



(12) 发明专利申请

(10) 申请公布号 CN 112364657 A

(43) 申请公布日 2021. 02. 12

(21) 申请号 201910668488.8

(22) 申请日 2019.07.23

(71) 申请人 北京京东尚科信息技术有限公司
地址 100086 北京市海淀区知春路76号8层
申请人 北京京东世纪贸易有限公司

(72) 发明人 王培英 申磊 陈蒙

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219
代理人 张一军 张效荣

(51) Int. Cl.

G06F 40/30 (2020.01)

G06F 40/284 (2020.01)

G06N 3/04 (2006.01)

权利要求书2页 说明书10页 附图5页

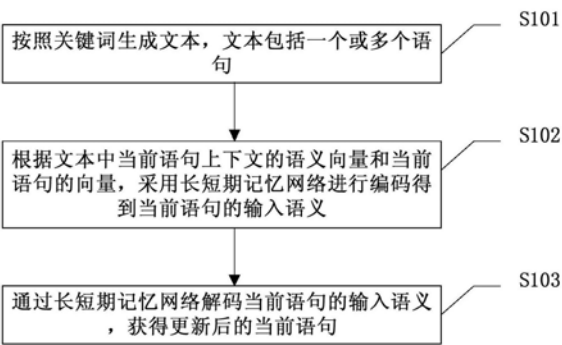
(54) 发明名称

生成文本的方法、装置、设备和计算机可读
介质

(57) 摘要

本发明公开了生成文本的方法、装置、设备和计算机可读介质,涉及计算机技术领域。该方法的一具体实施方式包括:按照关键词生成文本,所述文本包括一个或多个语句;根据所述语句的上下文的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。该实施方式能够提高生成文本的逻辑性,进而提高编辑文本的效率。

100



1. 一种生成文本的方法,其特征在于,包括:

按照关键词生成文本,所述文本包括一个或多个语句;

根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;

通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。

2. 根据权利要求1所述生成文本的方法,其特征在于,所述根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义,包括:

根据所述语句的上下句的语义向量、所述语句的语义向量和所述语句的预设词库向量,采用长短期记忆网络进行编码得到所述语句的输入语义。

3. 根据权利要求2所述生成文本的方法,其特征在于,所述根据所述语句的上下句的语义向量、所述语句的语义向量和所述语句的预设词库向量,采用长短期记忆网络进行编码得到所述语句的输入语义之前,还包括:

依据所述语句和所述语句对应的所述关键词,选取所述语句的相似语句;

选择所述语句中的词汇,按照词频逆文本频率排序构成所述语句的第一词库,选择所述相似语句中的词汇,按照词频逆文本频率排序构成所述语句的第二词库;

按照词频逆文本频率从高到低的顺序,在所述第二词库中筛选未在所述第一词库中的词汇,将前Y个筛选得到的词汇作为所述语句的润色词,Y是大于等于1的预设值;

将所述语句的润色词与所述语句对应的所述关键词共同编码,得到所述语句的预设词库向量。

4. 根据权利要求3所述生成文本的方法,其特征在于,所述选择所述语句中的词汇,构成所述语句的第一词库,选择所述相似语句中的词汇,构成所述语句的第二词库,包括:

选择所述语句中的名词、形容词和动词,构成所述语句的第一词库;

选择所述相似语句中的名词、形容词和动词,构成所述语句的第二词库。

5. 根据权利要求3所述生成文本的方法,其特征在于,所述依据所述语句和所述语句对应的所述关键词,选取所述语句的相似语句,包括:

依据所述语句和所述语句对应的所述关键词,在预设文本库中检索到前20个相关语句;

计算所述语句和所述相关语句的Jaccard相似度,选取Jaccard相似度在0.3至0.7之间最相似的所述相关语句作为所述相似语句。

6. 根据权利要求1所述生成文本的方法,其特征在于,所述按照关键词生成文本,包括:

将下一个语句的关键词和已生成的当前语句均转换为矢量,并基于长短期记忆网络进行编码得到下一个语句的输入语义;

根据所述下一个语句的输入语义、所述已生成的当前语句中的输出词和所述已生成的当前语句的隐藏状态,确定所述下一个语句的隐藏状态;

通过所述下一个语句的隐藏状态和所述下一个语句的关键词的矢量,解码得到所述下一个语句;

由所述已生成的当前语句和所述下一个语句生成所述文本。

7. 根据权利要求1所述生成文本的方法,其特征在于,所述通过长短期记忆网络解码所

述语句的输入语义,获得更新后的语句之后,还包括:

逐句更新所述文本中的一个或多个语句,生成包括所有更新后语句的文本。

8. 根据权利要求1所述生成文本的方法,其特征在于,所述方法还包括:

通过双向长短期记忆网络对所述语句编码,然后利用预设注意力模型,得到所述语句的语义向量。

9. 根据权利要求1所述生成文本的方法,其特征在于,所述方法还包括:

通过双向长短期记忆网络,对所述语句上文包括的 m 个语句和所述语句下文包括的 m 个语句,进行编码得到双向语义信息, m 是正整数;

利用预设注意力模型,结合所述双向语义信息,得到所述语句的上下句的语义向量。

10. 一种生成文本的装置,其特征在于,包括:

生成模块,用于按照关键词生成文本,所述文本包括一个或多个语句;

编码模块,用于根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;

解码模块,用于通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。

11. 一种生成文本的电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-9中任一所述的方法。

12. 一种计算机可读介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-9中任一所述的方法。

生成文本的方法、装置、设备和计算机可读介质

技术领域

[0001] 本发明涉及计算机技术领域,尤其涉及一种生成文本的方法、装置、设备和计算机可读介质。

背景技术

[0002] 从关键词到文本的自动生成,在很多技术领域应用,能够大大减少人工编辑文本的时间,进而提高编辑文本的效率。

[0003] 在实现本发明过程中,发明人发现现有技术中至少存在如下问题:按照关键词自动生成文本的逻辑性较差,从而降低编辑文本的效率。

发明内容

[0004] 有鉴于此,本发明实施例提供一种生成文本的方法、装置、设备和计算机可读介质,能够提高生成文本的逻辑性,进而提高编辑文本的效率。

[0005] 为实现上述目的,根据本发明实施例的一个方面,提供了一种生成文本的方法,包括:

[0006] 按照关键词生成文本,所述文本包括一个或多个语句;

[0007] 根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;

[0008] 通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。

[0009] 所述根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义,包括:

[0010] 根据所述语句的上下句的语义向量、所述语句的语义向量和所述语句的预设词库向量,采用长短期记忆网络进行编码得到所述语句的输入语义。

[0011] 所述根据所述语句的上下句的语义向量、所述语句的语义向量和所述语句的预设词库向量,采用长短期记忆网络进行编码得到所述语句的输入语义之前,还包括:

[0012] 依据所述语句和所述语句对应的所述关键词,选取所述语句的相似语句;

[0013] 选择所述语句中的词汇,按照词频逆文本频率排序构成所述语句的第一词库,选择所述相似语句中的词汇,按照词频逆文本频率排序构成所述语句的第二词库;

[0014] 按照词频逆文本频率从高到低的顺序,在所述第二词库中筛选未在所述第一词库中的词汇,将前Y个筛选得到的词汇作为所述语句的润色词,Y是大于等于1的预设值;

[0015] 将所述语句的润色词与所述语句对应的所述关键词共同编码,得到所述语句的预设词库向量。

[0016] 所述选择所述语句中的词汇,构成所述语句的第一词库,选择所述相似语句中的词汇,构成所述语句的第二词库,包括:

[0017] 选择所述语句中的名词、形容词和动词,构成所述语句的第一词库;

[0018] 选择所述相似语句中的名词、形容词和动词,构成所述语句的第二词库。

- [0019] 所述依据所述语句和所述语句对应的所述关键词,选取所述语句的相似语句,包括:
- [0020] 依据所述语句和所述语句对应的所述关键词,在预设文本库中检索到前20个相关语句;
- [0021] 计算所述语句和所述相关语句的Jaccard相似度,选取Jaccard相似度在0.3至0.7之间最相似的所述相关语句作为所述相似语句。
- [0022] 所述按照关键词生成文本,包括:
- [0023] 将下一个语句的关键词和已生成的当前语句均转换为矢量,并基于长短期记忆网络进行编码得到下一个语句的输入语义;
- [0024] 根据所述下一个语句的输入语义、所述已生成的当前语句中的输出词和所述已生成的当前语句的隐藏状态,确定所述下一个语句的隐藏状态;
- [0025] 通过所述下一个语句的隐藏状态和所述下一个语句的关键词的矢量,解码得到所述下一个语句;
- [0026] 由所述已生成的当前语句和所述下一个语句生成所述文本。
- [0027] 所述通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句之后,还包括:
- [0028] 逐句更新所述文本中的一个或多个语句,生成包括所有更新后语句的文本。
- [0029] 所述方法还包括:
- [0030] 通过双向长短期记忆网络对所述语句编码,然后利用预设注意力模型,得到所述语句的语义向量。
- [0031] 所述方法还包括:
- [0032] 通过双向长短期记忆网络,对所述语句上文包括的m个语句和所述语句下文包括的m个语句,进行编码得到双向语义信息,m是正整数;
- [0033] 利用预设注意力模型,结合所述双向语义信息,得到所述语句的上下句的语义向量。
- [0034] 根据本发明实施例的第二方面,提供了一种生成文本的装置,包括:
- [0035] 生成模块,用于按照关键词生成文本,所述文本包括一个或多个语句;
- [0036] 编码模块,用于根据所述语句的上下句的语义向量和所述语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;
- [0037] 解码模块,用于通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。
- [0038] 根据本发明实施例的第三方面,提供了一种生成文本的电子设备,包括:
- [0039] 一个或多个处理器;
- [0040] 存储装置,用于存储一个或多个程序,
- [0041] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如上述的方法。
- [0042] 根据本发明实施例的第四方面,提供了一种计算机可读介质,其上存储有计算机程序,所述程序被处理器执行时实现如上述的方法。
- [0043] 上述发明中的一个实施例具有如下优点或有益效果:因为按照关键词生成文本,

文本包括一个或多个语句;根据语句的上下句的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到语句的输入语义;通过长短期记忆网络解码语句的输入语义,获得更新后的语句,完成语句的更新。利用语句的上下句更新语句,能够提高生成文本的逻辑性,进而提高编辑文本的效率。

[0044] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0045] 附图用于更好地理解本发明,不构成对本发明的不当限定。其中:

[0046] 图1是根据本发明实施例的生成文本的方法的主要流程的示意图;

[0047] 图2是根据本发明实施例的生成文本的示意图;

[0048] 图3是根据本发明实施例的语句编码的示意图;

[0049] 图4是根据本发明实施例的另一个语句编码的示意图;

[0050] 图5是根据本发明实施例的构成预设词库向量的流程的示意图;

[0051] 图6是根据本发明实施例的生成文本的装置的主要结构的示意图

[0052] 图7是本发明实施例可以应用于其中的示例性系统架构图;

[0053] 图8是适于用来实现本发明实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0054] 以下结合附图对本发明的示范性实施例做出说明,其中包括本发明实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本发明的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0055] 以深度学习为代表的人工智能技术取得了越来越多的突破,可以将人工智能技术应用于生成文本。如:将文本生成过程作为一个机器翻译过程。将上一句作为源语句,下一句作为目标语句,根据预设关键词生成下一句,重复上述过程以生成文本。

[0056] 当按照上述方式生成文本,往往仅考虑前文信息无从考虑下文,上下文衔接松散,导致生成文本的逻辑性较差,降低编辑文本的效率。

[0057] 为了解决编辑文本的效率较低的技术问题,可以采用以下本发明实施例中的技术方案。

[0058] 参见图1,图1是根据本发明实施例的生成文本的方法主要流程的示意图,通过语句的上下句的语义向量更新语句,进而更新文本。如图1所示,具体包括以下步骤:

[0059] S101、按照关键词生成文本,文本包括一个或多个语句。

[0060] 在本发明实施例中,关键词是生成文本的基础。关键词可以预先设置。作为一个示例,可以根据一个关键词生成一个语句,由一个或多个语句构成文本。

[0061] 预先设置多个关键词: $K = \{k_i\}_{i=1}^n$, K 是关键词的集合, k_i 是第 i 个关键词,共有 n 个关键词。其中 i 的取值是大于等于1,且小于 n 。每个关键词对应生成一个语句。参见表1,表1中包括关键词,语句,以及关键词和语句的对应关系。

[0062] 表1

[0063]

关键词	语句
月色	月色迷离
相送	苦苦相送
沉重	情深心难重

[0064] 表1中包括三个关键词,每个关键词对应生成一个词句。关键词是月色,对应语句是月色迷离;关键词是相送,对应语句是苦苦相送;关键词是沉重,对应语句是情深心难重。

[0065] 在本发明的实施例中,可以利用Seq2Seq模型,将已生成的语句和下一个语句的关键词作为输入,将下一个语句作为目标语句输出。

[0066] 具体来说,将已生成的语句和下一个语句的关键词先转换为矢量,然后基于长短期记忆网络(Long Short-Term Memory,LSTM)的进行编码,得到下一个语句的输入语义 c_t 。

[0067] 解码端同样使用LSTM进行解码。在解码过程中,下一个语句的隐藏状态 s_t 与三个变量相关:已生成的语句的隐藏状态 s_{t-1} ,下一个语句的输入语义 c_t ;以及已生成的语句解码的输出词 $e_{y_{t-1}}$ 。具体如公式(1)所示。

$$[0068] \quad s_t = \text{LSTM} (s_{t-1}, c_t, e_{y_{t-1}}) \quad (1)$$

[0069] 其中, c_t 是通过注意力机制学习到的输入语义,表示解码器解码当前时刻的隐藏状态时,对输入向量采取不同的权重表示。

[0070] 基于下一个语句的输入语义 c_t ,结合公式(1),得到下一个语句的隐藏状态 s_t 。通过下一个语句的隐藏状态 s_t 和下一个语句的关键词的矢量,使用LSTM进行解码得到下一个语句。

[0071] 参见图2,图2是根据本发明实施例的生成文本的示意图。其中,图2中输入语义 c 的左侧是编码过程,输入语义 c 的右侧是解码过程。

[0072] 在编码过程中,将已生成的语句和下一个语句的关键词作为输入。示例性的,已生成的语句包括:吹散在风里,关键词:回头。那么,将已生成的语句分词,分词结果为:吹散,在,风里。其次,将词汇向量化,即以向量的方式表示词汇。作为一个示例,可以利用独热编码(OneHot),或词频逆文本频率(Term Frequency-Inverse Document Frequency,TF-IDF),将词汇向量化。将每个向量化的词汇输入LSTM中,经输入语义 c 后,进入解码过程。

[0073] 在解码过程中,由输入语义 c 输入至LSTM中输出下一个语句的第一个词汇,然后将第一词汇作为输入词汇,向量化后输入至LSTM中,输出下一句语句的第二个词汇:回头。按照上述方式,依次得到第三个词汇:你在,第四个词汇:那里。至此,输出下一个语句:猛然回头你在那里。

[0074] 按照图2中的示例,结合关键词,依次生成文本中的每个语句,最终基于生成的语句,构成文本。将文本标记为 $L = \{l_i\}_{i=1}^n$ 。其中,每个关键词对应一个语句,共有 n 个语句。

[0075] S102、根据语句的上下句的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到语句的输入语义。

[0076] 图2中生成语句是根据上一个语句,生成下一个语句,仅考虑到前文信息,并未参考下文信息,难以较好的做到上下文衔接。

[0077] 在本发明的一个实施例中,可以根据语句的上下句的语义向量和语句的语义向量,采用LSTM进行编码,得到语句的输入语义。

[0078] 文本包括一个或多个语句,按照文本中语句的先后顺序,依次采用LSTM进行编码。语句即按照文本中语句的先后顺序选择的。

[0079] 在编码过程中,不仅考虑历史生产内容,即文本中语句的前 m 个语句: $l_{i-m:i-1}$,而且已经生成的文本也提供语句的后 m 个语句 $l_{i+m:i+1}$,则可以同时考虑利用语句的上下句进行编码。其中, $l_{i-m:i-1}$, i 是第 i 个关键词,示例性的,每个关键词对应一个语句; m 是正整数。

[0080] 在本发明的一个实施例中,为了有效利用上下文信息,通过双向长短期记忆网络(Bi-LSTM),对文本中语句上文包括的 m 个语句和文本中语句下文包括的 m 个语句,进行编码得到双向语义信息。

$$[0081] \quad \vec{h}_j = \text{LSTM}(e_j, \vec{h}_{j-1}) \quad (2)$$

$$[0082] \quad \overleftarrow{h}_j = \text{LSTM}(e_j, \overleftarrow{h}_{j-1}) \quad (3)$$

[0083] 其中, e_h 是语句, $\vec{h}_j$ 表示前向隐藏状态, $\overleftarrow{h}_j$ 表示后向隐藏状态,可以利用 $\vec{h}_j$ 和 $\overleftarrow{h}_j$ 表示双向语义信息:

$$[0084] \quad \{h_j|h_j = \vec{h}_j \oplus \overleftarrow{h}_j\}_{j=1}^{2m} \quad (4)$$

[0085] 利用预设注意力模型,结合双向语义信息,得到语句的上下句的语义向量。其中,注意力模型是利用LSTM预先设置的模型。注意力模型在输出序列中的每个词汇时会专注考虑输入序列中的一些被认为比较重要的词。

[0086] 在上述实施例中,通过文本的上下句,得到语句上下句的语义向量,使得能够结合语句的上下句进行编码。

[0087] 语句是编码的基础。需要将语句进行编码,从而得到语句的语义向量。

[0088] 在本发明的一个实施例中,可以先通过Bi-LSTM对语句1编码,然后可以采用预设注意力模型获得语句的语义向量。

[0089] 参见图3,图3是根据本发明实施例的语句编码的示意图。图3主要包括两个部分,左部分和右部分。

[0090] 图3的左部分,将语句的前 m 个语句和语句的后 m 个语句,结合包括多个LSTM的Bi-LSTM进行编码,然后通过预设注意力模型输出语句的上下句的语义向量。

[0091] 图3的右部分,将语句输入多个LSTM,然后通过预设注意力模型输出语句的语义向量。

[0092] 根据语句的上下句的语义向量和语句的语义向量,确定输入语义。

[0093] 具体来说:语句的上下句的语义向量 $\oplus$ 语句的语义向量 $=c_t$,即得到公式(1)中的 c_t ,由于 s_{t-1} 是已知参数, e_{yt-1} 是已知参数,进而可以实现解码。

[0094] 参见图4,图4是根据本发明实施例的另一个更新语句的语义向量的示意图。图4中实施例与图3中实施例的不同之处在于,还可以在语句的上下句的语义向量和语句的语义向量的基础上,结合语句的预设词库向量,采用LSTM进行编码得到语句的输入语义。

[0095] 具体来说,图4主要包括三个部分,左部分、中间部分和右部分。

[0096] 图4的左部分,将语句的前 m 个语句和语句的后 m 个语句,结合包括多个LSTM的Bi-LSTM进行编码,然后通过预设注意力模型输出语句的上下句的语义向量。

[0097] 图4的中间部分,将语句输入多个LSTM,然后通过预设注意力模型输出语句的语义向量。

[0098] 图4的右部分,将 k_i 和 w_i 作为输入, k_i 是第 i 个关键词, w_i 预设词库中的词汇,预设词库 $W = \{w_i\}_{i=1}^{N_{iw}}$ 。将 w_i 和 k_i 一起通过预设的注意力模型进行编码得到:预设词库向量 iwv 。

$$[0099] \quad iwv = \sum_{w \in W \cup \{k_i\}} \alpha_w e_w \quad (5)$$

[0100] e_w 为每个预设词库中的词汇对应的嵌入(embedding)向量表示, α_w 为该词汇对应的注意力权重。

[0101] 根据语句的上下句的语义向量、语句的语义向量和语句的预设词库向量,确定输入语义。

[0102] 具体来说:语句的上下句的语义向量 $\oplus$ 语句的语义向量 $\oplus$ 语句的

预设词库向量 $= c_t$,即得到公式(1)中的 c_t ,由于 s_{t-1} 是已知参数, $e_{y_{t-1}}$ 是已知参数,进而可以实现解码。

[0103] 可见,图4中的实施例是在语句的上下句的语义向量和语句的语义向量的基础上,结合语句的预设词库向量,确定语句的输入语义。预设词库向量涉及的预设词库,是在按照关键词生成的文本的基础上预设建立的。这样,可以基于预设词库润色每个语句,进而提高每个语句的丰富用词。

[0104] 参见图5,图5是根据本发明实施例的构成预设词库向量的流程的示意图,具体包括:

[0105] S501、依据语句和语句对应的关键词,选取语句的相似语句。

[0106] 在预设的文本库中包括多个语句,语句是可以从名篇名作中选取的。按照S101中生成文本的语句1和该语句对应的关键词 k ,在预设文本库中检索以选取前 A 个相关语句,示例性的, A 等于20。

[0107] 分别计算语句1与相关语句的Jaccard相似度,示例性的,选取上述相似度在 $[0.3, 0.7]$ 之间最相似的一个相关语句作为相似语句,相似语句记为 r' 。Jaccard相似度用于比较有限样本集之间的相似性与差异性。Jaccard值越大,样本相似度越高。

[0108] S502、选择语句中的词汇,按照词频逆文本频率排序构成语句的第一词库,选择相似语句中的词汇,按照词频逆文本频率排序构成语句的第二词库。

[0109] 选择1中的词汇,按照TF-IDF排序构成1的第一词库;选择 r' 中的词汇,按照TF-IDF排序构成1的第二词库。可以理解的是,第一词库中词汇是1中的词汇;第二词库中词汇是与1相似语句中的词汇。示例性的,按照TF-IDF从高到低的排序,在第一词库或第二词库中排序越靠前的词汇重要程度越高。

[0110] 在本发明的一个实施例中,考虑到词性对于语义的影响,名词、形容词和动词往往具有更加丰富的语义信息。在构成词库的过程中,标注1中的词性,选择其中的名词、形容词和动词,构成1的第一词库。对应的,标注 r' 中的词性,选择其中的名词、形容词和动词,构成1的第二词库。

[0111] S503、按照词频逆文本频率从高到低的顺序,在第二词库中筛选未在第一词库中的词汇,将前Y个筛选得到的词汇作为语句的润色词,Y是大于等于1的预设值。

[0112] 第一词库与第二词库中可能存在相同的词汇。那么,为了选择可以润色语句的润色词,需要在第二词库中筛选未在第一词库中的词汇。具体来说,第二词库中的词汇已按照TF-IDF排序,则可以按照TF-IDF从高到低的顺序,在第二词库中筛选未在第一词库中的词汇,将前Y个筛选得到的词汇作为该语句的润色词。其中,Y是大于等于1的预设值。也就是说,润色词属于第二词库,且不属于第一词库。

[0113] 由润色词 w_i 构成预设词库 $W = \{w_i\}_{i=1}^{N_{iw}}$ 。其中, N_{iw} 是预设参数,Y等于 N_{iw} 。

[0114] S504、将润色词与语句对应的关键词共同编码,得到语句的预设词库向量。

[0115] 将 w_i 和 k_i 一起通过预设的注意力模型进行编码得到:1的预设词库向量 iwv 。

[0116] 文本中包括一个或多个语句,按照S501至S504,得到文本中每个语句的词库向量。进而由文本中所有语句的词库向量,构成预设词库向量。

[0117] 在上述实施例中,在文本中语句的基础上,通过相似语句,筛选出每个语句的润色词,进而构建语句的预设词库向量。从而能够实现对文本的润色。

[0118] S103、通过长短期记忆网络解码语句的输入语义,获得更新后的语句。

[0119] 根据语句的输入语义,通过LSTM就可以进行解码,获得更新后的语句。

[0120] 在本发明的一个实施例中,在获得更新后的语句之后,还可以逐句更新文本中的一个或多个语句,生成包括所有更新后语句的文本。

[0121] S101中生成的文本包括一个或多个语句,根据S102和S103,根据文本中语句的顺序,逐句更新文本的一个或多个语句。最终生成更新后的文本。更新后的文本可以包括所有更新后语句。

[0122] 在上述实施例中,根据语句的上下句的语义向量和语句的语义向量,采用LSTM进行编码得到语句的输入语义;然后,通过LSTM进行解码,获得更新后的语句。采用语句的上下句更新语句,更新后的当语句不仅考虑前句,还考虑到后句,因此文本的逻辑性得到增强,进而提高编辑文本的效率。

[0123] 此外,还可以通过预设词库润色语句。进一步提高文本用词的丰富性,以及上下文的连贯性。

[0124] 图6是根据本发明实施例的生成文本的装置的主要结构的示意图,生成文本的装置可以实现生成文本的方法,如图6所示,生成文本的装置具体包括:

[0125] 生成模块601,用于按照关键词生成文本,文本包括一个或多个语句。

[0126] 编码模块602,用于根据语句的上下句的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到语句的输入语义。

[0127] 解码模块603,用于通过长短期记忆网络解码语句的输入语义,获得更新后的语句。

[0128] 在本发明的一个实施例中,编码模块603,具体用于根据语句的上下句的语义向量、语句的语义向量和语句的预设词库向量,采用长短期记忆网络进行编码得到语句的输入语义。

[0129] 在本发明的一个实施例中,还包括词库模块605,用于依据语句和语句对应的所述关键词,选取语句的相似语句;

[0130] 选择语句中的词汇,按照词频逆文本频率排序构成语句的第一词库,选择相似语句中的词汇,按照词频逆文本频率排序构成语句的第二词库;

[0131] 按照词频逆文本频率从高到低的顺序,在第二词库中筛选未在第一词库中的词汇,将前Y个筛选得到的词汇作为语句的润色词,Y是大于等于1的预设值;

[0132] 将润色词与语句对应的关键词共同编码,得到语句的预设词库向量。

[0133] 在本发明的一个实施例中,词库模块605,具体用于选择语句中的名词、形容词和动词,构成语句的第一词库;

[0134] 选择相似语句中的名词、形容词和动词,构成语句的第二词库。

[0135] 在本发明的一个实施例中,词库模块605,具体用于依据语句和语句对应的关键词,在预设文本库中检索到前20个相关语句;

[0136] 计算语句和相关语句的Jaccard相似度,选取Jaccard相似度在0.3至0.7之间最相似的相关语句作为相似语句。

[0137] 在本发明的一个实施例中,生成模块601,具体用于将下一个语句的关键词和已生成的当前语句均转换为矢量,并基于长短期记忆网络进行编码得到下一个语句的输入语义;

[0138] 根据下一个语句的输入语义、已生成的当前语句中的输出词和已生成的当前语句的隐藏状态,确定下一个语句的隐藏状态;

[0139] 通过下一个语句的隐藏状态和下一个语句的关键词的矢量,解码得到下一个语句;

[0140] 由已生成的当前语句和下一个语句生成所述文本。

[0141] 在本发明的一个实施例中,还包括更新模块604,用于逐句更新文本中的一个或多个语句,生成包括所有更新后语句的文本。

[0142] 在本发明的一个实施例中,装置还包括语句模块606,用于通过双向长短期记忆网络对语句编码,然后利用预设注意力模型,得到语句的语义向量。

[0143] 在本发明的一个实施例中,装置还包括上下句模块607,用于通过双向长短期记忆网络,对文本中语句上文包括的m个语句和文本中语句下文包括的m个语句,进行编码得到双向语义信息,m是正整数;

[0144] 利用预设注意力模型,结合双向语义信息,得到语句的上下句的语义向量。

[0145] 图7示出了可以应用本发明实施例的生成文本的方法或生成文本的装置的示例性系统架构700。

[0146] 如图7所示,系统架构700可以包括终端设备701、702、703,网络704和服务器705。网络704用以在终端设备701、702、703和服务器705之间提供通信链路的介质。网络704可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0147] 用户可以使用终端设备701、702、703通过网络704与服务器705交互,以接收或发送消息等。终端设备701、702、703上可以安装有各种通讯客户端应用,例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0148] 终端设备701、702、703可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0149] 服务器705可以是提供各种服务的服务器,例如对用户利用终端设备701、702、703

所浏览的购物类网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的产品信息查询请求等数据进行分析等处理,并将处理结果(例如目标推送信息、产品信息—仅为示例)反馈给终端设备。

[0150] 需要说明的是,本发明实施例所提供的生成文本的方法一般由服务器705执行,相应地,生成文本的装置一般设置于服务器705中。

[0151] 应该理解,图7中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。

[0152] 下面参考图8,其示出了适于用来实现本发明实施例的终端设备的计算机系统800的结构示意图。图8示出的终端设备仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0153] 如图8所示,计算机系统800包括中央处理单元(CPU)801,其可以根据存储在只读存储器(ROM)802中的程序或者从存储部分808加载到随机访问存储器(RAM)803中的程序而执行各种适当的动作和处理。在RAM 803中,还存储有系统800操作所需的各种程序和数据。CPU 801、ROM 802以及RAM 803通过总线804彼此相连。输入/输出(I/O)接口805也连接至总线804。

[0154] 以下部件连接至I/O接口805:包括键盘、鼠标等的输入部分806;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分807;包括硬盘等的存储部分808;以及包括诸如LAN卡、调制解调器等网络接口卡的通信部分809。通信部分809经由诸如因特网的网络执行通信处理。驱动器810也根据需要连接至I/O接口805。可拆卸介质811,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器810上,以便于从其上读出的计算机程序根据需要被安装入存储部分808。

[0155] 特别地,根据本发明公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本发明公开的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分809从网络上被下载和安装,和/或从可拆卸介质811被安装。在该计算机程序被中央处理单元(CPU)801执行时,执行本发明的系统中限定的上述功能。

[0156] 需要说明的是,本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本发明中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于

由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0157] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0158] 描述于本发明实施例中涉及到的模块可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括发送单元、获取单元、确定单元和第一处理单元。其中,这些单元的名称在某种情况下并不构成对该单元本身的限定,例如,发送单元还可以被描述为“向所连接的服务端发送图片获取请求的单元”。

[0159] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该设备执行时,使得该设备包括:

[0160] 按照关键词生成文本,所述文本包括一个或多个语句;

[0161] 根据所述语句的上下句的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到所述语句的输入语义;

[0162] 通过长短期记忆网络解码所述语句的输入语义,获得更新后的语句。

[0163] 根据本发明实施例的技术方案,因为按照关键词生成文本,文本包括一个或多个语句;根据语句的上下句的语义向量和语句的语义向量,采用长短期记忆网络进行编码得到语句的输入语义;通过长短期记忆网络解码语句的输入语义,获得更新后的语句,完成语句的更新。利用语句的上下句更新语句,能够提高生成文本的逻辑性,进而提高编辑文本的效率。

[0164] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

100

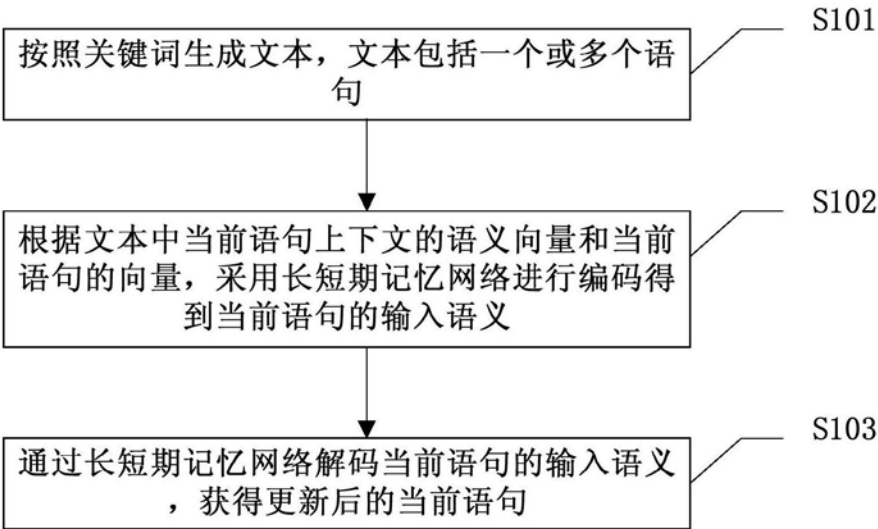


图1

200

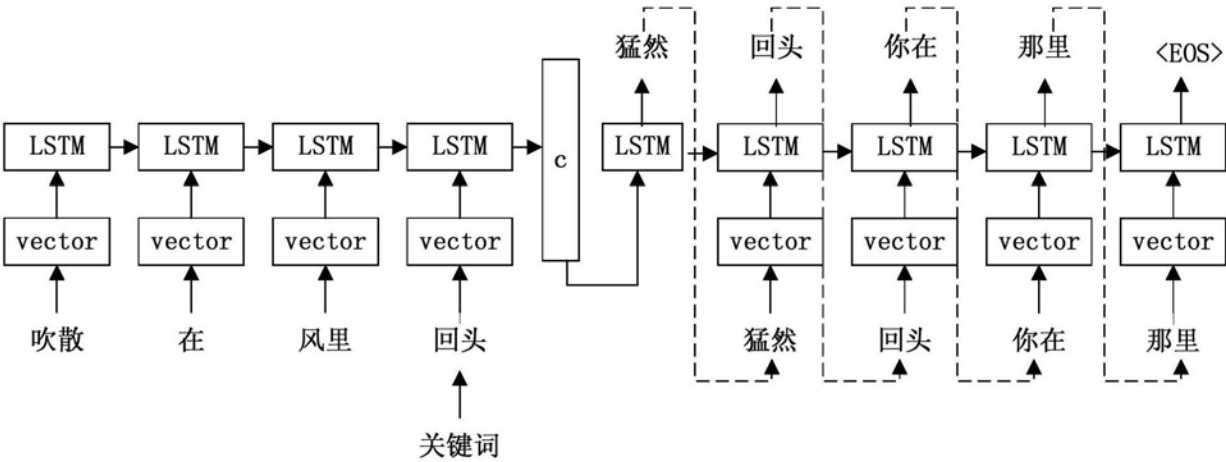


图2

300

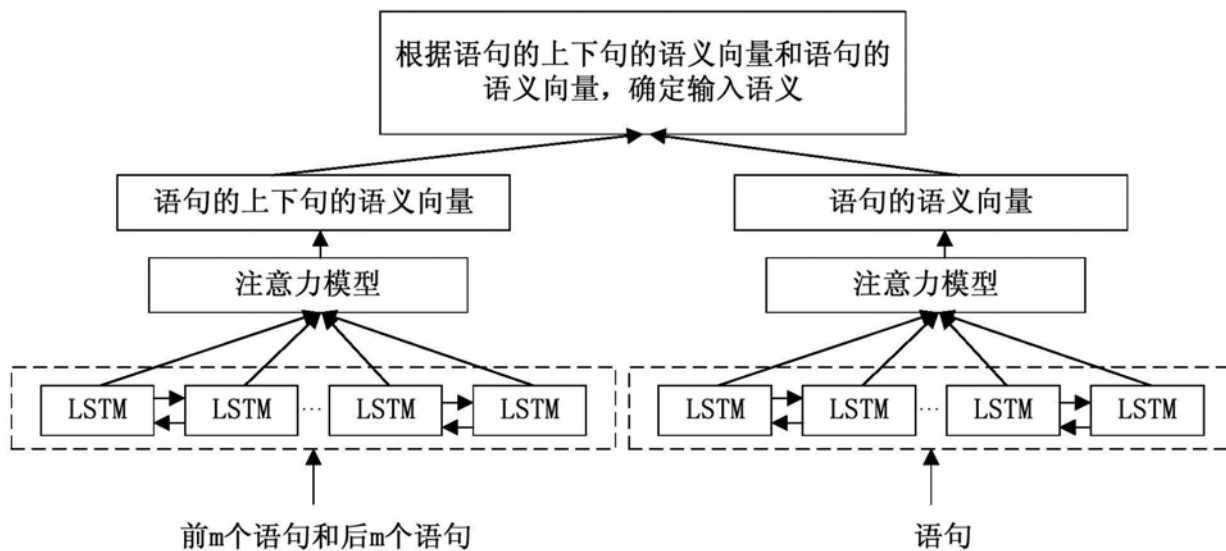


图3

400

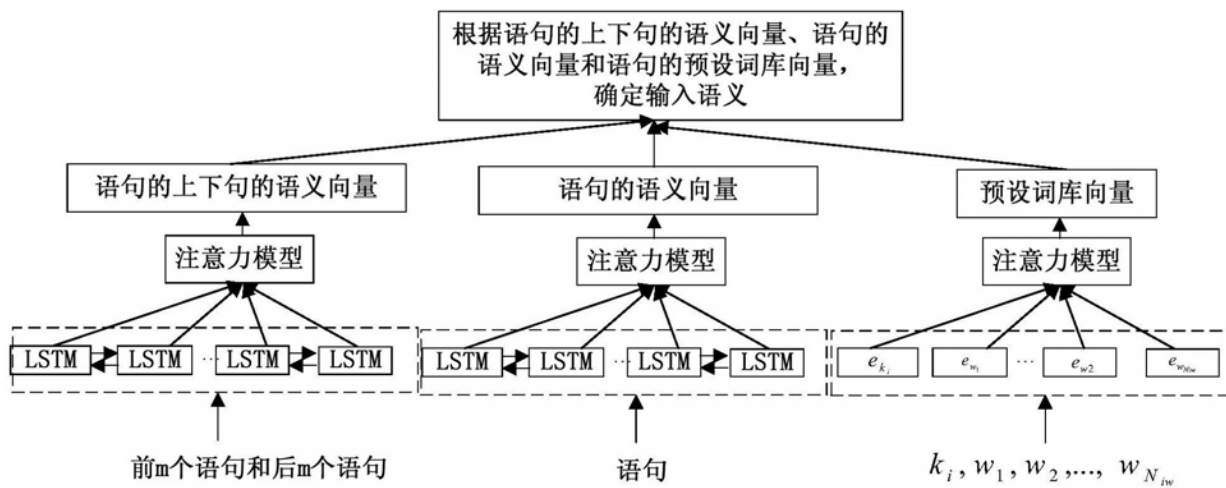


图4

500

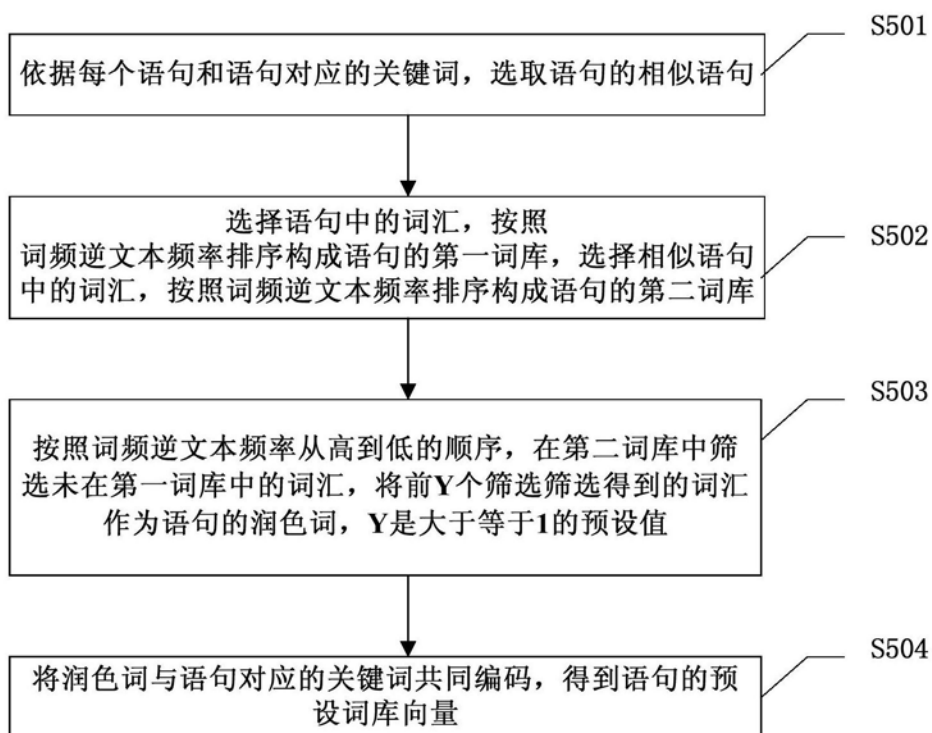


图5

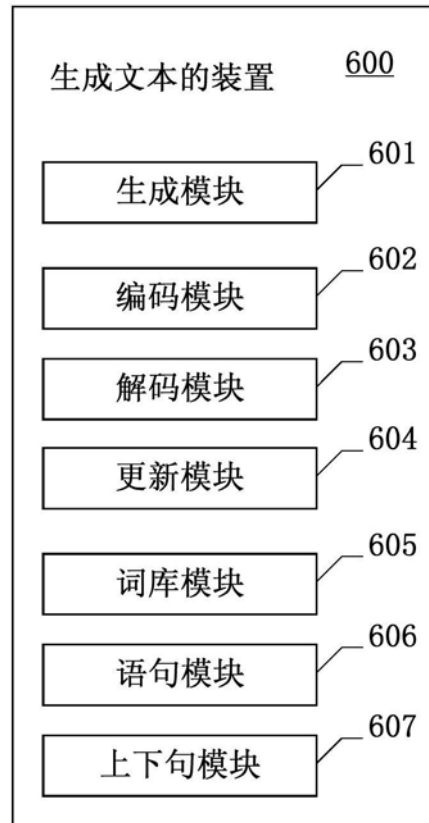


图6

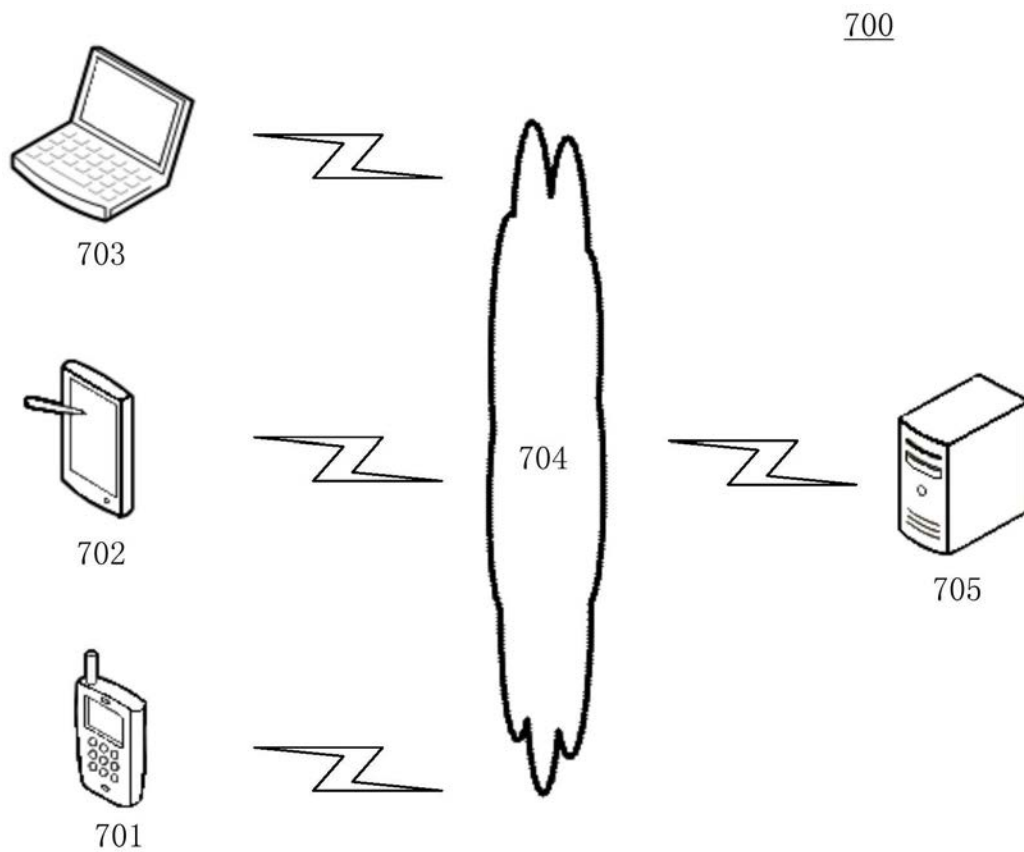


图7

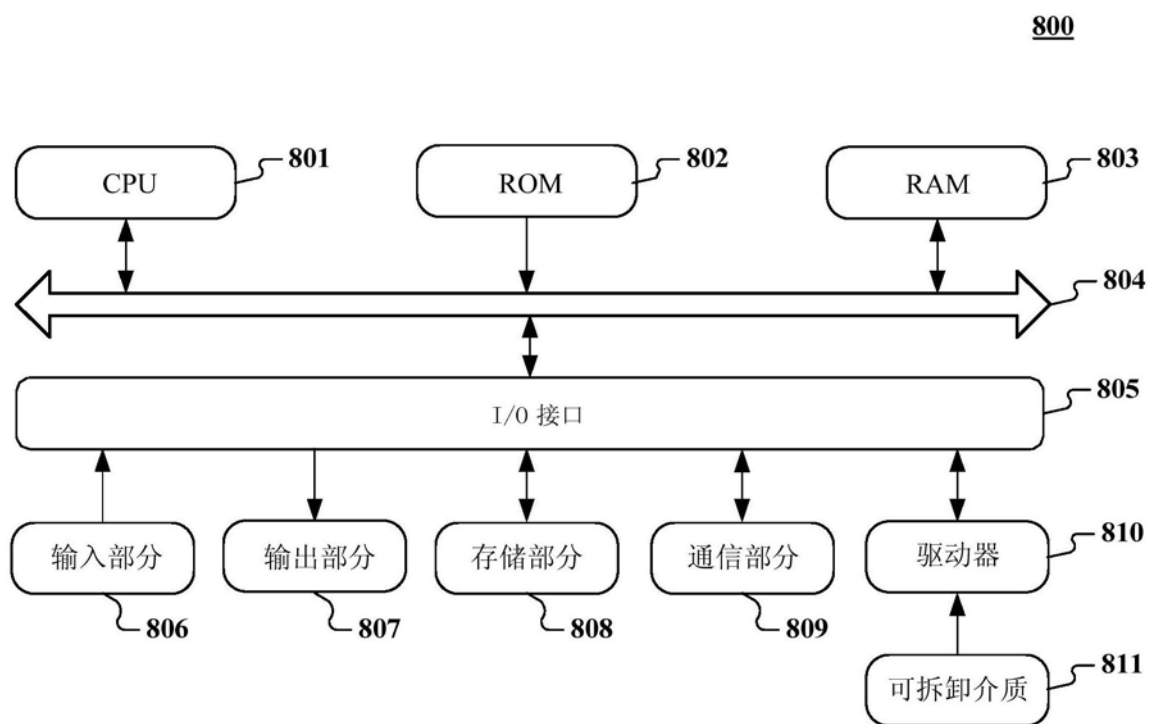


图8