·人工智能与模式识别· 文章编号: 1000-3428(2020)04-0040-06 文献标志码: A 中图分类号: TP391.1

基于 BERT 嵌入的中文命名实体识别方法

杨 飘,董文永

(武汉大学 计算机学院,武汉 430072)

摘 要: 在基于神经网络的中文命名实体识别过程中,字的向量化表示是重要步骤,而传统的词向量表示方法只是将字映射为单一向量,无法表征字的多义性。针对该问题,通过嵌入 BERT 预训练语言模型,构建 BERT-BiGRU-CRF 模型用于表征语句特征。利用具有双向 Transformer 结构的 BERT 预训练语言模型增强字的语义表示,根据其上下文动态生成语义向量。在此基础上,将字向量序列输入 BiGRU-CRF 模型中进行训练,包括训练整个模型和固定 BERT 只训练 BiGRU-CRF 2 种方式。在 MSRA 语料上的实验结果表明,该模型 2 种训练方式的 F1 值分别达到 95.43%和94.18%,优于 BiGRU-CRF、Radical-BiLSTM-CRF 和 Lattice-LSTM-CRF 模型。

关键词: 中文命名实体识别; BERT 模型; BiGRU 模型; 预训练语言模型; 条件随机场

开放科学(资源服务)标志码(OSID):

中文引用格式: 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(4):40-45,52.

英文引用格式: YANG Piao,DONG Wenyong. Chinese named entity recognition method based on BERT embedding[J]. Computer Engineering,2020,46(4):40-45,52.

Chinese Named Entity Recognition Method Based on BERT Embedding

YANG Piao,DONG Wenyong

(School of Computer Science,Wuhan University,Wuhan 430072,China)

【Abstract】 In Chinese Named Entity Recognition(NER) based on neural network,the vectorized representation of words is an important step. Traditional representation methods for word vectors only map a word to a single vector,and cannot represent the polysemy of a word. To address the problem,this paper introduces the BERT pretrained language model to build a BERT-BiGRU-CRF model for representation of sentence characteristics. The BERT pretrained language model with bidirectional Transformer structure is used to enhance the semantic representation of words and generate semantic vectors dynamically based on their context. On this basis,the word vector sequence is input into the BIGR-CRF model to train the whole model,or train the BIGR-CRF part only with BERT fixed. Experimental results on MSRA data show that the F1 value in the two training modes of this proposed model reaches 95.43% and 94.18% respectively,which is better than that of the BIGRU-CRF,the RADICAL-BILSTM-CRF and the GRAIN-LSTM-CRF models.

【Key words】 Chinese Named Entity Recognition(NER); BERT model; BiGRU model; pretrained language model; Conditional Random Field(CRF)

DOI: 10.19678/j.issn.1000-3428.0054272

0 概述

命名实体识别(Named Entity Recognition,NER)技术可用于识别文本中的特定实体信息,如人名、地名、机构名等,在信息抽取、信息检索、智能问答、机器翻译等方面都有广泛应用,是自然语言处理的基础方法之一。一般将命名实体识别任务形式化为序列标注任务,通过预测每个字或者词的标签,联合预测实体边界和实体类型。

随着神经网络的迅速发展,不依赖人工特征的端到端方案逐渐占据主流。文献[1]基于单向长短期记忆(Long-Short Term Memory,LSTM)模型和神经网络进行命名实体识别,提出 LSTM-CRF 模型。基于 LSTM 良好的序列建模能力,LSTM-CRF 成为命名实体识别的基础架构之一,很多方法都是以 LSTM-CRF 为主体框架,在其基础上融入各种相关

基金项目:国家自然科学基金(61672024);国家重点研发计划"智能电网技术与装备"重点专项(2018YFB0904200)。

作者简介:杨 飘(1995—),男,硕士研究生,主研方向为自然语言处理、深度学习;董文永(通信作者),教授、博士。

收稿日期:2019-03-18 修回日期:2019-04-29 E-mail:1724532024@qq.com

特征。例如文献 [2] 加入手工拼写特征，文献 [3-4] 使用一个字符 CNN 来抽取字符特征，文献 [5] 采用的是字符级 LSTM。也有基于 CNN 的命名实体识别方案，例如文献 [6] 提出的 CNN-CRF 结构，文献 [7] 在 CNN-CRF 基础上提出使用字符 CNN 来增强的模型。此后，文献 [8] 采用空洞卷积网络(IDCNN-CRF)进行命名实体识别，在提取序列信息的同时加快了训练速度，文献 [9] 在 BiLSTM-CRF 模型的基础上利用注意力机制获得词在全文范围内的上下文表示，并将该模型应用于化学药物实体识别任务，通过在生物文本上预训练词向量以及使用字符级 LSTM，获得了 90.77% 的 F1 值。文献 [10] 采用 GRU 计算单元，提出了基于双向 GRU 的命名实体识别方法，并将其应用于会议名称识别任务。文献 [11] 将 CNN-BiLSTM-CRF 模型应用于生物医学语料中，获得了较高的 F1 值。文献 [12] 针对裁判文书的实体抽取提出 SVM-BiLSTM-CRF 模型，主要抽取动产、不动产、知识财产 3 类实体。该模型利用 SVM 判断含有关键词的句子，并将其输入 BiLSTM-CRF 模型中进行抽取。文献 [13] 针对在线医疗网站的文本，提出 IndRNN-CRF 和 IDCNN-BiLSTM-CRF 模型，性能均优于经典的 BiLSTM-CRF 模型。

中文存在字和词的区分，因此，在中文领域存在基于字的命名实体识别、基于词的命名实体识别、基于字和词的联合命名实体识别 3 种方案。文献 [14-15] 通过字级别和词级别统计方法的对比，表明基于字符的命名实体识别方法一般有更好的表现。因此，一些研究者在基于神经网络的命名实体识别模型中采用基于字的命名实体识别方案[16-17]。另一些研究人员在字级别的命名实体识别方案中融入了词的信息，例如文献 [18-19] 将分词信息作为 soft feature 来增强识别效果，文献 [20] 则通过将分词和命名实体识别联合训练来融合分词信息。文献 [21] 提出的 Lattice LSTM 网络结构效果较好，其将传统的 LSTM 单元改进为网格 LSTM，在字模型的基础上显性利用词与词序信息，且避免了分词错误传递的问题，在 MSRA 语料上 F1 值达到 93.18%。

以上基于字的中文命名实体识别方法普遍存在的问题是无法表征字的多义性，例如在句子“这两批货物都打折出售，严重折本，他再也经不起这样折腾了”中，3 个“折”字表达的是不同的含义，但是在以往的字向量表示方法中，3 个字的向量表示完全一样，这与客观事实不符。较好的词表示应能包含丰富的句法和语义信息，并且能够对多义词进行建模，针对这个问题，研究人员提出使用预训练语言模型的方法来进行词表示。文献 [22] 使用一个词级别的语言模型来增强 NER 的训练，在大量原始语料上实现多任务学习。文献 [23-24] 采用 BiLSTM 网络结构，通过预训练一个字符语言模型来生成词上下文表示以增强词的表示。文献 [25] 建立的 BERT 模型则采用表义能力更强的双向 Transformer 网络结构来预训练语言模型。

由于 BERT 预训练语言模型具有较强的语义表达能力，因此本文通过嵌入该模型的双向 Transformer 编码结构，构建 BERT-BiGRU-CRF 模型，以提高中文命名实体识别准确率。

1　BERT-BiGRU-CRF 模型

BERT-BiGRU-CRF 模型整体结构如图 1 所示，整个模型分为 3 个部分，首先通过 BERT 预训练语言模型获得输入的语义表示，得到句子中每个字的向量表示，然后将字向量序列输入 BiGRU 中做进一步语义编码，最后通过 CRF 层输出概率最大的标签序列。

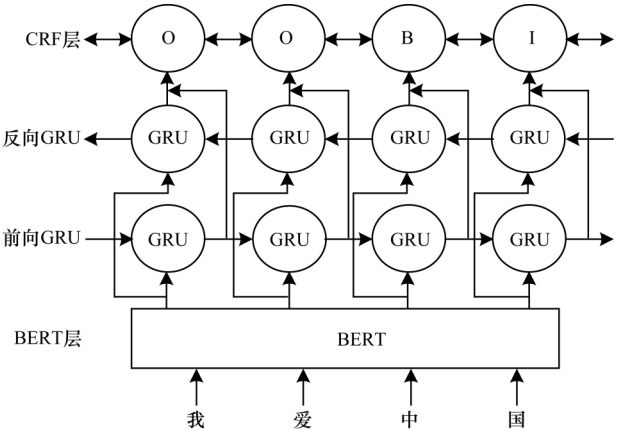

图 1　BERT-BiGRU-CRF 模型结构

Fig. 1　BERT-BiGRU-CRF model structure

与传统的命名实体识别模型相比，本文模型最主要的区别是加入了 BERT 预训练语言模型。BERT 预训练语言模型在大规模语料上学习所得，可以通过上下文计算字的向量表示，能够表征字的多义性，增强了句子的语义表示。该模型有 2 种训练方式：一种是训练整个 BERT-BiGRU-CRF 模型的参数；另一种是固定 BERT 参数，只训练 BiGRU-CRF 部分参数。第 2 种训练方式相对于第 1 种训练方式可以大幅减少训练参数，缩短训练时间。

1.1　BERT 预训练语言模型

近年来，研究人员将预训练深度神经网络作为语言模型，在此基础上以针对垂直任务进行微调的方式取得了很好的效果。比较典型的语言模型是从左到右计算下一个词的概率，如式(1)所示：

$$p(S)=p(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}p(w_i\mid w_1,w_2,\cdots,w_{i-1}) \tag{1}$$

在将预训练模型应用于垂直领域时，有时并不需要语言模型，而是需要一个字的上下文表示以表征字的多义性和句子的句法特征。针对该问题，文献 [22] 提出了 BERT 预训练语言模型，其结构如图 2所示。为融合字左右两侧的上下文，BERT 采用双向 Transformer 作为编码器。该文还提出了 Masked 语言模型和下一个句子预测 2 个任务，分别捕捉词级别和句子级别的表示，并进行联合训练。

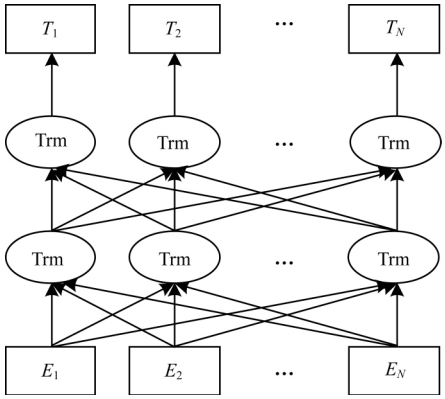

图 2 BERT 预训练语言模型结构

Fig. 2 BERT pretrained language model structure

Masked 语言模型用于训练深度双向语言表示向量，该方法采用一个非常直接的方式，即遮住句子里某些单词，让编码器预测这个单词的原始词汇。文献[22]随机遮住 15% 的单词作为训练样本，其中，80% 用 masked token 来代替，10% 用随机的一个词来替换，10% 保持这个词不变。

下一个句子预测是指预训练一个二分类的模型来学习句子之间的关系。很多 NLP 任务如 QA 和 NLI 都需要对 2 个句子之间关系的理解，而语言模型不能很好地直接产生这种理解。为理解句子关系，该方法同时预训练了一个下一个句子预测任务。具体做法是随机替换一些句子，然后利用上一句进行 IsNext/NotNext 的预测。

BERT 最重要的部分是双向 Transformer 编码结构，Transformer 舍弃了 RNN 的循环式网络结构，完全基于注意力机制对一段文本进行建模。Transformer 编码单元如图 3 所示。

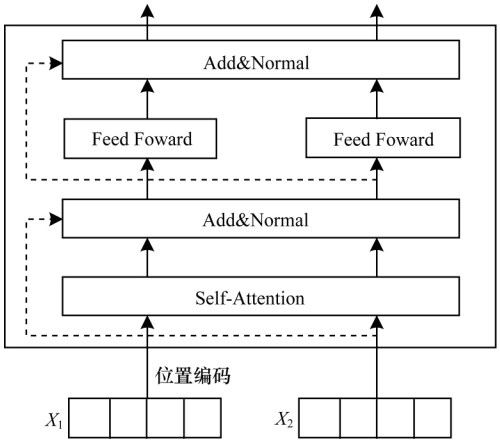

图 3 Transformer 编码单元

Fig. 3 Transformer coding unit

编码单元最主要的模块是自注意力(Self-Attention)部分，如式(2)所示：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 均是输入字向量矩阵，d_k 为输入向量维度。

上述方法的核心思想是计算一句话中的每个词对于这句话中所有词的相互关系，然后认为这些词与词之间的相互关系在一定程度上反映了这句话中不同词之间的关联性以及重要程度。在此基础上，利用这些相互关系来调整每个词的重要性(权重)即可获得每个词新的表达。这个新的表征不但蕴含了该词本身，还蕴含了其他词与这个词的关系，因此，其与单纯的词向量相比是一个更加全局的表达。

为扩展模型专注于不同位置的能力，增大注意力单元的表示子空间，Transformer 采用了“多头”模式，如式(3)和式(4)所示：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_h)\,W^o \tag{3}$$

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}W_i^Q,\boldsymbol{K}W_i^k,\boldsymbol{V}W_i^V) \tag{4}$$

此外，为解决深度学习中的退化问题，Transformer 编码单元中还加入了残差网络和层归一化，如式(5)和式(6)所示：

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

$$\text{FFN} = \max(0, xW_1 + b_1)\,W_2 + b_2 \tag{6}$$

在自然语言处理中一个很重要的特征是时序特征，针对自注意力机制无法抽取时序特征的问题，Transformer 采用了位置嵌入的方式来添加时序信息，如式(7)和式(8)所示。BERT 的输入是词嵌入、位置嵌入、类型嵌入之和。

$$\text{PE}(P_{\text{pos}},2i) = \sin(P_{\text{pos}}/10\,000^{2i/d_{\text{model}}}) \tag{7}$$

$$\text{PE}(P_{\text{pos}},2i+i) = \cos(P_{\text{pos}}/10\,000^{2i/d_{\text{model}}}) \tag{8}$$

与其他语言模型相比，BERT 预训练语言模型可以充分利用词左右两边的信息，获得更好的词分布式表示。

1.2 BiGRU 层

GRU(Gated Recurrent Unit)是一种特殊循环神经网络(Circulatory Neural Network，RNN)。在自然语言处理中，有很多数据前后之间具有关联性，传统前向神经网络无法对这种数据建模，由此出现了循环神经网络。

循环神经网络通过引入定向循环来处理序列化数据，其网络结构分为 3 层，分别为输入层、隐层、输出层。隐层之间可以前后相连，使得当前隐层的信息可以传递到下个节点，作为下个节点输入的一部分，这样使得序列中的节点能够“记忆”前文的信息，达到序列建模的目的。

RNN 神经网络理论上可以处理任意长度的序列信息，但是在实际应用中，当序列过长时会出现梯度消失的问题，且很难学到长期依赖的特征。针对这个问题，文献[26]改进了循环神经网络，提出了 LSTM 模型。LSTM 单元通过输入门、遗忘门和输出门来控制信息传递。

GRU[27]是 RNN 的另一种变体，其将遗忘门和输入门合成为一个单一的更新门，同时混合细胞状

态和隐藏状态。GRU 单元结构如图 4 所示，具体计算过程如式(9) ~式(12)所示。

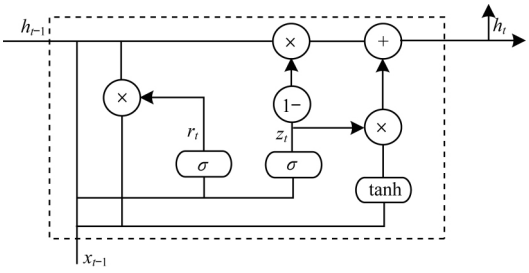

图 4　GRU 编码单元

Fig. 4　GRU coding unit

$$z_t = \sigma(W_i * [h_{t-1}, x_t]) \tag{9}$$

$$r_t = \sigma(W_r * [h_{t-1}, x_t]) \tag{10}$$

$$\tilde{h}_t = \tanh(W_c * [r_t \cdot h_{t-1}, x_t]) \tag{11}$$

$$h_t = (1 - z_t) \cdot c_{t-1} + z_t \cdot \tilde{h}_t \tag{12}$$

其中，σ 是 sigmoid 函数，$\cdot$ 是点积。$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 为时刻 t 的输入向量，$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 是隐藏状态，也是输出向量，包含前面 t 时刻所有有效信息。z_t 是一个更新门，控制信息流入下一个时刻，r_t 是一个重置门，控制信息丢失，z_t 和 r_t 共同决定隐藏状态的输出。

单向的 RNN 只能捕获序列的历史信息，对于序列标注任务而言，一个字的标签和该字的上下文都有关系。为充分利用上下文信息，文献[28]提出了双向循环神经网络(BRNN)，之后文献[29]提出了 BiLSTM 模型，将单向网络结构变为双向网络结构，该模型有效利用上下文信息，在命名实体识别等序列标注任务中得到广泛应用。

GRU 与 LSTM 相比结构更加简单，参数更少，可以缩短训练时间。由于 GRU 良好的序列建模能力，使得 GRU 在语音识别、命名实体识别和词性标注等方面都得到广泛应用。

1.3　CRF 层

GRU 只能考虑长远的上下文信息，不能考虑标签之间的依赖关系，如在命名实体识别中，有些标签不能连续出现，因此，模型不能独立地使用 $h^{(t)}$ 来做标签决策，而 CRF 能通过考虑标签之间的相邻关系获得全局最优标签序列，故使用 CRF 来建模标签序列。

CRF 对于给定序列 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 和对应的标签序列 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$，定义评估分数计算公式如式(13)所示：

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} (W_{y_{i-1}, y_i} + P_{i, y_i}) \tag{13}$$

其中，$W_{i,j}$ 表示标签转移分数，P_{i,y_i} 表示该字符的第 y_i 个标签的分数。P_i 定义如式(14)所示：

$$P_i = \boldsymbol{W} s h^{(t)} + b_s \tag{14}$$

其中，$\boldsymbol{W}$ 是转换矩阵，$h^{(t)}$ 是上一层 t 时刻输入数据 $x^{(t)}$ 的隐藏状态。

对 CRF 的训练采用的是最大条件似然估计，对训练集合$\{(x_i, y_i)\}$，其似然函数如式(15)所示，P 计算如式(16)所示，表示原序列到预测序列对应的概率。

$$L = \sum_{i=1}^{n} \log_a (P(y_i \mid x_i)) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \tag{15}$$

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{e^{s(\boldsymbol{x}, \boldsymbol{y})}}{\sum_{y \in Y_x} e^{s(\boldsymbol{x}, \boldsymbol{y})}} \tag{16}$$

2　实验结果与分析

2.1　实验数据

本文采用 MSRA 数据集，该数据集是微软公开的命名实体识别数据集，包含人名、机构名、地名3 类实体。数据集中包括训练集和测试集，训练集共包含 4.64×10^4 个句子、$2.169\ 9 \times 10^6$ 个字，测试集包括 4.4×10^3 个句子、1.726×10^5 个字。各类实体统计如表 1 所示。

表 1　数据集实体个数

Table 1　Number of entities in data set

数据集	地名	机构名	人名	共计
训练集	36 517	20 571	17 615	74 703
测试集	2 877	1 331	1 973	6 181

2.2　标注策略与评价指标

命名实体识别的标注策略有 BIO 模式、BIOE 模式、BIOES 模式。本文采用的是 BIO 标注策略，其中“B”表示实体开始，“I”表示实体非开始部分，“O”表示非实体的部分。因为在预测实体边界时需要同时预测实体类型，所以待预测的标签共有 7 种，分别为“O”“B-PER”“I-PER”“B-ORG”“I-ORG”“B-LOC”和“I-LOC”。在测试过程中，只有当一个实体的边界和实体的类型完全正确时，才能判断该实体预测正确。

命名实体识别的评价指标有精确率 P、召回率 R 和 F1 值 F_1。具体定义如式(17)所示。其中，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，F_N 为相关实体但是模型没有检测到的个数。

$$P = \frac{T_P}{T_P + F_P} \times 100\%$$

$$R = \frac{T_P}{T_P + F_N} \times 100\%$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{17}$$

2.3　实验环境

实验计算机配置如下：Ubuntu 操作系统，i7-6700HQ@2.60 GHz 的 CPU，Python 3.6，Tensorflow 1.12.0，16 GB 内存。

2.4　实验过程

BERT-BiGRU-CRF 模型有 2 种训练方式：一种是训练模型全部参数；另一种是固定 BERT 部分参数，只更新 BiGRU-CRF 参数。本文使用这两种方式分别进行实验。

为证明模型的有效性，将 BERT-BiGRU-CRF 模型与以下模型进行对比:

1) BiGRU-CRF 模型。该模型是序列标注经典模型，基于字的标注，采用预训练好的字向量，然后输入 BiGRU-CRF 模型中进行训练。

2) Radical-BiLSTM-CRF 模型[17]。该模型在 BiLSTM-CRF 的基础上融入笔画信息，将字的笔画序列输入 BiLSTM 中得到字的表示，然后以字的 Embedding 和笔画表示连接，作为该字新的语义表示输入上层 BiLSTM-CRF 中进行训练。

3) Lattice-LSTM-CRF 模型[21]。该模型在中文语料上达到了较好的效果，Lattice-LSTM 网络结构充分融合了字信息和该字的潜在词信息，可有效避免分词的错误传递。

2.5 参数设置

Google 提供的预训练语言模型分为 2 种: BERT-Base 和BERT-Large。2 种模型网络结构相同，只有部分参数不同。实验中采用的是 BERT-Base。BERT-Base 共 12 层，隐层为 768 维，采用 12 头模式，共 110M 个参数。最大序列长度采用 128，train_batch_size 为 16，learning_rate 为 5e－5，droup_out_rate 为 0.5，clip 为 5，BiGRU 隐藏层维数为 128。

2.6 实验结果

BERT-BiGRU-CRF 模型 F1 值随训练轮数的变化如图 5 所示，其中 BERT-BiGRU-CRF-f 模型表示在训练过程中更新整个模型的参数，BERT-BiGRU-CRF 表示固定 BERT 参数，只更新 BiGRU-CRF 部分参数。BERT-BiGRU-CRF-f 模型在训练 12 个 epoch 时达到最大 F1 值 95.43%; BERT-BiGRU-CRF 模型也是在训练 12 个 epoch 时达到最大 F1 值 94.18%; BiGRU-CRF 模型在第 14 个 epoch 达到最大 F1 值 87.97%。BERT-BiGRU-CRF 训练一轮的时间是 394 s，BiGRU-CRF 训练一轮的时间是 406 s，BERT-BiGRU-CRF-f 训练一轮的时间为2 044 s。另外测得 Lattice-LSTM-CRF 模型训练一轮的时间为 7 506 s，在第 37 个 epoch 才得到最优 F1 值，总体训练时间远超 BERT-BiGRU-CRF 模型。

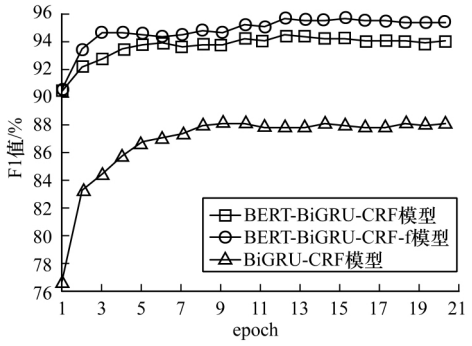

图 5 F1 值变化曲线

Fig. 5 Curve of F1 value changes

针对人名(PER)、地名(LOG)、机构名(ORG) 3 类实体的准确率、召回率、F1 值如表 2 所示，可以看出其中机构类实体预测准确率偏低，主要原因在于机构名中很多存在地名嵌套、缩略词、歧义等干扰信息，在没有其他充足的上下文时容易预测错误。

表 2 不同类型命名实体识别结果

Table 2 NER results for different types of entities %

模型	实体类型	P	R	F_1
BERT-BiGRU-CRF-f	LOC	96.58	95.31	95.94
	ORG	90.46	93.31	91.86
	PER	96.84	97.38	97.11
BERT-BiGRU-CRF	LOC	95.29	94.23	94.75
	ORG	88.31	90.83	89.56
	PER	96.70	96.31	96.51

部分错例如表 3 所示。可以看出: 例句 1 的机构名中嵌套了地名，类似的例子还有“中国政府陪同团”“中国东盟”等; 例句 2 中出现了“工商联”这一缩写，类似的还有“理事会”“委员会”等; 例句 3 中则出现了歧义的情形。这种情况下如果没有补充的上下文会导致难以预测。

表 3 预测错误实例

Table 3 Examples of wrong prediction

编号	句子	实体	预测实体
1	洛杉矶市民议政论坛	洛杉矶市民议政论坛-ORG	洛杉矶-LOC
2	工商联的任务更加繁重了	工商联-ORG	—
3	陆军及皇家空军法律服务处	陆军及皇家空军法律服务处-ORG	皇家空军法律服务处-ORG

本文模型与其他模型的对比如表 4 所示。可以看出，对比 BERT-BiGRU-CRF 模型和 BiGRU-CRF 模型，本文模型能提高 6.21% 的 F1 值，说明 BERT 预训练语言模型能更好地表示字的语义信息，这是因为 BERT 生成的字向量是上下文相关的，例如在句子“罗布汝信汤洪高安启元许其亮阮崇武”中，正确实体划分应该是“罗布|汝信|汤洪高|安启元|许其亮|阮崇武”，表示 6 个名字的并列，但在 BiGRU-CRF 模型中，“安启元”这个实体无法正确识别，而是将“汤洪高安启元”作为一个整体，主要原因是“安”字作为姓氏比较少见，在传统词向量中只能表义“平安”“安定”等，而在 BERT-BiGRU-CRF 模型中，生成的“安”字语义向量是上下文相关的，在该语句的上下文中包含有姓氏的含义，与“民族团结，社会安定”中的“安”字相比，生成的语义向量不同，语义不同。同样的例子还有“普”“亢”作为姓氏的情形。

表 4　不同模型命名实体识别结果

Table 4　NER results of different models　%

模型	P	R	F_1
BiGRU-CRF	88.80	87.16	87.97
Radical-BiLSTM-CRF	91.28	90.62	90.95
Lattice-LSTM-CRF	93.57	92.79	93.18
BERT-BiGRU-CRF-f	95.31	95.54	95.43
BERT-BiGRU-CRF	94.19	94.16	94.18

BERT-BiGRU-CRF 模型与 Radical-BiLSTM-CRF 模型、Lattice-LSTM 模型相比效果更好，说明 BERT 的特征抽取能力比较强，抽取的特征比单独训练笔画特征和字词融合特征更准确。对比 BERT-BiGRU-CRF-f 和 BERT-BiGRU-CRF 模型可以看出，BERT-BiGRU-f 效果更好，但训练参数量更大，所需要的训练时间更长。

3　结束语

针对传统词向量表示方法无法表征字多义性的问题，本文构建 BERT-BiGRU-CRF 模型，通过 BERT 预训练语言模型双向 Transformer 结构动态生成字的上下文语义表示。该模型性能优于 Lattice-CRF 模型，可有效提升中文命名实体识别的效果，但其缺点是当上下文信息不足且存在实体嵌套、缩写、歧义实体等情形时，无法实现对语句特征的正确抽取。下一步将在本文模型中融入潜在词特征，结合 BERT 与 Lattice LSTM 表征字的多义性，同时加入潜在词的特征，以应对上下文信息不足的情况。

参考文献

[1] HAMMERTON J. Named entity recognition with long short-term memory [C]//Proceedings of Conference on Natural Language Learning at HLT-NAACL. Edmonton, Canada: Association for Computational Linguistics, 2003: 1-4.

[2] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1508.01991v1.pdf.

[3] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1603.01354.pdf.

[4] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[5] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1603.01360.pdf.

[6] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[7] SANTOS C N D, VICTOR G. Boosting named entity recognition with neural character embeddings [EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1505.05008.pdf.

[8] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: [s.n.], 2017: 1-5.

[9] YANG Pei, YANG Zhihao, LUO Ling, et al. An attention-based approach for chemical compound and drug named entity recognition [J]. Journal of Computer Research and Development, 2018, 55(7): 1548-1556. (in Chinese)
杨培，杨志豪，罗凌，等. 基于注意机制的化学药物命名实体识别 [J]. 计算机研究与发展，2018，55(7)：1548-1556.

[10] WANG Jie, ZHANG Ruidong, WU Chensheng. Named entity recognition method based on GRU [J]. Computer Systems and Applications, 2018, 27(9): 18-24. (in Chinese)
王洁，张瑞东，吴晨生. 基于 GRU 的命名实体识别方法 [J]. 计算机系统应用，2018，27(9)：18-24.

[11] LI Lishuang, GUO Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF [J]. Journal of Chinese Information Processing, 2018, 32(1): 116-122. (in Chinese)
李丽双，郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别 [J]. 中文信息学报，2018，32(1)：116-122.

[12] ZHOU Xiaolei, ZHAO Xuejiao, LIU Liangtang, et al. Named entity recognition method of judgment documents with SVM-BiLSTM-CRF [J]. Computer Systems and Applications, 2019, 28(1): 245-250. (in Chinese)
周晓磊，赵薛蛟，刘堂亮，等. 基于 SVM-BiLSTM-CRF 模型的财产纠纷命名实体识别方法 [J]. 计算机系统应用，2019，28(1)：245-250.

[13] YANG Wenming, CHU Weijie. Named entity recognition of online medical question answering text [J]. Computer Systems and Applications, 2019, 28(2): 8-14. (in Chinese)
杨文明，褚伟杰. 在线医疗问答文本的命名实体识别 [J]. 计算机系统应用，2019，28(2)：8-14.

[14] LIU Zhangxun, ZHU Conghui, ZHAO Tiejun. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? [C]//Proceedings of International Conference on Intelligent Computing. Berlin, Germany: Springer, 2010: 634-640.

[15] LI H, HAGIWARA M, LI Q, et al. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese [C]//Proceedings of the 9th International Conference on Language Resources and Evaluation. [S.l.]: ELRA, 2014: 2532-2536.

[16] LU Yanan, ZHANG Yue, JI Donghong. Multi-prototype Chinese character embedding [C]//Proceedings of the 10th International Conference on Language Resources and Evaluation. [S.l.]: ELRA, 2016: 855-859.

[17] DONG Chuanhai, ZHANG Jiajun, ZONG Chengqing, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [C]//Proceedings of International Conference on Computer Processing of Oriental Languages. Berlin, Germany: Springer, 2016: 239-250.

[18] PENG N, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: [s.n.], 2015: 548-554.

（下转第 52 页）

蒋李鸣,吕佳宇,何哲华,等. 基于启发式的约束满足问题 AC 系列算法改进研究[J]. 软件工程,2018,21(2):30-34.

[8] WOODWARD R J,KARAKASHIAN S,CHOUEIRY B Y,et al. Revisiting neighbor-hood inverse consistency on binary CSPs[C]//Proceedings of International Conference on Principles and Practice of Constraint Programming. Berlin,Germany:Springer,2012:688-703.

[9] PALMIERI A,REGIN J C,SCHAUS P. Parallel strategies selection[C]//Proceedings of International Conference on Principles and Practice of Constraint Programming. Berlin,Germany:Springer,2016:388-404.

[10] LI Hongbo,LI Zhanshan,WANG Tao. Improving coarse-grained arc consistency algorithms in solving constraint catisfaction problems[J]. Journal of Software,2012,23(7):1816-1823.(in Chinese)
李宏博,李占山,王涛. 改进求解约束满足问题粗粒度弧相容算法[J]. 软件学报,2012,23(7):1816-1823.

[11] ZHANG Boyang,ZHU Yanguang,YANG Feng. Research on solution space contraction method for temporal constraint optimization problem[J]. Computer Engineering,2012,38(14):262-265.(in Chinese)
张博洋,朱延广,杨峰. 时间约束优化问题的解空间压缩方法研究[J]. 计算机工程,2012,38(14):262-265.

[12] LIU Junli,OUYANG Song. Improvement of workflow adaptability based on constraints theory[J]. Computer Engineering,2010,36(11):90-92.(in Chinese)
刘俊莉,欧阳松. 基于约束理论的工作流适应性改进[J]. 计算机工程,2010,36(11):90-92.

[13] YAPR H C,ZHANG Y. An optimal coarse-grained arc consistency algorithm[J]. Artificial Intelligence,2006,165(2):165-185.

[14] WANG R,XIA W,YAP R,et al. Optimizing simple table reduction with bitwise representation[C]//Proceedings of the 25th IEEE International Joint Conference on Artificial Intelligence. Washington D.C.,USA:IEEE Press,2016:1-9.

[15] DASYGENISM,STERGIOU K. Using parallelization to efficiently exploit the pruning power of strong local consistencies[C]//Proceedings of Hellenic Conference on Artificial Intelligence. Washington D.C.,USA:IEEE Press,2016:231-242.

[16] VERHAEGHE H,LECOUTRE C,DEVILLE Y,et al. Extending compact-table to basic smart tables[C]//Proceedings of International Conference on Principles and Practice of Constraint Programming. Berlin,Germany:Springer,2017:152-164.

[17] LI Hongbo,SHEN Haijiao,LI Zhanshan,et al. Reducing consistency checks in generating corrective explanations for interactive constraint satisfaction[J]. Knowledge-Based Systems,2013,43(2):103-111.

[18] YANG Mingqi,LI Zhanshan,ZHANG Jiachen. Simple tabular reduction algorithm based on time-stamp mechanism[J]. Journal of Software,2019,30(11):3355-3363.(in Chinese)
杨明奇,李占山,张家晨. 一种基于时间戳的简单表缩减弧相容算法[J]. 软件学报,2019,30(11):3355-3363.

[19] YANG Wei. Research on constraint propagation strategies and heuristics in CSP[D]. Changchun:Jilin University,2018.(in Chinese)
杨微. CSP 中的约束传播策略及启发式的研究[D]. 长春:吉林大学,2018.

[20] WOODWARDR J,CCHOUEIRY B Y. Weight-based variable ordering in the context of high-level consistencies[EB/OL]. [2019-03-21]. https://www.researchgate.net/publication/.

编辑 索书志

(上接第 45 页)

[19] HE Hangfeng,SUN Xu. F-score driven max margin neural network for named entity recognition in Chinese social media[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.]:ACL 2016:1-6.

[20] XU Yan,WANG Yining,LIU Tianren,et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries[J]. Journal of the American Medical Informatics Association,2013,21(e1):84-92.

[21] ZHANG Yue,YANG Jie. Chinese NER using lattice LSTM[EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1805.02023.pdf.

[22] REI M. Semi-supervised multitask learning for sequence labeling[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. [S.l.]:ACL,2017:2121-2130.

[23] PETERS M E,AMMAR W,BHAGAVATULA C,et al. Semi-supervised sequence tagging with bidirectional language models[EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1705.00108.pdf.

[24] PETERS M E,NEUMANN M,IYYER M,et al. Deep contextualized word representations[EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1810.04805.pdf.

[25] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2019-01-02]. https://arxiv.org/pdf/1810.04805.pdf.

[26] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]. Berlin,Germany:Springer,2008.

[27] CHO K,MERRIENBOER B,GULCEHRE C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing. Doha,Qatar:[s.n.],2014:1-15.

[28] SCHUSTER M,PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing,1997,45(11):2673-2681.

[29] GRAVES A,SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5):602-610.

编辑 金胡考