迭代次数为 10 次, 该模型的准确率为测试集在这 10 次迭代次数上的最高值, 损失率和迭代时间分别取与准确率所对应的值。

4.2 参数设置

实验参数设置为: BERT 预训练语言模型默认使用 12 头注意力机制的 Transformer, 预训练词向量长度 768 维。每次读取的序列长度 seq_length 是 128, 测试集的 batch_size 是 8, 训练集的 batch_size 是 64。训练 epoches 是 10。训练过程中使用 Adam 作为参数优化算法, 训练学习率为 2×10^{-5} , 实验结果显示较小的学习率有助于模型找到最优解。为缓解梯度消失和爆炸的影响, 还使用了梯度裁剪技术, clip 设置为 5。为避免过拟合, 在 BiGRU 的输入、输出以及 Attention 层中使用了 Dropout 率为 0.5 的正则化方法。

4.3 实验结果与分析

在数据集上, 采用多种模型进行性能对比分析。

实验结果如表 1 所示。

表 1 各模型的实验效果对比

模型	准确率/%	损失率	时间/s
BiLSTM-CRF	82.12	0.2856	1314
BiGRU-CRF	84.56	0.2817	1083
BiGRU-Attention-CRF	87.28	0.2635	1236
BERT-BiLSTM-Attention-CRF	90.78	0.2422	1641
BERT-BiGRU-Attention-CRF(新模型)	92.46	0.2414	1323

从整体来看,这 5 种模型在实验之后得到的准确率、损失率、迭代时间比较接近,随着模型复杂度的变化,准确率、损失率、迭代时间会跟随变化、上下波动。

由表 1 可以看出,基于 BiGRU 的模型在各方面的表现都优于基于 BiLSTM 的模型,说明 BiGRU 参数更少,计算更容易,拥有更好的收敛性,在降低模型训练时间的同时提高准确率降低损失率。对于 BiGRU-CRF 和 BiGRU-Attention-CRF,后者由于模型复杂度提升时间比前者多了约 200 s,但准确率提升了约 3%,损失率降低了约 0.02,说明使用注意力机制提高了隐藏层特征提取的质量。BERT-BiGRU-Attention-CRF 与 BiGRU-Attention-CRF 相比,模型复杂度增加导致时间多了约 100 s,但准确率提升约 5%,损失率降低了约 0.02,表明 BERT-BiGRU-Attention-CRF 模型使用的 BERT 词向量预训练模型在把握语义方面更加精准,对于实体识别等自然语言处理任务的性能提升有较大影响。再看 BiLSTM-CRF 与 BERT-BiGRU-Attention-CRF 的比较,文中模型在时间仅仅多了约 9 s 的情况下,准确率却有约 10% 的提升,损失率下降约 0.04。充分证明了新模型的有效性。

5 结束语

基于 BERT-BiGRU-Attention-CRF 的命名实体识别模型是提出的一种新型命名实体识别模型,将模型与最经典的 BiLSTM+CRF 命名实体识别模型相比,能够在模型复杂度增加的情况下却保证训练时间相当,并且提高准确率的同时降低损失率。一方面,说明 BiGRU 比 BiLSTM 更加简单,训练时间更短;另一方面,说明使用 BiGRU 替换经典模型中 BiLSTM,再与 BERT、Attention 模型结合的有效性。虽然基于 BERT-BiGRU-Attention-CRF 的模型在数据集上有较高准确率和较低损失率,但加入 BERT、Attention 机制模型增加了系统的计算和开销,时间代价也略微有所提升。

因此,接下来将研究在达到较高准确率、较低损失率的同时,是否有更小计算量、更小系统开销、更短训练时间的中文命名实体识别的神经网络模型。

参考文献:

[1] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493–2537.

[2] Graves A. Long Short-term memory [C]. Supervised sequence labelling with recurrent neural Networks. Springer berlin heidelberg, 2012: 1735–1780.

[3] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J]. Computation and Language, 2014, 12(3).

[4] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. Computation and Language, 2015, 3(4): 1508–1991.

[5] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. Computation and Language, 2018, 9(3): 1810–4805.

[6] 王月, 王孟轩, 张胜, 等. 基于 BERT 的警情文本命名实体识别 [J]. 计算机应用. 2019, 11(20): 1001–908.

[7] 古雪梅, 刘嘉勇, 程芃森, 等. 基于增强 BiLSTM-CRF 模型的推文恶意软件名称识别 [J]. 计算机科学, 2019, 10(29): 2096–4188.

[8] 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, 2019, 5(30): 36–41.

[9] 王宁, 李世林, 刘堂亮, 等. 基于注意力机制 BiGRU 判决结果倾向性分析 [J]. 计算机系统应用, 2019, 28(3): 191–195.

[10] 王伟, 孙玉霞, 齐庆杰, 等. 基于 BiGRU-attention 神经网络的文本情感分类模型 [J]. 计算机应用研究, 2019, 12(12): 3559–3564.

[11] 王子牛, 姜猛, 高建瓴, 等. 基于 BERT 的中文命名实体识别方法 [J]. 计算机科学, 2019, 11(3): 139–142.

[12] 冀相冰, 朱艳辉, 李飞, 等. 基于 Attention-BiLSTM 的中文命名实体识别 [J]. 湖南工业大学学报, 2019, 9(5): 74–78.

[13]

石丹春,秦岭. 基于 BGRU-CRF 的中文命名实体识别方法[J]. 计算机科学,2019,9(9):28-242.

[14]

李妮,关焕梅,杨飘,等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 山东大学学报(理学版),2020,1(2):1671-9352.

[15]

李扬,张伟,彭晨. 目标依赖的作者身份识算法[J]. 计算机应用,2019,11(20):2-7.

Research on Chinese Named Entity Recognition based on Deep Learning

WANG Xuemei, TAO Hongcai

(College of Information Science & Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Aiming at the problems of long training time of classic BiLSTM-CRF named entity recognition model, inability to resolve polysemy, and insufficient learning of text context semantic information, a Chinese named entity recognition model based on BERT-BiGRU-Attention-CRF is proposed. Firstly, the BERT language model is used to pre-train the word vector to make up for the problem that the traditional word vector model cannot solve the problem of polysemy. Secondly, the bi-directional gated recurrent unit (BiGRU) neural network layer is applied to extract the features of the deep information of the text, to calculate the predicted score of each label to get the hidden state sequence of the sentence. Thirdly, the attention layer is utilized to weight the representations of the words and mine the association between the words to get new predicted scores and new state sequences. Finally, the conditional random field (CRF) is used to calculate the global optimal solution for the new prediction score, so as to obtain the final prediction result of the model on the entity label. Through the experiments with MSRA corpus, the results show that the new model is effective.

Keywords: Chinese named entity recognition; BERT; BiGRU; Attention; CRF