(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 113065349 A
(43)申请公布日 2021.07.02

(21)申请号 202110274547.0

(22)申请日 2021.03.15

(71)申请人 国网河北省电力有限公司
地址 050022 河北省石家庄市富强大街32号
申请人 国网河北省电力有限公司雄安新区供电公司

(72)发明人 刘义江 李云超 姜琳琳 吴彦巧 姜敬 檀小亚 师孜晗 陈蕾 侯栋梁 池建昆 范辉 阎鹏飞 魏明磊 辛锐 陈曦 杨青 沈静文

(74)专利代理机构 石家庄新世纪专利商标事务所有限公司 13100
代理人 董金国 黄敬霞

(51)Int.Cl.
G06F 40/284(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/04(2006.01)

权利要求书2页 说明书7页 附图4页

(54)发明名称

基于条件随机场的命名实体识别方法

(57)摘要

本发明属于自然语言处理技术领域，涉及一种基于条件随机场的命名实体识别方法，该方法包括：接收包含中文文本的词语序列；词语序列中各词语按照其在原始语句的上下文顺序排列；使用命名实体识别网络的词向量模块将词语序列编码为词向量组；词向量组包含了各个词语的命名实体特征信息；使用命名实体识别网络的长短记忆网络模块提取词向量组中各个词向量的序列特征，并输出为命名实体分类空间的状态分数矩阵；使用命名实体识别网络的条件随机场模块查找状态分数矩阵中得分最高的分数路径作为词语序列中各词语的命名实体预测结果输出。本发明提供了一种高效的命名实体识别方法，特别是对财务票据图片中涉及的人名、组织机构名称和地点进行实体识别。

CN 113065349 A

1.一种基于条件随机场的命名实体识别方法,由处理器执行程序指令实现,该方法包括:

接收包含中文文本的词语序列;所述词语序列中各词语按照其在原始语句的上下文顺序排列;

使用命名实体识别网络的词向量模块将所述词语序列编码为词向量组;所述词向量组包含了各个词语的命名实体特征信息;

使用命名实体识别网络的长短记忆网络模块提取所述词向量组中各个词向量的序列特征,并输出为命名实体分类空间的状态分数矩阵;

使用命名实体识别网络的条件随机场模块查找所述状态分数矩阵中得分最高的分数路径作为所述词语序列中各词语的命名实体预测结果输出。

2.根据权利要求1所述的命名实体识别方法,其特征在于,对所述命名实体识别网络的训练包括:对所述词向量模块的训练;以及,对所述长短记忆网络模块和所述条件随机场模块的同时训练。

3.根据权利要求2所述的命名实体识别方法,其特征在于,对所述词向量模块的训练包括:使用one-hot编码的词库对所述词向量模块的神经网络进行基于预设命名实体分类的编码训练;使用语料库语句中连续固定长度的词语序列对所述词向量的神经网络进行基于判断所述词语序列中以词语是否相邻词语是否的编码训练。

4.根据权利要求3所述的命名实体识别方法,其特征在于,所述词语序列的固定长度为3。

5.根据权利要求4所述的命名实体识别方法,其特征在于,对所述词向量模块进行基于判断所述词语序列中以词语是否相邻词语是否的编码训练,损失L配置为由基于分类任务的交叉熵损失函数计算:

$$L = -[y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

其中,y代表真实值,$\hat{y}$代表预测值,两者取值范围都是{0,1}。

6.根据权利要求2所述的命名实体识别方法,其特征在于,对所述长短记忆网络模块和所述条件随机场模块的同时训练包括:使用由所述词向量模块输出的词向量组做为样本进行训练。

7.根据权利要求1至6任一项所述的命名实体识别方法,其特征在于,由长短记忆网络模块对所述词向量组的各个词向量进行基于上下文信息的分类打分,以获得各个词向量的分类向量;将所述分类向量组合为状态分数矩阵。

8.根据权利要求1至6任一项所述的命名实体识别方法,其特征在于,所述条件随机场模块配置为使用其转移矩阵将所述状态分数矩阵中各分类向量解码为包含全部分类路径的状态分数矩阵,并选择各个路径中得分和最大的路径为输出。

9.根据权利要求6所述的命名实体识别方法,其特征在于,所述长短记忆网络模块和所述条件随机场模块进行训练时,整体神经网络的损失函数Loss被设计为:

$$Loss = -log(\frac{P_{real-path}}{P_1 + P_2 + \cdots + P_n})$$

其中,P_1,P_2,…,P_n为根据在条件随机向量场计算得到各个路径,假设$P_{real-path}$为其中最

大路径。

10.根据权利要求1所述的命名实体识别方法，其特征在于，所述命名实体特征信息的分类维度配置为PER、ORG、LOC和O四个维度。

基于条件随机场的命名实体识别方法

技术领域

[0001] 本发明属于自然语言处理技术领域，尤其涉及一种命名实体识别系统和方法。

背景技术

[0002] 为了在数量庞大的文本内容中找出指定的信息是自然语言处理中要面对的一个重要的技术问题。由于海量数据的出现，人工查找定位信息的工作量太过庞大，已经不符合人们的要求，运用计算机技术从海量数据中迅速定位相关信息的技术应运而生，自然语言处理作为定位信息的基本处理方法，近年来迅速成为研究焦点。

[0003] 命名实体识别(NER)的主要任务是识别文本中有意义的专有名词及数量短语，如，人名、地名、组织名、时间、日期、货币等等，对于句法分析、语法分析、语义分析等都有着及其重要的影响；此外，命名实体识别又是信息抽取、信息过滤、信息检索、组块分析、问答系统、机器翻译等技术的重要基础。

[0004] 英文命名实体识别技术目前已经达到了较高的水平，有些系统已经实用化。和英文命名实体识别相比，中文命名实体识别就要落后和困难很多。主要包括但是不限于：中文词没有空格分词，人名构成形式多样，中文语料匮乏，文档中的标点，文本格式等都会对命名实体产生影响。

[0005] 与大多数的自然语言处理技术一样，命名实体识别的方法主要可以分为两大类：基于规则的方法和基于统计的方法。一般来说，基于规则的方法更接近人的思维方式，表示更直观自然，便于推理。但是规则的编写往往依赖于具体的语言和领域，其可移植性比较差，编写规则的过程也比较耗时耗力，而且规则很难编写完备，需要具体的领域专家及语言专家才能完成，总的性价比不高。和基于规则的方法相比，基于统计的方法就要灵活很多，其方法比较客观，不需要太多的人工干预及其领域知识，但在时空开销等性能方面存在问题。

发明内容

[0006] 本发明综合考虑了上述现有技术方案的缺点，目的在于提供一种高效的命名实体识别方法，特别是对混合中文文本的财务票据图片中主要涉及的人名、组织机构名称和地点进行实体识别。

[0007] 本发明提供的技术方案是一种基于条件随机场的命名实体识别方法，由处理器执行程序指令实现，所述程序指令包括由词向量(word2vec)模块、长短期记忆网络(LSTM)模块和条件随机场(CRF)模块组成的命名实体识别网络的实现指令。该方法包括：

[0008] 接收包含中文文本的词语序列；所述词语序列中各词语按照其在原始语句的上下文顺序排列；

[0009] 使用命名实体识别网络的词向量模块将所述词语序列编码为词向量组；所述词向量组包含了各个词语的命名实体特征信息；

[0010] 使用命名实体识别网络的长短记忆网络模块提取所述词向量组中各个词向量的

序列特征，并输出为命名实体分类空间的状态分数矩阵；

[0011] 使用命名实体识别网络的条件随机场模块查找所述状态分数矩阵中得分最高的分数路径作为所述词语序列中各词语的命名实体预测结果。

[0012] 优选的，对所述命名实体识别网络的训练包括：对所述词向量模块的训练；以及，对所述长短记忆网络模块和所述条件随机场模块的同时训练。

[0013] 优选的，对所述词向量模块的训练包括：使用one-hot编码的词库对所述词向量模块的神经网络进行基于预设命名实体分类的编码训练；使用语料库语句中连续固定长度的词语序列对所述词向量的神经网络进行基于判断所述词语序列中以词语是否相邻词语是否的编码训练；进一步的优选的，所述词语序列的固定长度为3。

[0014] 优选的，对所述词向量模块进行基于判断所述词语序列中以词语是否相邻词语是否的编码训练，损失L配置为由基于分类任务的交叉熵损失函数计算：

[0015] $$L = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$$

[0016] 其中，y代表真实值，$\hat{y}$代表预测值，两者取值范围都是{0,1}。

[0017] 优选的，对所述长短记忆网络模块和所述条件随机场模块的同时训练包括：使用由所述词向量模块输出的词向量组做为样本进行训练。

[0018] 优选的，由长短记忆网络模块对所述词向量组的各个词向量进行基于上下文信息的分类打分，以获得各个词向量的分类向量；将所述分类向量组合为状态分数矩阵。

[0019] 优选的，所述条件随机场模块配置为使用其转移矩阵将所述状态分数矩阵中各分类向量解码为包含全部分类路径的状态分数矩阵，并选择各个路径中得分和最大的路径为输出。

[0020] 优选的，所述长短记忆网络模块和所述条件随机场模块进行训练时，整体神经网络的损失函数Loss被设计为：

[0021] $$Loss = -log(\frac{P_{real-path}}{P_1 + P_2 + \cdots + P_n})$$

[0022] 其中，P_1，P_2，…，P_n为根据在条件随机向量场计算得到各个路径，假设$P_{real\text{-}path}$为其中最大路径。

[0023] 优选的，所述命名实体特征信息的分类维度配置为PER、ORG、LOC和O四个维度

[0024] 本发明一些实施例中，词向量模块配置为，假定其某个输入语句由n个词语[w_1，w_2，…，w_n]依序排列组成，则词向量模块主要用于对这n个词语进行编码，并得到编码后的各个词语在词空间中对应的向量[e_1，e_2，…，e_n]，其中词向量e_i对应于词语w_i，i∈n。一组词向量中各个向量依序作为长短记忆网络模块的输入，由长短记忆网络模块进行基于上下文信息的分类打分。本发明中一些实施例中，长短记忆网络模块基于双向LSTM网络(BiLSTM)结构配置，依次读取词向量模块输出的一组上下文词向量中的各个词向量，并考虑每个词的上下文信息，对这些编码后的词向量提取序列特征，对每一个词进行基于分类的打分，并输出一个分类空间中的分数向量。本发明另一些实施例中，基于任何现有技术中改进的可以处理上下文信息的长短记忆网络的配置长短记忆网络模块。本发明中，分类配置为实体命名，如人名、组织机构名称和地点即为三个方向上的分类，在一个分类的打分越高，代表长短记忆网络模块认为该词向量越有可能属于该分类。各个分类视为一个维度时，预配置的

全部分类形成本发明的分类空间。分数向量中最高得分分量所对应的类别代表该词属于哪类命名实体。为了综合考虑LSTM的输出作为一个整体序列的现实意义，本发明采用条件随机场模块对LSTM输出的打分进行进一步评价，选出得分最高的打分路径作为最后的预测结果。

[0025] 一些实施例中，特征生成与选择利用条件随机场模型生成并选择特征，将最终选择的特征保存在特征库里。由于命名实体本身的构成具有很强的随意性，仅仅依靠对命名实体本身结构和用字的分析很难取得较好的识别效果，因此，需要充分挖掘命名实体上下文的相关信息，条件随机场能够表达长距离的上下文依赖信息，并有效的将各种相关或不相关信息融合在一起。

[0026] 现有技术中，在条件随机场模型中，问题的特征空间一般都很大，对于训练语料，生成的特征往往成千上万，但并非所有的特征都是有用的，过多的特征反而会影响系统的运行效率，因此特征选择就是一个关键的问题。常见的有两种特征选择的方法:增量法和阈值法。增量法的基本思想是对每个特征计算其信息增益，如果加入的特征能够改善系统的性能，则保留此特征，否则删除此特征，增量法的特征选择效果较好，但是额外的时空开销也比较大。阈值法则是对特征出现的频度进行计数，如果某一特征的在语料中的总的出现次数小于某一设定阈值，则删除这一特征，否则保留此特征，阈值法的操作简单实用，但其不能保谨所选出的特征集合是最小完备的，很可能会存在一些冗余特征，这会降低系统的运行效率。本发明的一些优选实施例采用阈值法，且阈值设定为2。优选的，模型参数训练用特征库中的特征和训练语料采用L-BFGS方法进行参数训练得到模型参数；优选的，命名实体识别采用训练好的条件随机场模型进行命名实体识别并将标记命名实体的结果输出。

[0027] 本发明采取统计算法条件随机场进行中文命名实体识别，条件随机场继承了最大熵模型的优点，可以有效的整合各种相关或者不相关的上下文信息，将外部语义知识加入到条件随机场模型的特征选择之中，有效地提高了命名实体识别的准确率。本发明不必从有标注的语料中抽取一般的命名实体指示词，然后通过扩展来增加命名实体指示词的数量。

附图说明

[0028] 图1为本发明一实施例中实现基于条件随机场的命名实体识别方法的命名实体识别系统的结构示意图；

[0029] 图2为本发明一实施例中命名实体识别系统词向量模块的神经网络结构示意图；

[0030] 图3为本发明一实施例中命名实体识别系统长短记忆网络模块的神经网络结构示意图；

[0031] 图4为本发明一实施例中长短记忆网络模块中一个LSTM模块的神经网络结构示意图

[0032] 图5为本发明一实施例中命名实体识别系统条件随机场模块的神经网络工作原理示意图；

[0033] 图6为本发明一实施例中命名实体识别系统中数据处理过程示意图；

[0034] 图7为本发明一实施例中实施例中实现基于条件随机场的命名实体识别方法的流程图。

具体实施方式

[0035] 首先需要说明的是，现有技术中，用于命名实体识别的条件随机场模型，需要提起使用才有标注的外部语义库进行训练，条件随机场模型能够融合长距离的上下文相关信息，这种上下文信息用特征函数来描述，特征选择直接影响着条件随机场模型的性能。在中文中，语义特征是一类重要的上下文信息，但是中文中的这类语义信息往往隐含在上下文中需要深度挖掘，现有技术中使用中文中的语义信息并建立外部语义库以供条件随机场模型使用，这些建立供命名实体识别使用的外部语义库包括可能包含人名指示词库、中国人名姓氏表、常用人名表；地名指示词、常用地名表；组织名指示词、常见组织名列表、组织名特征后缀等。本发明考虑具体应用场景下的性能指标，在较少分类需求的情况下，通过训练进行词向量编码的神经网络，避免了对训练样本的标注处理。

[0036] 本文中，除非特别说明，否则，涉及深度学习网络模型、word2vec模型等所涉及术语“模型”指由计算机程序指令序列实现的一整套数学算法，这些计算机程序指令序列在由处理器读取并用于通过不同的配置参数以及限定的指定输入数据，可以用于实现对于计算机数据的处理，以实现指定的技术作用。本领域技术人员习惯于通过一些形象的可视化结构描述一类神经网络模型内部的配置有具体算法指令各个功能单元(数字神经元)之间逻辑上的输入输出关系，这些描述结构的图示虽然被称为神经网络，但本领域技术人员可以清楚明确理解到其实际要说明的数学算法的指令实现，在以下揭示本发明构思的具体实施方式中，一些具体的实现功能代码由于本领域技术人员在了解具体构思后可以借由本领域常识予以具体实施，因而在这些实施方式中不再赘述。

[0037] 参考图1至7，本实施例是一种基于条件随机场的命名实体识别方法，由执行于处理器的命名实体识别系统1000实现，所述处理器接收包含中文词汇的中文语句数据，并通过运行命名实体识别系统1000的程序。命名实体识别系统1000的运行的程序指令包括实现由词向量模块1001、长短期记忆网络模块1002和条件随机场模块1003等程序模块组成的命名实体识别网络的程序指令。

[0038] 示范的，词向量模块采用word2vec来对中文汉字和字母符号进行编码，本实施例中中文字库包括约7000个常用汉字，这种编码器的好处是可以对one-hot编码向量进行合理的降维操作。本实施例词向量模块训练时用到的词向量模型网络结构如图2所示，其中，Layer1网络层用于获得需词语序列中各个词语的词向量，词向量网络输入为：词语的one-hot编码，本实施例中，采用百度百科800w+条词条语句来作为训练使用的语料库，首先利用开源组件库“jieba分词”对整个语料库的每个句子完成分词。我们采用CBOW模式训练词向量网络模型。

[0039] 本实施例的一个优选方案是，在训练时，词向量模块，每次选取词语序列中3个词各自对应的one-hot向量作为词向量网络Layer1的输入，经过两层神经网络(Layer1、layer2)来进行判断中间的词是否可以作为上下两个词的中间填充词。训练中，损失函数采用分类任务常用的交叉熵损失函数：

[0040] $$L = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$$

[0041] 其中，y代表真实值，$\hat{y}$代表预测值，两者取值范围都是{0,1}，以便实现训练时图2中Layer2的二值分类。训练完成后，仅使用layer1输出本发明的词语序列中各词语的词向量，可见Layer1被训练为判断连续词语序列中各词语在本发明四维分类空间中的位置信

息,以及与前后相邻词语作为填充词的适配度信息。

[0042] 本实施例中,参考图3,长短记忆网络模块中的BiLSTM的输入是每个句子切分后对应的词向量,输出每个词的得分向量,该得分向量为4维向量,各维对应本实施中四个命名实体的分类,本实施例仅考虑对PER、ORG、LOC三种命名实体的特征提取,其余词可能的命名实体用O表示,即形成一个4维的命名实体分类空间。其中,对于输入的向量序列[x_1,x_2,…x_i],序随时间增大,在A层的LSTM模块循环考虑一个输入的与其上文的相关性信息,即,沿正序依次进行隐层状态s的迭代,A′层的LSTM模块考虑了一个输入与其下文的相关性信息,即,沿反序进行隐层状态s′的迭代。可以看出,实际上,由于LSTM类循环神经网络的特殊模型结构,其输入层对输入的序列长度本身没有限制。

[0043] 具体的,本实施例的单个LSTM模块如图4所示:整个模块在第t步接收来自上一步的细胞体状态C_{t-1},隐层装填H_{t-1},输出下一层的隐层H_t和细胞体状态C_t。图中M代表sigmoid激活函数:$M=\frac{1}{1+e^{-x}}$, $tanh=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$,其余部分计算公式如下:

[0044] $F_t=M(W_f\cdot[h_{t-1},x_t]+b_f)$

[0045] $I_t=M(W_i\cdot[h_{t-1},x_t]+b_i)$

[0046] $L_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c)$

[0047] $C_t=F_t*C_{t-1}+I_t*L_t$

[0048] $O_t=M(W_o\cdot[h_{t-1},x_t]+b_o)$

[0049] $H_t=O_t*\tanh(C_t)$

[0050] 其中,各个W和b为其下标线性关系模型中的系数和偏移,*为Hadamard积,x_t为t步的输入。

[0051] 示范的,本实施例中,条件随机场模块使用其条件随机场模型处理过程是:中文词语序列[w_1,w_2,w_3]经过训练好的词向量模块处理为一组词向量[v_1,v_2,v_3],该输入经过双向LSTM层的编码之后获得,会得到每个词向量的命名实体分类各个维度上的得分,即得分向量。条件随机场模块接收LSTM的得分向量,结合状态转移分数矩阵作为输入,计算全部路径得分,取最高得分的路径对应的序列作为最终的预测结果,并进一步从所有路径中选取综合概率最高的路径对应的分类结果。本发明利用到CRF作为解码器,条件随机场模块工作原理如图5所示。示范的,假设长短记忆模块的输出包含三个分类向量:a[0.1,0.2,0.8,0],b[0.4,0.2,0.2,0],c[0.6,0.1,0.2,0.1],每个分类向量中的四个值分别对应PER,ORG,LOC,O这四个分类。第一个路径(org,org,org)对应的得分为:a[0]+b[0]+c[0]+$FF_{start-org}$+$FF_{org-org}$+$FF_{org-org}$+$FF_{org-end}$。在分类空间维度为4时,容易算出来状态分数矩阵中的全部路径共有4×4×4×4种。

[0052] 容易理解,本发明中,当条件随机场模块接收来自长短记忆网络模块的输出[y_1,y_2,y_3,…,y_n]作为其状态分数,转移状态分数向量为一个6×6的状态分数矩阵,如下表示出本实施例的状态分数矩阵,其中START行、列和END行、列分别为其中长短记忆循环的开始和结束的编码标识,对训练结果不敏感,作为条件随机场选择路径的起点和终点。

[0053]

	Start	Per	Org	O	Loc	End
Start	…	…	…	…	…	…
Per	…	…	…	…	…	…

Org	…	…	…	…	…	…
O	…	…	…	…	…	…
Loc	…	…	…	…	…	…
End	…	…	…	…	…	…

[0054] 假设经过LSTM我们的得到的预测路径序列XL:[ORG,O,PER],该路径得分为:s＝$F_{org}+F_{o}+F_{per}+FF_{start-org}+FF_{org-o}+FF_{o-per}+FF_{per-end}$,其中F代表LSTM输出的分数,FF代表状态转移分数矩阵(即本发明中条件随机场的转移矩阵)。本实施例中,条件随机场模块配置为通过计算上述方法获得的最大得分路径得到最终的预测路径序列。

[0055] 本实施例中,在词向量模块训练完毕后,将其作为输入样本的固定预处理,使用其各个输出对后续的长短记忆模块和条件随机场模块同时进行训练,BILSTM需要训练的是各个细胞体的系数和偏移,CRF需要训练的参数是状态转移分数矩阵.训练时,在CRF计算得到每个路径$P_1,P_2,\cdots,P_n$之后,假设其中最大路径为$P_{real-path}$,训练时整体神经网络的损失函数Loss被设计为:

[0056] $$Loss = -log(\frac{P_{real-path}}{P_1 + P_2 + \cdots + P_n})$$

[0057] 最小化这个函数即可对条件随机场模块和长短记忆网络模块包含的神经网络整体进行监督训练,具体的,本实施例中,利用Adam优化器进行参数学习。

[0058] 在通过上述训练对本实施例命名实体识别网络中各个神经网络训练结束后,可通过本发明提供的命名实体识别方法获得词语序列的命名实体识别结果。下面通过本实施例一个完整的实施过程以充分揭示本发明提供技术方案,包括步骤100至400。

[0059] 步骤100,接收包含中文文本的词语序列;所述词语序列中各词语按照其在原始语句的上下文顺序排列。具体的,本实施例中词语序列通过以下方式获得:检测并识别票据图片中的文本内容,并转化为可用于计算机处理的第一文本字符串;使用外部分词工具对文本字符串进行分词,生成具备分词标识的第二字符串;根据分词表示提取第二字符串连续的三个词语组成词语序列。示范的,第一文本字符串为按标点符号切分的句子,以“张三在天安门。”为例,利用jieba分词开源工具对整个句子进行分词,得到“张三,在,天安门”三个词语,使用与训练词向量模块时相同的方式,对词语进行one-hot编码,获得形为$[w_1,w_2,w_3]$的词语序列。

[0060] 步骤200,使用命名实体识别网络的词向量模块将所述词语序列编码为词向量组;所述词向量组包含了各个词语的命名实体特征信息。利用已训练好的词向量模块对三个词语分别进行编码得到一组三个词向量$[v_1,v_2,v_3]$。

[0061] 步骤300,使用命名实体识别网络的长短记忆网络模块提取所述词向量组中各个词向量的序列特征,并输出为命名实体分类空间的状态分数矩阵。具体的,通过已训练好的长短记忆网络模块的BiLSTM得到每个输入词向量的一个得分向量,接着长短记忆网络模块的后处理部分将这个得分向量作为状态分数,与相应的转移矩阵一起发送给条件随机场模块。

[0062] 步骤400,使用命名实体识别网络的条件随机场模块查找所述状态分数矩阵中得分最高的分数路径作为所述词语序列中各词语的命名实体预测结果输出。训练好的条件随机场模块根据得分序列和训练获得的转移矩阵获得状态分数矩阵,取分数合最大的序列路

径即为最终的路径，如，路径序列XL:[PER,O,LOC]

[0063] 上述实施步骤仅为示范的，各步骤实施时间依赖于其前置条件而非步骤体现的时间顺，如对词向量模块的中神经网络的训练可以是提前做好的，并非必须在接词语序列之后实施。

[0064] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0065] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统和方法，可以通过其它的方式实现。例如，以上所描述的系统实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，如对外部神经网络单元的调用，可以是本地的，远程的或混合的资源配置形式。

[0066] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0067] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理设备中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个处理设备中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0068] 所述集成的模块如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM,Read-OnlyMemory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0069] 以上对本发明所提供的一种基于条件随机场的命名实体识别方法进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

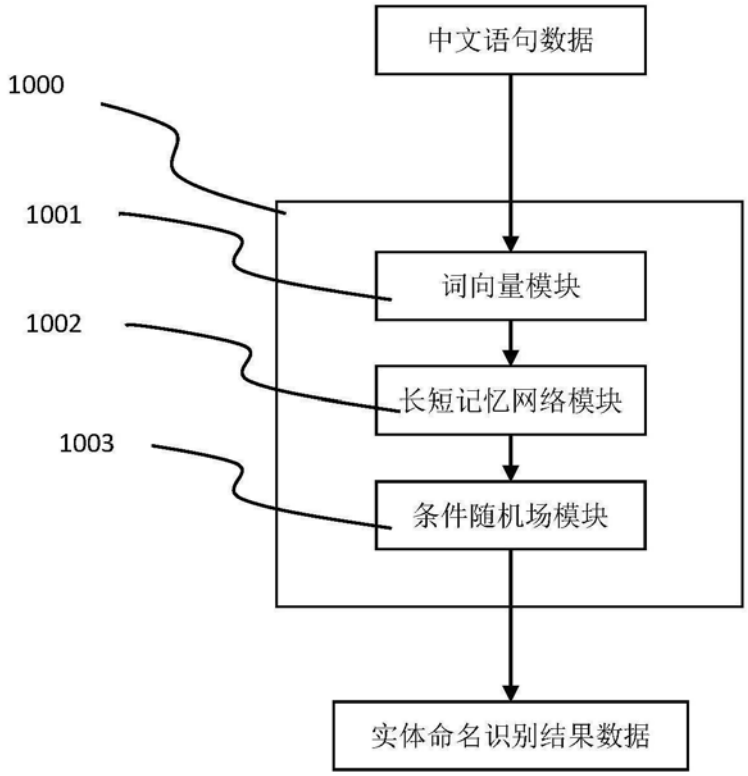
中文语句数据
1000
1001
词向量模块
1002
长短记忆网络模块
1003
条件随机场模块
实体命名识别结果数据

图1

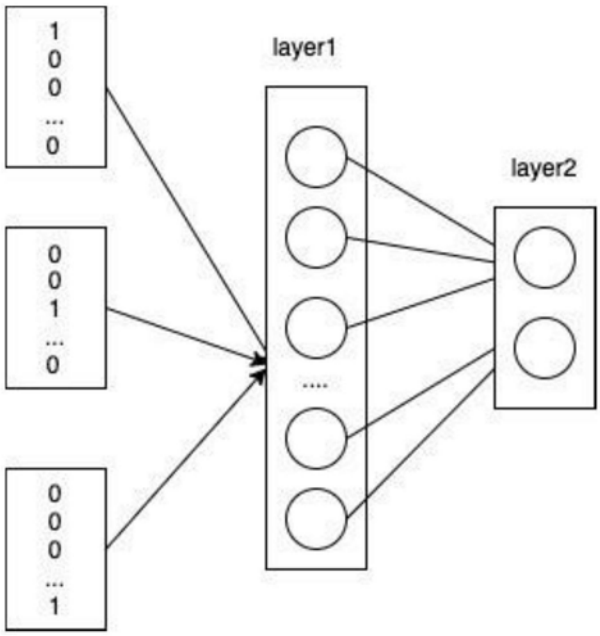
1
0
0
...
0
0
0
1
...
0
0
0
0
...
1
layer1
layer2

图2

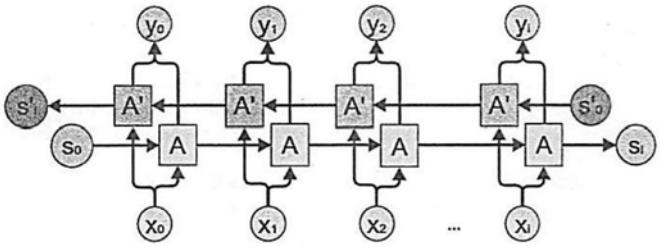
y0
y1
y2
yi
A'
A
s'
s0
si
x0
x1
x2
...
xi

图3

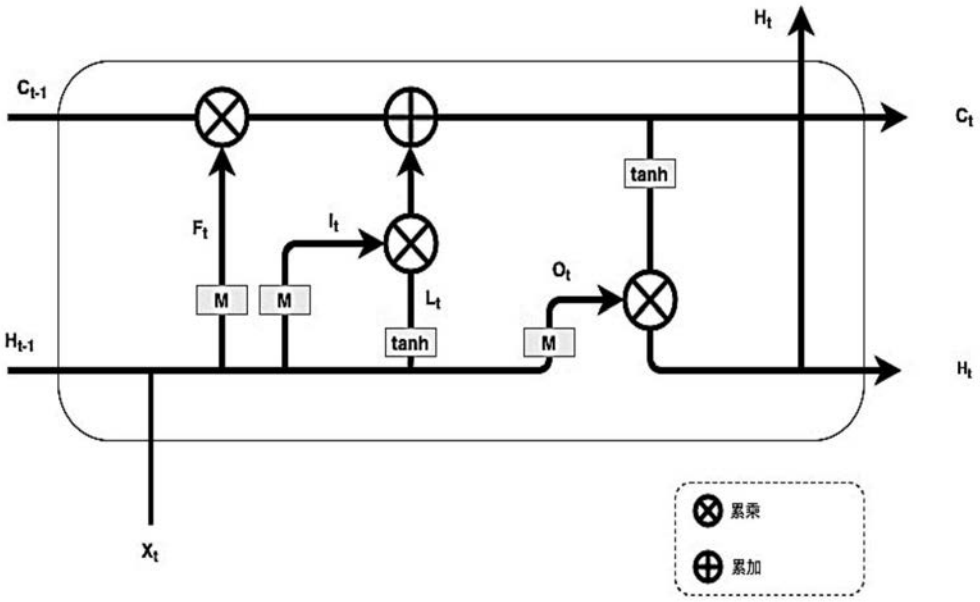
Ct-1
Ht
Ct
tanh
Ft
It
Ot
M
Lt
tanh
Ht-1
Ht
Xt
累乘
累加

图4

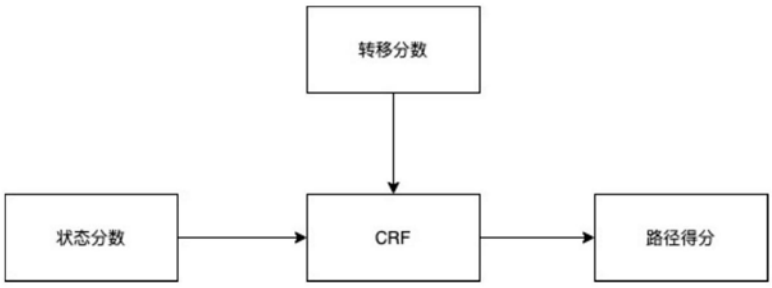
转移分数
状态分数
CRF
路径得分

图5

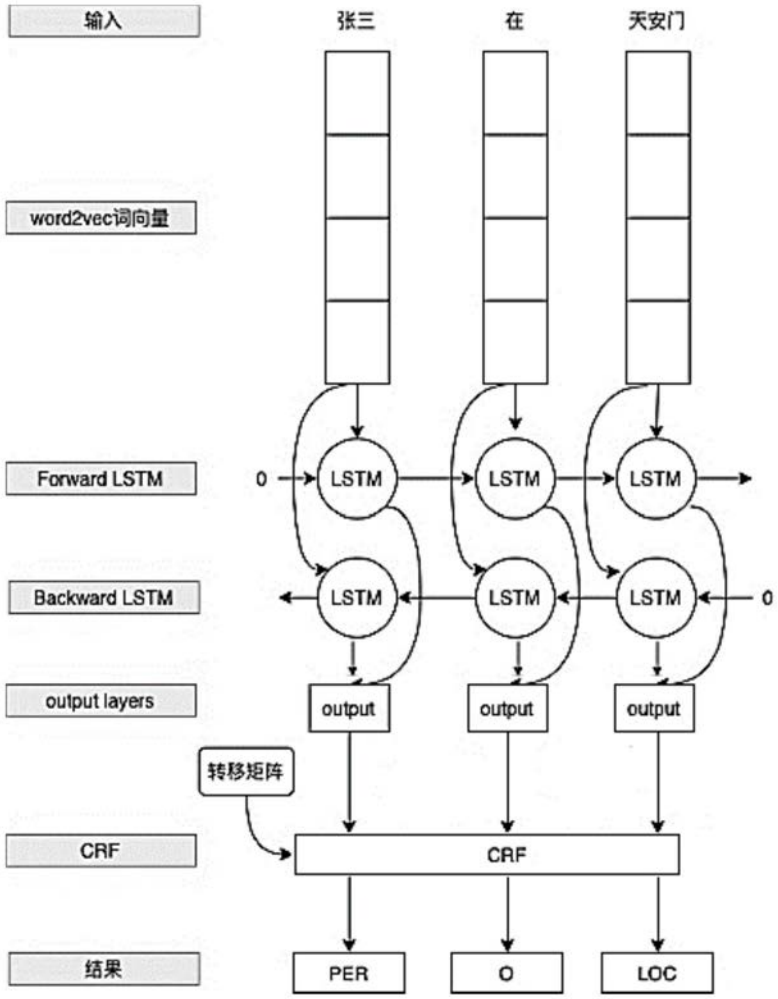
输入
张三
在
天安门
word2vec词向量
Forward LSTM
0
LSTM
LSTM
LSTM
Backward LSTM
LSTM
LSTM
LSTM
0
output layers
output
output
output
转移矩阵
CRF
CRF
结果
PER
O
LOC

图6

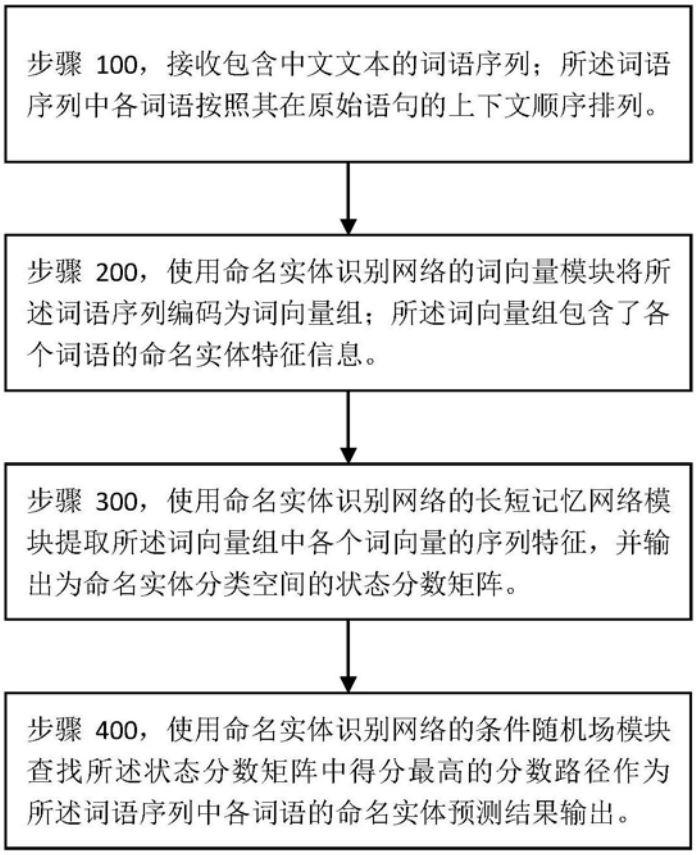
步骤 100，接收包含中文文本的词语序列；所述词语序列中各词语按照其在原始语句的上下文顺序排列。
步骤 200，使用命名实体识别网络的词向量模块将所述词语序列编码为词向量组；所述词向量组包含了各个词语的命名实体特征信息。
步骤 300，使用命名实体识别网络的长短记忆网络模块提取所述词向量组中各个词向量的序列特征，并输出为命名实体分类空间的状态分数矩阵。
步骤 400，使用命名实体识别网络的条件随机场模块查找所述状态分数矩阵中得分最高的分数路径作为所述词语序列中各词语的命名实体预测结果输出。

图7