(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111563383 A
(43)申请公布日 2020.08.21

(21)申请号 202010272320.8

(22)申请日 2020.04.09

(71)申请人 华南理工大学
地址 510640 广东省广州市天河区五山路381号

(72)发明人 蔡毅 郑煜佳

(74)专利代理机构 广州市华学知识产权代理有限公司 44245
代理人 裴磊磊

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/284(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

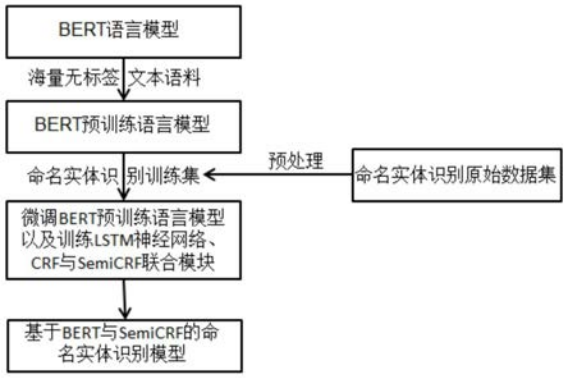
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于BERT与SemiCRF的中文命名实体识别方法

(57)摘要

本发明公开了一种基于BERT与SemiCRF的中文命名实体识别方法，构建命名实体识别模型，所述方法包括步骤：获取预训练好的BERT模型；对命名实体识别的原始语料数据进行预处理，构建命名实体识别的训练集；将构建的命名实体识别的训练集数据输入到预训练好的BERT语言模型；将BERT语言模型的输出依次输入到双向LSTM神经网络以及CRF与SemiCRF联合模块中，对双向LSTM神经网络及联合模块进行多次迭代训练；使用训练完成得到的完整命名实体识别模型，对中文文本进行命名实体识别。本发明解决了传统的word2vec无法区分多义词的问题，并通过引入的基于SemiCRF的方法，将传统的CRF方法往往会忽略掉的词级别信息与字级别的信息结合起来，在一定程度上提高了中文命名实体识别的效果。



CN 111563383 A

1.一种基于BERT与SemiCRF的中文命名实体识别方法,其特征在于,构建命名实体识别模型,所述模型包括BERT语言模型、双向LSTM以及CRF与SemiCRF联合模块,所述方法包括步骤:

获取预训练好的BERT模型;

对命名实体识别的原始语料数据进行预处理,构建命名实体识别的训练集;

将构建的命名实体识别的训练集数据输入到预训练好的BERT语言模型;

将BERT语言模型的输出依次输入到双向LSTM神经网络以及CRF与SemiCRF联合模块中,对双向LSTM神经网络及联合模块进行多次迭代训练;

使用训练完成得到的完整命名实体识别模型,对中文文本进行命名实体识别。

2.根据权利要求1所述的方法,其特征在于,预训练好的BERT模型的获取方式包括:下载谷歌开源的BERT源码,使用BERT源码在海量的无标签中文文本语料上自行预训练出BERT预训练语言模型;或者直接下载谷歌官方预训练好的中文BERT语言模型chinese_L-12_H-768_A-12。

3.根据权利要求1所述的方法,其特征在于,所述对命名实体识别的原始语料数据进行预处理,构建命名实体识别的训练集的步骤中,包括:

对于命名实体识别原始语料进行常规数据预处理;

根据实际的应用需求确定所要识别的实体类型,或者直接使用通用的实体类型;

采用BIOES的实体标注方法对原始语料进行标注;

根据实际的应用需求制定特定的标注规则,结合上述的标注规则,对于未标注过的原始语料进行人工标注,或对于已标注过的原始语料进行标注的转换和修正。

4.根据权利要求1所述的方法,其特征在于,所述将BERT语言模型的输出依次输入到双向LSTM神经网络以及CRF与SemiCRF联合模块中,对双向LSTM神经网络及联合模块进行多次迭代训练的步骤中,包括:

将BERT语言模型输出的序列输入到双向LSTM神经网络;

双向LSTM神经网络输出该序列中每个词被标注成所有实体类型的概率分布向量,即每个词在字级别的CRF特征;

将得到的CRF特征序列分别输入到CRF与SemiCRF联合模块中的CRF层和SemiCRF层;

CRF层采用bi-LSTM+CRF方法计算出CRF层的损失函数;

SemiCRF层根据句中各个词的CRF特征和ground true标签计算句中各个段的SemiCRF特征,进而计算最佳路径的分数;

SemiCRF层采用forward算法通过SemiCRF特征转移矩阵计算出所有路径的分数;

根据最佳路径分数和所有路径分数计算SemiCRF层的损失函数;

采用SGD用CRF层的损失函数与SemiCRF层的损失函数的加权和来更新命名实体识别模型的参数。

5.根据权利要求4所述的方法,其特征在于,所述SemiCRF层根据句中各个词的CRF特征和ground true标签计算句中各个段的SemiCRF特征,进而计算最佳路径的分数的步骤中,最佳路径分数的计算方法为:

$$score(s,w)=\prod_{i=1}^{|s|}\psi(l_{i-1},l_i,w,b_i,e_i)$$

$$\psi(l_{i-1},l_i,w,b_i,e_i)=exp\{m_i+b_{l_{i-1},l_i}\}$$

$$m_i=\sum_{k=b_i}^{e_i}\varphi_c(y_k,w'_k)=\sum_{k=b_i}^{e_i}a_{y_k}^{\mathrm{T}}w'_k$$

其中，s表示段级标注序列，w表示的是输入序列的词嵌入向量表示，l_i表示的是第i个段级标签，b_i和e_i分别表示第i个段级标签的开头和结尾在输入序列上的对应位置，m_i是第i个段本身的分数，$b_{i,j}$表示的是从类别i到类别j的段级转移分数，y_k是输入序列第k个字的ground true标签，a_{y_k}是与标签y_k相关的一个权重参数向量。w'_k表示的是第k个词的特征向量，其构建方式为：

$$w'_k=[w_k;w_{e_i}-w_{b_i};\emptyset(k-b_i+1)]$$

其中，$\emptyset(k-b_i+1)$是组成段的各个词在段内的索引所对应的嵌入向量。

6.根据权利要求4所述的方法，其特征在于，所述根据最佳路径分数和所有路径分数计算SemiCRF层的损失函数的步骤中，损失函数中需要对分数$P(\hat{s}|w)=\frac{score(\hat{s},w)}{\sum_{s'\in S}score(s',w)}$作负对数似然处理，所以损失函数最终表示为Loss＝$score_{all\_path}$-$score_{best\_path}$。

7.根据权利要求1所述的方法，其特征在于，所述使用训练完成得到的完整命名实体识别模型，对中文文本进行命名实体识别的步骤中，包括：

将需要命名实体识别的语句输入到训练完成的完整命名实体识别模型；

输入的序列在经过预训练好的BERT语言模型后依次通过双向LSTM神经网络和CRF与SemiCRF联合模块，计算输入语句每个词的CRF特征，进而计算CRF层的CRF特征矩阵和SemiCRF层的SemiCRF特征矩阵；

使用viterbi算法分别在CRF层与SemiCRF层解码出输入语句的最佳路径，即得到了CRF层标签序列以及CRF层标签序列在CRF层上的分数$score_{C-C}$和SemiCRF层标签序列以及SemiCRF层标签序列在SemiCRF层上的分数$score_{S-S}$；

计算CRF层解码所得的标签序列在SemiCRF层中的分数$score_{C-S}$，以及计算SemiCRF层解码所得的标签序列在CRF层中的分数$score_{S-C}$；

分别计算CRF层标签序列的总分$score_{C-C}$+$score_{C-S}$以及SemiCRF层标签序列的总分$score_{S-S}$+$score_{S-C}$，由于分数的计算经过负对数似然处理，因此取分数最小的那条标签序列作为命名实体识别的结果。

一种基于BERT与SemiCRF的中文命名实体识别方法

技术领域

[0001] 本发明涉及命名实体识别技术领域,尤其涉及一种基于BERT与 SemiCRF的中文命名实体识别方法。

背景技术

[0002] 命名实体识别(Named Entity Recognition,NER)是属于自然语言处理(Natural Language Processing,NLP)领域下的一个任务,该任务旨在从文本中识别出实体并将其分类到预定义好的实体类型,如人名、地名、机构名等。命名实体识别不仅可以单独作为用于信息提取的工具,还可以在自然语言处理领域的其他任务和应用中发挥重要的作用,如信息检索,自动文本摘要,问答,机器翻译和知识库构建等。

[0003] 现有的命名实体识别比较主流的方法是Bi-LSTM+CRF,其中所用到的 Bi-LSTM(双向长短期记忆网络)是深度学习中非常流行的一种深度神经网络,在命名实体识别中能够学习到长序列中的特征上下文关系;所用的 CRF(条件随机场)是一种传统的机器学习方法,在命名实体识别中能够学习到标签的上下文关系。

[0004] 上述基于Bi-LSTM+CRF的方法需要自行从命名实体识别数据集中学习词嵌入表示,这里存在的缺陷包括了:Bi-LSTM在学习词嵌入表示时无法应对一词多义的情况;命名实体识别数据集本身规模不算大,能够从中学习到的词嵌入表示的质量有限;Bi-LSTM不能并行处理数据,这导致其设置的词嵌入表示的规模大小受到限制不能太大,否则训练学习的时间成本将成倍增长。此外,在文本中存在着这一特性——命名实体大多以多个字组成的片段(segment)存在,而基于Bi-LSTM+CRF的方法中的CRF 条件随机场以字级别为单位没能利用到片段级别的信息。

发明内容

[0005] 本发明的目的在于克服现有技术的不足,提供一种基于BERT与 SemiCRF的中文命名实体识别方法。本发明能够解决词嵌入表示的学习质量有限和无法解决的一词多义的问题,并且能够避免CRF仅能利用字级别的信息而忽略了片段级别的信息这一问题。

[0006] 本发明的目的能够通过以下技术方案实现:

[0007] 一种基于BERT与SemiCRF的中文命名实体识别方法,构建命名实体识别模型,所述模型包括BERT语言模型、双向LSTM以及CRF与SemiCRF 联合模块,所述方法包括步骤:

[0008] 获取预训练好的BERT语言模型;

[0009] 对命名实体识别的原始语料数据进行预处理,构建命名实体识别的训练集;

[0010] 将得到的命名实体识别的训练集数据输入到预训练好的BERT语言模型;

[0011] 将BERT语言模型的输出依次输入到双向LSTM神经网络以及CRF与 SemiCRF联合模块中,对双向LSTM神经网络及联合模块进行多次迭代训练;

[0012] 使用训练完成得到的完整命名实体识别模型,对中文文本进行命名实体识别。

[0013] 进一步地,预训练好的BERT语言模型的获取方式包括:下载谷歌开源的BERT源码,

使用BERT源码在海量的无标签中文文本语料上自行预训练出BERT预训练语言模型;或者直接下载谷歌官方预训练好的中文BERT语言模型chinese_L-12_H-768_A-12。

[0014] 进一步地,所述对命名实体识别的原始语料数据进行预处理,构建命名实体识别的训练集的步骤中,包括:

[0015] 对于命名实体识别原始语料进行常规数据预处理;

[0016] 根据实际的应用需求确定所要识别的实体类型,或者直接使用通用的实体类型;

[0017] 采用BIOES的实体标注方法对原始语料进行标注;

[0018] 根据实际的应用需求制定特定的标注规则,结合上述标注规则,对未标注过的原始语料进行人工标注,或对于已标注过的原始语料进行标注的转换和修正。

[0019] 进一步地,所述将BERT语言模型的输出依次输入到双向LSTM神经网络以及CRF与SemiCRF联合模块中,对双向LSTM神经网络及联合模块进行多次迭代训练的步骤中,包括:

[0020] 将BERT语言模型输出的序列输入到双向LSTM神经网络;

[0021] 双向LSTM神经网络输出该序列中每个词被标注成所有实体类型的概率分布向量,即每个词在字级别(word level)的CRF特征;

[0022] 将得到的字级别的CRF特征序列分别输入到CRF与SemiCRF联合模块中的CRF层和SemiCRF层;

[0023] CRF层采用bi-LSTM+CRF方法计算出CRF层的损失函数;

[0024] SemiCRF层根据句中各个词的CRF特征和ground true标签计算句中各个段的SemiCRF特征,进而计算最佳路径的分数;

[0025] SemiCRF层采用forward算法通过SemiCRF特征转移矩阵计算出所有路径的分数;

[0026] 根据最佳路径分数和所有路径分数计算SemiCRF层的损失函数;

[0027] 采用SGD用CRF层的损失函数与SemiCRF层的损失函数的加权和来更新整个命名实体识别模型的参数。

[0028] 进一步地,所述使用训练完成得到的完整命名实体识别模型,对中文文本进行命名实体识别的步骤中,包括:

[0029] 将需要命名实体识别的语句输入到训练完成的完整命名实体识别模型;

[0030] 输入的序列在经过预训练好的BERT语言模型后依次通过双向LSTM 神经网络和CRF与SemiCRF联合模块,先计算输入语句每个词的CRF特征,再计算CRF层的CRF特征矩阵和SemiCRF层的SemiCRF特征矩阵;

[0031] 使用viterbi算法分别在CRF层与SemiCRF层解码出输入语句的最佳路径即得到了CRF层标签序列以及该序列在CRF层上的分数$score_{C-C}$、SemiCRF层标签序列以及该序列在SemiCRF层上的分数$score_{S-S}$;

[0032] 计算CRF层解码所得的标签序列在SemiCRF层中的分数$score_{C-S}$,以及计算SemiCRF层解码所得的标签序列在CRF层中的分数$score_{S-C}$;

[0033] 分别计算CRF层标签序列的总分$score_{C-C}+score_{C-S}$以及SemiCRF层标签序列的总分$score_{S-S}+score_{S-C}$,由于分数的计算经过负对数似然处理,因此取分数最小的那条标签序列作为命名实体识别的结果。

[0034] 本发明相较于现有技术,具有以下的有益效果:

[0035] 1、本发明所使用的BERT模型能够通过预训练和微调的的方式从规模庞大的中文

文本中学习到质量很好的词嵌入的表示，而不限于需要经过人工标注处理过的命名实体识别数据集，并且能够根据上下文场景来调整当前语义，从而解决一词多义的问题。

[0036] 2、本发明将半马尔可夫条件随机场(SemiCRF)引入命名实体识别中，相较于只利用到字级别信息的条件随机场(CRF)，能够更适应具有明显段级特征的命名实体，并且为了保证SemiCRF的命名实体识别效果，将其在一定程度上与CRF结合使得模型能够同时有效地利用字级别和片段级别的特征。

[0037] 3、本发明在训练和解码的过程同时考虑了CRF和SemiCRF两种方法，特别是解码时择优作为最后结果能够保证命名实体识别的精确率。

附图说明

[0038] 图1为本发明中一种基于BERT与SemiCRF的中文命名实体识别方法的流程图。

[0039] 图2为本发明实施例中基于BERT与SemiCRF的中文命名实体识别的命名实体识别模型的结构示意图。

具体实施方式

[0040] 下面结合实施例及附图对本发明作进一步详细的描述，但本发明的实施方式不限于此。

[0041] 实施例

[0042] 如图1所示为一种基于BERT与SemiCRF的中文命名实体识别方法的流程图，构建如图2所示的命名实体识别模型，所述模型包括BERT语言模型、双向LSTM以及CRF与SemiCRF联合模块，所述方法所述方法包括步骤：

[0043] S1、获取预训练好的BERT模型；

[0044] 具体地，获取方式包括：下载谷歌开源的BERT源码，使用BERT源码在海量的无标签中文文本语料上自行使用现有的预训练技术得出BERT 预训练语言模型；或者直接下载谷歌官方预训练好的中文BERT语言模型chinese_L-12_H-768_A-12。

[0045] S2、对命名实体识别的原始语料数据进行预处理，构建命名实体识别的训练集，包括步骤：

[0046] S21、对于命名实体识别原始语料进行常规数据预处理，包括对错别字进行修正以及对字符进行规范化等；所述原始语料数据为已经标注了的命名实体数据；

[0047] S22、根据实际的应用需求确定所要识别的实体类型，或者直接使用通用的实体类型如人名(PERSON)、地名(LOCATION)、机构名(ORGANIZATION)等；

[0048] S23、为了应对实体长度不一、难以区分实体边界的情况，采用BIOES 的实体标注方法：B标注长实体的开头、I标注长实体的内部、E标注长实体的尾部、S标注仅用一个字表示的实体、O标注非实体，比如“刘玄德”将被标注为(B-PER，I-PER，E-PER)；

[0049] S24、根据实际的应用需求制定特定的标注规则，结合步骤S22和S23 的标注规则，对于未标注过的原始语料进行人工标注，或对于已标注过的原始语料进行标注的转换和修正。

[0050] S3、将步骤S2预处理所得的命名实体识别的训练集数据输入到预训练好的BERT语言模型。

[0051] 具体地，所述训练集数据以句子为单位输入到预训练好的BERT语言模型中，BERT语言模型输出为词嵌入向量序列。

[0052] S4、将步骤S3中的BERT语言模型的输出，依次输入到双向LSTM 神经网络以及CRF与SemiCRF联合模块中，对双向LSTM神经网络及联合模块进行多次迭代训练，包括步骤：

[0053] S41、将BERT语言模型输出的序列输入到双向LSTM神经网络；

[0054] S42、双向LSTM神经网络输出该序列中每个词被标注成所有实体类型的概率分布向量，即每个词在字级别(word level)的CRF特征；

[0055] S43、将得到的CRF特征序列分别输入到CRF与SemiCRF联合模块中的CRF层和SemiCRF层；

[0056] S44、CRF层采用bi-LSTM+CRF方法计算出CRF层的损失函数；

[0057] 例如输入序列为“威尔逊医生来到加利福尼亚调查研究”，则ground true相应的标注序列为(B-PER,I-PER,E-PER,O,O,O,O,B-LOC, I-LOC,I-LOC,I-LOC,E-LOC,O,O,O,O)，对于其中的“利”字，CRF不仅考虑其本身这个位置被标注为I-LOC的分数，还会考虑其上下文“加”、“福”的标注结果，若“利”被标注为I-PER，显然从“加”的B-LOC 是不可能后接一个I-PER，从数据集中学习到标签上下文关系的CRF也因此会将“利”被标注为I-PER的分数给得很低；

[0058] S45、SemiCRF层根据句中各个词的CRF特征和ground true标签计算句中各个段的SemiCRF特征，进而计算最佳路径的分数；

[0059] 在本发明中，词嵌入向量中的词对应文本中的单字，段级别对应文本中的词语。同样地对于输入序列为“威尔逊医生来到加利福尼亚调查研究”，SemiCRF的段级标注序列为((1,3,PER)，(4,4,0)，(5,5,0)，(6,6,0)，(7,7,0)， (8,12,LOC)，(13,13,0)，(14,14,0)，(15,15,0)，(16,16,0))，计算最佳路径的分数可根据以下公式：

[0060] $$score(s,w)=\prod_{i=1}^{|s|}\psi(l_{i-1},l_i,w,b_i,e_i)$$

[0061] $$\psi(l_{i-1},l_i,w,b_i,e_i)=exp\{m_i+b_{l_{i-1},l_i}\}$$

[0062] $$m_i=\sum_{k=b_i}^{e_i}\varphi_c(y_k,w'_k)=\sum_{k=b_i}^{e_i}a_{y_k}^{\mathrm{T}}w'_k$$

[0063] 其中，s表示段级标注序列，w表示的是输入序列的词嵌入向量表示，l_i表示的是第i个段级标签，b_i和e_i分别表示第i个段级标签的开头和结尾在输入序列上的对应位置，m_i是第i个段本身的分数，$b_{i,j}$表示的是从类别 i到类别j的段级转移分数，y_k是输入序列第k个字的ground true标签，a_{y_k}是与标签y_k相关的一个权重参数向量。w'_k表示的是第k个词的特征向量，其构建方式为：

[0064] $$w'_k=\left[w_k;w_{e_i}-w_{b_i};\emptyset(k-b_i+1)\right]$$

[0065] 其中，$\emptyset(k-b_i+1)$是组成段的各个词在段内的索引所对应的嵌入向量；

[0066] S46、SemiCRF层采用forward算法通过SemiCRF特征转移矩阵计算出所有路径的分

数；

[0067] S47、根据最佳路径分数和所有路径分数计算SemiCRF层的损失函数；

[0068] 损失函数中需要对分数$P(\hat{s}|w)=\frac{score(\hat{s},w)}{\sum_{s'\in S}score(s',w)}$作负对数似然处理，所以表示为$Loss=score_{all_path}-score_{best_path}$；

[0069] S48、采用SGD(随机梯度下降法)用CRF层的损失函数与SemiCRF 层的损失函数的加权和来更新整个命名实体识别模型的参数，所述参数包括BERT语言模型、LSTM神经网络、CRF与SemiCRF联合模块在内的模型参数。加权的权重需要用控制变量法得出最佳的权重比，加权的权重会因为命名实体识别的训练数据的不同而有所变化。

[0070] S5、使用步骤S4训练完成得到的完整命名实体识别模型，对中文文本进行命名实体识别，包括步骤：

[0071] S51、将需要命名实体识别的语句输入到训练完成的完整命名实体识别模型；

[0072] S52、输入的序列在经过预训练好的BERT语言模型后依次通过双向 LSTM神经网络和CRF与SemiCRF联合模块，计算输入语句每个词的CRF 特征，进而计算CRF层的CRF特征矩阵和SemiCRF层的SemiCRF特征矩阵；

[0073] S53、使用viterbi算法分别在CRF层与SemiCRF层解码出输入语句的最佳路径，即得到了CRF层标签序列以及CRF层标签序列在CRF层上的分数$score_{C-C}$和SemiCRF层标签序列以及SemiCRF层标签序列在SemiCRF 层上的分数$score_{S-S}$；

[0074] S54、计算步骤S53中CRF层解码所得的标签序列在SemiCRF层中的分数$score_{C-S}$，以及计算步骤S53中SemiCRF层解码所得的标签序列在CRF 层中的分数$score_{S-C}$；

[0075] S55、分别计算步骤S53得到的两条标签序列的总分$score_{C-C}+score_{C-S}$、$score_{S-S}+score_{S-C}$，由于分数的计算经过负对数似然处理，因此取分数最小的那条标签序列作为命名实体识别的结果。

[0076] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

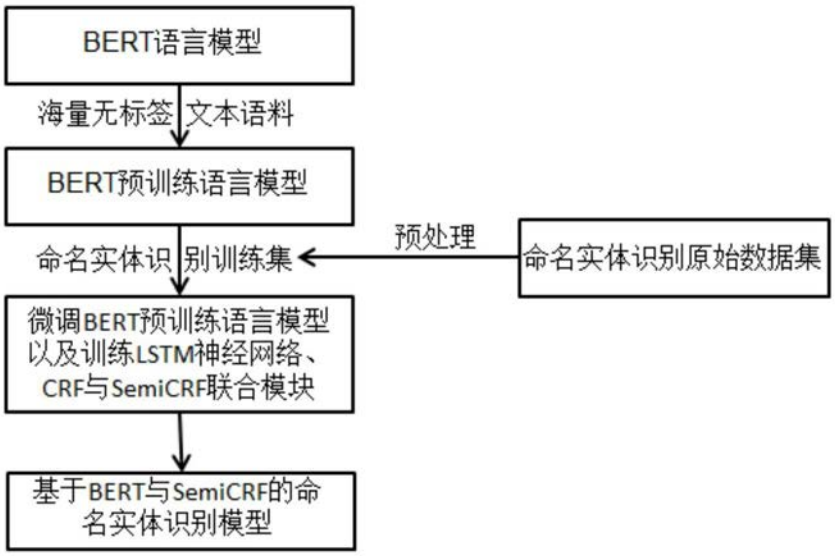

BERT语言模型
海量无标签文本语料
BERT预训练语言模型
预处理
命名实体识别训练集
命名实体识别原始数据集
微调BERT预训练语言模型以及训练LSTM神经网络、CRF与SemiCRF联合模块
基于BERT与SemiCRF的命名实体识别模型

图1

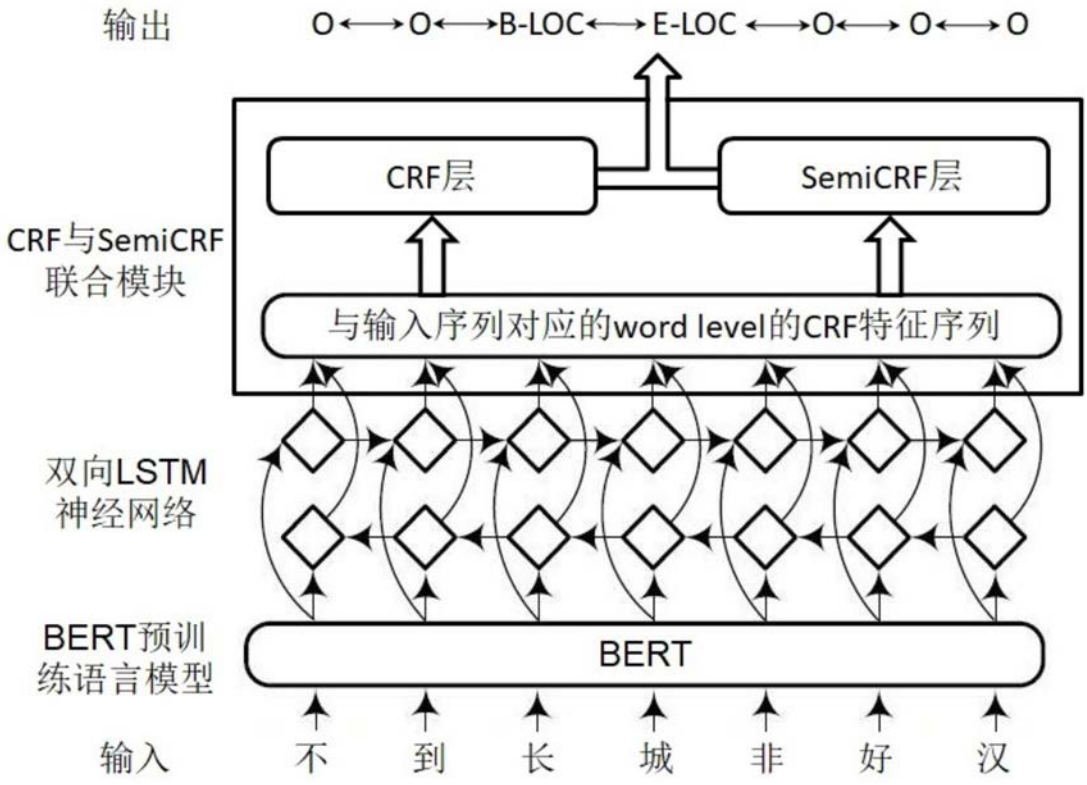

输出
O O B-LOC E-LOC O O O
CRF层
SemiCRF层
CRF与SemiCRF联合模块
与输入序列对应的word level的CRF特征序列
双向LSTM神经网络
BERT预训练语言模型
BERT
输入
不 到 长 城 非 好 汉

图2