(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 111428502 A
(43)申请公布日 2020.07.17

(21)申请号 202010102664.4

(22)申请日 2020.02.19

(71)申请人 中科世通亨奇(北京)科技有限公司
地址 100083 北京市海淀区学院路甲5号2幢平房北1102

(72)发明人 黄宇 冯洋

(74)专利代理机构 北京华际知识产权代理有限公司 11676
代理人 叶宇

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/169(2020.01)
G06N 3/04(2006.01)

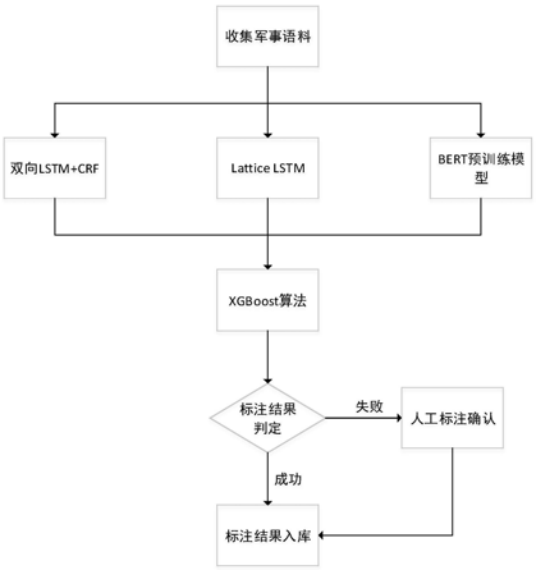
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种面向军事语料的命名实体标注方法

(57)摘要

本发明公开了一种面向军事语料的命名实体标注方法,分别使用基于双向LSTM与CRF结合的神经网络模型、基于Lattice LSTM神经网络模型和基于BERT预训练神经网络模型三种深度神经网络来进行机器命名实体识别自动标注;使用XGBoost方法将S1的三种算法获取的结果进行集成学习,获取标注成功的样本和标注失败的样本,其中成功样本的定义是三种机器实体识别中任意两种识别结果一致的样本,失败样本的定义三种机器实体识别结果都不一致的样本;使用人工标注的方式标注失败的样本;将所有样本标注结果以json的方式存入数据库管理。本发明可以显著提高军事语料中军事实体的标注准确率,同时以最小的人工代价达到最好的标注效果。



CN 111428502 A

1.一种面向军事语料的命名实体标注方法,其特征在于,所述方法包括以下步骤:

S1,分别使用基于双向LSTM与CRF结合的神经网络模型、基于Lattice LSTM神经网络模型和基于BERT预训练神经网络模型三种深度神经网络来进行机器命名实体识别自动标注;

S2,使用XGBoost方法将S1的三种算法获取的结果进行集成学习,获取标注成功的样本和标注失败的样本,其中成功样本的定义是三种机器实体识别中任意两种识别结果一致的样本,失败样本的定义三种机器实体识别结果都不一致的样本;

S3,使用人工标注的方式标注失败的样本;

S4,将所有样本标注结果以json的方式存入数据库管理。

2.根据权利要求1所述的一种面向军事语料的命名实体标注方法,其特征在于:将军事实体标注分为7种类型,包括人名实体、时间实体、地名实体、人员军职军衔实体、军事装备实体、军事设施实体、军事机构实体,分别记为person_entity、time_entity、location_entity、position_entity、weapon_entity、facility_entity、military_org_entity,将每个元素标注为“B-X”、“I-X”或者“O”。其中,“B-X”表示此元素所在的片段属于X类型并且此元素在此片段的开头,“I-X”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置,“O”表示不属于任何类型。

3.根据权利要求1所述的一种面向军事语料的命名实体标注方法,其特征在于:LSTM模型中长短时记忆模块计算过程如下:

(1)输入词Xt在t时刻通过输入门(Input Gate)进入网络,包含t时刻的输入以及与之相连的t-1时刻隐含层与细胞更新(cell)的输出,激活函数计算;

(2)通过遗忘门(Forget Gate)实现信息遗忘,与(1)相同,得到激活函数:

(3)细胞单元(cell)激活函数包括t时刻的输入与t-1时刻隐含层的输出;

(4)最终信息单元输出包括通过输出门Ot的向量输出及细胞单元输出,即前向推算的结果。

一种面向军事语料的命名实体标注方法

技术领域

[0001] 本发明涉及自然语言数据处理领域,具体涉及一种使用集成学习方法标注军事语料中的军事实体为命名实体识别在军事领域的应用提供训练语料,提高军事实体的识别准确度。

背景技术

[0002] 命名实体识别(Named Entity Recongition)是信息抽取和信息检索中一项重要的任务,其目的是识别出文本中表示命名实体的成分,并对其进行分类,因此有时也称为命名实体识别和分类。随着大数据时代的到来,互联网已经成为军事情报获取的重要来源。新闻专线、新闻杂志、军事报道、作战方案、演习报告、军报杂志、词典、政府公文、军事评论等途径都可以获得大量的军事文本信息,为了能够实现文本语义理解、语义表示、知识管理,需要提取面向军事领域内的军事实体,例如军政人物军职军衔、军用地名、军事装备名、军事设施名、军事机构名。为了达到计算机自动识别军事实体,需要大量高质量的军事实体标注语料,然而,在人力成本极高的当今时代,一方面,大量的标注语料将耗费不小的人力物力财力,另一方面,来自非专业人士的标注质量可能低于来自专家的标注质量,由此产生的低质量语料无法保证命名实体识别的准确性。因此,建立一种高效的面向军事语料的命名实体标注方法对于挖掘军事语料库潜在价值具有重要的价值和意义。

[0003] 目前语料标注常见的模式主要有3种,分别是传统标注模式、众包标注模式和团体标注模式。这三种标注模式其实都是通过人工标注的方式进行语料标注,传统标注模式是标注人员在标注规范的指导下进行标注在,众包标注模式利用网络,通过标注人员在线对同一篇语料进行标注,通过选票仲裁得到高质量的标注语料,团体标注则是利用大规模的标注团体进行标注获取语料。究其根本,还是通过标注人员的标注工作来获取标注语料。即便是具有高效的信息资源标引、组织和检索模式的社会标注和基于群体智慧语料标注方法,仍然摆脱不了这个缺点。利用了一些软件平台或者网络,还是需要我们的标注人员除了要统一标注规范之外,花费大量的时间去仲裁比对,决定最终采用最优的语料。

[0004] 发明中使用的Xgboost是目前最流行的一种集成学习方法。集成学习指的是利用多个弱监督模型以期得到一个更好更全面的强监督模型,集成学习潜在的思想是即便某一个弱分类器得到了错误的预测,其他的弱分类器也可以将错误纠正回来。Xgboost是华盛顿大学陈天奇于2016年提出的,兼具线性规模求解器和树学习的高效算法。它是传统的集成学习GBDT算法上的改进,更加高效。传统的GBDT方法只利用了一阶的导数信息,Xgboost则是对损失函数做了二阶的泰勒展开,并在目标函数之外加入了正则项,整体求最优解,用于权衡目标函数的下降和模型的复杂程度,避免过拟合,提高模型的求解效率,其步骤如下:

[0005] (1)给定数据集D＝{(xi,yi):i＝1,2,…,n,xi∈Rp,yi∈R},其中n为样本个数,每个样本有P个特征。假设我们给定k(k＝1,2,…,K)个回归树,xi表示第i个数据点的特征向量,fk是一个回归树,F是回归树的集合空间,模型可表示为:

[0006] $$\bar{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in F \qquad (6)$$

[0007] (2)目标函数定义如下:

[0008] $$Obj = \sum_{i=1}^{n} l(y_i, \bar{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (7)$$

[0009] 式中:yi为预测值,yi为真实值;为防止过拟合,定义正则化项,T和ω分别为树叶子节点数目和叶子权重值,γ为叶子树惩罚系数,λ为叶子权重惩罚系数。

[0010] (3)Xgboost使用梯度提升策略,保留已经有的模型,一次添加一个新的回归树到模型中,假设第i个样本在第t次迭代的预测结果为yi(t),ft(xi)为加入的新的回归树,可得如下推导过程:

[0011] $$\begin{aligned} &\bar{y}_i = 0 \\ &\bar{y}_i^{(1)} = f_1(x_i) = \bar{y}_i^{(0)} + f_1(x_i) \\ &\bar{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \bar{y}_i^{(1)} + f_2(x_i) \\ &\vdots \\ &\bar{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \bar{y}_i^{(t-1)} + f_t(x_i) \qquad (8) \end{aligned}$$

[0012] (4)将式(8)的结果代入式(7)中,可得:

[0013] $$Obj^{(t)} = \sum_{i=1}^{n} l\left[y_i, \bar{y}_i^{(t-1)} + f_t(x_i)\right] + \Omega(f_t) + C \qquad (9)$$

[0014] (5)将目标函数做二阶泰勒展开,并且引入正则项:

[0015] $$Obj^{(t)} \cong \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + C \qquad (10)$$

[0016] 式中:$g_i = \partial_{\bar{y}_i^{(t-1)}} l(y_i, \bar{y}_i^{(t-1)})$,$h_i = \partial^2_{\bar{y}_i^{(t-1)}} l(y_i, \bar{y}_i^{(t-1)})$ 。

[0017] XGBoost集成学习在各个规模的数据集上都有很好的表现,是目前提高算法准确率最稳定、效果最好的方法之一。

发明内容

[0018] 本发明目的在于提供一种为了解决海量互联网文本中包含的军事实体识别问题,为开源情报的发现和提取提供基础的面向军事语料的命名实体标注方法。

[0019] 为实现上述目的,采用了以下技术方案:本发明所述方法包括以下步骤:

[0020] S1,分别使用基于双向LSTM与CRF结合的神经网络模型、基于Lattice LSTM神经网络模型和基于BERT预训练神经网络模型三种深度神经网络来进行机器命名实体识别自动

标注；

[0021] S2,使用XGBoost方法将S1的三种算法获取的结果进行集成学习,获取标注成功的样本和标注失败的样本,其中成功样本的定义是三种机器实体识别中任意两种识别结果一致的样本,失败样本的定义三种机器实体识别结果都不一致的样本；

[0022] S3,使用人工标注的方式标注失败的样本；

[0023] S4,将所有样本标注结果以json的方式存入数据库管理。

[0024] 进一步的,将军事实体标注分为7种类型,包括人名实体、时间实体、地名实体、人员军职军衔实体、军事装备实体、军事设施实体、军事机构实体,分别记为person_entity、time_entity、location_entity、position_entity、weapon_entity、facility_entity、military_org_entity,将每个元素标注为“X-B”、“X-I”或者“O”。其中,“X-B”表示此元素所在的片段属于X类型并且此元素在此片段的开头,“X-I”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置,“O”表示不属于任何类型。例如“F-16战机于15日4时23分降落于安德森空军基地”,标注为“weapon_entity_B weapon_entity_I weapon_entity_I weapon_entity_I O time_entity_B time_entity_Itime_entity_I time_entity_I time_entity_I time_entity_I O O O location_entity_Blocation_entity_I location_entity_I location_entity_I location_entity_Ilocation_entity_I location_entity_I”。

[0025] 进一步的,LSTM模型中长短时记忆模块计算过程如下：

[0026] (1)输入词Xt在t时刻通过输入门(Input Gate)进入网络,包含t时刻的输入以及与之相连的t-1时刻隐含层与细胞更新(cell)的输出,激活函数计算；

[0027] (2)通过遗忘门(Forget Gate)实现信息遗忘,与(1)相同,得到激活函数；

[0028] (3)细胞单元(cell)激活函数包括t时刻的输入与t-1时刻隐含层的输出；

[0029] (4)最终信息单元输出包括通过输出门Ot的向量输出及细胞单元输出,即前向推算的结果。

[0030] 理论上来讲,后向推算是在前向推算的基础上逆推求导,过程与前向类似。双向LSTM针对已知的训练序列实行向前和向后两次LSTM特定训练,由此确保特征提取的全局性和完整性。

[0031] 与现有技术相比,本发明具有如下优点:可以显著提高军事语料中军事实体的标注准确率,同时以最小的人工代价达到最好的标注效果。

附图说明

[0032] 表1为本发明提出的军事实体标注规范。

[0033] 图1为本发明的基本流程图。

[0034] 图2为双向LSTM神经网络模型结构图。

[0035] 图3为基于Lattice LSTM神经网络模型结构图。

[0036] 图4为基于BERT预训练神经网络模型结构图。

具体实施方式

[0037] 下面结合附图对本发明做进一步说明：

[0038] 结合图1-图4，本发明所述方法包括以下步骤：

[0039] S1，分别使用基于双向LSTM与CRF结合的神经网络模型、基于Lattice LSTM神经网络模型和基于BERT预训练神经网络模型三种深度神经网络来进行机器命名实体识别自动标注；

[0040] S2，使用XGBoost方法将S1的三种算法获取的结果进行集成学习，获取标注成功的样本和标注失败的样本，其中成功样本的定义是三种机器实体识别中任意两种识别结果一致的样本，失败样本的定义三种机器实体识别结果都不一致的样本；

[0041] S3，使用人工标注的方式标注失败的样本；

[0042] S4，将所有样本标注结果以json的方式存入数据库管理。

[0043] 表1本发明中军事实体标注规范

[0044]

标注项目	标注内容	
	实体首字	实体非首字
人名实体	person_entity_B	person _entity_I
时间实体	time_entity_B	time_entity_I
地名实体	location_entity_B	location_entity_I
人员军职军衔实体	job_entity_B	job_entity_I
军事装备实体	weapon _entity_B	weapon _entity_I
军事设施实体	facility_entity_B	facility_entity_I
军事机构实体	org_entity_B	org_entity_I

[0045]

其他部分	O	O

[0046] 如表1所示，将军事实体标注分为7种类型，包括人名实体、时间实体、地名实体、人员军职军衔实体、军事装备实体、军事设施实体、军事机构实体，分别记为person_entity、time_entity、location_entity、position_entity、weapon_entity、facility_entity、military_org_entity，将每个元素标注为“X-B”、“X-I”或者“O”。其中，“X-B”表示此元素所在的片段属于X类型并且此元素在此片段的开头，“X-I”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置，“O”表示不属于任何类型。例如“F-16战机于15日4时23分降落于安德森空军基地”，标注为“weapon_entity_B weapon_entity_I weapon_entity_I weapon_entity_I weapon_entity_I O time_entity_B time_entity_I time_entity_I time_entity_I time_entity_I time_entity_I O O O location_entity_B location_entity_I location_entity_I location_entity_I location_entity_I location_entity_I location_entity_I”。

[0047] 进一步说明：

[0048] 1、军事命名实体词性标注规范制定

[0049] 2、军事文本导入与预处理

[0050] 对于语料标注平台而言，我们需要将许许多多的生语料进行标注处理，形成标注完全的语料库。生语料的获取途径无非是我们已有的文本数据，或者从网络上爬虫获得，所以我们对于文本载入部分而已，最基本的功能要求就是导入文本数据，和网路爬虫等载入方式，再加上人工输入的功能，来避免一些无法导入的文件内容无法标注的损失。在现有的基础上，以后如果想要更加完善强化该平台，可以考虑在文本载入功能上，加入图片文字识别输入等，现在随着网络和技术设备的发展，文本不仅仅记录于文本文件之中，图片，音频，

视频中其实都存在着大量的文字信息。当然我们做语料标注,无需对音视频进行分析,但是有些文本会在图片上记录下来,所以后期强化该平台可以考虑加入该功能。

[0051] 3、军事文本命名实体识别

[0052] 其中双向LSTM(Bi-LSTM)结合CRF的神经网络模型是命名实体识别中比较常用的提取算法,双向LSTM是循环神经网络的子类,最早由HOCHREITER等人提出,本质上也是复杂的非线性单元,其具备的显著特点是具有较强的记忆能力及对非线性关系的拟合能力。LSTM模型中长短时记忆模块计算过程如下:

[0053] (1)输入词Xt在t时刻通过输入门(Input Gate)进入网络,包含t时刻的输入以及与之相连的t-1时刻隐含层与细胞更新(cell)的输出,激活函数计算:

[0054] (2)通过遗忘门(Forget Gate)实现信息遗忘,与(1)相同,得到激活函数:

[0055] (3)细胞单元(cell)激活函数包括t时刻的输入与t-1时刻隐含层的输出:

[0056] (4)最终信息单元输出包括通过输出门Ot的向量输出及细胞单元输出,即前向推算的结果:

[0057] 理论上来讲,后向推算是在前向推算的基础上逆推求导,过程与前向类似。双向LSTM针对已知的训练序列实行向前和向后两次LSTM特定训练,由此确保特征提取的全局性和完整性。

[0058] 条件随机场(CRF)本质上是一种判别式无向图,理论基础是隐马尔科夫模型和最大熵模型,另有属于整个可观测向量的可观测符号X,主要用于词性标注和切分有序数据。条件随机场应用和发展至今仍保留了隐马尔科夫模型的部分特征,实际应用过程中的变量之间遵守马尔可夫假设,每个状态的转移概率取决于相邻变量的即时状态。以线性链随机场为例,假设随机变量序列,若两者满足马尔科夫性,即,则称P(Y X)为线性链条件随机场,其中X为输入观测序列,Y表示与之对应的输出标记序列(或状态序列)。条件随机场的特征函数包含转移特征和状态特征,转移特征函数限定的是前后词的词性,状态特征函数计算每个词所处每种状态的概率大小。

[0059] 4、标准标注语料入库

[0060] 在工作人员利用语料标注平台,对文本进行了实体的识别和属性的添加之后,就可以通过软件的语料生成功能进行语料的生成了,语料的生成功能,通过我们设计的符合语料规范的语料生成方案来自动生成语料,形成一个XML视图的语料编辑框,并且通过该框架进行调整修改。确认无误后可以通过点击生成XML生成语料,最后将其纳入标注完备的语料数据库中。

[0061] 以上所述的实施例仅仅是对本发明的优选实施方式进行描述,并非对本发明的范围进行限定,在不脱离本发明设计精神的前提下,本领域普通技术人员对本发明的技术方案做出的各种变形和改进,均应落入本发明权利要求书确定的保护范围内。

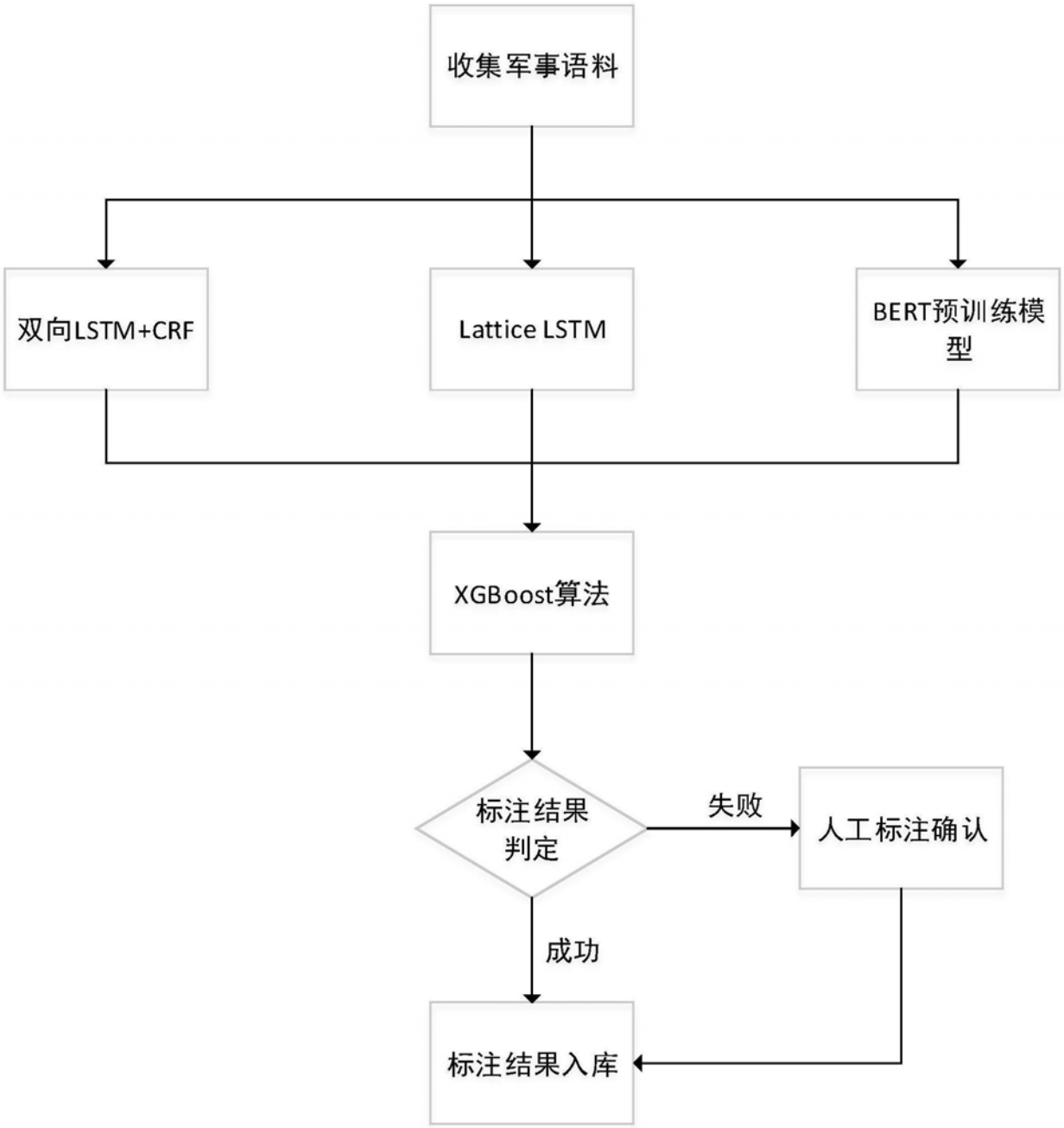

收集军事语料
双向LSTM+CRF
Lattice LSTM
BERT预训练模型
XGBoost算法
标注结果判定
失败
人工标注确认
成功
标注结果入库

图1

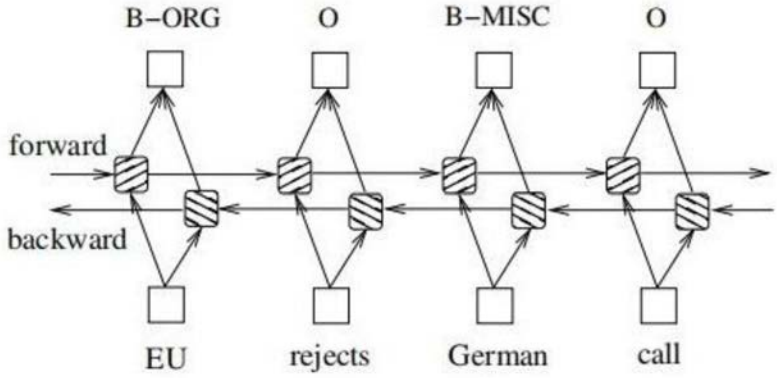

B-ORG
O
B-MISC
O
forward
backward
EU
rejects
German
call

图2

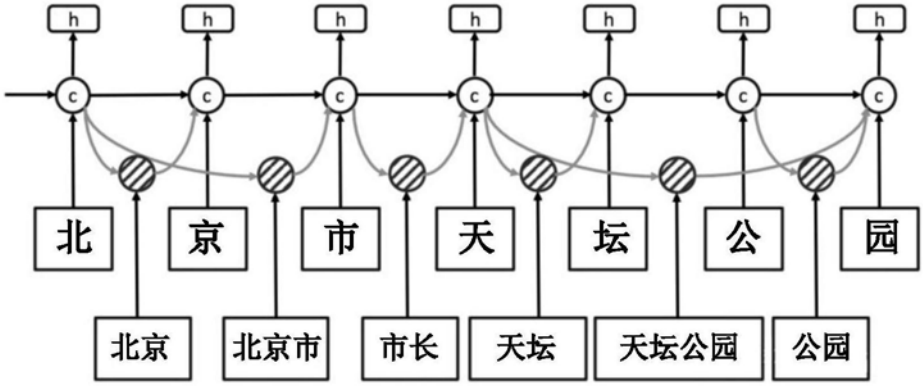

h
c
北
京
市
天
坛
公
园
北京
北京市
市长
天坛
天坛公园
公园

图3

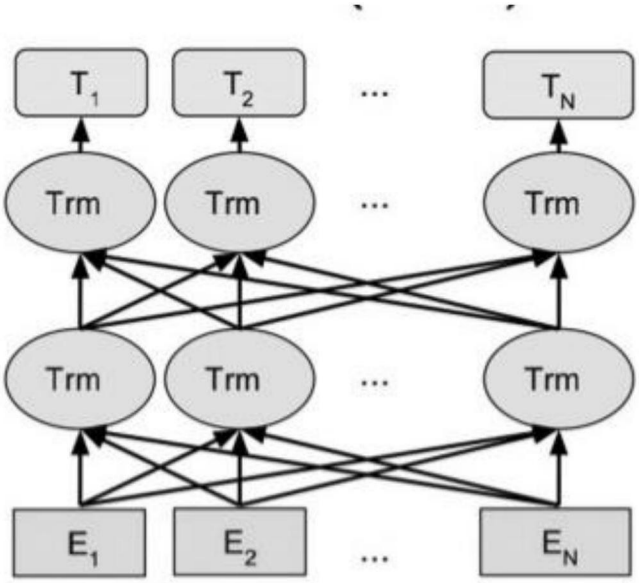

T1
T2
...
TN
Trm
E1
E2
EN

图4