



(12) 发明专利申请

(10) 申请公布号 CN 115238027 A

(43) 申请公布日 2022. 10. 25

(21) 申请号 202110448859.9

G06N 3/04 (2006.01)

(22) 申请日 2021.04.25

G06N 3/08 (2006.01)

(71) 申请人 京东科技控股股份有限公司

地址 100176 北京市大兴区经济技术开发区
科创十一街18号C座2层221室

(72) 发明人 王培英 陈蒙

(74) 专利代理机构 北京英赛嘉华知识产权代理
有限责任公司 11204

专利代理师 王达佐 马晓亚

(51) Int.Cl.

G06F 16/33 (2019.01)

G06F 16/332 (2019.01)

G06F 16/35 (2019.01)

G06F 40/295 (2020.01)

G06K 9/62 (2022.01)

权利要求书2页 说明书12页 附图4页

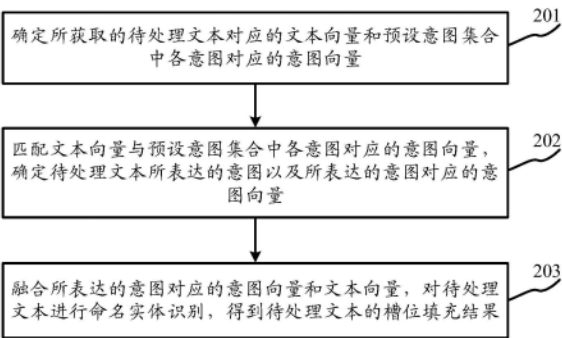
(54) 发明名称

文本数据的处理方法及装置

(57) 摘要

本申请公开了一种文本数据的处理方法及装置。方法的一具体实施方式包括：确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量；匹配文本向量与预设意图集合中各意图对应的意图向量，确定待处理文本所表达的意图以及所表达的意图对应的意图向量；融合所表达的意图对应的意图向量和文本向量，对待处理文本进行命名实体识别，得到待处理文本的槽位填充结果。本申请提供了一种进行意图识别和槽位填充的文本数据的处理方法，并在文本向量的基础上融合意图向量确定槽位填充结果，提高了槽位填充的准确度。

200



1. 一种文本数据的处理方法,包括:

确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;

匹配所述文本向量与所述预设意图集合中各意图对应的意图向量,确定所述待处理文本所表达的意图以及所表达的意图对应的意图向量;

融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果。

2. 根据权利要求1所述的方法,其中,所述融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果,包括:

基于门控机制,融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果。

3. 根据权利要求2所述的方法,其中,所述基于门控机制,融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果,包括:

基于门控机制,融合所表达的意图对应的意图向量和所述文本向量,确定所述待处理文本中的槽值;

针对于所述待处理文本中的每个槽值,确定该槽值所属的槽位类型。

4. 根据权利要求1所述的方法,其中,所述文本数据的处理方法通过文本处理模型执行,其中,所述文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系;以及

所述文本处理模型通过如下方法训练得到:

获取训练样本集,其中,所述训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型;

基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到所述文本处理模型。

5. 根据权利要求4所述的方法,其中,所述基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到所述文本处理模型,包括:

针对于所述训练样本集中的训练样本,执行如下训练操作,直至确定所述文本处理模型训练完成:确定该训练样本中的文本样本对应的文本向量;匹配该文本样本对应的文本向量与所述预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定所述文本处理模型是否训练完成。

6. 一种文本数据的处理装置,包括:

确定单元,被配置成确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;

匹配单元,被配置成匹配所述文本向量与所述预设意图集合中各意图对应的意图向量,确定所述待处理文本所表达的意图以及所表达的意图对应的意图向量;

融合单元,被配置成融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果。

7. 根据权利要求6所述的装置,其中,所述融合单元,进一步被配置成:

基于门控机制,融合所表达的意图对应的意图向量和所述文本向量,对所述待处理文本进行命名实体识别,得到所述待处理文本的槽位填充结果。

8. 根据权利要求7所述的装置,其中,所述融合单元,进一步被配置成:

基于门控机制,融合所表达的意图对应的意图向量和所述文本向量,确定所述待处理文本中的槽值;针对于所述待处理文本中的每个槽值,确定该槽值所属的槽位类型。

9. 根据权利要求6所述的装置,其中,所述文本数据的处理方法通过文本处理模型执行,其中,所述文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系;以及

还包括,训练单元,被配置成通过如下方法训练得到所述文本处理模型:

获取训练样本集,其中,所述训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型;基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到所述文本处理模型。

10. 根据权利要求9所述的装置,其中,所述训练单元,进一步被配置成:

针对于所述训练样本集中的训练样本,执行如下训练操作,直至确定所述文本处理模型训练完成:确定该训练样本中的文本样本对应的文本向量;匹配该文本样本对应的文本向量与所述预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定所述文本处理模型是否训练完成。

11. 一种计算机可读介质,其上存储有计算机程序,其中,所述程序被处理器执行时实现如权利要求1-5中任一所述的方法。

12. 一种电子设备,包括:

一个或多个处理器;

存储装置,其上存储有一个或多个程序,

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1-5中任一所述的方法。

文本数据的处理方法及装置

技术领域

[0001] 本申请实施例涉及计算机技术领域,具体涉及一种文本数据的处理方法及装置。

背景技术

[0002] 自然语言处理(Natural Language Processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。意图识别和槽位填充是自然语言处理中的两个重要的子任务,在聊天机器人、智能语音助手等方面有着广泛的应用。其中,意图识别表征识别待处理文本所表达的意图,槽位填充表征识别待处理文本中的命名实体(槽值)以及对应的实体类型(槽位)。

发明内容

[0003] 本申请实施例提出了一种文本数据的处理方法及装置。

[0004] 第一方面,本申请实施例提供了一种文本数据的处理方法,包括:确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0005] 在一些实施例中,上述融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果,包括:基于门控机制,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0006] 在一些实施例中,上述基于门控机制,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果,包括:基于门控机制,融合所表达的意图对应的意图向量和文本向量,确定待处理文本中的槽值;针对于待处理文本中的每个槽值,确定该槽值所属的槽位类型。

[0007] 在一些实施例中,文本数据的处理方法通过文本处理模型执行,其中,文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系;以及文本处理模型通过如下方法训练得到:获取训练样本集,其中,训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型;基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到文本处理模型。

[0008] 在一些实施例中,上述基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到文本处理模型,包括:针对于训练样本集中的训练样本,执行如下训练操作,直至确定文本处理模型训练完成:确定该训练样本中的文本样本对应的文本向量;匹配该文本样本对应的文本向

量与预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定文本处理模型是否训练完成。

[0009] 第二方面,本申请实施例提供了一种文本数据的处理装置,包括:确定单元,被配置成确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配单元,被配置成匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合单元,被配置成融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0010] 在一些实施例中,融合单元,进一步被配置成:基于门控机制,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0011] 在一些实施例中,融合单元,进一步被配置成:基于门控机制,融合所表达的意图对应的意图向量和文本向量,确定待处理文本中的槽值;针对于待处理文本中的每个槽值,确定该槽值所属的槽位类型。

[0012] 在一些实施例中,文本数据的处理方法通过文本处理模型执行,其中,文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系;以及上述装置还包括,训练单元,被配置成通过如下方法训练得到文本处理模型:获取训练样本集,其中,训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型;基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到文本处理模型。

[0013] 在一些实施例中,训练单元,进一步被配置成:针对于训练样本集中的训练样本,执行如下训练操作,直至确定文本处理模型训练完成:确定该训练样本中的文本样本对应的文本向量;匹配该文本样本对应的文本向量与预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定文本处理模型是否训练完成。

[0014] 第三方面,本申请实施例提供了一种计算机可读介质,其上存储有计算机程序,其中,程序被处理器执行时实现如第一方面任一实现方式描述的方法。

[0015] 第四方面,本申请实施例提供了一种电子设备,包括:一个或多个处理器;存储装置,其上存储有一个或多个程序,当一个或多个程序被一个或多个处理器执行,使得一个或多个处理器实现如第一方面任一实现方式描述的方法。

[0016] 本申请实施例提供的文本数据的处理方法及装置,通过确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配文本向量与预设意图集合

中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果,从而提供了一种进行意图识别和槽位填充的文本数据的处理方法,并在文本向量的基础上融合意图向量确定槽位填充结果,提高了槽位填充的准确度。

附图说明

[0017] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述,本申请的其它特征、目的和优点将会变得更明显:

[0018] 图1是本申请的一个实施例可以应用于其中的示例性系统架构图;

[0019] 图2是根据本申请文本数据的处理方法的一个实施例的流程图;

[0020] 图3是根据本实施例的文本数据的处理方法的应用场景的示意图;

[0021] 图4是根据本申请的文本数据的处理方法的又一个实施例的具体示意图;

[0022] 图5是根据本申请的训练文本处理模型的一个实施例的流程图;

[0023] 图6是根据本申请的文本数据的处理装置的一个实施例的结构图;

[0024] 图7是适于用来实现本申请实施例的计算机系统的结构示意图。

具体实施方式

[0025] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅仅用于解释相关发明,而非对该发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与有关发明相关的部分。

[0026] 需要说明的是,在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0027] 图1示出了可以应用本申请的文本数据的处理方法及装置的示例性架构100。

[0028] 如图1所示,系统架构100可以包括终端设备101、102、103,网络104和服务器105。终端设备101、102、103之间通信连接构成拓扑网络,网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104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。

[0029] 终端设备101、102、103可以是支持网络连接从而进行数据交互和数据处理的硬件设备或软件。当终端设备101、102、103为硬件时,其可以是支持网络连接,信息获取、交互、显示、处理等功能的各种电子设备,包括但不限于智能手机、平板电脑、电子书阅读器、膝上型便携计算机和台式计算机等等。终端设备上可以设置各种类型的应用,包括但不限于即时通讯应用、购物类应用、办公类应用。当终端设备101、102、103为软件时,可以安装在上述所列举的电子设备中。其可以实现成例如用来提供分布式服务的多个软件或软件模块,也可以实现成单个软件或软件模块。在此不做具体限定。

[0030] 服务器105可以是提供各种服务的服务器,例如获取终端设备101、102、103发送的待处理文本,对待处理文本进行意图识别和槽位填充的后台处理服务器。服务器105可以将意图识别和槽位填充的结果信息反馈至终端设备。作为示例,服务器105可以是云端服务器。

[0031] 需要说明的是,服务器可以是硬件,也可以是软件。当服务器为硬件时,可以实现

成多个服务器组成的分布式服务器集群,也可以实现成单个服务器。当服务器为软件时,可以实现成多个软件或软件模块(例如用来提供分布式服务的软件或软件模块),也可以实现成单个软件或软件模块。在此不做具体限定。

[0032] 还需要说明的是,本公开的实施例所提供的文本数据的处理方法可以由服务器执行,也可以由终端设备执行,还可以由服务器和终端设备彼此配合执行。相应地,文本数据的处理装置包括的各个部分(例如各个单元)可以全部设置于服务器中,也可以全部设置于终端设备中,还可以分别设置于服务器和终端设备中。

[0033] 应该理解,图1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。当文本数据的处理方法运行于其上的电子设备不需要与其他电子设备进行数据传输时,该系统架构可以仅包括文本数据的处理方法运行于其上的电子设备(例如服务器或终端设备)。

[0034] 继续参考图2,示出了文本数据的处理方法的一个实施例的流程200,包括以下步骤:

[0035] 步骤201,确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量。

[0036] 本实施例中,文本数据的处理方法的执行主体(例如图1中的服务器)可以确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量。

[0037] 其中,待处理文本可以是自然语言所表征的任意文本,其可以通过任何现有或将来的获取形式所获取,包括但不限于是输入即时文本、导入的已有文本、语音形式所转化的文本、基于字符识别所识别的文本。预设意图集合中包括各种各样的意图。具体的,当本实施例对应于所有应用场景时,预设意图集合中包括各种应用场景中待处理文本所涉及的意图;当本实施例对应于特定的应用场景时,预设意图集合中包括特定应用场景中待处理文本所涉及的意图。

[0038] 作为示例,上述执行主体可以通过预先训练的词向量模型确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量。其中,词向量模型例如是Skip-Gram模型、CBOW (Bag-of-words model,词袋模型)、GloVe (Global Vectors,全局向量)模型、BERT (Bidirectional Encoder Representations from Transformers,来自变压器的双向编码器表示)模型。

[0039] 步骤202,匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量。

[0040] 本实施例中,上述执行主体可以匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量。

[0041] 作为示例,上述执行主体确定文本向量与预设意图集合中每一个意图对应的意图向量之间的匹配度,将与文本向量匹配度最高的意图向量确定为对应于待处理文本的意图向量,将所确定的意图向量对应的意图确定为待处理文本所表达的意图。

[0042] 具体的,首先,上述执行主体计算待处理文本中每个字符与预设意图集合中每个意图对应的意图向量之间的匹配度分数。在这里,可以采用点积作为交互函数来计算待处理文本中的字符*i*和预设意图集合中的意图*k*之间的匹配度分数:

$$[0043] \quad l_{ki} = l_k \cdot h_i^T$$

[0044] 其中, I_{ki} 表征字符 i 和意图 k 之间的匹配度分数, l_k 表征预设意图集合中各个意图对应的意图向量, k 为 1 至预设意图集合所包括的意图的数量之间的整数, h_i^T 表征字符 i 转置后的字符向量, i 为 1 至待处理文本所包括的字符的数量之间的整数。

[0045] 进而, 可以确定:

$$[0046] \quad I = LH^T$$

[0047] 其中, L 表征预设意图集合中各意图的意图向量, 由预设意图集合中各意图的意图向量组成; H^T 表征待处理文本对应的文本向量的转置矩阵, 由待处理文本所包括的各字符对应的字符向量组成; I 为交互矩阵, I 中的每一行 I_k 表示意图 k 与待处理文本中的各个字符之间的交互特征。

[0048] 然后, 将每个意图 k 与待处理文本中的各个字符之间的交互特征 I_k 通过两个全连接层进行处理, 得到意图 k 与待处理文本之间的匹配度分数 s^k :

$$[0049] \quad s^k = W_2 \text{ReLU}(I_k W_1 + b)$$

[0050] 其中, ReLU 为第一层全连接层的激活函数, W_1 、 W_2 表征参数矩阵, b 表征偏差。

[0051] 最后, 将各意图的匹配得分 $s = [s^1, s^2, \dots, s^c]$ 进行归一化, 得到预测意图的概率分布:

$$[0052] \quad \hat{y}^I = \text{softmax}(s)$$

[0053] 并选取概率最高的意图作为待处理文本所表达的意图。其中, c 为预设意图集合中的意图的数量。

[0054] 步骤 203, 融合所表达的意图对应的意图向量和文本向量, 对待处理文本进行命名实体识别, 得到待处理文本的槽位填充结果。

[0055] 本实施例中, 上述执行主体可以融合所表达的意图对应的意图向量和文本向量, 对待处理文本进行命名实体识别, 得到待处理文本的槽位填充结果。

[0056] 在自然语言处理领域, 意图识别与槽位填充是息息相关的处理任务。在确定待处理文本所表达的意图后, 可以将待处理文本所表达的意图融入槽位填充任务, 以提高针对待处理文本的槽位填充结果的准确度。

[0057] 本实施例中, 上述执行主体可以根据任何融合方式融合待处理文本所表达的意图对应的意图向量和待处理文本对应的文本向量。作为示例, 可以基于预设权重矩阵, 融合所表达的意图对应的意图向量和文本向量, 对待处理文本进行命名实体识别, 得到待处理文本的槽位填充结果。作为示例, 可以通过 RNN 模型及其变体模型, 如 LSTM (Long Short-Term Memory, 长短期记忆) 模型、GRU (Gated Recurrent Unit, 门控循环单元) 模型预测待处理文本的槽位填充结果。

[0058] 其中, 命名实体识别 (Named Entity Recognition, NER) 任务本质是序列标注任务, 其数据标注方式也遵照序列标注问题的方式, 主要是 BIO 和 BIOES 两种。其中, B, 即 Begin, 表示开始; I, 即 Intermediate, 表示中间; O, 即 Other, 表示其他, 用于标记无关字符; S, 即 Single, 表示单个字符; E, 即 End, 表示结尾。命名实体识别的过程, 就是根据输入的待处理文本, 预测出其槽值标注序列的过程。

[0059] 作为示例, 针对于待处理文本的槽值标注序列如下表所示:

[0060]	槽值标注序列	槽位	槽值
--------	--------	----	----

[0061]	今天天气怎么样 B I O O O O O	Date (日期)	今天
	我要去万达广场 O O O B I I I	Location (位置)	万达广场
	推荐一款 1000 元左右的手机 O O O O B I O O O B I	Price (价格) Type (类型)	1000 元 手机

[0062] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行上述步骤203:

[0063] 基于门控机制,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0064] 上述执行主体可以通过激活函数(例如,sigmoid函数)充当门控信号,来控制意图向量和文本向量的融合,以在命名实体识别过程中有效利用意图的指导信息,同时避免错误意图造成的干扰。然后,根据融合信息对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0065] 在本实施例的一些可选的实现方式中,上述执行主体将槽位填充任务划分为两个子任务:槽值提取和槽位类型识别。

[0066] 上述执行主体可以通过如下方式执行上述步骤203:

[0067] 第一,基于门控机制,融合所表达的意图对应的意图向量和文本向量,确定待处理文本中的槽值。

[0068] 具体的,为了在命名实体识别过程中有效利用意图的指导信息,同时避免错误意图造成的干扰,本实现方式引入一个门函数 g_i 来控制意图向量的融入:

[0069] $g_i = \sigma(W_3 h_i + W_4 l_{int})$

[0070] 其中, σ 为sigmoid函数,该函数可以将数据转换到0-1范围内,从而可以用来充当门控信号, W_3 、 W_4 为权重矩阵, h_i 表征待处理文本中的字符i的字符向量, l_{int} 表征待处理文本所表达的意图对应的意图向量。

[0071] 然后,将向量 h_i 与意图向量 l_{int} 共同输入到分类层,得到预测到的槽值标注序列:

$\hat{y}^v = \{\hat{y}_1^v, \hat{y}_2^v, \dots, \hat{y}_n^v\}$:

[0072] $\hat{y}_i^v = \text{softmax}(W_5(h_i + g_i \odot l_{int}))$

[0073] 其中, W_5 表征权重矩阵, $\hat{y}_i^v$ 表征待处理文本中第i个字符的标注标签(例如,B、I、O)。

[0074] 第二,针对于待处理文本中的每个槽值,确定该槽值所属的槽位类型。

[0075] 预设的槽位类型集合为 $y^s = (y_1^s, y_2^s \dots y_m^s)$ 。M表征槽位类型集合中槽位类型的数量。

[0076] 根据上一步预测出来的标签序列,槽值可表示为 $h_{ent} = \frac{1}{r}(h^B + h^I)$,其中, h^B 是标签为“B”的字符对应的向量, h^I 为后续标签为“I”的字符(这里可能有多个)对应的向量,r

为槽值的长度。将各字符向量相加并取平均值得到槽值的编码表示,最后将其输入到一个全连接层,得到该槽值其归属于各槽位的概率分布:

$$[0077] \quad \hat{y}^s = \text{softmax}(W_6 h_{\text{ent}} + b)$$

[0078] 其中, W_6 表征权重矩阵, b 表征偏差。

[0079] 继续参见图3,图3是根据本实施例的文本数据的处理方法的应用场景的一个示意图300。在图3的应用场景中,目标用户301发出“我要去万达广场”的语音,终端设备302接收“我要去万达广场”的语音,并将其转化为文本形式的待处理文本,传输至服务器303。服务器303首先确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;然后,匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图“导航去万达广场”以及所表达的意图对应的意图向量。最后,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0080] 本公开的上述实施例提供的方法,通过确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果,从而提供了一种进行意图识别和槽位填充的文本数据的处理方法,并在文本向量的基础上融合意图向量确定槽位填充结果,提高了槽位填充的准确度。

[0081] 在本实施例的一些可选的实现方式中,上述执行主体可以通过文本处理模型执行上述文本数据的处理方法。其中,文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系。

[0082] 文本处理模型通过如下方法训练得到:

[0083] 第一,获取训练样本集。

[0084] 其中,训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型。

[0085] 第二,基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到文本处理模型。

[0086] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行上述第二步骤:针对于训练样本集中的训练样本,执行如下训练操作,直至确定文本处理模型训练完成:

[0087] 首先,确定该训练样本中的文本样本对应的文本向量;然后,匹配该文本样本对应的文本向量与预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;然后,融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;最后,根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定文本处理模型是否训练完成。

[0088] 作为示例,当该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值小于损失值阈值,则判定文本处理模型训练完成;当该

训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值不小于损失值阈值则判定文本处理模型未训练完成,从训练样本集中选取未经训练的训练样本继续进行文本处理模型的训练。

[0089] 本实现方式中,针对于意图、槽值标注序列、槽位类型,上述执行主体可以分别计算实际信息与预测信息之间的损失值。

[0090] 具体的,上述执行主体可以通过如下公式计算训练样本中的意图与所预测的意图之间的损失值:

$$[0091] \quad Loss_I = CrossEntropy(y^I, \hat{y}^I)$$

[0092] 其中, $Loss_I$ 表示训练样本中的意图 y^I 与所预测的意图 $\hat{y}^I$ 之间的损失值, $CrossEntropy()$ 表示交叉熵损失函数。

[0093] 上述执行主体可以通过如下公式计算训练样本中的槽值标注序列与所预测的槽值标注序列之间的损失值:

$$[0094] \quad Loss_v = CrossEntropy(y^v, \hat{y}^v)$$

[0095] 其中, $Loss_v$ 表示训练样本中的槽值标注序列 y^v 与所预测的槽值标注序列 $\hat{y}^v$ 之间的损失值。

[0096] 上述执行主体可以通过如下公式计算训练样本中的槽位类型与所预测的槽位类型之间的损失值:

$$[0097] \quad Loss_s = CrossEntropy(y^s, \hat{y}^s)$$

[0098] 其中, $Loss_s$ 表示训练样本中的槽位类型 y^s 与所预测的槽位类型 $\hat{y}^s$ 之间的损失值。

[0099] 在确定上述各损失值之后,上述执行主体通过如下公式计算总损失值:

$$[0100] \quad Loss = \lambda_1 Loss_I + \lambda_2 Loss_v + \lambda_3 Loss_s$$

[0101] 其中, $Loss$ 表示总损失值, λ_1 、 λ_2 、 λ_3 分别表示权重参数。

[0102] 继续参见图4,示出了根据本申请的文本数据的处理方法的一个实施例的具体示意图400。首先,采用预训练的BERT模型(文本处理模型)401,对待处理文本“我想要变频空调”进行编码表示,得到待处理文本的向量表示: $H = \{h_1, h_2, \dots, h_7\}$ 。然后,将预设意图集合402中的各个意图标签编码为意图向量: $L = \{l_1, l_2, \dots, l_c\} \in \mathbb{R}^{c \times d}$, 其中 c 表示意图的类别个数, d 为意图向量尺寸。然后,计算待处理文本中的每个字符的字符向量与预设意图集合中的每个意图的意图向量之间的匹配得分。在这里,采用如下公式计算字符 i 和意图 k 之间的匹配分数:

$$[0103] \quad I_{ki} = l_k \cdot h_i^T$$

$$[0104] \quad I = LH^T$$

[0105] 其中, I 中的每一行 I_k 表示意图 k 与待处理文本中的各个字符之间的交互特征。

[0106] 然后,将每个意图 k 的交互特征 I_k 通过两个全连接层,通过如下公式,计算得到该意图与待处理文本之间的匹配得分 s^k :

$$[0107] \quad s^k = W_2 \text{ReLU}(I_k W_1 + b)$$

[0108] 然后,将各意图的匹配得分 $s = [s^1, s^k, \dots, s^c]$ 进行归一化,得到预测类别的概率

分布：

$$[0109] \quad \hat{\mathbf{y}}^l = \text{softmax}(\mathbf{s})$$

[0110] 然后,选取概率最高的意图作为预测的意图,其对应的意图向量表示为 $\mathbf{l}_{\text{int}}$ 。

[0111] 然后,将意图向量 $\mathbf{l}_{\text{int}}$ 通过一个门控机制403,输入到槽位填充的任务中,通过如下公式融合意图向量 $\mathbf{l}_{\text{int}}$ 和文本向量:

$$[0112] \quad \mathbf{g}_i = \sigma(\mathbf{W}_3 \mathbf{h}_i + \mathbf{W}_4 \mathbf{l}_{\text{int}})$$

[0113] 然后,将向量 $\mathbf{h}_i$ 与意图向量 $\mathbf{l}_{\text{int}}$ 共同输入到分类层,得到预测到的槽值标注序列:

$$[0114] \quad \hat{\mathbf{y}}_i^p = \text{softmax}(\mathbf{W}_5(\mathbf{h}_i + \mathbf{g}_i \odot \mathbf{l}_{\text{int}}))$$

[0115] 最后,通过如下公式,确定槽值归属于各槽位的概率分布,得到槽位填充结果:

$$[0116] \quad \hat{\mathbf{y}}^s = \text{softmax}(\mathbf{W}_6 \mathbf{h}_{\text{ent}} + \mathbf{b})$$

[0117] 继续参考图5,示出了根据本申请的文本处理模型的一个实施例的示意性流程500,包括以下步骤:

[0118] 步骤501,获取训练样本集。

[0119] 其中,训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型

[0120] 步骤502,针对于训练样本集中的训练样本,执行如下训练操作,直至确定文本处理模型训练完成:

[0121] 步骤5021,确定该训练样本中的文本样本对应的文本向量。

[0122] 步骤5022,匹配该文本样本对应的文本向量与预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量。

[0123] 其中,预设意图集合中各意图对应的意图向量可以是预先确定的意图向量。

[0124] 步骤5023,融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型。

[0125] 步骤5024,根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定文本处理模型是否训练完成。

[0126] 从本实施例中可以看出,与图2对应的实施例相比,本实施例中的文本处理模型的训练流程500具体说明了文本处理模型的训练过程,通过文本处理模型执行上述文本数据的处理方法,进一步提高了意图识别和槽位填充的处理效率。

[0127] 继续参考图6,作为对上述各图所示方法的实现,本公开提供了一种文本数据的处理装置的一个实施例,该装置实施例与图2所示的方法实施例相对应,该装置具体可以应用于各种电子设备中。

[0128] 如图6所示,文本数据的处理装置包括:确定单元601,被配置成确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配单元602,被配置成匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合单元603,被配置成融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0129] 在本实施例的一些可选的实现方式中,融合单元603,进一步被配置成:基于门控

机制,融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0130] 在本实施例的一些可选的实现方式中,融合单元603,进一步被配置成:基于门控机制,融合所表达的意图对应的意图向量和文本向量,确定待处理文本中的槽值;针对于待处理文本中的每个槽值,确定该槽值所属的槽位类型。

[0131] 在本实施例的一些可选的实现方式中,文本数据的处理方法通过文本处理模型执行,其中,文本处理模型表征待处理文本与待处理文本所表达的意图、对应的槽位填充结果之间的对应关系;以及上述装置还包括,训练单元(图中未示出),被配置成通过如下方法训练得到文本处理模型:获取训练样本集,其中,训练样本集中的训练样本包括文本样本,以及对应于文本样本的意图、槽值标注序列和槽位类型,槽值标注序列表征文本样本中的实体,槽位类型表征槽值标注序列中的槽值所属的槽位的类型;基于机器学习算法,以训练样本中的文本样本作为输入,以所输入的训练样本中的意图、槽值标注序列和槽位类型作为期望输出,训练得到文本处理模型。

[0132] 在本实施例的一些可选的实现方式中,训练单元(图中未示出),进一步被配置成:针对于训练样本集中的训练样本,执行如下训练操作,直至确定文本处理模型训练完成:确定该训练样本中的文本样本对应的文本向量;匹配该文本样本对应的文本向量与预设意图集合中各意图对应的意图向量,预测该文本样本所表达的意图以及所表达的意图对应的意图向量;融合该文本样本所表达的意图对应的意图向量和该文本样本对应的文本向量,对该文本样本进行命名实体识别,预测该文本样本的槽值标注序列和槽位类型;根据该训练样本中的意图、槽值标注序列、槽位类型与所预测的意图、槽值标注序列、槽位类型之间的损失值,确定文本处理模型是否训练完成。

[0133] 本实施例中,文本数据的处理装置中的确定单元确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配单元匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合单元融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果,从而提供了一种进行意图识别和槽位填充的文本数据的处理装置,并在文本向量的基础上融合意图向量确定槽位填充结果,提高了槽位填充的准确度。

[0134] 下面参考图7,其示出了适于用来实现本申请实施例的设备(例如图1所示的设备101、102、103、105)的计算机系统700的结构示意图。图7示出的设备仅仅是一个示例,不应对本申请实施例的功能和使用范围带来任何限制。

[0135] 如图7所示,计算机系统700包括处理器(例如CPU,中央处理器)701,其可以根据存储在只读存储器(ROM)702中的程序或者从存储部分708加载到随机访问存储器(RAM)703中的程序而执行各种适当的动作和处理。在RAM703中,还存储有系统700操作所需的各种程序和数据。处理器701、ROM702以及RAM703通过总线704彼此相连。输入/输出(I/O)接口705也连接至总线704。

[0136] 以下部件连接至I/O接口705:包括键盘、鼠标等的输入部分706;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707;包括硬盘等的存储部分708;以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分709。通信部分709经由诸如因

特网的网络执行通信处理。驱动器710也根据需要连接至I/O接口705。可拆卸介质711,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710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708。

[0137] 特别地,根据本公开的实施例,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709从网络上被下载和安装,和/或从可拆卸介质711被安装。在该计算机程序被处理器701执行时,执行本申请的方法中限定的上述功能。

[0138] 需要说明的是,本申请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。在本申请中,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。而在本申请中,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。

[0139] 可以以一种或多种程序设计语言或其组合来编写用于执行本申请的操作的计算机程序代码,程序设计语言包括面向目标的程序设计语言——诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言——诸如“C”语言或类似的设计语言。程序代码可以完全地在客户计算机上执行、部分地在客户计算机上执行、作为一个独立的软件包执行、部分在客户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到客户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0140] 附图中的流程图和框图,图示了按照本申请各种实施例的装置、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,该模块、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,框图和/或流程图中的每个方框、以及框图和/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。

[0141] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中,例如,可以描述为:一种处理器,包括确定单元、匹配单元和融合单元。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,融合单元还可以被描述为“融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果的单元”。

[0142] 作为另一方面,本申请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该装置执行时,使得该计算机设备:确定所获取的待处理文本对应的文本向量和预设意图集合中各意图对应的意图向量;匹配文本向量与预设意图集合中各意图对应的意图向量,确定待处理文本所表达的意图以及所表达的意图对应的意图向量;融合所表达的意图对应的意图向量和文本向量,对待处理文本进行命名实体识别,得到待处理文本的槽位填充结果。

[0143] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解,本申请中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

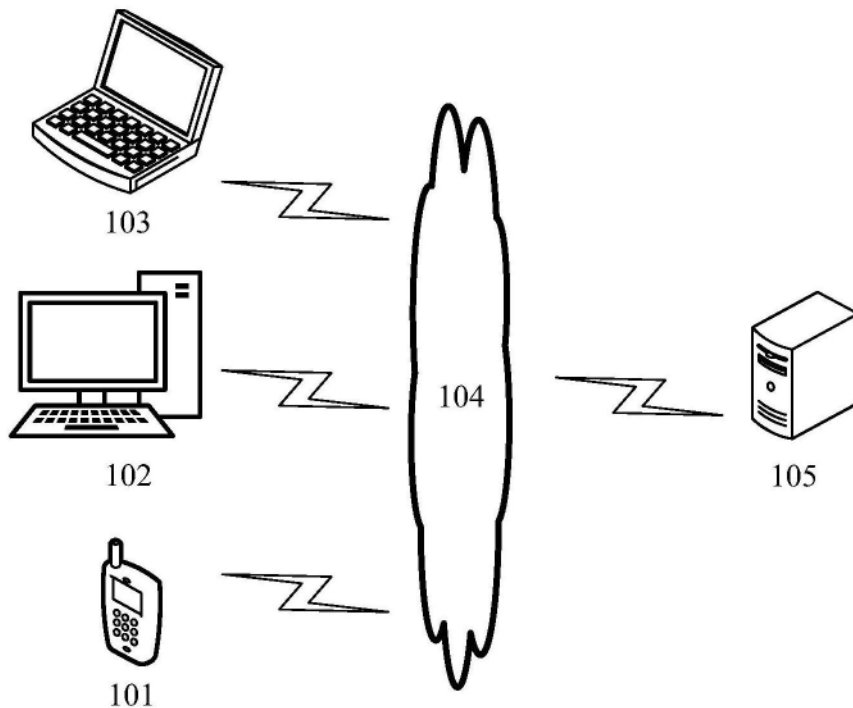
100

图1

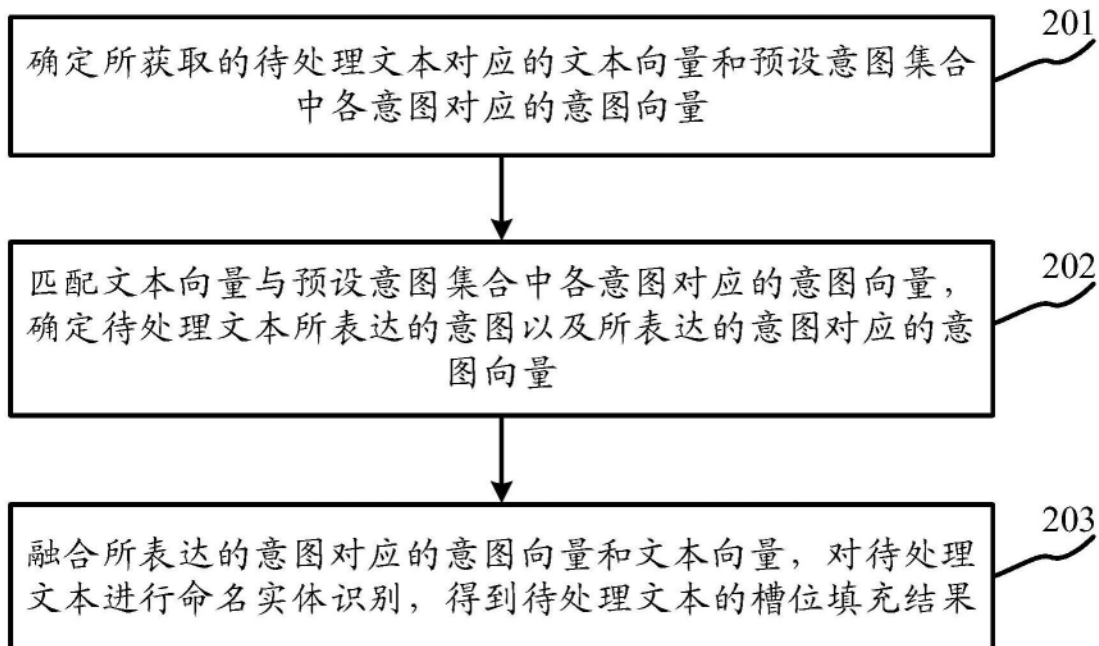
200

图2

300

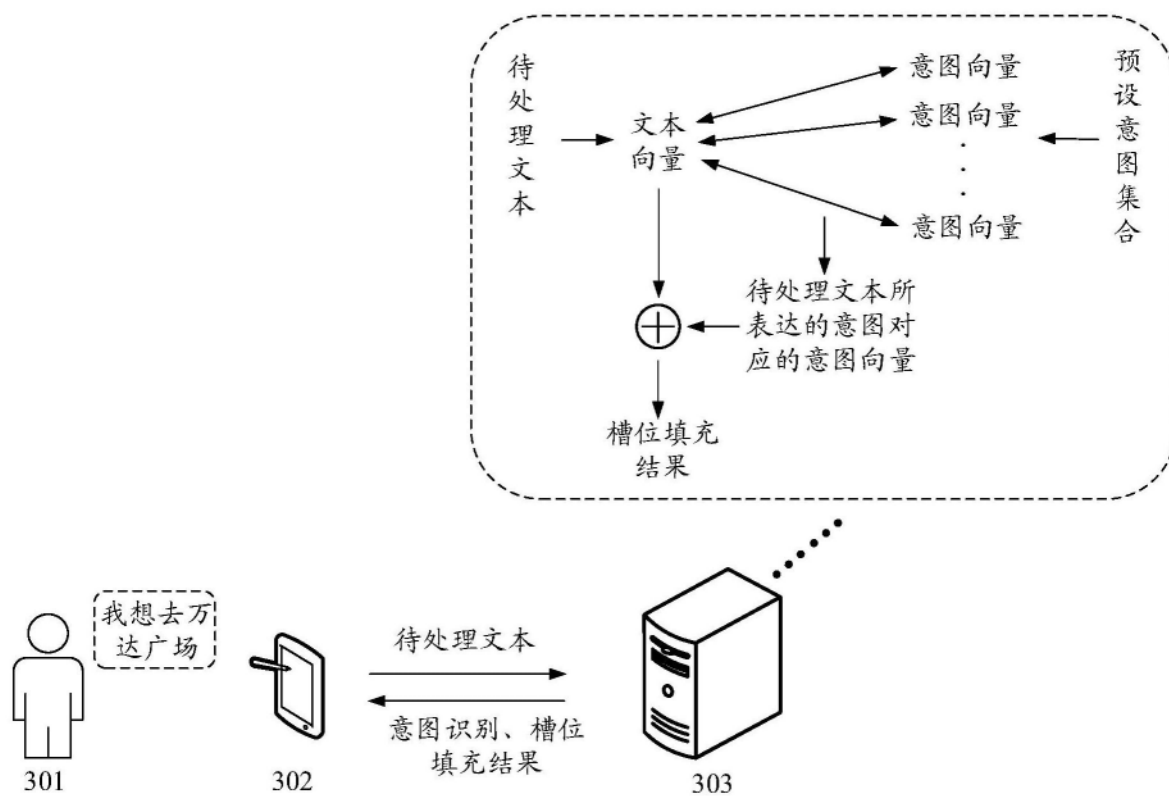


图3

400

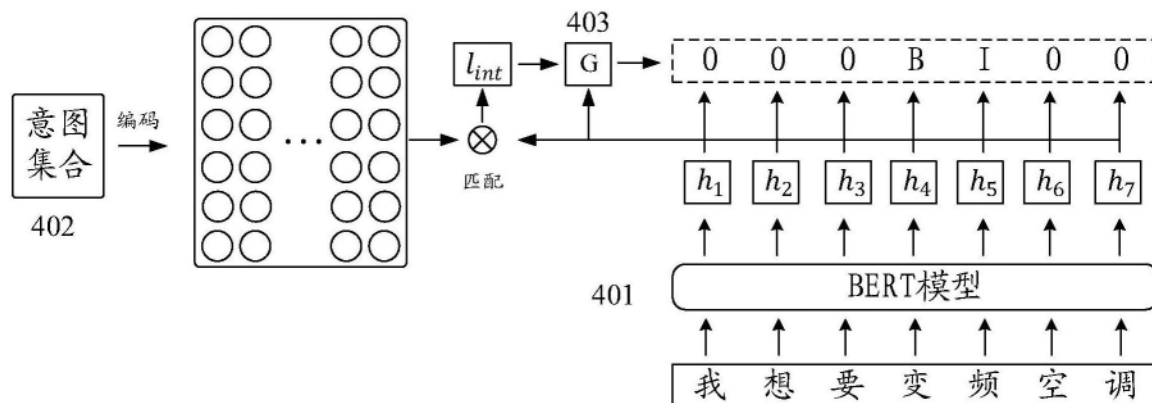


图4

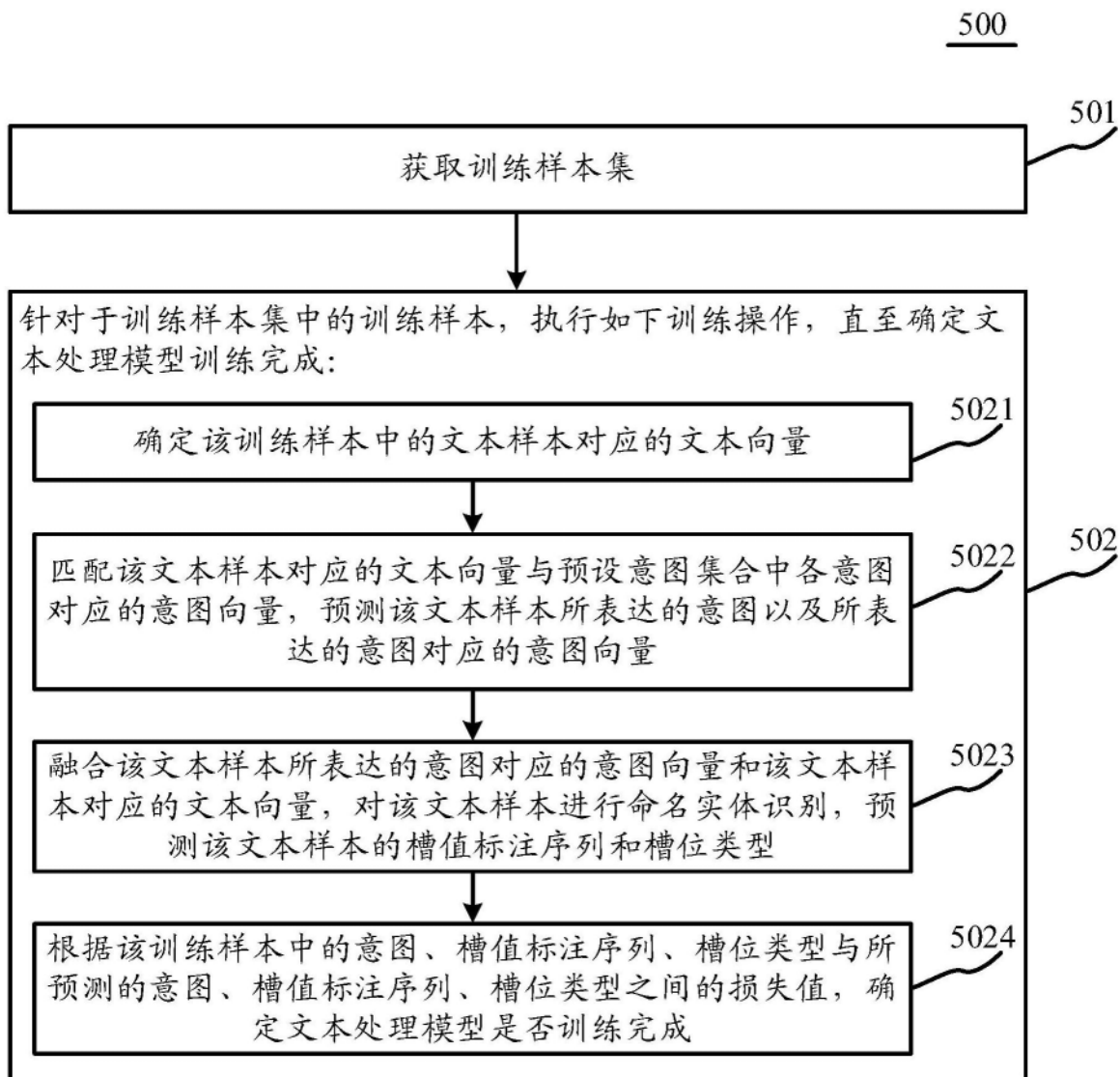


图5

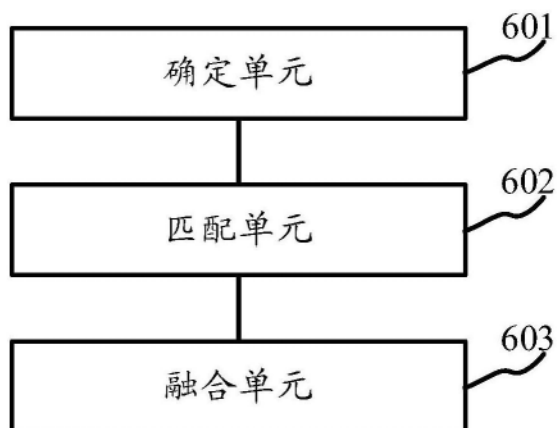
600

图6

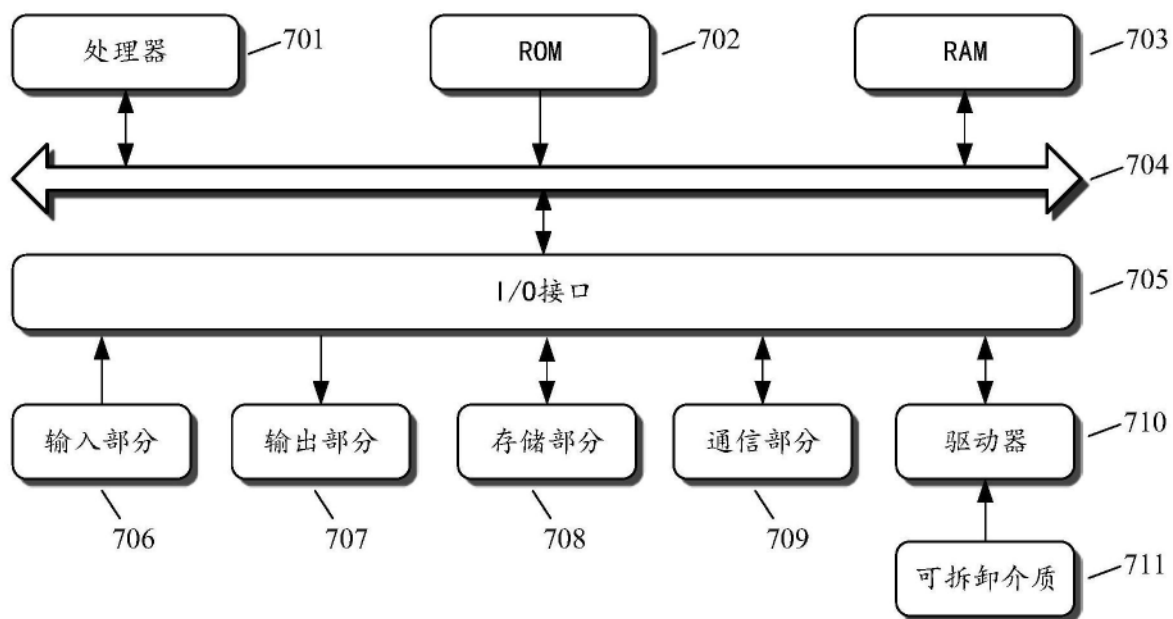
700

图7