

基于 BGRU-CRF 的中文命名实体识别方法

石春丹 秦 岭

(南京工业大学计算机科学与技术学院 南京 211816)

**摘 要** 针对传统的命名实体识别方法存在严重依赖大量人工特征、领域知识和分词效果,以及未充分利用词序信息等问题,提出了一种基于双向门控循环单元(BGRU)神经网络结构的命名实体识别模型。该模型利用外部数据,通过在大型自动分词文本上预先训练词嵌入词典,将潜在词信息整合到基于字符的 BGRU-CRF 中,充分利用了潜在词的信息,提取了上下文的综合信息,并更加有效地避免了实体歧义。此外,利用注意力机制来分配 BGRU 网络结构中特定信息的权重,从句子中选择最相关的字符和单词,有效地获取了特定词语在文本中的长距离依赖关系,识别信息表达的分类,对命名实体进行识别。该模型明确地利用了词与词之间的序列信息,并且不受分词错误的影响。实验结果表明,与传统的序列标注模型以及神经网络模型相比,所提模型在数据集 MSRA 上实体识别的总体 F1 值提高了 3.08%,所提模型在数据集 OntoNotes 上的实体识别的总体 F1 值提高了 0.16%。

**关键词** 命名实体识别,双向门控循环单元,注意力机制

中图法分类号 TP391 文献标识码 A DOI 10.11896/j.issn.1002-137X.2019.09.035

Chinese Named Entity Recognition Method Based on BGRU-CRF

SHI Chun-dan QIN Lin

(School of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

**Abstract** Aiming at the problem that the traditional named entity recognition method relies heavily on plenty of hand-crafted features, domain knowledge, word segmentation effect, and does not make full use of word order information, an named entity recognition model based on BGRU(bidirectional gated recurrent unit) was proposed. This model utilizes external data and integrates potential word information into character-based BGRU-CRF by pre-training words into dictionaries on large automatic word segmentation texts, making full use of the information of potential words, extracting comprehensive information of context, and more effectively avoiding ambiguity of entity. In addition, attention mechanism is used to allocate the weight of specific information in BGRU network structure, which can select the most relevant characters and words from the sentence, effectively obtain long-distance dependence of specific words in the text, recognize the classification of information expression, and identify named entities. The model explicitly uses the sequence information between words, and is not affected by word segmentation errors. Compared with the traditional sequence labeling model and the neural network model, the experimental results on MSRA and OntoNotes show that the proposed model is 3.08% and 0.16% higher than the state-of-art complaint models on the overall F1 value respectively.

**Keywords** Named entity recognition, Bidirectional gated recurrent unit, Attention mechanism

1 引言

命名实体识别(Named Entity Recognition, NER)是自然语言处理应用程序的基本技术,主要用于查找文本中的命名实体(人名、地名、组织名等),在关系抽取、信息检索、机器翻译等任务中有着重要作用。早期的命名实体识别是采用基于词典和规则的方法,通过人工构建规则,不仅费时费力,而且可移植性差。在面对众多领域的复杂文本时,传统命名实体识别方法不再适用。随着技术的不断发展,基于机器学习的NER方法被提出。在该类方法中,中文NER的任务通常被

认为是序列标注问题,对序列中的每一个字标注一个标签,并根据最终的标签判定命名实体的边界和类型。通常,序列标注的标签模式为 BIOES,其中 B 表示命名实体的开始, I 表示命名实体的中间, E 表示命名实体的结尾, S 表示单个字符, O 表示其他(用于标记无关字符)。主流的序列标注模型包括隐马尔可夫模型(Hidden Markov Model, HMM)、最大熵(Maximum Entropy, ME)、支持向量机(Support Vector Machine, SVM)和条件随机场模型(Conditional Random Fields, CRF)<sup>[1]</sup>等。基于传统的机器学习方法面临着特征选取要求高的问题。针对以上问题,学者们提出了混合的方法,即将机

到稿日期:2018-08-13 返修日期:2018-11-06

石春丹(1994—),女,硕士生,主要研究领域为机器学习与深度学习;秦岭(1980—),男,硕士,讲师,主要研究领域为面向流程工业的机器学习, E-mail: ql@njtech.edu.cn(通信作者)。

器学习和人工知识结合起来识别命名实体。Duan 等<sup>[1]</sup>将中文实体的前后缀特征与 CRFs 结合,对《人民日报》语料中的人名、地名和机构名进行识别;Zhou 等<sup>[2]</sup>提出使用 4 种不同的特征来提高 HMM 在 NER 任务中的性能这类方法的主要优势体现在机器学习的方法可以减少人工构建规则的成本,人工知识的加入又降低了对大规模语料的依赖性;但总体来说,人工特征的成本代价偏高,特征学习缺乏领域自适应性,识别方法的泛化能力欠缺。

随着深度学习的快速发展,神经网络在 NER 任务中的表现越来越突出,其性能也越来越优于 CRF 等常用的统计算法。但是,常见的神经网络对训练样本的学习只考虑了训练样本的输入,并没有考虑训练样本的输出之间的关系。因此,通常采用神经网络与 CRF 相结合的方式弥补神经网络模型的不足。最近的研究大多利用了 LSTM-CRF 架构。Huang 等<sup>[3]</sup>提出了双向 LSTM-CRF 模型,该模型在序列标注任务上是稳健的,且对词向量的依赖较少;Ma 等<sup>[4]</sup>以及 Chiu 等<sup>[5]</sup>提出了一种使用 CNN 来模拟字符级信息的 BLSTM-CNNs-CRF 架构;Lample 等<sup>[6]</sup>提出了一个 Bi-LSTM-CRF 体系结构,该结构使用基于字符的 LSTM 层从监督语料库中学习拼写特征,并且除了使用大量的无标记语料库中无监督学习预训练的词嵌入外,没有使用任何额外的资源或地名词典。上述方法不再依赖人工特征,而是直接使用字符向量或者词向量就可以得到很好的命名实体识别效果。因此,本文也采用神经网络的方法来实现 NER。

目前,很多关于 NER 的研究都利用了外部信息来源,词典特征已被广泛使用。Peters 等<sup>[7]</sup>预先训练一个字符语言模型来增强单词表示。Yang 等<sup>[8]</sup>通过多任务学习,利用跨域和跨语言知识进行 NER 的研究。另外,很多研究已经对比了 NER 基于词级和基于字符级的方法,研究结果表明后者在经验上是一个更好的选择<sup>[9]</sup>。但是基于字符级的方法无法理解上下文的语义关系,如何更好地利用单词信息进行中文 NER 受到了研究者的持续关注。针对以上问题,本文提出利用外部信息,通过在大型自动分割文本上预先训练词嵌入词典,将句子与大型自动分割获得的词嵌入词典进行匹配,生成的单词序列消除了上下文中潜在的相关命名实体的歧义,更好地利用了明确的单词和潜在的有用单词序列信息。

本文通过将句子与大型自动分割获得的词嵌入词典进行匹配来构造词-字符结构,使用基于词-字符结构的 GRU 表示句子中的词典单词,将潜在单词信息整合到基于字符的 GRU-CRF 中。此结构自动控制了从句子开头到结尾的信息流。加入了注意力机制的门控单元动态地将来自不同路径的不同权重的信息路由到每个字符。通过训练 NER 数据,基于词-字符,GRU 可以从上下文中自动找到更有用的单词以获得更好的 NER 性能。与基于字符和基于词的 NER 方法相比,本文的模型利用显式词信息而不是字符序列标记,且不受分割错误的影响。

2 相关工作

2.1 BGRU

在基于机器学习的方法中,命名实体识别被认为是典型的序列标注问题。区别于其他神经网络,循环神经网络

(RNN)是一种能有效解决序列标注问题且可以模拟序列元素之间依赖关系的神经网络模型。循环神经网络的隐藏层之间的节点不是无连接的,隐藏层的值不仅仅取决于当前的输入,还取决于上一时刻隐藏层的值,具有一定的记忆功能。但实践中,RNN 更偏向于序列中最近的输入,无法很好地处理长距离依赖问题,并且在训练时存在梯度消失或梯度爆炸问题,在实际应用中网络层数较少,最终命名实体识别的效果甚微。

为了解决这个问题,更复杂的非线性激活函数长短期记忆(Long Short-Term Memory, LSTM)和门控循环单元(Gated Recurrent Unit, GRU)被提出。LSTM 由输入门、遗忘门、输出门以及一个 cell 单元组成。而 GRU 作为 LSTM 的变体,将忘记门和输入门合成了一个单一的更新门,同时还混合了细胞状态和隐藏状态。与 LSTM 相比,GRU 模型参数更少,结构更简单,但能够取得与 LSTM 模型相当的结果<sup>[10]</sup>。因此,本文选用 GRU 来学习给定句子的结构信息。GRU 的内部结构如图 1 所示。GRU 的定义如下:

$$z_j^\epsilon = \sigma(W_{zx}x_j^\epsilon + W_{zh}h_{j-1}^\epsilon + b^\epsilon) \tag{1}$$

$$r_j^\epsilon = \sigma(W_{rx}x_j^\epsilon + W_{rh}h_{j-1}^\epsilon + b^\epsilon) \tag{2}$$

$$\tilde{h}_j^\epsilon = \tanh(W_{cx}x_j^\epsilon + W_{ch}(r_j^\epsilon \odot h_{j-1}^\epsilon + b^\epsilon)) \tag{3}$$

$$h_j^\epsilon = (1 - z_j^\epsilon) \odot h_{j-1}^\epsilon + z_j^\epsilon \odot \tilde{h}_j^\epsilon \tag{4}$$

其中, $z_j^\epsilon$ , $r_j^\epsilon$ 和 $\tilde{h}_j^\epsilon$ 分别表示一组更新门、重置门和候选隐藏状态。 $x_j^\epsilon$ 表示当前神经网络的输入, $h_{j-1}^\epsilon$ 表示上一隐藏节点输出的激活值。 $\sigma$ 表示 sigmoid 激活函数。 $W$ 和 $b^\epsilon$ 是每个单元的模型参数。 $\odot$ 表示 Hadamard 乘积。更新门 $z_j^\epsilon$ 决定之前时刻的信息是否被忽略,其值越大,之前时刻隐藏节点提供的信息越多。重置门 $r_j^\epsilon$ 决定来自先前隐藏状态的需要重置的信息量,当其值接近 0 时,前一时刻的隐藏状态会被忽略,隐藏状态重置为当前时刻的输入,此机制可以抛弃某些无用的信息,降低计算复杂度。

GRU 只在一个方向上对信息流进行建模,通常使用一个反向的额外的递归网络来获取下文信息。更具体地说,我们使用 GRU 的扩展,即双向门控循环单元,其中给定一个长度为  $T$  的序列,一个 GRU 从 1 到  $T$ 、另一个 GRU 从  $T$  到 1 建模不同长度的依赖关系。设 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 分别表示正向 GRU 和反向 GRU 在位置  $t$  的隐藏状态。连接两个隐藏状态,形成最终的隐藏状态  $h_t = [\vec{h}_t, \overleftarrow{h}_t]$ ,并将该状态作为输出,以获取上下文信息。

2.2 注意力机制

注意力机制(Attention Mechanism)最初被应用在图像领域,随后被广泛使用在语音识别、自然语言处理等各种类型的深度学习任务中。注意力机制可以通俗地理解为利用有限的注意力资源从大量信息中快速筛选出高价值的信息。

虽然 GRU 可以对序列数据进行编码,但它仍然很难学习长期依赖关系。受近期众多自然语言处理领域注意机制研究的启发,我们通过引入新的注意机制<sup>[11]</sup>来规避这个问题。通过合并注意机制,模型可以关注并选择输入的一部分特定信息识别信息表达的分类,使模型行为更易于解释。

2.3 条件随机场

条件随机场(Conditional Random Field, CRF)是一种典型的判别式模型,最早由 Lafferty 于 2001 年提出<sup>[12]</sup>,通常在

观测序列的基础上对目标序列进行条件概率建模,重点解决序列化标注的问题。具体地,若令  $X = \{x_1, x_2, \dots, x_n\}$  为观测序列,  $Y = \{y_1, y_2, \dots, y_n\}$  为与之对应的标注序列,则条件随机场的目标是构建条件概率模型  $P(X|Y)$ 。条件随机场是近几年自然语言处理领域常用的算法之一,常用于句法分析、命名实体识别、词性标注等。

### 3 基于 BGRU-CRF 的命名实体识别模型

本文设计了一种可以用于命名实体识别的,在基于字符模型的基础上扩展了基于单词的单元并添加了注意力机制的双向 GRU-CRF 模型结构,如图 1 所示。

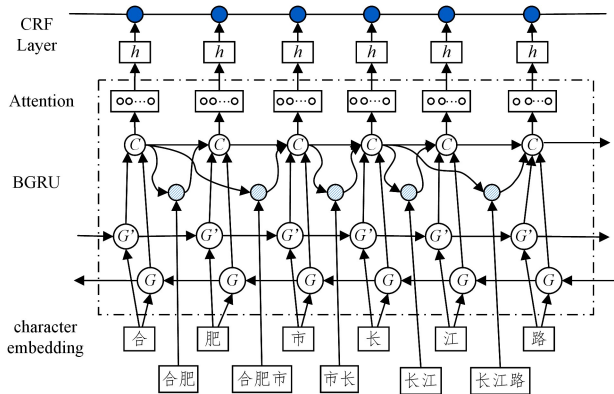


图 1 基于双向 GRU-CRF 的命名实体识别模型

Fig. 1 Named entity recognition model based on bidirectional GRU-CRF

该结构的本质在于将字符序列送入 BGRU 中,再将大型自动分割的词典中的单词与所有字符子序列进行匹配,充分利用了明确的单词和潜在的有用单词序列信息,然后将 BGRU 的输出送入注意力机制层,最后进行 CRF 序列标注,对标签转移概率进行建模,从而获取全局最优的标签序列,以实现命名实体的识别。字符单元向量包含了句子中大量的词典子序列,记录了从句子开头到该字符的循环信息流,使得我们能更加有效地利用单词信息实现实体消歧。BGRU 有效地包含了上下文中字词的表示,注意力机制分配了 BGRU 网络结构中特定信息的权重,可以有效获取特定词语在文本中的长距离依赖关系;CRF 则是对 BGRU 获得的信息进行再利用。

#### 3.1 字词结合特征

本文采用字符和词单元结合的方法,如图 2 所示。

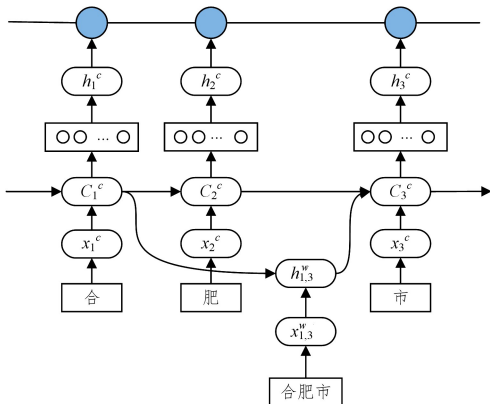


图 2 字词结合模型

Fig. 2 Character and word combination model

该方法可以看作是基于字符的模型的扩展,集成了基于单词的单元,并控制信息权重,贯穿了整个信息流,避免了训练数据分割错误的影响。

模型的输入是原字符序列  $c_1, c_2, \dots, c_m$  和大型自动分割的词典中词语匹配的所有字符子序列。使用  $w_{b,e}^w$  来表示以字符索引  $b$  开头、以字符索引  $e$  结尾的子序列,图 1 中  $w_{1,2}^w$  为“合肥”,而  $w_{5,6}^w$  为“长江”。

模型中涉及 4 种类型的向量,即输入向量、输出隐藏向量、单元向量和门向量。 $x_j^c$  表示每个字符  $c_j$  的字符输入向量:

$$x_j^c = e^c(c_j) \quad (5)$$

其中,  $e^c$  表示字符嵌入查找表。

$\tilde{h}_j^c$  的隐藏状态需要考虑句子中的词典子序列  $w_{b,e}^w$ , 因此每个词典子序列  $w_{b,e}^w$  的向量表示为  $x_{b,e}^w$ :

$$x_{b,e}^w = e^w(w_{b,e}^w) \quad (6)$$

其中,  $e^w$  表示词嵌入查找表。

词单元  $\tilde{h}_{b,e}^w$  用于表示从句子开头起  $x_{b,e}^w$  的循环状态。 $\tilde{h}_{b,e}^w$  的值由下式计算:

$$z_{b,e}^w = \sigma(W_{zx}x_{b,e}^w + W_{zh}h_{b,e}^c + b^w) \quad (7)$$

$$r_{b,e}^w = \sigma(W_{rx}x_{b,e}^w + W_{rh}h_{b,e}^c + b^w) \quad (8)$$

$$\tilde{h}_{b,e}^w = \tanh(W_{cx}x_{b,e}^w + W_{ch}(r_j^c \odot h_b^c)) \quad (9)$$

$$h_{b,j}^w = (1 - z_{b,j}^w) \odot h_b^c + z_{b,j}^w \odot \tilde{h}_{b,j}^w \quad (10)$$

其中,  $z_{b,e}^w$  和  $r_{b,e}^w$  是一组更新门和重置门。因为序列标记是以字符为级别的,所以词单元没有输出门,单元状态即为词汇信息。

#### 3.2 Attention 层

双向 GRU 中的词汇信息不会全部融入当前字符单元中,因此需要对这些词汇信息进行取舍,故采用注意力机制的方法控制当前字符的字符向量、当前词汇的单元状态和权重矩阵信息的贡献值,并采用归一化的方法求出当前字符单元的各输入权重,类似于 softmax 函数。

$$r_{b,e}^c = \sigma(W_{lx}X_e^c + W_{lh}h_{b,e}^w + b^l) \quad (11)$$

$$\alpha_{b,j}^c = \frac{\exp(r_{b,j}^c)}{\exp(r_{b,j}^c) + \sum_{b' \in \{b' | w_{b',j}^d \in D\}} \exp(r_{b',j}^c)} \quad (12)$$

$$\alpha_j^c = \frac{\exp(r_j^c)}{\exp(r_j^c) + \sum_{b' \in \{b' | w_{b',j}^d \in D\}} \exp(r_{b',j}^c)} \quad (13)$$

最后,将当前词汇和字符隐藏值进行加权,以获得当前字符的单元状态:

$$c_j^c = \sum_{b \in \{b' | w_{b',j}^d \in D\}} \alpha_{b,j}^c \odot h_{b,e}^w + \alpha_j^c \odot \tilde{h}_j^w \quad (14)$$

#### 3.3 CRF 层

隐藏的上下文向量  $h = \{h_1, h_2, \dots, h_t\}$  可以直接用作特征,以便为标签序列  $y = \{y_1, y_2, \dots, y_t\}$  的每个输出  $y_t$  做出独立的标记决策。但在中文命名实体识别中,输出标签之间存在很强的依赖关系。CRF 层可以为最后预测的标签添加一些约束来保证预测的标签是合法的。例如, I-PER 不能后接 B-ORG,它限制了 B-ORG 之后可能的输出标签。因此,我们使用 CRF 来共同模拟整个句子的输出,在词级 GRU 输出的隐藏状态  $h$  上应用条件随机场层。

设  $Y(h)$  为  $h$  的标记序列空间。给定隐藏状态序列  $h$  的



标签序列  $y$  的条件对数概率为:

$$\log(p(y|h))=f(h,y)-\log(\sum_{y'\in Y(h)}e^{f(h,y')})$$

(15)

其中,  $f$  为每对  $h$  和  $y$  的得分函数。

为了定义函数  $f(h,y)$ , 对于每个位置  $t$ , 将隐藏状态  $h_t^c$  与由标注  $y_t$  对应的参数向量  $w_{y_t}^T$  相乘, 以获得在位置  $t$  处 BG-RU 网络输出的分数矩阵。由于我们还需要考虑标注之间的相关性, 因此通过在位置  $t$  处添加过渡分数矩阵  $A$  来建立一阶依赖关系,  $A$  是定义不同标签对之间的相似性得分的参数矩阵。对于一系列预测  $y=\{y_1,y_2,\cdots,y_l\}$ , 我们把它的分数定义为:

$$f(h,y)=\sum_{t=1}^lA_{y_{t-1},y_t}+\sum_{t=1}^lw_{y_t}^Th_t^c$$

(16)

其中,  $A$  是过渡分数矩阵, 用于模拟从标签  $i$  到标签  $j$  的过渡。我们将开始标记  $y_0$  和结束标记  $y_n$  添加到标记集中。因此,  $A$  是大小为  $k+2$  的方阵。

在所有可能的标签序列上应用 softmax 层之后, 序列  $y$  的概率为:

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

(17)

在训练期间, 最大化正确标签序列的对数概率:

$$\log(p(y|h))=f(h,y)-\log(\sum_{y'\in Y(h)}e^{f(h,y')})$$
$$=f(h,y)-\log_{y'\in Y(h)}addf(h,y')$$

(18)

其中,  $Y(h)$  表示所有可能的标记序列, 包括那些不遵守 IOB 格式约束的标记序列。模型在预测过程(解码)时, 使用一阶 Viterbi 算法在基于词或基于字符的输入序列上找到得分最高的标签序列, 求解最优路径:

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

(19)

将双向 GRU 网络和 CRF 组合在一起, 形成 BGRU-CRF 模型。该网络可以通过 BGRU 层有效地使用过去和未来的输入特征, 并通过 CRF 层有效地使用句子级标注信息。CRF 层连接连续输出层的线性表示。将 CRF 层状态转移矩阵作为参数。通过 CRF 层, 可以有效地使用过去和将来的标注来预测当前标注, 该标注类似于通过双向 GRU 网络使用过去和未来的输入特征。

本文模型的算法描述如算法 1 所示。

**算法 1** 基于 BGRU-CRF 的中文命名实体识别算法

输入:  $C=\{c_1,c_2,\cdots,c_m\}$ ,  $Y=\{y_1,y_2,\cdots,y_m\}$  (其中  $C$  是原始字符,  $Y$  为实体的最终标签,  $w_{b,e}^w$  为与词典中词语匹配的字符子序列)

输出: 命名实体结果

Encoder(编码层):

1. forward=GRU( $c_1,w_{b,e}^w$ )//前向 GRU
2. backward=GRU( $c_1,w_{b,e}^w$ )//后向 GRU
3. h=concat(forward,backward)//双向 GRU

Decoder(解码层):

4. 使用式(11)一式(13)计算 Attention 权重;
5. 使用式(14)得到字符单元状态;
6. 通过式(15)一式(19)加入 CRF 模型, 计算实体名之间的依赖关系。

**3.4 训练过程**

本文需要一个能够以最小的复杂度很好地解释训练数据的优化模型。正则化具有选择经验风险和模型复杂性较小的特点, 并且可以更好地避免模型过度拟合。因此, 使用 L2 正则化的句子级对数似然损失函数来训练模型:

$$L=\sum_{i=1}^N\log(P(y_i|h_i))+\frac{\lambda}{2}\|w\|^2$$

(20)

其中,  $\lambda$  是 L2 正则化参数,  $w$  表示参数集。

**4 实验结果及分析**

本节将介绍本文所提模型在各数据集上获得的结果以及网络配置对模型性能的影响, 并将其与其他方法进行比较。

**4.1 基准数据集**

1) MSRA: 2006 年 SIGHAN 命名实体识别语料库, SIGHAN(Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics)是国际计算语言学会(ACL)中文语言处理小组的简称。

2) OntoNotes-5.0: 由 1 745 000 条英语、900 000 条中文和 300 000 条阿拉伯语文本数据组成, 其数据来源多样, 包括电话对话、广播新闻、广播对话、新闻通讯和博客。本文只使用 OntoNotes-5.0 中文数据。

3) 《人民日报》1998 年 1 月份的标注语料: 以 1998 年《人民日报》的语料为对象, 是北京大学计算语言学研究所和富士通研究开发有限公司共同制作的标注语料库。该语料库对 600 多万字节的中文文章进行了分词及词性标注, 其被作为原始数据应用于大量的研究和论文中。

**4.2 实验设置**

《人民日报》是公开数据集, 并且已经给出了分词结果和词性标注, 因此不需要做任何预处理, 可以直接用于实验。本文使用 word2vec<sup>[13]</sup> 对 3 个数据集进行词向量训练, 非线性函数选取的是 tanh 函数, 生成字向量和词向量, 最终的词典中包含 813 400 个词汇。在使用 word2vec 进行预训练时, 对词嵌入进行微调。分词采用 Yang 等<sup>[14]</sup> 提出的分词方法。表 1 列出了本文实验模型的超参数值, 这些值根据文献中的工作进行了修正, 没有针对单独数据集进行调整。embedding 大小设置为 50, GRU 模型的隐藏大小设置为 200。Dropout 应用于词语和字嵌入, 速率为 0.5。随机梯度下降(SGD)用于模型参数的优化, 初始学习率为 0.015, 学习率的更新采用逐步降低(step decay)法, 衰减率为 0.05。

表 1 参数配置

Table 1 Configuration of parameters

参数名	参数值
字向量维度	50
Word-char 向量维度	50
GRU 隐藏节点数	200
Char dropout	0.5
Word-char dropout	0.5
GRU 层	1
正则化强度	$1\times10^{-8}$
学习率	0.015
衰减率	0.05

**4.3 对比实验**

为了验证本文提出的基于 word-char 的 BGRU-CRF 的中文命名实体识别方法的有效性, 将基于神经网络的 char 级和 word 级中文命名实体方法作为对比基线。基线采用的是 Bi-LSTM-CRF 模型。对 MRSA、OntoNotes 和《人民日报》数据集进行模型训练, 并利用测试集进行测试。将本文提出的 BGRU-CRF 模型与其他模型以及基线模型进行对比, 结果如表 2—表 4 所列。

表 2 MSRA 数据集上的对比结果

模型	精确率/%	召回率/%	F1/%	Sec/Epoch
Chen(2006)	91.22	81.71	86.20	16
Zhang(2006)	92.20	90.18	91.18	23
Zhou(2012)	88.94	84.20	86.51	32
Zhou(2013)	91.86	88.75	90.28	53
Dong(2016)	91.28	90.62	90.95	80
Word baseline	90.57	83.06	86.65	76
Char baseline	90.74	86.96	88.81	72
Ours	94.65	93.87	94.26	56

表 3 OntoNotes 数据集上的对比结果

模型	精确率/%	召回率/%	F1/%	Sec/Epoch
Wang(2013)	76.43	72.32	74.32	19
Che(2013)	77.71	72.51	75.02	21
Yang(2016)	65.59	71.84	68.57	46
Yang(2017)	72.98	80.15	76.40	62
Word baseline	72.84	59.72	65.63	76
Char baseline	68.79	60.35	64.30	73
Ours	77.82	75.34	76.56	55

表 4 《人民日报》数据集上的对比结果

模型	人名/%			机构名/%			地名/%			Sec/Epoch
	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1	
CRF	95.40	85.70	90.30	95.70	89.30	92.40	93.70	87.60	90.50	16
张(2017) <sup>[23]</sup>	94.12	95.12	94.61	87.22	89.96	88.57	97.32	96.28	96.80	20
冯(2018) <sup>[24]</sup>	98.23	89.49	93.66	97.52	89.34	93.25	93.74	87.98	90.77	85
Word baseline	90.27	89.18	89.72	88.94	87.49	88.21	96.17	91.79	93.93	75
Char baseline	90.18	88.65	89.41	90.35	88.75	89.49	92.77	89.19	90.95	72
Ours	97.35	91.83	94.50	98.12	89.68	93.71	97.25	97.56	97.40	56

在数据集 MSRA 上,Chen 等<sup>[15]</sup>使用基于字符的 CRF 模型,F1 值为 86.20%;Zhang 等<sup>[16]</sup>使用 ME(Maximum Entropy)模型并结合多方面综合知识,其 F1 值达到了 91.18%;Zhou 等<sup>[17]</sup>在 MSRA 数据集上使用基于单词的 CRF 模型和人工特征获得的 F1 为 86.51%;Zhou 等<sup>[18]</sup>采用了一种更细粒度的标注方法,F1 值达到了 90.28%。Dong 等<sup>[19]</sup>提出的 BLSTM-CRF,利用基本特征,预训练字符嵌入,F1 值有所提高,达到了 90.95%,取得了很好的性能。本文所提模型的 F1 值比表 2 中其他最高的结果高出了 3.08%。

在数据集 OntoNotes 上,Wang 等<sup>[20]</sup>有效利用了大量双语文本来获取实体信息;Che 等<sup>[21]</sup>通过双语约束将实体标签统一起来,提高了 F1 值;Yang 等<sup>[22]</sup>在 Che 的基础上结合了离散和神经特征的序列标注,大大提高了命名实体的结果。本文模型在 OntoNotes 上的 F1 值比 Yang 等的结果高出了 0.16%。

在《人民日报》数据集上,将本文模型与目前实验结果较好的模型进行了对比。对比结果显示:本文模型在机构名和地名的识别上都超越了他模型,但是在人名的 F1 值方面还稍微欠缺,这主要是因为人名形式多变,用字灵活。

从表 2—表 4 可以得出,与现有方法相比,本文提出的基于单词-字符的 BGRU-CRF 模型更具有竞争力,明显优于其他方法,在此标准的基准测试中获得了最佳结果。因此,该方法被证明是十分有效的,将潜在的词语信息融合到基于字符的 BGRU-CRF 中,充分利用了潜在词的信息,提取了上下文的综合信息,并更加有效地避免了实体歧义。

比较各表格中最后一列的训练时间可以看出:基本的 CRF 模型的训练时间最短,这说明 CRF 对训练有更优的时间性能。本文提出的模型在基于 CRF 的基础上加入了 BGRU,相比于 character-based Bi-LSTM-CRF(Char baseline)模型和 word-based Bi-LSTM-CRF(Word baseline)模型,多了注意力机制层,但因为 BGRU 结构更简单,且计算复杂度更低,而且我们将潜在词语融合到了基于字符的双向 GRU-CRF 中,所以训练时间也减少了 18 s 左右,这说明结合了 BGRU

和 CRF 的网络模型能有效减少训练时间,从而验证了本文所提模型有着更优的训练性能。

句子长度也影响着模型的 F1 值,图 3 分别显示了在 MRSA 和 OntoNotes 数据集上,本文提出的 BGRU-CRF 模型在不同句子长度下的 F1 值。可以看出,不管在何种数据集中,短句中都拥有着充分高的 F1 值,但是随着句子长度的增加,BGRU-CRF 的 F1 值会稍微降低,这与词语组合的复杂度有关;但同时,本文模型对句子长度的增加表现出了很强的鲁棒性,证明了其对单词信息的使用更有效。

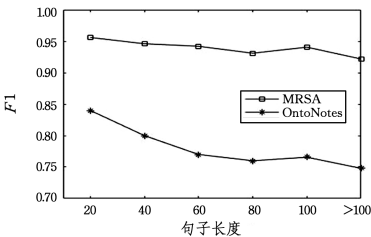


图 3 不同句子长度下的 F1 值

Fig. 3 F1 values for different sentence lengths

**结束语** 本文提出了一种利用 BGRU-CRF 模型解决中文命名实体识别问题的有效方法。与基于字符的方法相比,本文模型明确地利用了单词和单词序列信息。与基于单词的方法相比,word-char BGRU 不会受到分割错误的影响;同时由于添加了注意力机制,所提模型能从句子中选择最相关的字符和单词,以获得更好的 NER 结果。实验表明,该模型在上下文中更自由地选择词典单词,其完全独立于分词,更有效地使用了单词信息,实现了命名实体识别的消歧。

根据实验发现,由于存在噪声词,不同质量的词典会导致本文 NER 的准确性发生变化。因此,后续我们要研究词典质量带来的影响,以及如何在词典质量有限的情况下,通过其他的神经网络模型或者模型组合的方法来提高 NER 的性能。

**参考文献**

[1] DUAN H,ZHENG Y. A Study on Features of the CRFs-based

Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence Paradigms, 2011, 3(2): 287-294.

[2] ZHOU G D, SU J. Named Entity Recognition Using an HMM-based Chunk Tagger[C]//Proceedings of the 40<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL). 2002; 473-480.

[3] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [J/OL]. <https://arxiv.org/abs/1508.01991>.

[4] MA X Z, HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J/OL]. <http://adsabs.harvard.edu/abs/2016arXiv160301354M>.

[5] CHIU J P C, NICHOLS E. Named Entity Recognition with Bidirectional LSTM-CNNs [J/OL]. <https://arxiv.org/abs/1511.08308>.

[6] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2016). 2016; 260-270.

[7] PETERS M E, 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. 2017; 1756-1765.

[8] YANG Z, SALAKHUTDINOV R, COHEN W W. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks[C]//International Conference on Learning Representations (ICLR 2017). 2017.

[9] LIU Z, ZHU C, ZHAO T. Chinese Named Entity Recognition with a Sequence Labeling Approach: Based on Characters, or Based on Words? [M]. Berlin: Springer Berlin Heidelberg, 2010; 634-640.

[10] YANG Z, SALAKHUTDINOV R, COHEN W. Multi-Task Cross-Lingual Sequence Tagging from Scratch[J]. arXiv: 1603.06270, 2016.

[11] SHIMAOKA S, STENETORP P, INUI K, et al. An Attentive Neural Architecture for Fine-grained Entity Type Classification [C]//Proceedings of the 5th Workshop on Automated Knowledge Base Construction. 2016.

[12] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001). 2001; 282-289.

[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in Neural Information Processing Systems. Berlin: Springer, 2013; 3111-3119.

[14] YANG J, ZHANG Y, DONG F. Neural word segmentation with rich pretraining[EB/OL]. [https://www.researchgate.net/publication/316598949\\_Neural\\_Word\\_Segmentation\\_with\\_Rich\\_Pretraining](https://www.researchgate.net/publication/316598949_Neural_Word_Segmentation_with_Rich_Pretraining).

[15] CHEN A, PENG F, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models[EB/OL]. <https://www.semanticscholar.org/paper/Chinese-Named-Entity-Recognition-with-Conditional-Chen-Peng/7c3c13060b7101816a11566eda4fa21d2a82af9e>.

[16] ZHANG S, WEN J, WANG X. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006; 158-161.

[17] ZHOU J, HE L, DAI X, et al. Chinese Named Entity Recognition with a Multi-Phase Model[EB/OL]. <http://www.docin.com/p-195138504.html>.

[18] ZHOU J, QU W, ZHANG F. Chinese Named Entity Recognition via Joint Identification and Categorization[J]. Chinese Journal of Electronics, 2013, 22(2): 225-230.

[19] DONG C, ZHANG J, ZONG C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[C]//International Conference on Computer Processing of Oriental Languages. Springer International Publishing, 2016; 239-250.

[20] WANG M, CHE W, MANNING C D. Effective bilingual constraints for semi-supervised learning of named entity recognizers [C]//Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press, 2013; 919-925.

[21] CHE W X, WANG M Q, MANNING C D, et al. Named entity recognition with bilingual constraints[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013). 2013; 52-62.

[22] YANG J, ZHANG Y, DONG F. Neural Word Segmentation with Rich Pretraining [EB/OL]. [http://www.researchgate.net/profile/Jie\\_Yang126/publication/318740993\\_Neural\\_Word\\_Segmentation\\_with\\_Rich\\_Pretraining/links/59a4ff84a6fdcc773a389875/Neural-Word-Segmentation-with-Rich-Pretraining.pdf](http://www.researchgate.net/profile/Jie_Yang126/publication/318740993_Neural_Word_Segmentation_with_Rich_Pretraining/links/59a4ff84a6fdcc773a389875/Neural-Word-Segmentation-with-Rich-Pretraining.pdf).

[23] ZHANG H N, WU D Y, LIU Y, et al. Chinese Named Entity Recognition Based on Deep Neural Network[J]. Journal of Chinese Information Processing, 2017, 31(4): 28-35. (in Chinese) 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.

[24] FENG Y H, YU H, SUN G, et al. Named Entity Recognition Method Based on BLSTM[J]. Computer Science, 2018, 45(2): 261-268. (in Chinese) 冯艳红, 于红, 孙庚, 等. 基于 BLSTM 的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261-268.