Computer Science and Application 计算机科学与应用, 2022, 12(11), 2512-2525
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211257

Hans汉斯

基于BERT的中文计算机实体识别

王君仙，武国宁*

中国石油大学(北京)理学院，北京

收稿日期：2022年10月17日；录用日期：2022年11月15日；发布日期：2022年11月22日

摘　要

针对中文命名实体识别过程中由于中、英文数字混合导致的文本特征学习不彻底、实体识别边界模糊、对不断涌现新的实体识别不准确等问题，本文提出了一种关于字典的实体识别方法。首先，通过字典进行数据预处理，以减少中、英文和数字符号混合对实体识别的影响，再使用BERT模型获取文本特征，将得到的特征作为双向长短时记忆网络的输入进行训练，然后，利用随机条件场进行解码，得到标注序列，最终获取得到相应实体。该模型在人民日报语料、MSRA语料和计算机领域语料上分别取得了95.10%、95.09%和99.45%的F1值，实验结果表明，本文方法能够有效提升命名实体识别效果。

关键词

计算机实体，BERT，序列标注，双向长短时记忆网络，随机条件场

Chinese Computer Entity Recognition Based On BERT

Junxian Wang, Guoning Wu*

School of Science, China University of Petroleum (Beijing), Beijing

Received: Oct. 17th, 2022; accepted: Nov. 15th, 2022; published: Nov. 22nd, 2022

Abstract

Focused on the problems of incomplete text feature learning, the fuzzy boundary of entity recognition and the inaccurate recognition of emerging new entities caused by the mixing of Chinese and English numbers in the process of Chinese-named entity recognition, this paper proposes a method based on the dictionary entity recognition method. Firstly, the data is preprocessed through the dictionary to reduce the impact of the mixing of Chinese, English and digital symbols on entity recognition; secondly, the BERT model is used for data preprocessing to obtain text features, and use the

*通讯作者。

文章引用: 王君仙, 武国宁. 基于 BERT 的中文计算机实体识别[J]. 计算机科学与应用, 2022, 12(11): 2512-2525.
DOI: 10.12677/csa.2022.1211257

features as the input of bidirectional long short-term memory for training; thirdly, the conditional random field is used to decode, and the annotated sequence is obtained. Finally, the corresponding entity is obtained. The model achieved F1-score values of 95.10%, 95.09% and 99.45% on the People's Daily data set, MSRA data set and computer field data set respectively. The experimental results show that the method in this paper can effectively improve the effect of named entity recognition.

Keywords

Computer Entity, Bidirectional Encoder Representation from Transformers (BERT), Sequence Labeling, Bidirectional Long Short-Term Memory, Conditional Random Field

1. 引言

命名实体识别(Named Entity Recognition, NER)是自然语言处理(Natural Language Processing, NLP)中的一项基础任务，主要目的是从非结构化文本中识别出固有实体。NER 是文本分类、情感分析、自动摘要和机器翻译等功能实现的基础，随着计算机技术的发展和大数据时代的来临，越来越多的研究者投入到从海量的文本信息中提取有用的实体知识，并挖掘出实体之间存在的潜在价值的问题中来。

针对命名实体识别中由于中英文数字混合导致的文本特征学习不彻底、实体识别边界模糊、对不断涌现新的实体识别不准确等问题，本文提出了一种预处理方法，该方法将中、英文和数字混合通过一种“字典”进行编码，将待识别的文本通过该字典转换为全中文文本。这种预处理方法增加了字典的冗余性。

本文提出方法在人民日报语料、MSRA 语料和计算机领域语料上进行实验，实验结果充分验证了本文模型实体识别的有效性。

2. 相关工作

第六届 MUC 会议(The Sixth Message Understanding Conferences, MUC-6)首次提出命名实体概念，但并未对命名实体进行明确的定义，只是简要说明需要标注的实体是“实体唯一标识符(Unique Identifiers of Entities)”[1] [2]。CoNLL-2002、CoNLL-2003 会议将对命名实体识别重新定义为句子中所包含的短语，主要包括人名、地名、机构名和时间等特定领域专有名词，大体沿用了 MUC 会议的定义和分类规则[3]。

随着命名实体识别的不断发展，实体的类型和数量也在不断地进行扩充。Alfonseca 等[4]从本体构建组成角度出发，将无监督方法应用于不同语言的命名实体，用来自动扩充具有特定领域知识的实体。并且提出命名实体定义：对于与问题相关的实体对象都可以被称为命名实体。Sekine 等[5]研究发现对于信息提取、问答系统、摘要和信息检索等方面的应用，MUC 会议提出的 7 种命名实体类别不能满足实际需要。某些特定领域需要细分，故提出了一种命名实体层次结构，该结构包含大约 150 种实体类型，并在后续又对实体类别又进行扩充[6]。Marrero 等[7]从理论和实践的角度仔细分析了命名实体领域的演变和发展历程，发现采取适当的措施和标准方法仔细划分命名实体，有助于提升 NER 解决问题的能力。关于

命名实体，目前也没有一个较为官方的、得到普遍认可的定义。但 NLP 问题的研究特点以实用性为首位，纵观整个 NER 研究历史，所谓命名实体识别实质上就是识别无序文本中的人名、地名、机构名和特定领域的专有名词。

早期识别方法主要是基于规则的方法。此类方法应用时间集中在 MUC-6 会议前后，主要识别三个实体：人名、地名和机构名。基于规则方法运用的原理主要是基于字符规则[8]和短语规则[9]。具体说来，为根据字符前后信息、实体前后提示词以及前后文语境得到匹配规则，再根据规则进行实体识别。MUC-6 结果显示，对于人名实体的识别效果明显高于其他实体，分类准确率均高于 91%，且 F1 值更是高达 96.42% [10]。基于规则的方法主要通过制定有限的规则和字典，然后从文本找寻含有这些规则的字符或字典中存在的字符串，将其识别出来并标注为各类别实体。但就解决 NER 问题而言，由于无法制定出所有规则或穷举出包含全部实体的字典，故对于大型语料库的 NER 而言该方法存在较大的局限性。

基于统计机器学习的方法。MUC-7 会议中研究者将最大熵(ME) [11] [12]和隐马尔可夫(HMM) [13]等机器学习方法应用于实体识别问题中去。实验结果表明，实体的识别不再严重依赖于字典的规模，通过合理的应用文本信息，使用较小的字典也能得到很好的精度和召回率[14]。基于机器学习方法的 NER 实质上是将其转化为分类问题，主要包括：确定实体边界；对命名实体进行分类。其中，最著名的方法为序列标注方法，此类方法主要是将 NER 问题当作序列标注问题处理，利用大规模语料来学习标注序列。标注方法可分为基于短语模型和基于字符模型两种方法，基于对短语模型存在短语边界模糊导致分词准确度不稳定，并且出现过多未登录词时，模型不能很好地对未登陆词标签进行权重赋值[15]。在实际应用中，基于字符的序列标注方法应用更广泛，这种方法可有效减少基于词模型标注所带来的问题，但是该模型也存在一定的局限性，由于只利用到单个字符，所以不能对标注的实体进行级别衡量。字符标注规则是对文本中的每个字符分别赋予一个标签，不同标签对应不同类别的命名实体位置。此时，NER 任务就简化为首先对每个字符进行序列化自动标注，然后对标注结果进行统计分类，最终就可以确定实体类别。这种序列化标注方法的提出对于解决 NER 问题具有里程碑式的意义，随后 AdaBoost [16]、支持向量机[17]和随机条件场(Conditional Random Field, CRF) [18]等方法都依次被成功应用到序列化标注中，极大提升了 NER 的效果。目前，序列化标注方法仍是解决 NER 问题的最有效方法。机器学习方法的出现使得基于规则提出的方法不具有的广泛性得到改善和解决。但这种方法在特征提取方面严重依赖于大规模人工标注语料，识别效果并不是十分理想。

随着神经网络的不断发展，深度学习方法也被应用于解决 NER 问题中来。尤其是利用词向量来表示文本的方法，不仅解决了高维度大型稀疏矩阵带来的高难度计算问题，还解决了人工筛选特征对应信息不完整等问题。该方法利用统一的向量维度进行特征表示，结合序列标注方法为 NER 问题带来了极大的希望，其中最广泛的方法为使用词向量作为特征[19]。CoNLL-2002 和 CoNLL-2003 两次会议研究者多采用深度学习方法解决 NER 问题[20]，一些感知机模型[21]和双向长短时记忆网络(Bidirectional Long Short-Term Memory, BiLSTM) [22]也得到了尝试。研究者主要将现有的深度学习模型和方法进行改进，将其应用于 NER 问题上，如长短时记忆网络(Long Short-Term Memory, LSTM)和 CRF 相结合的模型，该模型使得 F1 值提高了 5% [23]。Lample 等提出了两种新的神经网络结构，一种是基于 BiLSTM 和 CRF 结合，另一种是基于转换方法构造标记片段。模型同时从标注语料和未标注语料训练得到特征，并在四种不同的语料上获得了前所未有的实体识别效果[24]。相较于传统机器学习方法，深度学习方法通过自主学习从非结构化数据中获得更深层次和更抽象的文本特征，较好地解决了机器学习方法特征选取不准确和对非结构化数据噪声造成的干扰等问题。除了 LSTM，其他深度学习方法也被成功应用于解决此类 NER 问题，如卷机神经网络 CNN [25]、混合神经网络 HNN [26]和循环神经网络 RNN [27]等。

近期监督学习被引进应用解决 NER 的数据预处理环节当中，以用来提升模型的特征学习能力[28]。

Bidirectional Encoder Representation from Transformers (BERT)是由谷歌 2018 年提出的一种基于深度学习的语言表示模型。BERT 在 11 种不同的自然语言处理测试任务中取得最佳效果，是 NLP 领域近期重要的研究成果[29]。BERT 内部机制采用 Transformer 的编码器和解码器结构，其中，BERT-base 和 BERT-large 分别采用 12 层和 24 层 Transformer 编码器作其基本网络结构，相比于传统的深度学习网络，Transformer 具有更强大的文本编码能力，能够完成大型数据语料的训练[30]。当前，BERT 模型已经成为一个基础性工具，经过预训练－微调手段可广泛应用于各种 NLP 领域[31]。

3. BERT-BiLSTM-CRF 模型

BERT-BiLSTM-CRF 模型结构如图 1 所示，该模型共分为三个部分，首先利用 BERT 进行数据预处理学习文本特征，再将学习到的特征作为 BiLSTM 的输入进行训练对文本进行双向编码，最后利用 CRF 解码输出概率最大的标签序列。

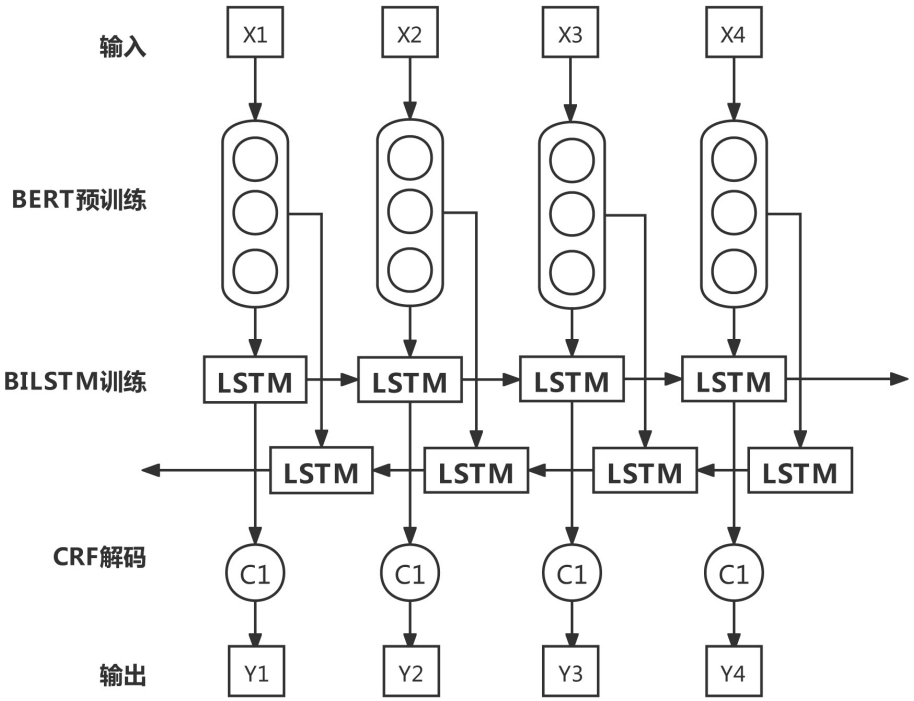

Figure 1. BERT-BilSTM-CRF model
图 1. BERT-BilSTM-CRF 模型

3.1. BERT 预训练

最早的神经语言模型 Bengio 提出的，是一种关于计算概率的方法，具体说来为从左到右计算下一个词出现的概率[32]。主要是由词 $w_1, w_2, \cdots, w_m$ 构成的句子组成的训练集，再由神经网络训练得到出现概率的语言模型。

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

传统的语言模型是静态的，无法上下文表征字的多义性和语法特征等。故根据此类问题，本文采用 BERT 预训练模型，结构如图 2 所示，其中 $E_1, E_2, \cdots, E_n$ 为输入向量，$T_1, T_2, \cdots, T_n$ 为该模型的输出向量，Trm 表示 Transformer 结构。

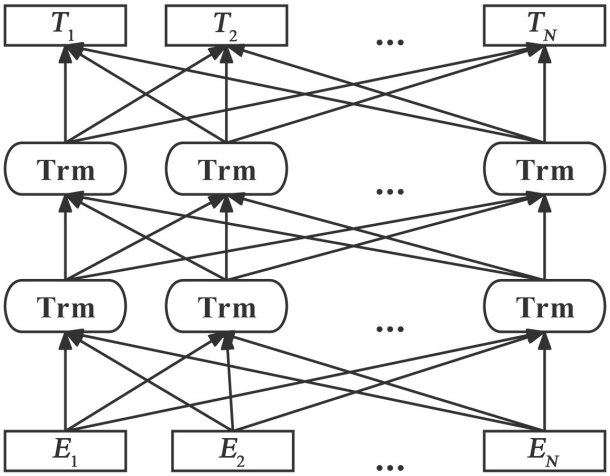

Figure 2. BERT pre-training
图 2. BERT 预训练

BERT 内部机制主要是基于 Transformer 的 Encoder 结构，其模型结构比 Transformer 更深，这种机制可以更好地获取上下文信息，极大地提升了模型抽取特征的能力。BERT 的训练主要分为两个阶段：预训练阶段和微调阶段。BERT 的预训练有两个无监督学习任务，分别是 Masked Language Model (MLM)和 Nest Sentence Prediction (NSP)。微调阶段是后续用于一些下游任务的时候进行的微调，根据要实现的功能进行调整。

Table 1. Examples of MLM task
表 1. MLM 任务举例

概率	文本
80%	小明每天 8 点[Mask] [Mask]
10%	小明每天 8 点起床
10%	小明每天 8 点时间

BERT 的第一个预训练任务 MLM，在句子中随机选择 10%的单词进行 Mask，在选择为 Mask 的单词中，有 80%真的使用[Mask]进行替换，10%不进行替换，剩下 10%使用一个随机单词替换，具体举例见表 1。然后利用上下文的信息预测被遮盖的单词，这样可以更好地根据全文理解单词的意思。由于本文训练集均为中文数据集，MLM 采用符合中文语法规则的全词 Mask，具体 NSP 预测任务举例如表 2 所示。

Table 2. Examples of NSP prediction task
表 2. NSP 预测任务举例

输入	[CLS]标签
[CLS]小明每天 8 点[Mask] [Mask] [SEP] 9 点去上学[SEP]	IsNest
[CLS]小明每天 8 点[Mask] [Mask] [SEP]今天天气很好[SEP]	NotNest

BERT 的特征提取主要基于 Transformer 的特征提取器，Transformer 是 NLP 研究者热衷的模型结构，自注意力机制(Self-Attantion)和前馈神经网络模型(Feed Forward Network)组成的基础，自注意力机制能帮助当前节点不仅仅只关注当前的词，从而能更好地获取到上下文语义信息[30]。Transformer 具体结构如图 3 所示。

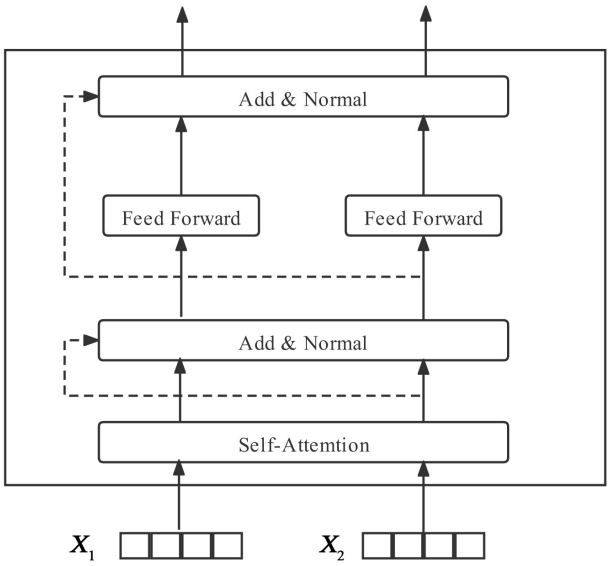

Figure 3. Transformer model
图 3. Transformer 模型

Transformer 的关键部分为自注意力机制，模型采用了 Encoder-Decoder 架构，但其结构相比于 Attention 更加复杂。首先，自注意力机制会计算出三个新的向量，分别为 Query($\boldsymbol{Q}$)、Key($\boldsymbol{K}$)、Value($\boldsymbol{V}$)，向量维度相同。自注意力机制主要通过句子中词和词之间的关联程度调整权重系数矩阵，以此来获取词的表征。

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

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 为字符的输入向量矩阵，d_k 为 Embedding 维度。具体自注意力机制模型如图 4 所示，其中 MatMul 表示矩阵相乘运算。

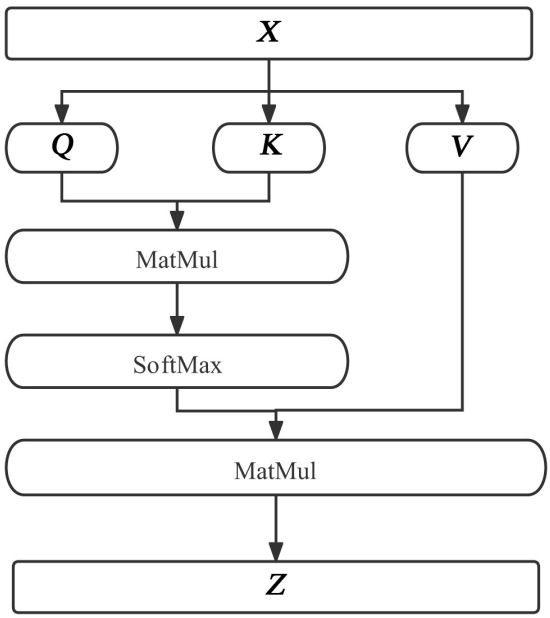

Figure 4. Self-attention mechanism model
图 4. 自注意力机制模型

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)，该机制主要是通过线性变换对 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 投影，最后将不同的 Attention 结构加和起来，从而获取句子的语义信息。

$$\text{Multihead} = \text{concat}\left(\text{head}_1,\text{head}_2,\cdots,\text{head}_n\right)\cdot\boldsymbol{W}^O$$

$$\text{head}_i = \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^k,\boldsymbol{V}\boldsymbol{W}_i^v\right)$$

其中，$\boldsymbol{W}_i^Q \in R^{d_{model}\times d_k},\boldsymbol{W}_i^k \in R^{d_{model}\times d_k},\boldsymbol{W}_i^v \in R^{d_{model}\times d_v},\boldsymbol{W}^O \in R^{hd_v\times d_{model}}$，在本文 BERT 中使用 $h=8$ 个平行自注意力机制，对每一个自注意力机制使用 $d_k = d_v = \frac{d_{model}}{h} = 64$，由于每个自注意力机制头部数量的减少，多头注意力机制与全维度的单头注意力机制计算成本相似。

多头注意力机制的好处是允许模型在不同的表示子空间里学习到相关的信息，使得模型特征学习更加准确。

3.2. BiLSTM

LSTM 是循环神经网络的变体，主要是为了解决长序列训练过程中出现的梯度消失或梯度爆炸的问题。LSTM 可以非常有效地利用文本上下文信息，并深度挖掘文本潜在的语义信息，减少工作量的同时，确定约束输出模型，从而达到提高实体识别精度，该方法广泛应用于文本上下文语义信息提取的问题中。

LSTM 的结构可由如下公式表示：

$$f_t = \sigma\left(W_f\cdot\left[h_{t-1},x_t\right]+b_f\right), \tag{3}$$

$$i_t = \sigma\left(W_i\cdot\left[h_{t-1},x_t\right]+b_i\right), \tag{4}$$

$$o_t = \sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right), \tag{5}$$

$$c_t = \tan h\left(W_c\cdot\left[h_{t-1},x_t\right]+b_c\right), \tag{6}$$

$$C_t = f_t\times C_{t-1}+i_t\times c_t, \tag{7}$$

$$h_t = o_t\times\tanh\left(C_t\right) \tag{8}$$

其中，f_t 表示遗忘门；i_t 表示输入门；o_t 表示输出门；c_t 表示上一时刻隐藏层状态；C_t 表示当前时刻隐藏层状态；x_t 表示当前输入；h_t 表示每一步的输出；b_f,b_i,b_o,b_c 为偏执项。

LSTM 内部主要分为三个阶段：

第一阶段为忘记阶段。这个阶段主要是对上一个节点传进来的输入进行选择性忘记，即通过 f_t，来控制忘记门对上一状态的 C_{t-1} 哪些需要遗忘，其中，f 表示 forget。该门会读取 h_{t-1} 和 x_t，输出一个在 0 到 1 之间的数值给每个在细胞状态 C_{t-1} 中的数字。1 表示完全保留，0 表示完全舍弃。

第二阶段选择记忆阶段。该阶段主要为对输入进行有选择性地进行记忆，当前的输入内容为 C_t 表示，选择的输入门控制信号为 i_t。将上面两步得到的结果相加即可得到传输给下一状态的 C_t。我们把旧状态 C_{t-1} 与 f_t 相乘，丢弃掉我们确定需要丢弃的信息。接着加上 $i_t\times c_t$ 得到新的候选值，再根据预期更新状态的程度进行变化。

第三阶段为输出阶段。该阶段决定哪些结果将会被当成当前状态的输出，输出门控制信号为 o_t。并且还可以对上一阶段得到的 C_t 通过一个 tanh 激活函数进行处理。

以上就是 LSTM 的内部结构。单向的 LSTM 无法同时处理上下文信息，故双向的 BiLSTM 更多地应用于文本处理当中。

BiLSTM 模型通过门控状态来控制传输状态，记住需要记忆的并忘记不重要的信息。由于不同于普通的 RNN 那样只能单一的记忆叠加方式，该网络结构能动态地捕获数据信息，对信息具有更强记忆能力。

由于 BiLSTM 利用记忆单元和控制门限制，实现了对长距离信息的有效利用，解决了梯度消失问题，故该方法对信息检索、自动问答和知识图谱构建等领域有着重要的应用价值。

3.3. CRF

CRF 模型是给定一组输入随机变量条件下，求另一组输出随机变量的条件概率分布模型，该模型特点是假设输出随机变量构成马尔可夫随机场。CRF 常用于命名实体识别的序列标注问题。

设 $X=\{X_1,X_2,\cdots,X_n\}$，$Y=\{Y_1,Y_2,\cdots,Y_n\}$ 均为线性链表示的随机变量序列，若在给定随机变量序列 X 的条件下，随机变量序列 Y 的条件概率分布 $P(Y\mid X)$ 构成条件随机场，即：

$$P(Y_i\mid X,Y_1,\cdots,Y_{i-1},Y_{i+1},\cdots,Y_n)=P(Y_i\mid X,Y_{i-1},Y_{i+1}) \tag{9}$$

其中，$i=1,2,\cdots,n$ (在 $i=1$ 或 n 时只考虑单边)，称 $P(Y\mid X)$ 为线性链条件场，在标注问题中，X 表示输入的观测序列，Y 表示的对应的输出标记序列或状态序列。CRF 与马尔可夫随机场相比，不仅考虑了当前时刻观测状态的信息，也考虑了上一时刻的隐藏状态信息。因此，在带有时序关系的场合，CRF 的效果要更好一些。线性链条件场如图 5 所示。

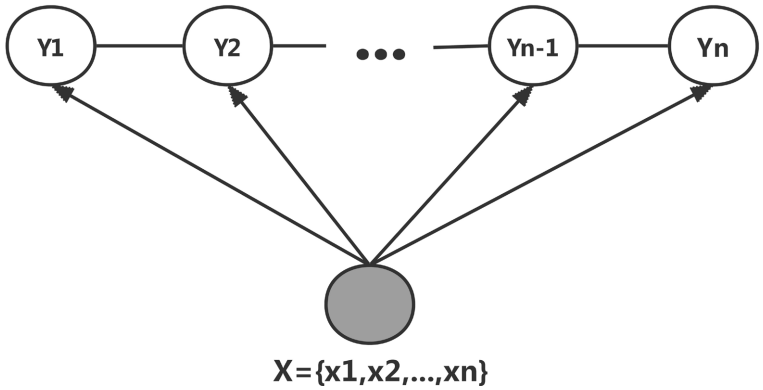

Figure 5. Conditional field of linear chain
图 5. 线性链条件场

本文将采用 CRF 进行状态序列预测，CRF 能够有效考虑字符与标签之间关系，主要利用单个字符标签的分数与字符标签之间的迁移矩阵计算不同种类标签的概率，从而得到最大概率序列即为所要寻找的状态序列。对于输入数列 $X=\{x_1,x_2,\cdots,x_n\}$，经 CRF 层被训练来输出预测最大概率标签序列为 $Y=\{y_1,y_2,\cdots,y_n\}$，则定义分数为：

$$s(X,Y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{x_i,y_i} \tag{10}$$

其中，A 矩阵是标签转移矩阵，$A_{y_i,y_{i+1}}$ 表示标签 y_i 转移到 y_{i+1} 的转移概率；P 矩阵是 BiLSTM 输出矩阵，P_{x_i,y_i} 代表字符 x_i 映射到标签 y_i 的非归一化概率。

模型训练的时候，对于每个序列 Y 优化对数损失函数，调整矩阵 A 的值，利用 Softmax 函数，为每一个正确的标签序列定义一个转移概率值：

$$p(Y\mid X)=\frac{e^{s(X,Y)}}{\sum_{\tilde{Y}\in Y_X}s(X,\tilde{Y})} \tag{11}$$

其中，Y_X 代表所有的标签序列，$\tilde{Y}$ 表示真实标注序列。故训练过程中，我们只需要最大化似然概率 $p(Y\mid X)$ 即可，这里我们利用对数似然函数：

$$\ln\left(p\left(Y\mid X\right)\right)=\ln\left(\frac{\mathrm{e}^{s(X,Y)}}{\sum_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right)}\right)=s\left(X\mid Y\right)-\log\left(\sum_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right)\right) \tag{12}$$

然后，我们将损失函数定义为$\log\left(p\left(Y\mid X\right)\right)$，就可以利用梯度下降法进行网络学习。

$$Loss=\ln\left(\sum_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right)\right)-s\left(X\mid Y\right) \tag{13}$$

当模型完成训练，进行预测时按如下策略寻找最优路径：

$$y^{*}=\arg\max_{\tilde{Y}\in Y_X}s\left(X,\tilde{Y}\right) \tag{14}$$

其中，y^{*}表示集合中使得$s\left(X,\tilde{Y}\right)$函数最大的序列。

3.4. 中文转换

为了减少数字、英文字母、符号和中文混合的数据对中文实体识别的影响，我们首先将数据集中的数字、英文字符和符号转化为中文，得到更规范的数据集。人民日报为主要摘自报纸数据集，MSRA 数据集是收录中文普通话的数据集，这两个数据集中夹杂的数字、英文字母和符号等较少，对识别结果影响较小，故不对这两个数据集进行规范化处理。

由于计算机领域数据集中的实体信息如计算机中央处理器、型号等多为英文和数字的组合，这种情况给中文实体识别带来了极大挑战，故本文拟对该中英文混合数据集进行数据预处理，具体中文转换规则对照如表 3 和表 4 所示。

Table 3. Roman numbers Chinese comparison table
表 3. 罗马数字中文对照表

数字	中文	数字	中文	数字	中文
0	零	4	四	8	八
1	一	5	五	9	九
2	二	6	六		
3	三	7	七		

Table 4. English letters and symbols Chinese comparison table
表 4. 英文字母及符号中文对照表

字母	中文	字母	中文	字母	中文
Q	子	A	戌	X	丁
W	丑	S	亥	C	戊
E	寅	D	金	V	己
R	卯	F	木	B	庚
T	辰	G	水	N	辛
Y	巳	H	火	M	壬
U	午	J	土	+	正
I	未	K	甲	-	反
O	申	L	乙	/	兑
P	酉	Z	丙	.	点

具体数据转化为全中文举例如表 5 所示。

Table 5. Example of converting data into Chinese
表 5. 数据转化为全中文举例

原始数据	联想 Lenovo 昭阳 E43-80031 笔记本电脑/I5-8250U/4G/500G/2G 独显/无光驱/14 寸/DOS
转化为中文数据	联想乙寅辛申己申昭阳寅四三反八零零三一笔记本电脑兑未五反八二五零午兑四水兑五零零水兑二水独显兑无光驱兑一四寸兑金申亥

4. 实验过程与结果

4.1. 数据集

本文有三个实验数据集分别是 1998 年人民日报语料、MSRA 语料和自行标注的计算机信息数据集。数据采用 BIO 标注策略，具体为：B 表示 Beginning，为标注实体的开始；I 表示 Inside，表示实体除开始剩余的部分；O 代表 Other，表示无用信息。

人民日报和 MSRA 数据集主要标记三个实体：人名(PER)、地名(LOC)和组织机构名(ORG)，数据集标签包含 7 个，分别是“O”、“B-PER”、“I-PER”、“B-LOC”、“I-LOC”、“B-ORG”、“I-ORG”。自行标注计算机数据集主要识别八个实体：品牌(BRD)、中央处理器(CPU)、硬盘(DSK)、图形处理器(GPU)、缓存(MEM)、尺寸(SCR)、操作系统(SYS)、型号(TYP)，该数据集共包含 17 个标签“O”、“B-BRD”、“I-BRD”、“B-CPU”、“I-CPU”、“B-DSK”、“I-DSK”、“B-GPU”、“I-GPU”、“B-MEM”、“I-MEM”、“B-SCR”、“I-SCR”、“B-SYS”、“I-SYS”、“B-TYP”、“I-TYP”。语料规模具体情况如表 6 所示。

Table 6. Corpus dataset (type: character)
表 6. 语料数据集(类型：字符)

	人民日报	MSRA	计算机领域
训练集	3.2×10^7	2.2×10^6	5.8×10^4
测试集	6.6×10^5	1.7×10^5	5.8×10^4

4.2. 评价指标

在 NLP 领域评价指标主要为以下四个：准确度(Accuracy)、精度(Precision)、召回率(Recall)、F1 值(F1-score)。在具体介绍指标之前，需要明确混淆矩阵的概念，我们定义 TP、FN、FN 和 FP 如下。

	正类	负类
预测为正类	Ture Positive (TP)	Ture Negative (TN)
预测为负类	False Positive (FP)	False Negative (FN)

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precison} + \text{recall}} \times 100\%$$

4.3. 实验参数

本文模型采用 Tensorflow 环境进行搭建，Python 版本 3.6，具体训练主要超参数如表 7 所示。

Table 7. Corpus dataset (type: character)
表 7. 语料数据集(类型：字符)

参数	数值	参数	数值
Transformer 层数	12	Max_seq_length	128
Hidden_dim	768	Optimizer	Adam
Learning_rate	10^{-5}	LSTM_size	128
Batch_size	16	Dropout_rate	0.5
Gradient_clip	0.5	Epoch	16

4.4. 实验结果

为了验证模型的优劣，本文在三个数据集上进行对比实验，分别采用 BiLSTM-CRF [33]、BERT-CRF [34]和 BERT-BiLSTM-CRF 模型进行实验。具体实验结果如表 8~10 所示。

Table 8. Comparison of experimental results of human name daily corpus (unit: %)
表 8. 人名日报语料实验结果对比(单位：%)

方法	准确度	精度	召回率	F1 值
BiLSTM-CRF	98.36	88.95	87.78	88.36
BERT-CRF	99.37	93.83	94.94	94.38
BERT-BiLSTM-CRF	99.39	94.64	95.56	95.10

Table 9. Comparison of experimental results of MSRA corpus (unit: %)
表 9. MSRA 语料实验结果对比(单位：%)

方法	准确度	精度	召回率	F1 值
BiLSTM-CRF	98.38	88.57	86.61	87.58
BERT-CRF	99.33	95.38	94.74	95.06
BERT-BiLSTM-CRF	99.34	95.16	95.01	95.09

Table 10. Comparison of experimental results of computer field corpus (unit: %)
表 10. 计算机领域语料实验结果对比(单位：%)

方法	准确度	精度	召回率	F1 值
BiLSTM-CRF	99.19	96.39	97.09	96.74
BERT-CRF	99.56	98.30	98.86	98.62
BERT-BiLSTM-CRF	99.84	99.38	99.61	99.45

Table 11. Comparison of BERT-BiLSTM-CRF entity recognition results of types in computer field (unit: %)
表 11. 计算机领域各类别 BERT-BiLSTM-CRF 实体识别结果对比(单位：%)

实体类别	精度	召回率	F1 值
品牌(BRD)	99.99	99.99	99.99
中央处理器(CPU)	99.73	99.73	99.73
硬盘(DSK)	99.79	99.79	99.79
图形处理器(GPU)	99.70	99.99	99.85
缓存(MEM)	99.38	99.45	99.41
尺寸(SCR)	99.62	99.49	99.56
系统(SYS)	99.74	99.74	99.74
型号(TYP)	96.56	98.69	99.63

比较 BiLSTM-CRF 和 BERT-CRF 两个模型在三个数据集上的结果，BERT-CRF 较 BiLSTM-CRF 的 F1 值分别提升了 6.02%、7.51%和 1.88%。从实验结果可知，BERT 模型具有更强的特征提取能力，充分证明 BERT 中多头注意力机制使得评价指标大幅度提升。BERT-CRF 和 BERT-BiLSTM-CRF 两个模型在人民日报语料上的实验结果得到的 F1 值，后者较前者提升了 0.72%，这种小幅度的提升主要由于 BiLSTM 学习中的上下文信息提取的能力，它能充分提取字符序列特征，这使得模型学习效果更好。

综合上述实验结果我们可以看到，使用 BERT 预训练的模型命名实体识别评价指标结果明显大幅度提升。BERT-BILSTM-CRF 模型在不同实验语料中评价指标中均表现最为优秀，F1 值分别取得了 95.1%、95.09%和 99.45%的斐然成绩。对于大型 MSRA 语料而言，BERT-BiLSTM-CRF 较 BiLSTM-CRF 模型的 F1 值更是有 12.41%的惊人提升。对于计算机领域语料而言，即使识别实体种类达到 8 个，各个实体 F1 值均大于 99%，其中，品牌实体识别准确率更是高达 100%。在计算机领域数据集 BERT-BiLSTM-CRF 模型各类别实体识别结果如表 11 所示。

通过以上实验结果充分表明，具有双向 Transformer 结构的 BERT 预训练模型具有更强的语义表征能力。对于计算机领域语料 BERT-BiLSTM-CRF 模型，除型号实体类别外，其他实体类别 F1 值均接近 100%。由于计算机型号实体构成复杂，且不同品牌的命名型号规则区别很大的原因，导致模型识别效果不理想。该模型还存在一定的进步空间，需要对复杂实体的命名实体识别进一步进行探究。

5. 结束语

针对于中文命名实体识别问题，本文构造了一种不依赖于人工特征的神经网络 BERT-BiLSTM-CRF 模型。该模型通过基于双层 Transformer 的 BERT 模型进行预训练，得到具有更强表征能力的词向量，并以此作为 BiLSTM 神经网络的输入，充分利用文本的上下文信息对其进行进一步处理，最后，利用 CRF 模块解码，计算相邻标签的关联性，进而获得全局最优标签序列。与传统模型相比，该模型具有较大的优势，且在多个语料上的实验结果均有较好的表现效果。但对于构成复杂实体识别效果表现稍有欠缺，容易出现识别不全现象。如何对上述问题进行优化是进一步需要解决的问题。

参考文献

[1] Chinchor, N. (1995) MUC-6 Named Entity Task Definition (Version 2.1). 6*th Message Understanding Conference*, Columbia, 6-8 November 1995, 317-332.

[2] Chinchor, N. and Robinson, P. (1997) MUC-7 Named Entity Task Definition. *Proceedings of the 7th Conference on Message Understanding*, Vol. 29, 1-21.

[3] Sang, E.F. and De Meulder, F. (2003) Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Edmonton, 31 May 2003, 142-147.

[4] Alfonseca, E. and Manandhar, S. (2002) An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. *Proceedings of the* 1*st International Conference on General WordNet*, Mysore, 21-25 January 2002, 34-43.

[5] Ekine, S., Sudo, K. and Nobata, C. (2002) Extended Named Entity Hierarchy. *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, May 2002, 1818-1824.

[6] Sekine, S. and Nobata, C. (2004) Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, May 2004, 1977-1980.

[7] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., *et al*. (2013) Named Entity Recognition: Fallacies, Challenges and Opportunities. *Computer Standards & Interfaces*, **35**, 482-489. https://doi.org/10.1016/j.csi.2012.09.004

[8] 赵佳. 基于字符增强的命名实体识别方法研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2020.

[9] Aberdeen, J., Burger, J.D., Day, D., *et al*. (1995) MITRE: Description of the Alembic System Used for MUC-6. *Sixth Message Understanding Conference* (*MUC*-6): *Proceedings of a Conference*, Columbia, 6-8 November 1995, 141-155. https://doi.org/10.3115/1072399.1072413

[10] Krupka, G. (1995) SRA: Description of the SRA System as Used for MUC-6. *Sixth Message Understanding Conference* (*MUC*-6): *Proceedings of a Conference*, Columbia, 6-8 November 1995, 221-235. https://doi.org/10.3115/1072399.1072419

[11] Borthwick, A., Sterling, J., Agichtein, E., *et al*. (1998) NYU: Description of the MENE Named Entity System as Used in MUC-7. *Seventh Message Understanding Conference* (*MUC*-7): *Proceedings of a Conference*, Fairfax, 29 April-1 May 1998.

[12] 陆铭, 康雨洁, 俞能海. 简约语法规则和最大熵模型相结合的混合实体识别[J]. 小型微型计算机系统, 2012, 33(3): 537-541.

[13] 冯静, 李正武, 张登云, 等. 基于隐马尔可夫模型的桥梁检测文本命名实体识别[J]. 交通世界, 2020, 8(3): 32-33.

[14] 焦凯楠, 李欣, 朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用, 2021, 57(16): 1-15.

[15] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[16] Carreras, X., Marquez, L. and Padró, L. (2002) Named Entity Extraction Using Adaboost. *COLING*-02: *The* 6*th Conference on Natural Language Learning* (*CoNLL*-2002), Volume 20, 1-4. https://doi.org/10.3115/1118853.1118857

[17] 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.

[18] McCallum, A. and Li, W. (2003) Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL* 2003, Edmonton, 31 May-1 June 2003, 188-191. https://doi.org/10.3115/1119176.1119206

[19] Cherry, C. and Guo, H. (2015) The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. *Proceedings of the* 2015 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Denver, May-June 2015, 735-745. https://doi.org/10.3115/v1/N15-1075

[20] 陈曙东, 欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术, 2020, 46(3): 251-260.

[21] 姜文斌, 赵海兴, 刘群. 基于感知机模型藏文命名实体识别[J]. 计算机工程与应用, 2014(15): 172-176.

[22] Hammerton, J. (2003) Named Entity Recognition with Long Short-Term Memory. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, Volume 4, 172-175. https://doi.org/10.3115/1119176.1119202

[23] Peng, N. and Dredze, M. (2016) Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Volume 2, 149-155. https://doi.org/10.18653/v1/P16-2025

[24] Lample, G., Ballesteros, M., Subramanian, S., *et al*. (2016) Neural Architectures for Named Entity Recognition. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[25] Dong, X., Qian, L., Guan, Y., *et al*. (2016) A Multiclass Classification Method Based on Deep Learning for Named Entity Recognition in Electronic Medical Records. 2016 *New York Scientific Data Summit* (*NYSDS*) *IEEE*, New York, 14-17 August 2016, 1-10. https://doi.org/10.1109/NYSDS.2016.7747810

[26] Shao, Y., Hardmeier, C. and Nivre, J. (2016) Multilingual Named Entity Recognition Using Hybrid Neural Networks. *The Sixth Swedish Language Technology Conference* (*SLTC*).

[27] Yadav, V. and Bethard, S. (2019) A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, August 2018, 2145-2158.

[28] 王子牛, 姜猛, 高建瓴, 陈娅先. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学, 2019, 46(z2): 138-142.

[29] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, June 2019, 4171-4186.

[30] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 5998-6008.

[31] 庄穆妮, 李勇, 谭旭, 等. 基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真[J]. 系统仿真学报, 2021, 33(1): 24-36.

[32] Bengio, Y., Ducharme, R., Vincent, P., *et al*. (2003) A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, **3**, 1137-1155.

[33] 顾溢. 基于 BiLSTM-CRF 的复杂中文命名实体识别研究[D]: [硕士学位论文]. 南京: 南京大学, 2019.

[34] 田梓函, 李欣. 基于 BERT-CRF 模型的中文事件检测方法研究[J]. 计算机工程与应用, 2021, 57(11): 135-139.