计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：　基于 BERT 嵌入的中文命名实体识别方法
作者：　杨飘，董文永
DOI：　10.19678/j.issn.1000-3428.0054272
网络首发日期：　2019-05-30
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/OL]．计算机工程．https://doi.org/10.19678/j.issn.1000-3428.0054272

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：命名实体识别是自然语言处理的基础之一。在基于神经网络的中文命名实体识别方法中，字的向量化表示是重要一步，传统的词向量表示将字映射为单一向量，这种方法无法表征字的多义性。针对这个问题，提出了基于 BERT 嵌入的中文命名实体识别方法。该方法通过 BERT(Bidirectional Encoder Representations from Transformers)预训练语言模型增强字的语义表示，根据字的上下文动态生成语义向量，然后再将字向量序列输入 BiGRU-CRF 中进行训练，训练时可以训练整个 BERT-BiGRU-CRF 模型，也可以固定 BERT，只训练 BiGRU-CRF 部分。实验表明，该模型的两种训练方法在 MSRA 语料上分别达到 95.43%F1 值和 94.18% F1 值，均优于目前最优的 Lattice-LSTM 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

开放科学标识码：

Chinese NER based on BERT Embedding

Yang Piao, Dong Wenyong

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

【Abstract】Named entity recognition (NER) is one of the foundations of natural language processing(NLP). In the method of Chinese named entity recognition based on neural network, the vector representation of words is an important step. Traditional word embedding method map words or chars into a single vector, which can not represent the polysemy of words. To solve this problem, a named entity recognition method based on BERT Embedding model is proposed. The method enhances the semantic representation of words by BERT(Bidirectional Encoder Representations from Transformers) pre-trained language model. BERT can generates the semantic vectors dynamically according to the context of the words, and then inputs the word vectors into BiGRU-CRF for training. The whole model can be trained during training. It is also possible to fix the BERT and train only the BiGRU-CRF part. Experiments show that the two training methods of the model reach 95.43% F1 and 94.18% F1 in MSRA corpus, respectively, which are better than the current optimal Lattice-LSTM model.

【Key words】Chinese NER； BERT； BiGRU； Pre-trained Language Model； CRF

DOI:10.19678/j.issn.1000-3428.0054272.

0 概述

命名实体识别(NER)旨在识别文本中特定实体信息，如人名、地名、机构名等，在信息抽取，信息检索，智能问答，机器翻译中都有广泛应用，是自然语言处理的基础之一。一般将命名实体识别任务形式化序列标注任务，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。

随着神经网络的飞速发展，不依赖人工特征的端到端方案越来越占据主流。在英文领域，第一个采用神经网络进行命名实体识别的是 Hammerton 等人[1]，使用的网络结构是单向 LSTM。由于 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架构之一，很多方法都是以 LSTM-CRF 为主体框架，在此之上融入各种相关特征，Huang 等人[2]加入的是手工拼写特征；Ma 和 Hovy[3]以及 Chiu 等人[4]则使用了一个字符 CNN 来抽取字符特征；Lample 等人[5]采用的是字符级 LSTM。也有基于 CNN 的命名实体识别方案，2011 年 Collobert 等人[6]提出了一个 CNN-CRF 的结构，获得了不错的结果。在此之上，2015 年 Santos 等人[7]提出了使用字符 CNN 来增强 CNN-CRF 模型。2017 年，Strubell 等人[8]提出采用空洞卷积网络(IDCNN-CRF)进行命名实体识别，提取序列信息的同时加快

基金项目：国家自然科学基金资助项目（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，湖北天门人，硕士研究生，主要研究方向为自然语言处理、深度学习；董文永，男（汉族）（通信作者），教授，博士，主要研究方向为人工智能、机器学习、智能计算 **E-mail：1724532024@qq.com**

训练速度。2018 年，买买提阿依甫等人[9]根据维吾尔语的特点，提出了 BiLSTM-CNN-CRF 模型。杨培等人在 BiLSTM-CRF 模型的基础上引入了注意力机制，注意力机制可以获得词在全文范围下的上下文表示，该模型应用在化学药物实体识别的任务上，通过在生物文本上预训练词向量以及字符级 LSTM，最终获得了 90.77%的 F1 值。王洁等人[11]采用 GRU 计算单元，提出了基于双向 GRU 的命名实体识别方法，应用在会议名称识别的任务上。李丽双等人[12]将 CNN-BiLSTM-CRF 模型应用在生物医学语料上，获得了当时最高 F1 值。周晓磊等人[13]则针对裁判文书的实体抽取提出了 SVM-BiLSTM-CRF 模型，主要抽取动产，不动产，知识财产三类实体，该模型首先利用 SVM 判断含有关键字的句子，然后再将句子输入 BiLSTM-CRF 模型中进行抽取。杨文明等人[14]针对在线医疗网站的文本，提出了 IndRNN-CRF 和 IDCNN-BiLSTM-CRF 两个模型，实验表明这两个模型在该数据集均优于经典的 BiLSTM-CRF 模型。

中文存在字和词的区分，所以在中文领域存在基于字的的命名实体识别，基于词的命名实体识别，基于字和词的联合命名实体识别三种方案。He and Wang[15]、Liu[16]、 Li 等人[17]进行了字级别和词级别的统计方法对比，研究表明基于字符的命名实体识别方法一般有更好的表现。所以在基于神经网络的中命名实体识别模型中，一些研究人员采用基于字的命名实体识别方案，如 Chen[18]、Lu[19]、 Dong 等人[20]均采用基于字的方案；另外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，其中 Zhao[21]、Peng[22]、He 等人[23]将分词信息作为 soft feature 来增强识别效果；Xu 等人[24]则通过将分词和命名实体识别联合训练来融合分词信息。其中效果最好的是 Zhang 等人[25]提出的 Lattice LSTM 网络结构，该方法将传统的 LSTM 单元改进为网格 LSTM，在字模型的基础之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 语料上达到了 93.18%F1 值。

以上基于字的中文命名实体识别方法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，三个“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，三个字的向量表示完全一样，这与客观事实不符。更好的词表示应该能够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，研究人员提出了使用预训练语言模型的方法来进行词表示。Rei[26]使用了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。Peters[27-28]预训练了一个字符语言模型来生成词上下文表示来增强词的表示，采用的是 BiLSTM 网络结构；2018 年 Devlin 等人[29]提出的 BERT(Bidirectional Encoder Representations from Transformers)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。

鉴于 BERT 预训练语言模型强大的表义能力，本文在中文命名实体识别任务上引入 BERT 预训练语言模型，在此基础上，提出了 BERT-BiGRU-CRF 网络结构，该模型在训练过程中可以训练整个模型的参数，也可以固定 BERT 参数，只训练 BiGRU 部分参数，其中第二种训练方式相对于第一种训练方式能够减少训练参数，缩短训练时间。实验表明，该模型在 MSRA 中文语料上可以获得较好的效果，并且超过了此前效果最好的 Lattice LSTM 模型，训练整个模型时，F1 值达到 95.43%，比 LatticeLSTM 模型高出 2.25%；只训练 BiGRU 部分时，F1 值为 94.18%，比 LatticeLSTM 模型高出 0.96%。

1 BERT-BiGRU-CRF 模型

BERT-BiGRU-CRF 模型整体结构如图 1，整个模型分为三个部分，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之后，再将字向量序列输入 BiGRU 之中进行进一步语义编码，最后通过 CRF 层输出概率最大标签序列。

与传统的命名实体识别模型相比，该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；该模型有两种训练方式，第一种方式是训练整个 BERT-BiGRU-CRF 模型的参数；第二种方式是固定 BERT 参数，只训练 BiGRU-CRF 部分参数。第二种训练方式相对于第一种训练方式可以大幅度减少训练参数，缩短训练时间。

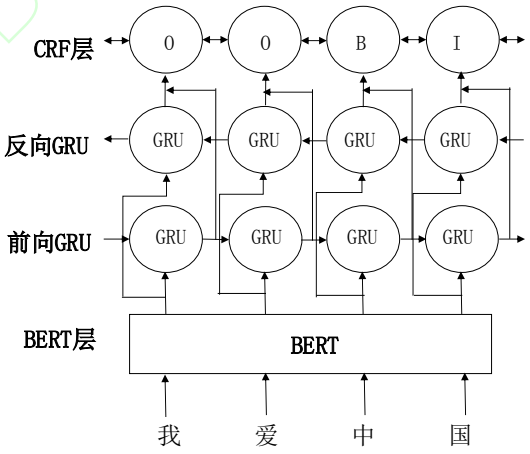

图 1 基于 BERT-BiGRU-CRF 的中文命名实体识别模型

1.1 BERT 预训练语言模型

近几年来，研究人员通过预训练深度神经网络作为语言模型，然后在此基础上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，如公式(1)。但是很多时候在将预训练模型应用到垂直领域的时候，并不需要语言模型，而是需要一个字的上下文表示，能够表征字的多义性，句子的句法特征等。针对这个问题，在 2018 年 Devlin 等人[26]提出了 BERT 预训练语言模型。

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

BERT 模型的全称是 Bidirectional Encoder Representations

from Transformers，结构如图 2，为了融合字左右两侧的上下文，BERT 采用双向 Transformer 作为编码器；该模型还创新性的提出了“Masked 语言模型”和“下一个句子预测”两个任务，分别捕捉词级别和句子级别的表示，并进行联合训练。

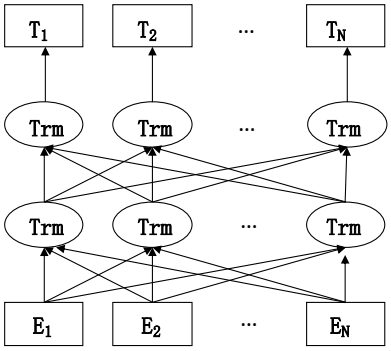

图 2 BERT 预训练语言模型

“Masked 语言模型”是为了训练深度双向语言表示向量，该方法采用了一个非常直接的方式，遮住句子里某些单词，让编码器预测这个单词的原始词汇。作者随机遮住 15%的单词作为训练样本。

（1）其中 80%用 masked token 来代替。

（2）10%用随机的一个词来替换。

（3）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 任务比如 QA 和 NLI 都需要对两个句子之间关系的理解，而语言模型不能很好的直接产生这种理解。为了理解句子关系，该方法同时预训练了一个“下一个句子预测”的任务。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来对一段文本进行建模。Transformer 编码单元如图 3：

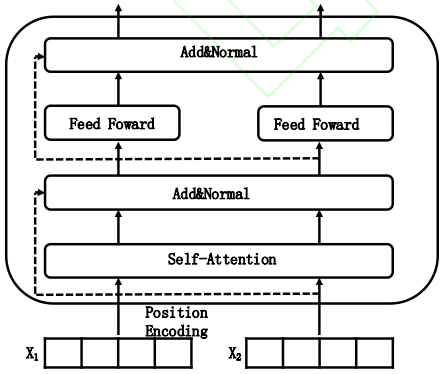

图 3 Transformer 编码单元

编码单元最主要模块的是自注意力部分，如公式(2)，其中 Q，K，V 均是输入字向量矩阵，dk 为输入向量维度。其核心思想是去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应了这句话中不同词之间的关联性以及重要程度。因此再利用这些相互关系来调整每个词的重要性（权重）就可以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，因此和单纯的词向量相比是一个更加全局的表达。

$$Attention(Q,K,V)=soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

为了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，如公式(3)(4)：

$$MultiHead(Q,K,V)=Concat(head_1,...,head_h)W^o \tag{3}$$

$$head_i=Attention(QW_i^Q,KW_i^k,VW_i^V) \tag{4}$$

此外，为了解决深度学习中的退化问题，Transformer 编码单元中加入了残差网络和层归一化，如公式：

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

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

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，Transforme 采用了位置嵌入的方式来添加时序信息，如公式(7)(8)所示。BERT 的输入是词嵌入，位置嵌入，类型嵌入之和。

$$PE(pos,2i)=\sin(pos/10000^{2i/d_{model}}) \tag{7}$$

$$PE(pos,2i+i)=\cos(pos/10000^{2i/d_{model}}) \tag{8}$$

BERT 预训练语言模型与其它语言模型相比，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(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，因此，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，其网络结构分为三层，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，在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，Graves 等人[30]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 是 RNN 的另一种变体，2014 年由 Cho 等人[31]提出，GRU 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，其单元结构如图 4 所示：

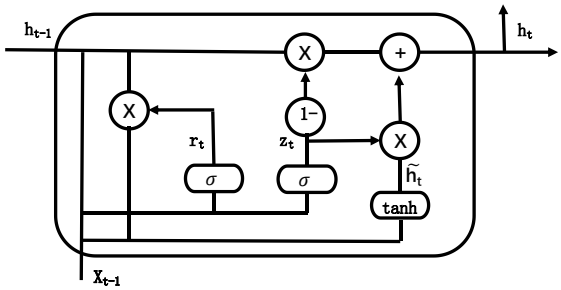

图 4　GRU 编码单元

具体计算过程如公式(9)(10)(11)(12)所示：

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

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

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

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

其中 σ 是 sigmoid 函数，$\bullet$ 是点积。x_t 为时刻 t 的输入向量，h_t 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；二者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。为了充分利用上下文信息，Schuster 等人[32]提出了双向循环神经网络(BRNN)，之后 Graves 等人[33]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，可以加快训练时间。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，词性标注等方面都有广泛应用。

1.3　CRF 层

GRU 只能能够考虑长远的上下文信息，不能考虑标签之间的依赖关系。比如在命名实体识别中，有些标签不能连续出现，因此，模型不能独立的使用 $h^{(t)}$来做标签决策，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，因此使用 CRF 来建模标签序列。

CRF 对于给定序列 $x=(x_1,x_2,x_3...x_n)$ 和对应的标签序列 $y=(y_1,y_2,y_3...y_n)$，定义评估分数为公式(13)：

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

其中 W 是转换矩阵，$W_{i,j}$ 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 定义如公式(14)：

$$P_i=W_sh^{(t)}+b_s \tag{14}$$

其中，$h^{(t)}$ 是上一层 t 时刻输入数据 $x^{(t)}$ 的隐藏状态，参数分别为权值矩阵和参数。

对 CRF 的训练采用的是最大条件似然估计，对训练集合 $\{(x_i,y_i)\}$，其似然函数如公式(15)：

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

其中 P 如公式（16）：

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

表示序列原序列到预测序列对应的概率。

2　实验及结果分析

2.1　实验数据

本文采用 MSRA 数据集，MSRA 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，地名三类实体。该数据集包括训练集合测试集，其中训练集共包含 46.4k 个句子，2169.9k 个字；测试集包括 4.4k 个句子，172.6k 个字。各类实体统计如表 1。

表 1 实体个数统计

	地名	机构名	人名	共计
训练集	36517	20571	17615	74703
测试集	2877	1331	1973	6181

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 表示不是实体的部分。在预测实体边界的时候需要同时预测实体类型，所以待预测的标签一共有七种，分别是 "O", "B-PER", "I-PER", "B-ORG", "I-ORG", "B-LOC", "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 F 值。具体定义如公式(17)：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.

$$\begin{aligned}P&=\frac{T_p}{T_p+F_P}\times100\%\\R&=\frac{T_p}{T_p+F_n}\times100\%\\F1&=\frac{2PR}{P+R}\times100\%\end{aligned} \tag{17}$$

2.3　实验环境

所有实验采用的环境如表 2：

表 2 实验环境

操作系统	Ubuntu
CPU	i7-6700HQ@2.60GHz
GPU	GTX 1070 (8 GB)
Python	3.6
Tensorflow	1.12.0
内存	16G

2.4　实验过程

BERT-BiGRU-CRF 模型有两种训练方式，一种是训练模

型全部参数；另外一种方法是固定 BERT 部分参数，只更新 BiGRU-CRF 参数，实验过程中使用这两种方式分别进行实验。

为了证明模型的有效性，和以下模型做对比：

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，该模型是 Dong 等人[20]在 2016 年提出的中文命名实体模型。该模型是在 BiLSTM-CRF 的基础之上融入了笔画信息。将字的笔画序列输入 BiLSTM 中得到字的表示，然后将字的 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，该模型由 Zhang 等人[25]在 2018 年提出，在此之前,Lattice-LSTM-CRF 模型在中文语料上达到了最好的效果。Lattice-LSTM 网络结构充分融合了字信息和该字的潜在词信息，有效避免了分词的错误传递。

2.5 参数设置

Google 提供的预训练语言模型分为两种：BERT-Base 和 BERT-Large。两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 一共 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 模型随着训练轮数 F1 值变化如图 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，最大 F1 值为 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 394s，BiGRU-CRF 训练一轮的时间是 406s，BERT-BiGRU-CRF-f 训练一轮的时间为 2044s。另外测得 Lattice-LSTM-CRF 模型训练一轮的时间为 7506s，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。

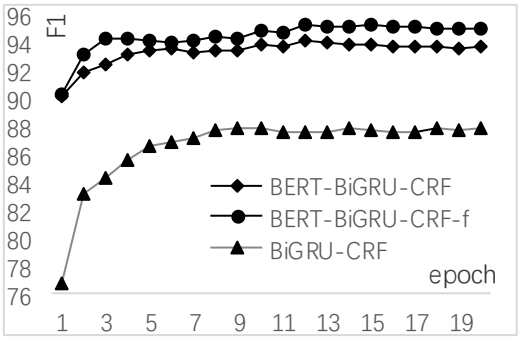

图 5 F1 值变化图

人名，地名，机构名三类实体准确率，召回率，F1 值如表 3 所示：

表 3 不同类型命名实体识别结果

Models	Type	P	R	F1
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

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；在没有其它充足的上下文时容易预测错误。部分错例如表 4，在例句 1 中，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；在例句 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 4 预测错误实例

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

与其它相关工作对比如表 5 所示：

表 5 不同模型命名实体识别结果

Models	P	R	F1
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**

对比 BERT-BiGRU-CRF 模型和 BiGRU-CRF 模型，BERT 预训练语言模型相对于传统的词向量表示能提高 6.21%F1 值，说明 BERT 预训练语言模型能更好的表示字的语义信息。BERT 表义效果更好是因为 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”，表示六个名字的并列，但是在 BiGRU-CRF 模型中，“安启元”这个实体无法正确识别，它将“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。

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，字词融合特征要好。

对比 BERT-BiGRU-CRF-f 和 BERT-BiGRU-CRF 模型，BERT-BiGRU-f 效果更好，但是训练参数量更大，所需要的训练时间更长。

3 结束语

针对传统词向量表示无法表征字的多义性问题，提出了 BERT-BiGRU-CRF 模型。BERT 预训练语言模型通过双向 Transformer 结构动态生成字的上下文语义表示，比传统的词向量表示更能表征语句特征，该模型优于目前最优的 Lattice-CRF 模型，提升了中文命名实体识别的效果。

该模型存在的问题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，无法正确抽取，还有待进一步研究。下一步的研究方向是在模型中显性融入潜在词特征，将 BERT 与 Lattice LSTM 相结合，表征字的多义性同时加入潜在词的特征，当上下文信息不足时，潜在词特征可以起到主要作用。

4 参考文献

[1] Hammerton J . Named Entity Recognition with Long Short-Term Memory[C]// Conference on Natural Language Learning at Hlt-naacl. Association for Computational Linguistics, 2003.

[2] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[3] Ma X , Hovy E . End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.

[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[J]. arXiv preprint arXiv:1603.01360, 2016.

[6] Collobert R , Weston J , Bottou L , et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011.

[7] Santos C N D , Guimarães, Victor. Boosting Named Entity Recognition with Neural Character Embeddings[J]. Computer Science, 2015.

[8] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.

[9] 买买提阿依甫，吾守尔·斯拉木，帕丽旦·木合塔尔，等。基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别[J].计算机工程，2018,44(8):230-236.

[10] 杨培,杨志豪,罗凌,林鸿飞,王健.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(07):1548-1556.

[11] 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(09):18-24.

[12] 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.

[13] 周晓磊，赵薛蛟，刘堂亮，宗子潇，王其乐，里剑桥．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(01):245-250.

[14] 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(02):8-14.

[15] He J, Wang H. Chinese named entity recognition and word segmentation based on character[C]//Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

[16] Liu Z , Zhu C , Zhao T . Chinese Named Entity Recognition with a Sequence Labeling Approach: Based on Characters, or Based on Words?[M]// Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Springer Berlin Heidelberg, 2010.

[17] Li H, Hagiwara M, Li Q, et al. Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese[C]//LREC. 2014: 2532-2536.

[18] Chen W, Zhang Y, Isahara H. Chinese named entity recognition with conditional random fields[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 118-121.

[19] YananLu, YueZhang, andDong-HongJi.2016. Multi-prototype Chinese character embedding. In LREC.

[20] Dong C, Zhang J, Zong C, et al. Character-based lstm-crf with radical-level features for chinese named entity recognition[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 239-250.

[21] Zhao H, Kit C. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition[C]//Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

[22] Peng N, Dredze M. Named entity recognition for chinese social media with jointly trained embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 548-554.

[23] He H, Sun X. F-score driven max margin neural network for named entity recognition in chinese social media[J]. arXiv preprint arXiv:1611.04234, 2016.

[24] Xu Y, Wang Y, Liu T, 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): e84-e92.

[25] Zhang Y, Yang J. Chinese NER Using Lattice LSTM[J]. arXiv preprint

arXiv:1805.02023, 2018.

[26] Rei M. Semi-supervised multitask learning for sequence labeling[J]. arXiv preprint arXiv:1704.07156, 2017.

[27] Peters M E, Ammar W, Bhagavatula C, et al. Semi-supervised sequence tagging with bidirectional language models[J]. arXiv preprint arXiv:1705.00108, 2017.

[28] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[29] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[30] Graves A . Supervised Sequence Labelling with Recurrent Neural Networks[J]. Studies in Computational Intelligence, 2008, 385.

[31] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[32] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. 1997, 45(11):2673-2681.

[33] Graves A , Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Netw, 2005, 18(5):602-610.