(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 111209751 A
(43)申请公布日 2020.05.29

(21)申请号 202010095159.1

(22)申请日 2020.02.14

(71)申请人 全球能源互联网研究院有限公司
地址 102209 北京市昌平区未来科技城滨河大道18号
申请人 国家电网有限公司
国网浙江省电力有限公司

(72)发明人 宋博川 张强 柴博 贾全烨 戴铁潮

(74)专利代理机构 北京三聚阳光知识产权代理有限公司 11250
代理人 罗啸

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 40/211(2020.01)
G06F 40/216(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

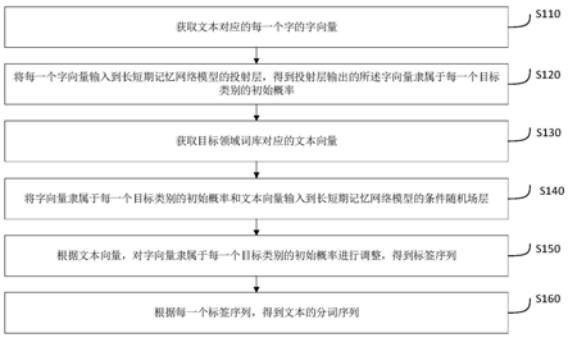
权利要求书2页 说明书8页 附图2页

(54)发明名称

一种中文分词方法、装置及存储介质

(57)摘要

本发明提供了一种中文分词方法、装置及存储介质，其中，方法包括：获取文本对应的每一个字的字向量；将每一个字向量输入到长短期记忆网络模型的投射层，得到投射层输出的所述字向量隶属于每一个目标类别的初始概率；获取目标领域词库对应的文本向量；将字向量隶属于每一个目标类别的初始概率和文本向量输入到长短期记忆网络模型的条件随机场层；根据文本向量，对字向量隶属于每一个目标类别的初始概率进行调整，得到标签序列；根据标签序列，得到文本的分词序列。通过实施本发明，利用长短期记忆网络模型和目标领域词库对字向量隶属于每一个目标类别的初始概率进行计算与调整，得到中文分词序列，提高了分词结果的准确性。



CN 111209751 A

1.一种中文分词方法,其特征在于,包括如下步骤:

获取文本对应的每一个字的字向量;

将每一个字向量输入到长短期记忆网络模型的投射层,得到所述投射层输出的所述字向量隶属于每一个目标类别的初始概率;

获取目标领域词库对应的文本向量;

将所述字向量隶属于每一个目标类别的初始概率和所述文本向量输入到所述长短期记忆网络模型的条件随机场层;

根据所述文本向量,对所述字向量隶属于每一个目标类别的初始概率进行调整,得到标签序列;

根据每一个所述标签序列,得到所述文本的分词序列。

2.根据权利要求1所述的方法,其特征在于,所述获取文本对应的每一个字的字向量,包括:

将所述文本输入到所述长短期记忆网络模型的第一编码层,得到所述文本对应的每一个字的初始字向量;

将所述对应的每一个字的初始字向量输入到第二编码层,得到表征上下文关系的字向量,将所述表征上下文关系的字向量作为所述文本对应的每一个字的字向量。

3.根据权利要求1所述的方法,其特征在于,所述目标类别包括多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。

4.根据权利要求1所述的方法,其特征在于,根据所述文本向量,对所述字向量隶属于每一个目标类别的初始概率进行调整,得到所述字向量的标签,包括:

获取转移概率矩阵;

根据所述转移概率矩阵,对所述字向量隶属于每一个目标类别的初始概率进行调整,得到所述字向量的标签。

5.一种中文分词装置,其特征在于,包括:

字向量获取模块,用于获取文本对应的每一个字的字向量;

初始概率获取模块,用于将每一个字向量输入到长短期记忆网络模型的投射层,得到所述投射层输出的所述字向量隶属于每一个目标类别的初始概率;

文本向量获取模块,用于获取目标领域词库对应的文本向量;

条件随机场层输入模块,用于将所述字向量隶属于每一个目标类别的初始概率和所述文本向量输入到所述长短期记忆网络模型的条件随机场层;

标签获取模块,用于根据所述文本向量,对所述字向量隶属于每一个目标类别的初始概率进行调整,得到所述字向量的标签;

分词序列获取模块,用于根据每一个所述字向量的标签,得到所述文本的分词序列。

6.根据权利要求5所述的装置,其特征在于,所述字向量获取模块,包括:

初始字向量获取模块,用于将所述文本输入到所述长短期记忆网络模型的第一编码层,得到所述文本对应的每一个字的初始字向量;

字向量获取子模块,用于将所述对应的每一个字的初始字向量输入到第二编码层,得到表征上下文关系的字向量,将所述表征上下文关系的字向量作为所述文本对应的每一个字的字向量。

7.根据权利要求5所述的装置，其特征在于，所述目标类别包括多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。

8.根据权利要求5所述的装置，其特征在于，标签获取模块，包括：

转移概率矩阵获取模块，用于获取转移概率矩阵；

标签获取子模块，用于根据所述转移概率矩阵，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到所述字向量的标签。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现权利要求1-4任一项所述的中文分词方法的步骤。

10.一种存储介质，其上存储有计算机指令，其特征在于，该指令被处理器执行时实现权利要求1-4任一项所述的中文分词方法的步骤。

一种中文分词方法、装置及存储介质

技术领域

[0001] 本发明涉及自然语言处理领域，具体涉及一种中文分词方法、装置及存储介质。

背景技术

[0002] 中文分词是将连续的字序列按照一定的规范重新组合成词序列的过程。在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独对词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂得多、困难得多。

[0003] 相关技术中，分词方法为基于传统的统计学习的分词方法，但需要人工设计规则模板，而且面临严重的数据稀疏问题，导致分词结果的准确率低。

发明内容

[0004] 因此，本发明要解决的技术问题在于克服现有技术中的分词结果的准确率低缺陷，从而提供一种中文分词方法、装置及存储介质。

[0005] 根据第一方面，本发明实施例提供一种中文分词方法，包括如下步骤：

[0006] 获取文本对应的每一个字的字向量；将每一个字向量输入到长短期记忆网络模型的投射层，得到所述投射层输出的所述字向量隶属于每一个目标类别的初始概率；获取目标领域词库对应的文本向量；将所述字向量隶属于每一个目标类别的初始概率和所述文本向量输入到所述长短期记忆网络模型的条件随机场层；根据所述文本向量，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到标签序列；根据所述标签序列，得到所述文本的分词序列。

[0007] 结合第一方面，在第一方面第一实施方式中，所述获取文本对应的每一个字的字向量，包括：将所述文本输入到所述长短期记忆网络模型的第一编码层，得到所述文本对应的每一个字的初始字向量；将所述对应的每一个字的初始字向量输入到第二编码层，得到表征上下文关系的字向量，将所述表征上下文关系的字向量作为所述文本对应的每一个字的字向量。

[0008] 结合第一方面，在第一方面第二实施方式中，所述目标类别包括多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。

[0009] 结合第一方面，在第一方面第三实施方式中，根据所述文本向量，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到所述字向量的标签，包括：获取转移概率矩阵；根据所述转移概率矩阵，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到所述字向量的标签。

[0010] 根据第二方面，本发明实施例提供一种中文分词装置，包括：字向量获取模块，用于获取文本对应的每一个字的字向量；初始概率获取模块，用于将每一个字向量输入到长短期记忆网络模型的投射层，得到所述投射层输出的所述字向量隶属于每一个目标类别的初始概率；文本向量获取模块，用于获取目标领域词库对应的文本向量；条件随机场层输入

模块，用于将所述字向量隶属于每一个目标类别的初始概率和所述文本向量输入到所述长短期记忆网络模型的条件随机场层；标签获取模块，用于根据所述文本向量，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到标签序列；分词序列获取模块，用于根据每一个所述标签序列，得到所述文本的分词序列。

[0011] 结合第二方面，在第二方面第一实施方式中，所述字向量获取模块，包括：初始字向量获取模块，用于将所述文本输入到所述长短期记忆网络模型的第一编码层，得到所述文本对应的每一个字的初始字向量；字向量获取子模块，用于将所述对应的每一个字的初始字向量输入到第二编码层，得到表征上下文关系的字向量，将所述表征上下文关系的字向量作为所述文本对应的每一个字的字向量。

[0012] 结合第二方面，在第二方面第二实施方式中，所述目标类别包括多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。

[0013] 结合第二方面，在第二方面第三实施方式中，标签获取模块，包括：转移概率矩阵获取模块，用于获取转移概率矩阵；标签获取子模块，用于根据所述转移概率矩阵，对所述字向量隶属于每一个目标类别的初始概率进行调整，得到所述字向量的标签。

[0014] 根据第三方面，本发明实施例提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现第一方面或第一方面任一实施方式所述的中文分词方法的步骤。

[0015] 根据第四方面，本发明实施例提供一种存储介质，其上存储有计算机指令，该指令被处理器执行时实现第一方面或第一方面任一实施方式所述的中文分词方法的步骤。

[0016] 本发明技术方案，具有如下优点：

[0017] 1.本发明提供了一种中文分词方法/装置，通过长短期记忆网络模型以及外部引入目标领域词库对输入的文本信息中的字向量隶属于每一个目标类别的初始概率进行计算与调整，从而得到文本的中文分词序列，提高了分词结果的准确性。

[0018] 2.本发明提供了一种中文分词方法/装置，通过将字向量输入第二编码层，得到包含上下文信息的隐层向量，使得在后续计算字向量隶属于每一个目标类别的初始概率的准确率更高，进一步提高了分词结果的准确性。

[0019] 3.本发明提供了一种中文分词方法/装置，通过转移概率矩阵对初始概率进行约束，并通过约束条件调整初始概率，从而调整字向量标签，进一步提高了中文分词的准确率。

附图说明

[0020] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案，下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施方式，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0021] 图1为本发明实施例中中文分词方法的一个具体示例的流程图；

[0022] 图2为本发明实施例中中文分词装置的一个具体示例的原理框图；

[0023] 图3为本发明实施例中电子设备的一个具体示例的原理框图。

具体实施方式

[0024] 下面将结合附图对本发明的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0025] 在本发明的描述中，需要说明的是，术语“中心”、“上”、“下”、“左”、“右”、“竖直”、“水平”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0026] 在本发明的描述中，需要说明的是，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，还可以是两个元件内部的连通，可以是无线连接，也可以是有线连接。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0027] 此外，下面所描述的本发明不同实施方式中所涉及的技术特征只要彼此之间未构成冲突就可以相互结合。

[0028] 本申请实施例提供一种中文分词方法，如图1所示，包括如下步骤：

[0029] S110，获取文本对应的每一个字的字向量。

[0030] 示例性地，文本对应的每一个字的字向量的获取方式可以是将字符进行编码，每个字符有唯一的id值。然后根据id值，选择预先设定的字嵌入矩阵中对应的字向量，得到神经网络可以处理的字向量。假设字符“一”的id值是239，则选择字嵌入矩阵中编号为239的字向量v_{239}，v_{239}即为神经网络可以处理的字向量，还可以是利用字嵌入的方法，将文本中的每个字转换成对应字向量的表达式，以供神经网络读取，其具体的方法可以是利用word2vec工具中的跳字模型和连续词袋模型，对模型进行训练，以使跳字模型和连续词袋模型将文本转换成每个字的字向量。本实施例对文本对应的每一个字的字向量的获取方式不做限定，本领域技术人员可以根据需要确定。

[0031] S120，将每一个字向量输入到长短期记忆网络模型的投射层，得到投射层输出的字向量隶属于每一个目标类别的初始概率。

[0032] 示例性地，预先根据大量的训练样本训练得到用于进行分词操作的长短期记忆网络模型，长短期记忆网络模型包括投射层和条件随机场层，其中，投射层用于计算每一个字向量属于每一个目标类别的初始概率。其初始概率的计算方式可以是以每一个字的字向量为v，初始概率为s，s是由v通过一个线性变换得到：$s=Wv+b$。W是一个(4,h)的矩阵，h是每一个字的字向量的维度，b是一个任意数值。将初始概率组成初始概率矩阵(t,4)，其中t是输入句子的字符数，即每个字向量和句子中的字符一一对应。目标类别可以是多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。以输入文本为“南京市长江大桥建成通车”为例，分别计算出“南”、“京”、“市”、“长”、“江”、“大”、“桥”、“建”、“成”、“通”、“车”对应的字向量分别隶属于多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语的概率，将得到的概率作为初始概率。比如，“长”对应的字向量分别隶属于多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语计算得到的概率分别为0.3、0.1、0.4、

0.2。本实施例对目标类别不做限定，本领域技术人员可以根据需要确定。

[0033] S130，获取目标领域词库对应的文本向量。

[0034] 示例性地，目标领域词库表征与文本内容领域对应的词库，词库中可以包含该领域的常用词汇、专业词汇、最新词汇等词汇组合方式。目标领域词库对应的文本向量的获取方式可以仍然按照上述步骤S110表述的获取文本对应的每一个字的字向量的方式，在此不做赘述。本实施例对目标领域词库对应的文本向量的获取方式不做限定，本领域技术人员可以按照需要确定。

[0035] S140，将字向量隶属于每一个目标类别的初始概率和文本向量输入到长短期记忆网络模型的条件随机场层。

[0036] 示例性地，条件随机场层是一种概率无向图模型，通过输入的字向量隶属于每一个目标类别的初始概率和文本向量调整输出的字向量隶属于每一个目标类别的概率。

[0037] S150，根据文本向量，对字向量隶属于每一个目标类别的初始概率进行调整，得到标签序列。

[0038] 示例性地，字向量的标签可以是“B”，“M”，“E”，“S”，其中，“B”代表一个多字词语中的第一个字，“M”代表一个多字词语中除去第一个字和最后一个字的其他字，“E”代表一个多字词语中的最后一个字，“S”代表一个单字词语，在本实施例中，字向量的标签对应上述步骤S120的目标类别。根据文本向量，对字向量隶属于每一个目标类别的初始概率进行调整的方式可以是将目标领域词库中存在的文本向量与文本对应的每一个字的字向量进行比对，当该字向量与相邻字向量在目标领域词库中的文本向量中存在，则增加该字向量在对应的目标类别的初始概率或者增加该字向量在对应的目标类别的权重偏置；当该字向量与相邻字向量在目标领域词库中的文本向量中不存在，则减小该字向量在对应的目标类别的初始概率或者减小该字向量在对应的目标类别的权重偏置，从而得到调整后的每一个字向量对每一个目标类别的初始概率得到最终概率，或者利用维特比算法计算得到最终的分词结果标签序列。其中，权重偏置的计算方式可以是一个预设值，在整个计算过程中不发生变化，也可以是计算当前输入句子隐层向量矩阵的平均值，还可以是计算当前输入句子的隐层向量矩阵中所有元素的平均值，且用0替代所有负数。

[0039] 比如，仍以输入文本为“南京市长江大桥建成通车”为例，“南京市长江大桥”存在两种划分方式：“南京市长/江大桥”、“南京市/长江大桥”，而在上述步骤S120中得到的“长”在多字词语的尾位的初始概率最高，表明句子被初步划分为“南京市长/江大桥”，目标领域词库对应的文本向量中存在“长江大桥”，而不存在“江大桥”，所以，增加长在多字词语的首位的初始概率，增加为0.4，降低长在多字词语的尾位的概率，降低为0.3。此时得到“长”在各个目标类别概率为0.4、0.1、0.3、0.2，选择字向量在各个目标类别中初始概率最高目标类别对应的标签，即0.4对应的目标类别，标签对应为B。

[0040] 再如，以输入文本为“悠然见南山”为例，权重偏置为a，假设每一个字向量隶属于每一个目标类别的初始概率，建立初始概率矩阵如表1所示。

[0041] 表1

[0042]

	B	M	E	S
悠	0.3	-0.4	0.5	0.6
然	0.3	-0.4	0.5	0.6

见	0.3	-0.4	0.5	0.6
南	0.3	-0.4	0.5	0.6
山	0.3	-0.4	0.5	0.6

[0043] 此时，“悠然”和“南山”是存在于目标领域词库中的。并且，“悠”和“南”是一个词的开始，则对矩阵中的“悠”和“南”字的“B”标签加上一个权重偏置a；“然”和“山”是一个词的结尾，则对矩阵中的“然”、“山”的“E”标签加上一个权重偏置a。加入权重偏置的矩阵变为如表2所示：

[0044] 表2

[0045]

	B	M	E	S
悠	0.3+a	-0.4	0.5	0.6
然	0.3	-0.4	0.5+a	0.6
见	0.3	-0.4	0.5	0.6
南	0.3+a	-0.4	0.5	0.6
山	0.3	-0.4	0.5+a	0.6

[0046]

[0047] 权重偏置a的大小在本实施例中，可以是设定一个预设值，比如0.2，整个计算过程中不发生变化；也可以通过计算当前输入句子隐层向量矩阵的平均值，这个例子中为：((0.3+0.5+0.6)*5+(-0.4)*5)/20＝0.25；还可以通过计算当前输入句子的隐层向量矩阵中所有元素的平均值，但是用0替代所有负数，这个例子中为：((0.3+0.5+0.6)*5+0*5)/20＝0.35。

[0048] 如果记一个句子的标签序列为：$y=(y_1,2,...,n)$，y_1、$y_2...y_n$分别表示一个句子的标签，标签可以是“B”，“M”，“E”，“S”中任意一个。对于输入句子中的任一字向量x的标签y的分数为：

[0049] $$score(x,y)=\sum E_{i,y_i}+\sum T_{y_{i-1},y_i}$$

[0050] 其中E是初始概率矩阵，T是加入权重偏置或者修改字向量在对应的目标类别的初始概率的矩阵。最终的概率可以通过softmax确定：

[0051] $$P(y|x)=\frac{\exp\left(score(x,y)\right)}{\sum_i \exp\left(score(x,y_i)\right)}$$

[0052] 通过上式计算出整个句子对应的标签序列y_i。

[0053] 本实施例对根据文本向量，调整字向量隶属于每一个目标类别的初始概率的方式不做限定，本领域技术人员可以根据需要确定。

[0054] S160，根据标签序列，得到文本的分词序列。

[0055] 示例性地，仍以输入文本为“南京市长江大桥建成通车”和“悠然见南山”为例，最终得到的标签序列分别为BMEBMMEBEBE和BESBE，根据标签序列，得到的文本分词序列则分别为“南京市/长江大桥/建成/通车”和“悠然/见/南山”。

[0056] 本实施例提供了一种中文分词方法，通过长短期记忆网络模型以及外部引入目标领域词库对输入的文本信息中的字向量隶属于每一个目标类别的初始概率进行计算与调

整，从而得到文本的中文分词序列，提高了分词结果的准确性。

[0057] 作为本实施例一种可选的实施方式，上述步骤S110，包括：

[0058] 首先，将文本输入到长短期记忆网络模型的第一编码层，得到文本对应的每一个字的初始字向量。

[0059] 示例性地，第一编码层可以是字符编码层，实现将输入文本编码为长短期记忆网络可以处理的初始字向量，其具体的编码方式参见上述步骤S110，此处不再赘述。本实施例对第一编码层不做限定，本领域技术人员可以根据需要确定。

[0060] 其次，将对应的每一个字的初始字向量输入到第二编码层，得到表征上下文关系的字向量，将表征上下文关系的字向量作为文本对应的每一个字的字向量。

[0061] 示例性地，第二编码层可以是长短期记忆网络编码层，实现对第一编码层得到的字向量进行编码，得到隐层向量，每个隐层向量都与输入句子中的每个字符一一对应，将该隐层向量作为文本对应的每一个字的字向量。假设输入的句子中有13个字符，则对应的隐层向量有13个。其中每个隐层向量不仅包含单个字符的信息，同时也包含字符在句子中的上下文信息。

[0062] 本实施例提供的中文分词方法，通过将字向量输入第二编码层，得到包含上下文信息的隐层向量，使得在后续计算字向量隶属于每一个目标类别的初始概率的准确率更高，进一步提高了分词结果的准确性。

[0063] 作为本实施例一种可选的实施方式，上述步骤S150，包括：

[0064] 首先，获取转移概率矩阵。

[0065] 示例性地，转移概率矩阵用于对初始概率的计算进行约束。比如，输入文本的第一个字向量是B的概率最大，而B后面的字向量为M或者E比为S的概率大，此时，建立一个转移概率矩阵实现约束。转移概率矩阵的获取方式可以是随机初始化转移概率矩阵，转移概率矩阵随着长短期记忆网络模型的训练进行迭代更新。本实施例对转移概率矩阵的获取方式不做限定，本领域技术人员可以根据需要确定。

[0066] 其次，根据转移概率矩阵，对字向量隶属于每一个目标类别的初始概率进行调整，得到字向量的标签。

[0067] 示例性地，根据转移概率矩阵，对字向量隶属于每一个目标类别的初始概率进行调整的方式可以是将转移概率矩阵中的参数作为计算字向量隶属于每一个目标类别的初始概率的权重，通过权重对每一个字向量隶属于每一个目标类别的初始概率进行调整，得到调整后的每一个字向量隶属于每一个目标类别的概率，选择字向量在各个目标类别中调整后的初始概率最高的目标类别对应的标签，将该标签作为字向量的标签，得到标签序列。

[0068] 本实施例提供的中文分词方法，通过转移概率矩阵对初始概率进行约束，并通过约束条件调整初始概率，从而调整字向量标签，进一步提高了中文分词的准确率。

[0069] 本申请实施例提供一种中文分词装置，如图2所示，包括：

[0070] 字向量获取模块210，用于获取文本对应的每一个字的字向量；具体实现方式见本实施例方法步骤S110对应部分，在此不再赘述。

[0071] 初始概率获取模块220，用于将每一个字向量输入到长短期记忆网络模型的投射层，得到投射层输出的字向量隶属于每一个目标类别的初始概率；具体实现方式见本实施例方法步骤S120对应部分，在此不再赘述。

[0072] 文本向量获取模块230,用于获取目标领域词库对应的文本向量;具体实现方式见本实施例方法步骤S130对应部分,在此不再赘述。

[0073] 条件随机场层输入模块240,用于将字向量隶属于每一个目标类别的初始概率和文本向量输入到长短期记忆网络模型的条件随机场层;具体实现方式见本实施例方法步骤S140对应部分,在此不再赘述。

[0074] 标签获取模块250,用于根据文本向量,对字向量隶属于每一个目标类别的初始概率进行调整,得到字向量的标签;具体实现方式见本实施例方法步骤S150对应部分,在此不再赘述。

[0075] 分词序列获取模块260,用于根据每一个字向量的标签,得到文本的分词序列。具体实现方式见本实施例方法步骤S160对应部分,在此不再赘述。

[0076] 本实施例提供了一种中文分词装置,通过长短期记忆网络模型以及外部引入目标领域词库对输入的文本信息中的字向量隶属于每一个目标类别的初始概率进行计算与调整,从而得到文本的中文分词序列,提高了分词结果的准确性。

[0077] 作为本申请一种可选的实施方式,字向量获取模块210,包括:

[0078] 初始字向量获取模块,用于将文本输入到长短期记忆网络模型的第一编码层,得到文本对应的每一个字的初始字向量;具体实现方式见本实施例方法对应部分,在此不再赘述。

[0079] 字向量获取子模块,用于将对应的每一个字的初始字向量输入到第二编码层,得到表征上下文关系的字向量,将表征上下文关系的字向量作为文本对应的每一个字的字向量。具体实现方式见本实施例方法对应部分,在此不再赘述。

[0080] 作为本实施例一种可选的实施方式,目标类别包括多字词语的首位、多字词语的中间位、多字词语的尾位和单字词语。具体实现方式见本实施例方法对应部分,在此不再赘述。

[0081] 作为本实施例一种可选的实施方式,标签获取模块250,包括:

[0082] 转移概率矩阵获取模块,用于获取转移概率矩阵;具体实现方式见本实施例方法对应部分,在此不再赘述。

[0083] 标签获取子模块,用于根据转移概率矩阵,对字向量隶属于每一个目标类别的初始概率进行调整,得到字向量的标签。具体实现方式见本实施例方法对应部分,在此不再赘述。

[0084] 本申请实施例还提供一种电子设备,如图3所示,处理器310和存储器320,其中处理器310和存储器320可以通过总线或者其他方式连接。

[0085] 处理器310可以为中央处理器(Central Processing Unit,CPU)。处理器310还可以为其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等芯片,或者上述各类芯片的组合。

[0086] 存储器320作为一种非暂态计算机可读存储介质,可用于存储非暂态软件程序、非暂态计算机可执行程序以及模块,如本发明实施例中的中文分词方法对应的程序指令/模块。处理器通过运行存储在存储器中的非暂态软件程序、指令以及模块,从而执行处理器的

各种功能应用以及数据处理。

[0087] 存储器320可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储处理器所创建的数据等。此外，存储器可以包括高速随机存取存储器，还可以包括非暂态存储器，例如至少一个磁盘存储器件、闪存器件、或其他非暂态固态存储器件。在一些实施例中，存储器320可选包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至处理器。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0088] 所述一个或者多个模块存储在所述存储器320中，当被所述处理器310执行时，执行如图1所示实施例中的中文分词方法。

[0089] 上述电子设备的具体细节可以对应参阅图1所示的实施例中对应的相关描述和效果进行理解，此处不再赘述。

[0090] 本实施例还提供了一种计算机存储介质，所述计算机存储介质存储有计算机可执行指令，该计算机可执行指令可执行上述任意方法实施例中中文分词方法。其中，所述存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)、随机存储记忆体(Random Access Memory，RAM)、快闪存储器(Flash Memory)、硬盘(Hard Disk Drive，缩写：HDD)或固态硬盘(Solid-State Drive，SSD)等；所述存储介质还可以包括上述种类的存储器的组合。

[0091] 显然，上述实施例仅仅是为清楚地说明所作的举例，而并非对实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。而由此所引伸出的显而易见的变化或变动仍处于本发明创造的保护范围之中。

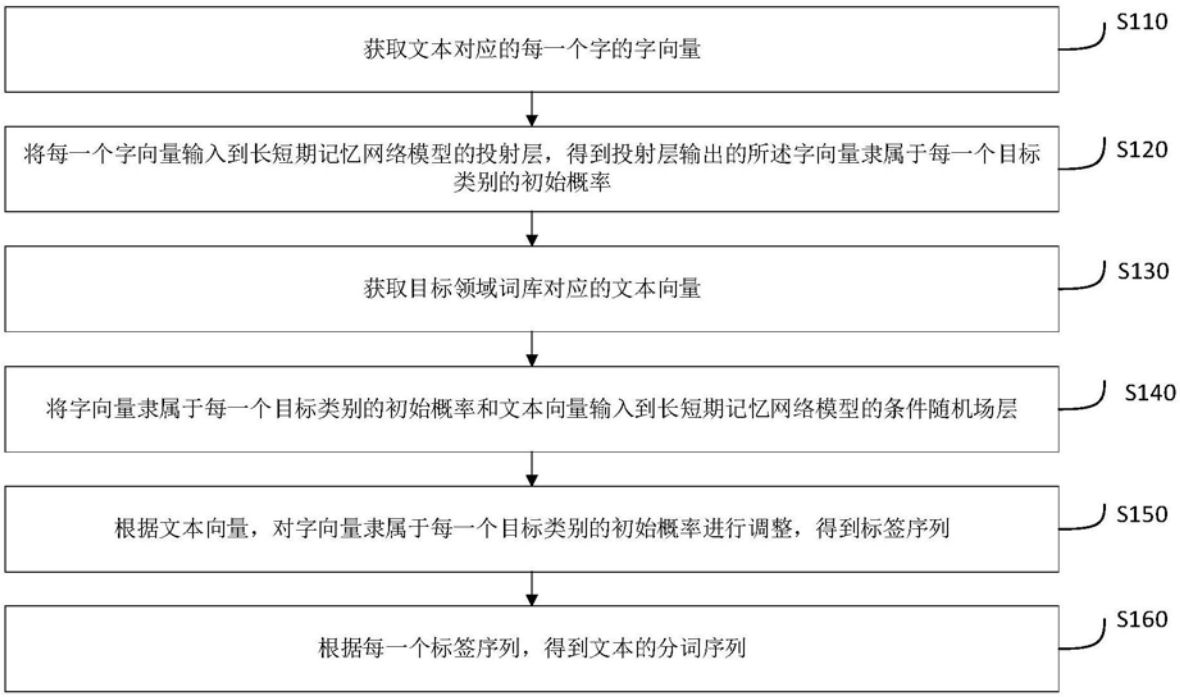

获取文本对应的每一个字的字向量
S110
将每一个字向量输入到长短期记忆网络模型的投射层，得到投射层输出的所述字向量隶属于每一个目标类别的初始概率
S120
获取目标领域词库对应的文本向量
S130
将字向量隶属于每一个目标类别的初始概率和文本向量输入到长短期记忆网络模型的条件随机场层
S140
根据文本向量，对字向量隶属于每一个目标类别的初始概率进行调整，得到标签序列
S150
根据每一个标签序列，得到文本的分词序列
S160

图1


字向量获取模块210
初始概率获取模块220
文本向量获取模块230
条件随机场层输入模块240
标签获取模块250
分词序列获取模块260

图2


处理器310
存储器320

图3