(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 109992770 A
(43)申请公布日 2019.07.09

(21)申请号 201910159512.5

(22)申请日 2019.03.04

(71)申请人 昆明理工大学
地址 650093 云南省昆明市五华区学府路253号

(72)发明人 周兰江 李炫达 张建安 满志博

(51)Int.Cl.
G06F 17/27(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

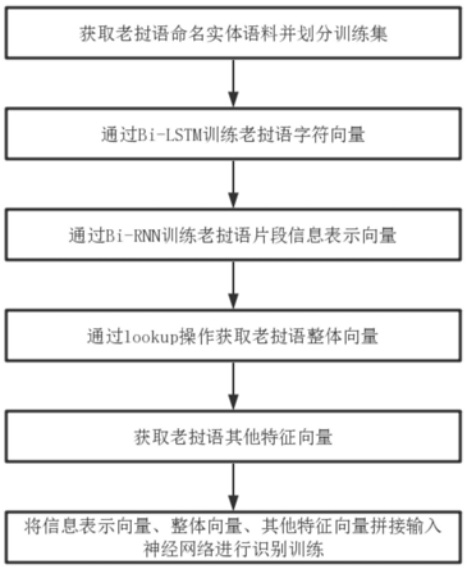
权利要求书2页 说明书4页 附图3页

(54)发明名称

一种基于组合神经网络的老挝语命名实体识别方法

(57)摘要

本发明公开了一种基于组合神经网络的老挝语命名实体识别方法,属于自然语言处理中小语种识别领域。首先利用Bi-LSTM(双向长短期记忆模型)将老挝语句子序列进行编码,输出字符向量。之后将字符向量进行切片分段,输入到Bi-RNN(双向循环神经网络)模型中,获得片段内部单元的信息表示向量。在此基础上使用lookup操作获取片段的整体向量表示,然后将获得的片段信息表示向量、整体向量、其他特征向量拼接作为特征输入到神经网络模型中,进行老挝语命名实体识别训练。本发明识别效果明显优于传统统计学习方法,并获得与当前其他最优的老挝语命名实体识别系统相当的识别性能。



CN 109992770 A

1.一种基于组合神经网络的老挝语命名实体识别方法,其特征在于:包括如下步骤:

Step1、将老挝语命名实体语料预处理后进行数据集划分,训练集占90%,测试集占10%;

Step2、将Step1预处理好的老挝语句子序列中的字符通过Bi-LSTM模型进行编码,输出字符向量;

Step3、将获得的字符向量序列进行切片分段,作为Bi-RNN的初始输入,通过前向循环神经网络和后向循环神经网络获得输出向量,连接构成片段内部单元的信息表示向量;

Step4、对片段通过lookup操作,从片段向量表中获得该片段的整体向量表示;

Step5、片段相关的其他特征向量包含片段上文切片分段相关信息和片段本身长度信息的特征,处理当前片段时,对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示;

Step6、将片段的内部信息表示向量、整体向量和其他特征向量拼接作为特征输入到神经网络模型中,进行老挝语命名实体识别训练。

2.根据权利要求1所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述方法的具体步骤为:所述步骤Step1中,数据通过老挝语留学生手动标注的老挝语语料进行训练,在30万字老挝语语料中,将90%作为训练集,10%作为测试集,其中将训练集的句子转化为“训练集句子-片段标记序列”作为模型输入的训练数据集,测试集是不包含任何切分信息的老挝语句子。

3.根据权利要求2所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述步骤Step2中,使用的Bi-LSTM模型来自google开发的tensorflow深度学习框架,将Step1预处理好的句子序列x中的字符x_i输入Bi-LSTM模型,输出向量C_i。

4.根据权利要求3所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述步骤Step3中,使用的Bi-RNN模型来自google开发的tensorflow深度学习框架,切分片段s_j获得的字符向量序列C进行切片分段,作为Bi-RNN模型的初始输入,通过模型的前向循环神经网络和后向循环神经网络获得输出向量,连接构成片段s_j内部单元的信息表示向量E_{unitj}。

5.根据权利要求4所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述步骤Step4中使用的lookup操作为lookup函数,对于当前片段s_j,通过lookup操作从片段向量表中获得该片段整体向量表示E_{segj},若当前片段在片段向量表中不存在,则选取特殊符号“UNKSEG”的向量表示,“UNKSEG”的初始值选取随机值。

6.根据权利要求5所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述步骤Step5中,片段相关的其他特征向量表示$F_{extendj}$主要是包含片段上文切分片段相关信息和片段本身长度信息的特征,处理当前片段时,对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示,若片段向量表中不存在查询的片段,则选用特殊符号“UNKPSEG”的向量表示,“UNKPSEG”取随机值初始化,片段长度特征信息通过查询片段长度特征表获得,每个长度值对应唯一的长度表示向量,初始向量值为随机值。

7.根据权利要求6所述基于组合神经网络的老挝语命名实体识别方法,其特征在于:所述步骤Step6中,使用Step3得到的片段内部信息表示向量、Step4获得的整体向量、Step5获得的其他特征向量进行拼接整合,作为提取出的老挝语命名实体识别特征,将特征输入到

神经网络中进行训练，设置神经网络模型的超参数，初始学习率设置为η_0，并使用SGD算法进行优化，训练过程中的正则化方法采用dropout技术，最终使用测试集对完成训练的模型进行测试，完成老挝语的命名实体识别。

一种基于组合神经网络的老挝语命名实体识别方法

技术领域

[0001] 本发明涉及一种基于组合神经网络的老挝语命名实体识别方法，属于自然语言处理中小语种识别领域。

背景技术

[0002] 命名实体识别(NER)是指从文本中识别出人名、地名和机构名等专有名词，是自然语言处理的关键技术之一，也是信息抽取、问答系统、句法分析、机器翻译等应用的重要基础工作。传统基于统计学习模型的命名实体识别方法严重依赖特征工程，特征设计需要大量人工参与和专家知识，而且已有的方法通常大多将中文命名实体识别任务看作一个字符序列标注问题，需要依赖局部字符标记区分实体边界。

发明内容

[0003] 本发明要解决的技术问题是提供一种基于组合神经网络的老挝语命名实体识别方法，用于解决老挝语命名实体识别准确率不高等问题。

[0004] 本发明采用的技术方案是：一种基于组合神经网络的老挝语命名实体识别方法，包括如下步骤：

[0005] Step1、将老挝语命名实体语料预处理后进行数据集划分，训练集占90％，测试集占10％；

[0006] Step2、将Step1预处理好的老挝语句子序列中的字符通过Bi-LSTM模型(双向长短期记忆模型)进行编码，输出字符向量；

[0007] Step3、将获得的字符向量序列进行切片分段，作为Bi-RNN(双向循环神经网络)的初始输入，通过前向循环神经网络和后向循环神经网络获得输出向量，连接构成片段内部单元的信息表示向量；

[0008] Step4、对片段通过lookup操作，从片段向量表中获得该片段的整体向量表示；

[0009] Step5、片段相关的其他特征向量包含片段上文切片分段相关信息和片段本身长度信息的特征，处理当前片段时，对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示；

[0010] Step6、将片段的内部信息表示向量、整体向量和其他特征向量拼接作为特征输入到神经网络模型中，进行老挝语命名实体识别训练。

[0011] 具体地，所述步骤Step1中，数据通过老挝语留学生手动标注的老挝语语料进行训练，在30万字老挝语语料中，将90％作为训练集，10％作为测试集，其中将训练集的句子转化为“训练集句子-片段标记序列”作为模型输入的训练数据集，测试集是不包含任何切分信息的老挝语句子。

[0012] 具体地，所述步骤Step2中，使用的Bi-LSTM模型来自google开发的tensorflow深度学习框架，将Step1预处理好的句子序列x中的字符x_i输入Bi-LSTM模型，输出向量C_i。

[0013] 具体地，所述步骤Step3中，使用的Bi-RNN模型来自google开发的tensorflow深度

学习框架，切分片段s_j获得的字符向量序列C进行切片分段，作为Bi-RNN模型的初始输入，通过模型的前向循环神经网络和后向循环神经网络获得输出向量，连接构成片段s_j内部单元的信息表示向量E_{unitj}。

[0014] 具体地，所述步骤Step4中使用的lookup操作为lookup函数，对于当前片段s_j，通过lookup操作从片段向量表中获得该片段整体向量表示E_{segj}，若当前片段在片段向量表中不存在，则选取特殊符号“UNKSEG”的向量表示，“UNKSEG”的初始值选取随机值。

[0015] 具体地，所述步骤Step5中，片段相关的其他特征向量表示$F_{extendj}$主要是包含片段上文切分片段相关信息和片段本身长度信息的特征，处理当前片段时，对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示，若片段向量表中不存在查询的片段，则选用特殊符号“UNKPSEG”的向量表示，“UNKPSEG”取随机值初始化，片段长度特征信息通过查询片段长度特征表获得，每个长度值对应唯一的长度表示向量，初始向量值为随机值。

[0016] 具体地，所述步骤Step6中，使用Step3得到的片段内部信息表示向量、Step4获得的整体向量、Step5获得的其他特征向量进行拼接整合，作为提取出的老挝语命名实体识别特征，将特征输入到神经网络中进行训练，设置神经网络模型的超参数，初始学习率设置为η_0，并使用SGD算法进行优化，训练过程中的正则化方法采用dropout技术，最终使用测试集对完成训练的模型进行测试，完成老挝语的命名实体识别。

[0017] 本发明的有益效果是：

[0018] 1、该基于组合神经网络的片段级老挝语命名实体识别方法中，相比神经网络的字符级老挝语命名实体识别方法识别精度有明显提高。

[0019] 2、该基于组合神经网络的片段级老挝语命名实体识别方法中，使用了Bi-LSTM、Bi-RNN和准神经层构成的组合神经网络，相比一般的Bi-LSTM方法，可以有效避免向量填充(padding)，减少人工设置参数对系统的影响和限制，并在特征提取效果上有了比较不错的提高。

[0020] 3、该基于组合神经网络的片段级老挝语命名实体识别方法中，识别效果明显优于传统统计学习方法，并获得与当前其他最优的老挝语命名实体识别系统相当的识别性能。

附图说明

[0021] 图1为本发明中的总体流程图；

[0022] 图2为组合神经网络获取片段内部单元的信息表示向量模型结构图；

[0023] 图3为获取片段表示向量的组合神经网络结构图。

具体实施方式

[0024] 下面结合附图和具体实施例对本发明作进一步的说明。

[0025] 实施例1：如图1-3所示，一种基于组合神经网络的老挝语命名实体识别方法，包括如下步骤：

[0026] Step1、将老挝语命名实体语料预处理后进行数据集划分，训练集占90％，测试集占10％；

[0027] Step2、将Step1预处理好的老挝语句子序列中的字符通过Bi-LSTM模型进行编码，

输出字符向量；

[0028] Step3、将获得的字符向量序列进行切片分段，作为Bi-RNN的初始输入，通过前向循环神经网络和后向循环神经网络获得输出向量，连接构成片段内部单元的信息表示向量；

[0029] Step4、对片段通过lookup操作，从片段向量表中获得该片段的整体向量表示；

[0030] Step5、片段相关的其他特征向量包含片段上文切片分段相关信息和片段本身长度信息的特征，处理当前片段时，对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示；

[0031] Step6、将片段的内部信息表示向量、整体向量和其他特征向量拼接作为特征输入到神经网络模型中，进行老挝语命名实体识别训练。

[0032] 进一步地，所述步骤Step1中，数据通过老挝语留学生手动标注的老挝语语料进行训练，在30万字老挝语语料中，将90％作为训练集，10％作为测试集，其中将训练集的句子转化为“训练集句子-片段标记序列”作为模型输入的训练数据集，测试集是不包含任何切分信息的老挝语句子。

[0033] 进一步地，所述步骤Step2中，使用的Bi-LSTM模型来自google开发的tensorflow深度学习框架，将Step1预处理好的句子序列x中的字符x_i输入Bi-LSTM模型，输出向量C_i，如图2所示。

[0034] 进一步地，所述步骤Step3中，使用的Bi-RNN模型来自google开发的tensorflow深度学习框架，切分片段s_j获得的字符向量序列C进行切片分段，作为Bi-RNN模型的初始输入，通过模型的前向循环神经网络和后向循环神经网络获得输出向量，连接构成片段s_j内部单元的信息表示向量E_{unitj}，如图2所示。

[0035] 进一步地，所述步骤Step4中使用的lookup操作为lookup函数，对于当前片段s_j，通过lookup操作从片段向量表中获得该片段整体向量表示E_{segj}，若当前片段在片段向量表中不存在，则选取特殊符号“UNKSEG”的向量表示，“UNKSEG”的初始值选取随机值。

[0036] 进一步地，所述步骤Step5中，片段相关的其他特征向量表示$F_{extendj}$主要是包含片段上文切分片段相关信息和片段本身长度信息的特征，处理当前片段时，对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示，若片段向量表中不存在查询的片段，则选用特殊符号“UNKPSEG”的向量表示，“UNKPSEG”取随机值初始化，片段长度特征信息通过查询片段长度特征表获得，每个长度值对应唯一的长度表示向量，初始向量值为随机值。

[0037] 进一步地，所述步骤Step6中，使用Step3得到的片段内部信息表示向量、Step4获得的整体向量、Step5获得的其他特征向量进行拼接整合，作为提取出的老挝语命名实体识别特征，将特征输入到神经网络中进行训练，设置神经网络模型的超参数，初始学习率设置为η_0，并使用SGD算法进行优化，训练过程中的正则化方法采用dropout技术，最终使用测试集对完成训练的模型进行测试，完成老挝语的命名实体识别。

[0038] 本发明提出一种基于组合神经网络的片段级老挝语命名实体识别方法，减弱了系统对人工特征设计的依赖，避免字符序列化标注方法的不足，通过采用深度学习片段神经网络结构，实现特征的自动学习，并通过获取片段信息对片段整体分配标记，同时完成实体边界识别和分类。

[0039] 以上结合附图对本发明的具体实施方式作了详细说明，但是本发明并不限于上述实施方式，在本领域普通技术人员所具备的知识范围内，还可以在不脱离本发明宗旨的前提下作出各种变化。

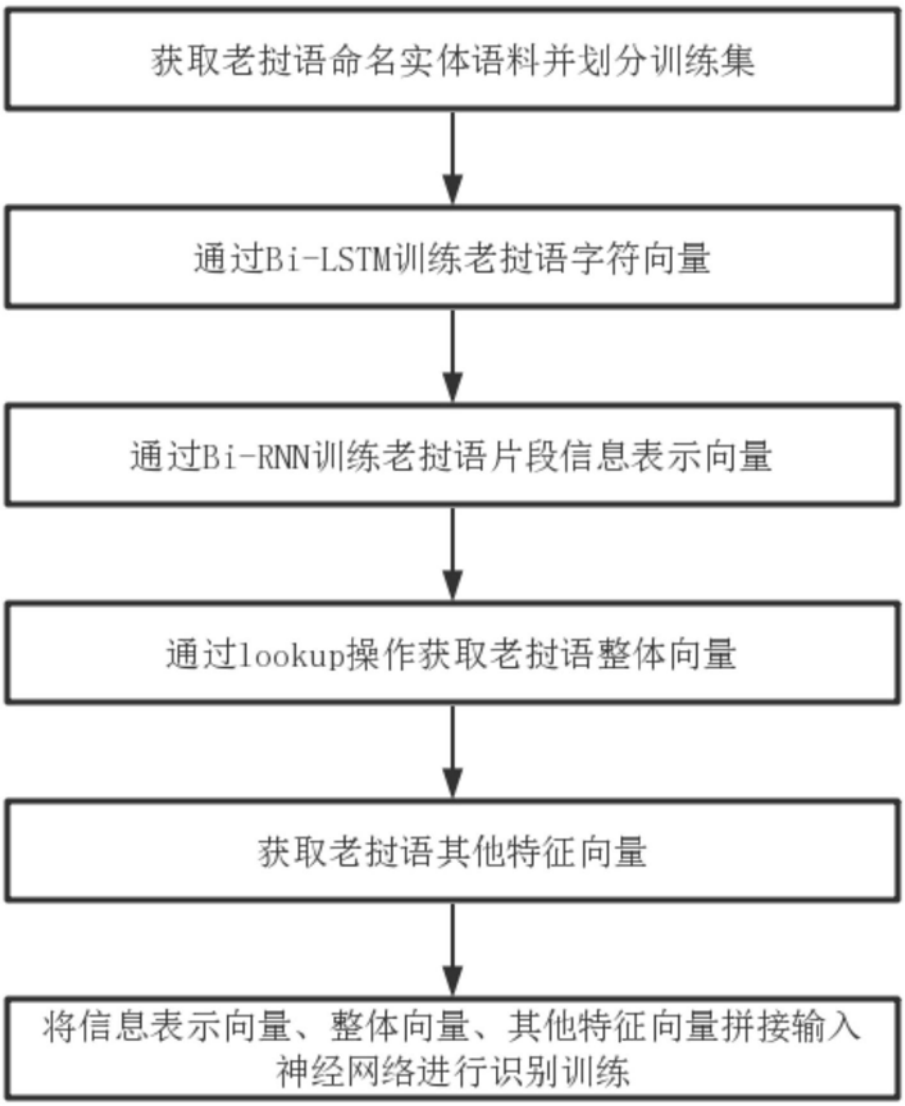

获取老挝语命名实体语料并划分训练集
通过Bi-LSTM训练老挝语字符向量
通过Bi-RNN训练老挝语片段信息表示向量
通过lookup操作获取老挝语整体向量
获取老挝语其他特征向量
将信息表示向量、整体向量、其他特征向量拼接输入神经网络进行识别训练

图1

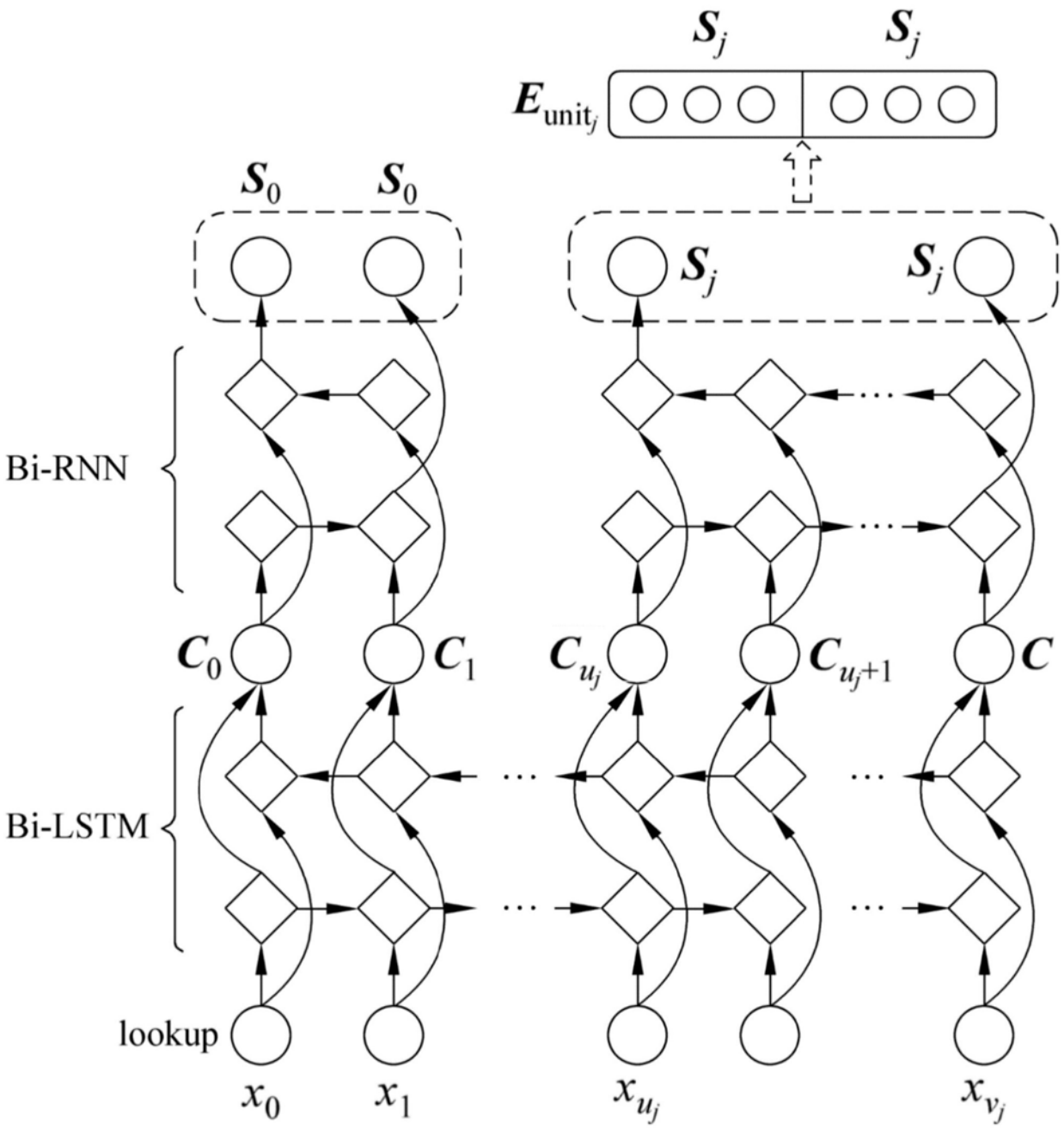

S_j
S_j
E_{unit_j}
S_0
S_0
S_j
S_j
Bi-RNN
C_0
C_1
C_{u_j}
C_{u_j+1}
C
Bi-LSTM
lookup
x_0
x_1
x_{u_j}
x_{v_j}

图2

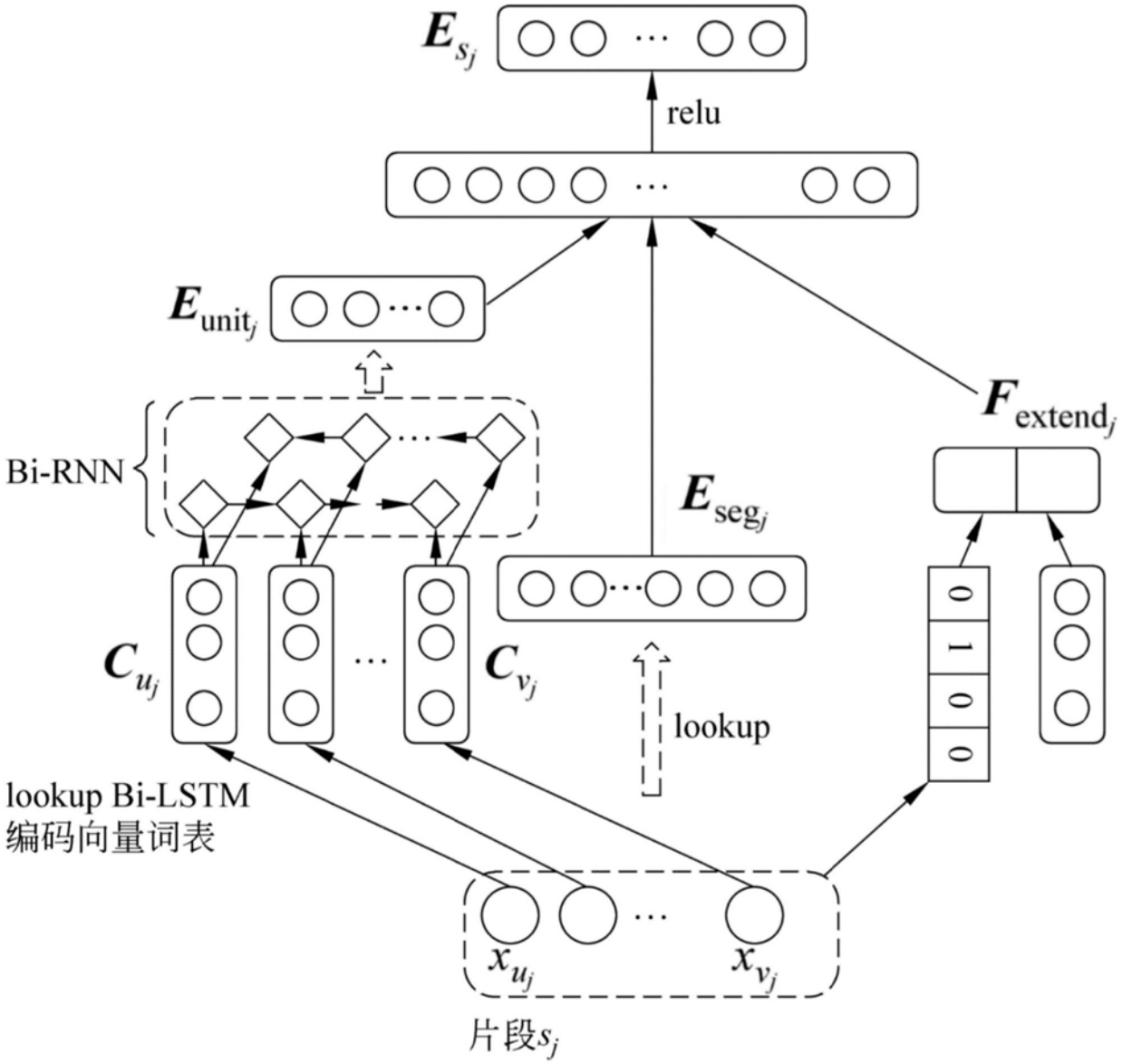

E_{s_j}
relu
E_{unit_j}
Bi-RNN
E_{seg_j}
F_{extend_j}
C_{u_j}
C_{v_j}
lookup
0
1
0
0
lookup Bi-LSTM
编码向量词表
x_{u_j}
x_{v_j}
片段s_j

图3