



(12) 发明专利申请

(10) 申请公布号 CN 116431808 A

(43) 申请公布日 2023. 07. 14

(21) 申请号 202310377762.2

G06N 3/08 (2023.01)

(22) 申请日 2023.04.10

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 覃杰 王泽勋 冯明超 陈蒙

(74) 专利代理机构 北京路浩知识产权代理有限
公司 11002

专利代理师 刘亚平

(51) Int. Cl.

G06F 16/35 (2019.01)

G06F 18/214 (2023.01)

G06F 18/241 (2023.01)

G06N 3/0455 (2023.01)

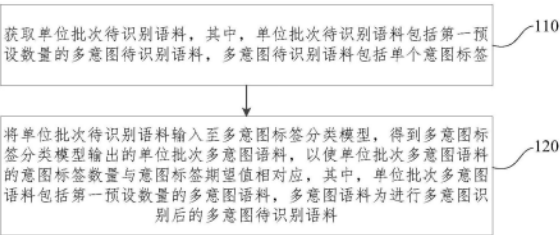
权利要求书2页 说明书11页 附图4页

(54) 发明名称

多意图识别方法、装置、电子设备及存储介
质

(57) 摘要

本公开提供一种多意图识别方法、装置、电子设备及存储介质,所述方法包括:获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签;将单位批次待识别语料输入至多意图标签分类模型得到单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料,多意图标签分类模型基于损失函数通过预训练得到,损失函数包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定。通过本公开实现了可以自动、准确得对多意图标签进行全面标注。



1. 一种多意图识别方法,其特征在于,所述多意图识别方法包括:

获取单位批次待识别语料,其中,所述单位批次待识别语料包括第一预设数量的多意图待识别语料,所述多意图待识别语料包括单个意图标签;

将所述单位批次待识别语料输入至多意图标签分类模型,得到所述多意图标签分类模型输出的单位批次多意图语料,以使所述单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,所述单位批次多意图语料包括第一预设数量的多意图语料,所述多意图语料为进行多意图识别后的多意图待识别语料,

其中,所述多意图标签分类模型基于损失函数通过预训练得到,其中,所述损失函数至少包括平均期望损失函数,所述平均期望损失函数根据所述意图标签期望值确定。

2. 根据权利要求1所述的多意图识别方法,其特征在于,所述多意图标签分类模型基于损失函数通过预训练得到,具体包括:

获取第二预设数量的种子语料,其中,所述种子语料包括目标意图标签,所述种子语料的对话应用场景与所述单位批次待识别语料的对话应用场景相同;

基于所述损失函数对所述多意图标签分类模型进行多轮次迭代训练,直至所述损失函数中的所述平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中所述多意图标签分类模型的输出作为下一轮次训练过程中所述多意图标签分类模型的输入,所述第二预设数量的种子语料作为首轮次训练过程中所述多意图标签分类模型的输入。

3. 根据权利要求2所述的多意图识别方法,其特征在于,所述种子语料还包括附加意图标签,其中,所述附加意图标签为除所述目标意图标签之外的其他意图标签;

所述平均期望损失函数采用以下方式确定:

基于所述目标意图标签的数量和所述附加意图标签的数量,确定各所述种子语料的意图标签数量;

基于各所述种子语料的意图标签数量的平均值,确定所述意图标签期望值;

基于所述意图标签期望值和预测意图标签数量,确定所述平均期望损失函数,其中,所述预测意图标签数量为每轮次训练过程中所述多意图标签分类模型的输出中包括的意图标签数量。

4. 根据权利要求3所述的多意图识别方法,其特征在于,所述基于所述意图标签期望值和预测意图标签数量,确定所述平均期望损失函数,具体包括:

基于所述第二预设数量与所述意图标签期望值之积,得到意图期望参数;

基于所述意图期望参数与所述预测意图标签数量的差值,确定所述平均期望损失函数。

5. 根据权利要求1所述的多意图识别方法,其特征在于,包括单个意图标签的多意图待识别语料采用以下方式获取:

将未标注单个意图标签的初始多意图待识别语料输入至单意图分类模型,得到所述单意图分类模型输出的标注单个意图标签的多意图待识别语料,其中,所述单意图分类模型基于种子语料预训练得到,所述种子语料的对话应用场景与所述多意图待识别语料的对话应用场景相同。

6. 一种多意图识别装置,其特征在于,所述多意图识别装置包括:

第一模块,用于获取单位批次待识别语料,其中,所述单位批次待识别语料包括第一预

设数量的多意图待识别语料,所述多意图待识别语料包括单个意图标签;

第二模块,用于将所述单位批次待识别语料输入至多意图标签分类模型,得到所述多意图标签分类模型输出的单位批次多意图语料,以使所述单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,所述单位批次多意图语料包括第一预设数量的多意图语料,所述多意图语料为进行多意图识别后的多意图待识别语料,

其中,所述多意图标签分类模型基于损失函数通过预训练得到,其中,所述损失函数至少包括平均期望损失函数,所述平均期望损失函数根据所述意图标签期望值确定。

7.根据权利要求6所述的多意图识别装置,其特征在于,所述第二模块采用以下方式基于损失函数通过预训练得到多意图标签分类模型:

获取第二预设数量的种子语料,其中,所述种子语料包括目标意图标签,所述种子语料的对话应用场景与所述单位批次待识别语料的对话应用场景相同;

基于所述损失函数对所述多意图标签分类模型进行多轮次迭代训练,直至所述损失函数中的所述平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中所述多意图标签分类模型的输出作为下一轮次训练过程中所述多意图标签分类模型的输入,所述第二预设数量的种子语料作为首轮次训练过程中所述多意图标签分类模型的输入。

8.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述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如权利要求1至5任一项所述的多意图识别方法。

9.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至5任一项所述的多意图识别方法。

10.一种计算机程序产品,包括计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至5任一项所述的多意图识别方法。

多意图识别方法、装置、电子设备及存储介质

技术领域

[0001] 本公开涉及语言处理技术领域,尤其涉及一种多意图识别方法、装置、电子设备及存储介质。

背景技术

[0002] 意图识别是指通过分类技术将意图待识别语料分配到相应意图类别中的一种处理方式。

[0003] 传统的意图识别主要是针对单意图识别,即对一句话中出现的主要意图进行识别,其他次要意图忽略处理,因此只完成单个意图识别判断,其智能化程度低。在对话智能水平不断提高的今天,不再只是对单意图进行识别,还需要对意图待识别语料中表达的多个意图均进行识别。

[0004] 当前,在对意图待识别语料进行多意图识别时,往往需要人工参与意图标签的标注,导致在进行多意图标签标注时工作量繁杂,且容易出现标注不全面的问题。

发明内容

[0005] 本公开提供一种多意图识别方法、装置、电子设备及存储介质,用以解决现有技术中在进行多意图标签标注时工作量繁杂,且容易出现标注不全面的问题,实现了可以自动、准确得对多意图标签进行全面标注,减少了人力成本,提高了标注准确性。

[0006] 本公开提供一种多意图识别方法,所述多意图识别方法包括:获取单位批次待识别语料,其中,所述单位批次待识别语料包括第一预设数量的多意图待识别语料,所述多意图待识别语料包括单个意图标签;将所述单位批次待识别语料输入至多意图标签分类模型,得到所述多意图标签分类模型输出的单位批次多意图语料,以使所述单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,所述单位批次多意图语料包括第一预设数量的多意图语料,所述多意图语料为进行多意图识别后的多意图待识别语料,其中,所述多意图标签分类模型基于损失函数通过预训练得到,其中,所述损失函数至少包括平均期望损失函数,所述平均期望损失函数根据所述意图标签期望值确定。

[0007] 根据本公开提供的一种多意图识别方法,所述多意图标签分类模型基于损失函数通过预训练得到,具体包括:获取第二预设数量的种子语料,其中,所述种子语料包括目标意图标签,所述种子语料的对话应用场景与所述单位批次待识别语料的对话应用场景相同;基于所述损失函数对所述多意图标签分类模型进行多轮次迭代训练,直至所述损失函数中的所述平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中所述多意图标签分类模型的输出作为下一轮次训练过程中所述多意图标签分类模型的输入,所述第二预设数量的种子语料作为首轮次训练过程中所述多意图标签分类模型的输入。

[0008] 根据本公开提供的一种多意图识别方法,所述种子语料还包括附加意图标签,其中,所述附加意图标签为除所述目标意图标签之外的其他意图标签;所述平均期望损失函数采用以下方式确定:基于所述目标意图标签的数量和所述附加意图标签的数量,确定各

所述种子语料的意图标签数量;基于各所述种子语料的意图标签数量的平均值,确定所述意图标签期望值;基于所述意图标签期望值和预测意图标签数量,确定所述平均期望损失函数,其中,所述预测意图标签数量为每轮次训练过程中所述多意图标签分类模型的输出中包括的意图标签数量。

[0009] 根据本公开提供的一种多意图识别方法,所述基于所述意图标签期望值和预测意图标签数量,确定所述平均期望损失函数,具体包括:基于所述第二预设数量与所述意图标签期望值之积,得到意图期望参数;基于所述意图期望参数与所述预测意图标签数量的差值,确定所述平均期望损失函数。

[0010] 根据本公开提供的一种多意图识别方法,包括单个意图标签的多意图待识别语料采用以下方式获取:将未标注单个意图标签的初始多意图待识别语料输入至单意图分类模型,得到所述单意图分类模型输出的标注单个意图标签的多意图待识别语料,其中,所述单意图分类模型基于种子语料预训练得到,所述种子语料的对话应用场景与所述多意图待识别语料的对话应用场景相同。

[0011] 本公开还提供一种多意图识别装置,所述多意图识别装置包括:第一模块,用于获取单位批次待识别语料,其中,所述单位批次待识别语料包括第一预设数量的多意图待识别语料,所述多意图待识别语料包括单个意图标签;第二模块,用于将所述单位批次待识别语料输入至多意图标签分类模型,得到所述多意图标签分类模型输出的单位批次多意图语料,以使所述单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,所述单位批次多意图语料包括第一预设数量的多意图语料,所述多意图语料为进行多意图识别后的多意图待识别语料,其中,所述多意图标签分类模型基于损失函数通过预训练得到,其中,所述损失函数至少包括平均期望损失函数,所述平均期望损失函数根据所述意图标签期望值确定。

[0012] 根据本公开提供的一种多意图识别装置,所述第二模块采用以下方式基于损失函数通过预训练得到多意图标签分类模型:获取第二预设数量的种子语料,其中,所述种子语料包括目标意图标签,所述种子语料的对话应用场景与所述单位批次待识别语料的对话应用场景相同;基于所述损失函数对所述多意图标签分类模型进行多轮次迭代训练,直至所述损失函数中的所述平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中所述多意图标签分类模型的输出作为下一轮次训练过程中所述多意图标签分类模型的输入,所述第二预设数量的种子语料作为首轮次训练过程中所述多意图标签分类模型的输入。

[0013] 本公开还提供一种电子设备,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如上述任一种所述的多意图识别方法。

[0014] 本公开还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如上述任一种所述的多意图识别方法。

[0015] 本公开还提供一种计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现如上述任一种所述的多意图识别方法。

[0016] 本公开提供的多意图识别方法、装置、电子设备及存储介质,通过基于损失函数预训练多意图标签分类模型,其中,损失函数包括平均期望损失函数,平均期望损失函数根据

意图标签期望值确定,可以确保将单位批次待识别语料输入至多意图标签分类模型,能够得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,即确保单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。实现了可以自动、准确得对多意图标签进行全面标注,进而减少了人力成本,提高了标注准确性。

附图说明

[0017] 为了更清楚地说明本公开或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图是本公开的一些实施例,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。

[0018] 图1是本公开提供的多意图识别方法的流程示意图之一;

[0019] 图2是本公开提供的基于损失函数通过预训练得到多意图标签分类模型的流程示意图;

[0020] 图3是本公开提供的多意图标签分类模型的应用场景示意图;

[0021] 图4是本公开提供的确定平均期望损失函数的流程示意图;

[0022] 图5是本公开提供的单意图分类模型的应用场景示意图;

[0023] 图6是本公开提供的多意图识别方法的流程示意图之二;

[0024] 图7是本公开提供的多意图识别装置的结构示意图;

[0025] 图8是本公开提供的电子设备的结构示意图。

具体实施方式

[0026] 为使本公开的目的、技术方案和优点更加清楚,下面将结合本公开中的附图,对本公开中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公开一部分实施例,而不是全部的实施例。基于本公开中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本公开保护的范围。

[0027] 本公开提供多意图识别方法,通过在损失函数的基础上增加平均期望损失函数,可以鼓励多意图标签分类模型输出的预测意图的期望数值逼近项目中意图数量(又称真实意图数量)的期望数值。根据构建好的多标签分类基础数据(对应单位批次待识别语料)就可以基于多意图标签分类模型的迭代训练过程和预测过程,使输入的单位批次待识别语料中的意图数量的期望值不断逼近项目中意图数量的期望值,从而可以自动、准确的获取带有多意图标签的多意图语料。

[0028] 图1是本公开提供的多意图识别方法的流程示意图之一。

[0029] 为了进一步介绍本公开提供的多意图识别方法,下面将结合图1进行说明。

[0030] 在本公开一示例性实施例中,结合图1可知,多意图识别方法可以包括步骤110和步骤120,下面将分别介绍各步骤。

[0031] 在步骤110中,获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签。

[0032] 在一种实施例中,可以获取单位批次待识别语料。可以理解的是,单位批次待识别

语料可以一个batch的待识别语料,在一示例中,单位批次待识别语料中可以包括第一预设数量的多意图待识别语料。其中,多意图待识别语料可以是缺失意图标签的待识别语料,即多意图待识别语料可以包括单个意图标签。在应用过程中,为了获取完整的多意图语料,需要补全多意图待识别语料中缺失的意图标签。

[0033] 在一种实施例中,第一预设数量可以是一个多意图待识别语料。可以理解的是,此场景下的单位批次待识别语料为一个多意图待识别语料。

[0034] 需要说明的是,单个多意图待识别语料往往具有独立性,若以单个多意图待识别语料作为处理对象,无法确保每一个多意图待识别语料均可以准确的补充完整缺失的意图标签。在应用过程中,若以单位批次待识别语料(包括第一预设数量的多意图待识别语料)作为处理对象,可以确保补充完整单位批次待识别语料所缺失的意图标签。

[0035] 在本公开又一示例中,第一预设数量可以是多个,在此场景下,将以单位批次待识别语料作为处理对象,通过将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应。

[0036] 其中,第一预设数量可以根据实际情况进行调整,在本实施例中不对第一预设数量作具体限定。

[0037] 在本公开又一示例性实施例中,包括单个意图标签的多意图待识别语料采用以下方式获取:将未标注单个意图标签的初始多意图待识别语料输入至单意图分类模型,得到单意图分类模型输出的标注单个意图标签的多意图待识别语料,其中,单意图分类模型基于种子语料预训练得到,种子语料的对话应用场景与多意图待识别语料的对话应用场景相同。在本实施例中,基于单意图分类模型获取包括单个意图标签的多意图待识别语料,可以实现多意图识别的全程自动化处理,进而有效降低人力成本,提高识别效率。

[0038] 需要说明的是,初始多意图待识别语料可以理解为是未标注单个意图标签的多意图待识别语料。

[0039] 在又一实施例中,多意图待识别语料的单个意图标签可以为多意图待识别语料的意图标签中的主要意图标签。

[0040] 需要说明的是,在通过分类的方法获取多意图待识别语料的各个意图标签的概率值时,其中,最大概率值所对应的意图标签可以认为是主要意图标签。

[0041] 在又一实施例中,包括单个意图标签的多意图待识别语料还可以直接通过人工标注获得。

[0042] 在步骤120中,将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料。

[0043] 在一种实施例中,多意图标签分类模型基于损失函数通过预训练得到,其中,损失函数可以至少包括平均期望损失函数(Average Expectation Loss,又称AEL),平均期望损失函数可以根据意图标签期望值确定。

[0044] 在又一种实施例中,为了确保将单位批次待识别语料输入至多意图标签分类模型,可以得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料

的意图标签数量与意图标签期望值相对应,需要基于平均期望损失函数对多意图标签分类模型进行迭代训练。其中,单位批次多意图语料的意图标签数量与意图标签期望值相对应可以理解为,单位批次多意图语料的意图标签数量与意图标签期望值和第一预设数量的乘积相同。

[0045] 在应用过程中,意图标签期望值是根据与单位批次待识别语料的对话应用场景相同的种子语料确定的,因此,基于种子语料得到的意图标签期望值可以认为是单位批次待识别语料需要达到的期望值。

[0046] 通过前述实施例,可以确保单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。从而可以实现自动、准确得对多意图标签进行全面标注,进而减少了人力成本,提高了标注准确性。

[0047] 本公开提供的多意图识别方法,通过基于损失函数预训练多意图标签分类模型,其中,损失函数包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定,可以确保将单位批次待识别语料输入至多意图标签分类模型,能够得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,确保单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。实现了可以自动、准确得对多意图标签进行全面标注,进而减少了人力成本,提高了标注准确性。

[0048] 为了进一步介绍本公开提供的多意图识别方法,下面将结合图2进行说明。

[0049] 在本公开一示例性实施例中,结合图2可知,基于损失函数通过预训练得到多意图标签分类模型可以包括步骤210和步骤220,下面将分别介绍各步骤。

[0050] 在步骤210中,获取第二预设数量的种子语料,其中,种子语料包括目标意图标签,种子语料的对话应用场景与单位批次待识别语料的对话应用场景相同。

[0051] 在一种实施例中,种子语料可以是语料数据库中获取并借助人工标注而得到的。种子语料也是一种具有多意图标签的语料,其中,种子语料可以包括目标意图标签,目标意图标签可以是种子语料的多个意图标签中的主要意图标签,目标意图标签也可以是除主要意图标签之外的其他意图标签(对应本公开的附加意图标签)。

[0052] 在又一实施例中,种子语料还可以通过单意图分类模型处理得到。在应用过程中,将为标注有目标意图标签的种子语料输入至单意图分类模型,可以直接输出标注有目标意图标签的种子语料。

[0053] 需要说明的是,种子语料的对话应用场景与单位批次待识别语料的对话应用场景是相同的。由于种子语料的对话应用场景与单位批次待识别语料的对话应用场景相同。因此,基于种子语料预训练得到的多意图标签分类模型能够确保将单位批次待识别语料输入至多意图标签分类模型,可以得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,从而可以使得单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。

[0054] 在又一种实施例中,可以构建符合项目(对应单位批次待识别语料)中真实意图分布情况的少量的种子语料。其中,种子语料可以通过单意图对话系统中已标注的单意图分类语料加工后获得。在应用过程中,可以通过多分类模型预测每条query对应的各个意图的得分,将预测得分低于一定阈值的训练数据提取出来进行二次标注完成种子语料的构建。

[0055] 进一步的,可以计算第二预设数量的种子语料中的意图分布情况,以获取第二预

设数量的种子语料的平均意图期望值,这个期望值可以代表这个场景下意图数量的期望大小(对应意图标签期望值)。其中,第二预设数量可以根据实际情况进行调整,在本实施例中,不对第二预设数量作具体限定。

[0056] 在又一种实施例中,目标意图标签可以为种子语料的意图标签中的主要意图标签。种子语料可以采用以下方式实现:获取训练语料,其中,训练语料包括多个意图标签;将训练语料输入至单意图分类模型,得到单意图分类模型输出的种子语料,其中,种子语料包括主要意图标签,单意图分类模型通过预训练得到。可以理解的是,训练语料是未标注主要意图标签的种子语料。

[0057] 在步骤220中,基于损失函数对多意图标签分类模型进行多轮次迭代训练,直至损失函数中的平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中多意图标签分类模型的输出作为下一轮次训练过程中多意图标签分类模型的输入,第二预设数量的种子语料为首轮次训练过程中多意图标签分类模型的输入。

[0058] 在一种实施例中,可以根据损失函数对多意图标签分类模型进行多轮次迭代训练直至损失函数中的平均期望损失函数小于或等于损失值阈值。在一示例中,损失函数可以是二分类交叉熵损失函数。通过在损失函数中增加基于意图标签期望值构建的平均期望损失函数,可以迫使多意图标签分类模型在训练过程中不断调整一个batch预测结果中意图标签的数量,反复迭代训练和预测,使预测的标签数量接近真实的项目意图分布数量(对应意图标签期望值),并得到训练好的多意图标签分类模型。

[0059] 需要说明的是,平均期望损失函数主要是针对多意图标签分类模型在一个batch中,对未预测到的标签且应该命中的标签进行惩罚,迫使多意图标签分类模型对当前未知且应该命中的标签进行预测。

[0060] 在又一实施例中,多意图标签分类模型的初始输入为意图标签信息为单意图标签的多意图待识别语料。换句话说,首轮次训练过程中多意图标签分类模型的输入为第二预设数量的种子语料。在应用过程中,不断增加预测的多意图待识别语料的意图标签数量,最终使预测到的所有多意图待识别语料的期望值符合项目中真实的意图数量的期望值。通过此种方式,可以得到训练好的多意图标签分类模型,以使训练好的多意图标签分类模型可以根据输入至该模型地单位批次待识别语料,得到单位批次多意图语料,其中,单位批次多意图语料的意图标签数量与意图标签期望值相对应。

[0061] 图3是本公开提供的多意图标签分类模型的应用场景示意图。

[0062] 结合图3可知,在对多意图标签分类模型进行训练的过程中,模型的输入为种子语料中各个字 $W(1) \cdots W(n)$,以及意图标签占位符 $L(1) \cdots L(n)$,即多意图标签分类模型输入部分主要是由用户query和意图占位标签拼接的文本。在BERE模型特征化句子和标签信息后,得到种子语料的隐藏表征 $h(1) \cdots h(n)$, $l(1) \cdots l(n)$,再采用交叉注意力网络进行融合聚合文本特征和标签特征信息后,得到各个意图标签的概率值 $P(1) \cdots P(n)$,用以预测多意图。在应用过程中,当意图标签的概率值超过概率阈值时,说明该意图标签成立。此时,再判断得到的意图标签的数量是否满足意图标签期望值(可以通过平均期望损失函数是否小于或等于损失值阈值进行判断),若不满足,则反复迭代训练和预测,不断增加预测query的意图标签数量,即将当前轮次训练过程中多意图标签分类模型的输出作为下一轮次训练过程中多意图标签分类模型的输入,直至平均期望损失函数小于或等于损失值阈值。

[0063] 通过前述实施例,可以完成对多意图标签分类模型的训练,进而可以确保将单位批次待识别语料输入至多意图标签分类模型,能够得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应。换句话说,使得单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。实现了可以自动、准确得对多意图标签进行全面标注,进而减少了人力成本,提高了标注准确性。

[0064] 图4是本公开提供的确定平均期望损失函数的流程示意图。

[0065] 下面将结合图4对确定平均期望损失函数的过程进行说明。

[0066] 在本公开一示例性实施例中,结合图4可知,确定平均期望损失函数可以包括步骤410至步骤430,下面将分别介绍各步骤。

[0067] 在步骤410中,基于目标意图标签的数量和附加意图标签的数量,确定各种子语料的意图标签数量。

[0068] 在一种实施例中,种子语料还可以包括附加意图标签,其中,附加意图标签可以是除目标意图标签之外的其他意图标签。可以理解的是,目标意图标签的数量和附加意图标签的数量之和,可以定义为是各个种子语料的意图标签数量。

[0069] 在步骤420中,基于各种子语料的意图标签数量的平均值,确定意图标签期望值。

[0070] 由于种子语料的对话应用场景与单位批次待识别语料的对话应用场景相同,因此,基于种子语料预训练得到的多意图标签分类模型,能够确保将单位批次待识别语料输入至多意图标签分类模型,可以得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应。从而可以使得单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。

[0071] 在一示例中,可以基于各种子语料的意图标签数量的平均值,确定意图标签期望值。

[0072] 在步骤430中,基于意图标签期望值和预测意图标签数量,确定平均期望损失函数,其中,预测意图标签数量为每轮次训练过程中多意图标签分类模型的输出中包括的意图标签数量。

[0073] 在本公开一示例性实施例中,基于意图标签期望值和预测意图标签数量,确定平均期望损失函数可以采用以下方式实现:

[0074] 基于第二预设数量与意图标签期望值之积,得到意图期望参数;

[0075] 基于意图期望参数与预测意图标签数量的差值,确定平均期望损失函数。

[0076] 在一示例中,可以将意图期望参数与预测意图标签数量的差值的平方,确定平均期望损失函数。

[0077] 在一示例中,基于意图标签期望值和预测意图标签数量,确定平均期望损失函数,还可以采用以下公式(1)实现:

$$[0078] \quad AEL = \left(\frac{F-k}{L*B} \right)^2 \quad (1)$$

[0079] 其中,AEL表示平均期望损失函数;F表示预测意图标签数量;k表示第二预设数量与意图标签期望值之积,又称为意图期望参数;L表示意图标签的分类总数;B表示第二预设数量。可以理解的是, $\left(\frac{F-k}{L*B} \right)^2$ 可以对应意图期望参数与预测意图标签数量的差值的平方。

[0080] 根据公式(1)可知,当AEL越小时,说明预测意图标签数量越接近第二预设数量与意图标签期望值之积,即预测意图标签数量越符合真实意图标签数量的期望,此时,表示多意图标签分类模型接近训练完成。

[0081] 图6是本公开提供的多意图识别方法的流程示意图之二。

[0082] 为了进一步介绍本公开提供的多意图识别方法,下面将结合图6进行说明。

[0083] 在本公开一示例性实施例中,结合图6可知,多意图识别方法可以包括步骤610至步骤690,下面将分别介绍各步骤。

[0084] 在步骤610中,获取种子语料。

[0085] 在步骤620中,基于种子语料,计算意图期望值。

[0086] 在步骤630中,基于种子语料,得到单意图分类模型的训练数据。

[0087] 在步骤640中,基于训练数据训练单意图分类模型,得到训练后的单意图分类模型。

[0088] 在一种实施例中,可以准备模型训练所需数据(对应种子语料)。在一示例中,种子语料的来源主要可以是实际项目中真实对话记录随机抽样标注的语料。在应用过程中,可以提取项目中预设数量,例如随机10000条对话数据进行标注。标注数据可以将抽取的日志按文本长度从小到大排序,慢慢增加标注难度,保证数据标签的准确性。标注完成后可以计算标注数据中所有query(对应种子语料)的平均意图数量。并将种子语料的平均意图数量作为该项目意图的固定期望数值(对应意图期望值,又称意图标签期望值)。

[0089] 进一步的,可以将标注完成的种子语料中的多意图数据拆分成单意图样本,构建多分类训练数据(对应单意图分类模型的训练数据)。在一示例中,可以采用Bert模型训练单意图多分类模型,以得到训练后的单意图分类模型。该过程主要是保证单意图多分类模型能准确命中query中的任意一个单意图。

[0090] 在步骤650中,将项目日志输入至训练后的单意图分类模型。

[0091] 在步骤660中,基于项目日志,得到训练后的单意图分类模型输出的单意图预测语料。

[0092] 在步骤670中,将单意图预测语料输入至多意图标签分类模型。

[0093] 在一种实施例中,可以提取项目日志(又可以称为query),并将项目日志输入至训练后的单意图分类模型,以得到训练后的单意图分类模型预测得到的每条query的单意图(对应单意图预测语料),并将含有一个单意图标签的query作为多意图标签分类模型的输入语料输入至多意图标签分类模型。

[0094] 在步骤680中,判断意图预测值是否逼近意图期望值。

[0095] 在步骤690中,若意图预测值逼近意图期望值,则多意图标签分类模型输出意图预测值。

[0096] 由于多意图标签分类模型是基于平均期望损失函数训练得到。平均期望损失函数可以理解为针对多意图标签分类模型在一个batch(对应项目日志)中,对未预测的标签且应该命中的标签进行惩罚,迫使模型对当前未知且应该命中的标签进行预测。

[0097] 对于多意图标签分类模型,模型初始输入query对应的标签信息仅单意图标签的query,通过多意图多标签分类模型的反复迭代训练和预测过程,不断增加预测query的意图标签数量,最终使预测到的所有query的期望值符合项目中真实的意图数量的期望值(对

应意图期望值),进而使得多意图标签分类模型输出意图预测值,以获取更多的多意图语料。

[0098] 根据上述描述可知,本公开提供的多意图识别方法,通过基于损失函数预训练多意图标签分类模型,其中,损失函数包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定,可以确保将单位批次待识别语料输入至多意图标签分类模型,能够得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,确保单位批次多意图语料的意图标签数量符合真实意图标签数量的期望。实现了可以自动、准确得对多意图标签进行全面标注,进而减少了人力成本,提高了标注准确性。

[0099] 基于相同的构思,本公开还提供一种多意图识别装置。

[0100] 下面对本公开提供的多意图识别装置进行描述,下文描述的多意图识别装置与上文描述的多意图识别方法可相互对应参照。

[0101] 图7是本公开提供的多意图识别装置的结构示意图。

[0102] 在本公开一示例性实施例中,结合图7可知,多意图识别装置可以包括第一模块710和第二模块720,下面将分别介绍各模块。

[0103] 第一模块710,可以被配置为用于获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签;

[0104] 第二模块720,可以被配置为用于将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料,其中,多意图标签分类模型基于损失函数通过预训练得到,其中,损失函数至少包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定。

[0105] 在本公开一示例性实施例中,第二模块720可以采用以下方式基于损失函数通过预训练得到多意图标签分类模型:

[0106] 获取第二预设数量的种子语料,其中,种子语料包括目标意图标签,种子语料的对话应用场景与单位批次待识别语料的对话应用场景相同;

[0107] 基于损失函数对多意图标签分类模型进行多轮次迭代训练,直至损失函数中的平均期望损失函数小于或等于损失值阈值,其中,在当前轮次训练过程中多意图标签分类模型的输出作为下一轮次训练过程中多意图标签分类模型的输入,第二预设数量的种子语料为首轮次训练过程中多意图标签分类模型的输入。

[0108] 在本公开一示例性实施例中,种子语料还可以包括附加意图标签,其中,附加意图标签为除目标意图标签之外的其他意图标签;

[0109] 第二模块720可以采用以下方式确定平均期望损失函数:

[0110] 基于目标意图标签的数量和附加意图标签的数量,确定各种子语料的意图标签数量;

[0111] 基于各种子语料的意图标签数量的平均值,确定意图标签期望值;

[0112] 基于意图标签期望值和预测意图标签数量,确定平均期望损失函数,其中,预测意图标签数量为每轮次训练过程中多意图标签分类模型的输出中包括的意图标签数量。

[0113] 在本公开一示例性实施例中,第二模块720可以采用以下方式实现基于意图标签期望值和预测意图标签数量,确定平均期望损失函数:

[0114] 基于第二预设数量与意图标签期望值之积,得到意图期望参数;

[0115] 基于意图期望参数与预测意图标签数量的差值,确定平均期望损失函数。

[0116] 在本公开一示例性实施例中,第一模块710可以采用以下方式获取包括单个意图标签的多意图待识别语料:

[0117] 将未标注单个意图标签的初始多意图待识别语料输入至单意图分类模型,得到单意图分类模型输出的标注单个意图标签的多意图待识别语料,其中,单意图分类模型基于种子语料预训练得到,种子语料的对话应用场景与多意图待识别语料的对话应用场景相同。

[0118] 图8示例了一种电子设备的实体结构示意图,如图8所示,该电子设备可以包括:处理器(processor)810、通信接口(Communications Interface)820、存储器(memory)830和通信总线840,其中,处理器810,通信接口820,存储器830通过通信总线840完成相互间的通信。处理器810可以调用存储器830中的逻辑指令,以执行多意图识别方法,所述多意图识别方法包括:获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签;将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料,其中,多意图标签分类模型基于损失函数通过预训练得到,其中,损失函数至少包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定。

[0119] 此外,上述的存储器830中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用,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,本公开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)执行本公开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-Only Memory)、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。

[0120] 另一方面,本公开还提供一种计算机程序产品,所述计算机程序产品包括计算机程序,计算机程序可存储在非暂态计算机可读存储介质上,所述计算机程序被处理器执行时,计算机能够执行上述各方法所提供的多意图识别方法,所述多意图识别方法包括:获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签;将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料,其中,多意图标签分类模型基于损失函数通过预训练得到,其中,损失函数至少包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定。

[0121] 又一方面,本公开还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各方法提供的多意图识别方法,所述多意图识别方法包括:获取单位批次待识别语料,其中,单位批次待识别语料包括第一预设数量的多意图待识别语料,多意图待识别语料包括单个意图标签;将单位批次待识别语料输入至多意图标签分类模型,得到多意图标签分类模型输出的单位批次多意图语料,以使单位批次多意图语料的意图标签数量与意图标签期望值相对应,其中,单位批次多意图语料包括第一预设数量的多意图语料,多意图语料为进行多意图识别后的多意图待识别语料,其中,多意图标签分类模型基于损失函数通过预训练得到,其中,损失函数至少包括平均期望损失函数,平均期望损失函数根据意图标签期望值确定。

[0122] 以上所描述的装置实施例仅仅是示意性的,其中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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0123] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,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)执行各个实施例或者实施例的某些部分所述的方法。

[0124] 进一步可以理解的是,本公开实施例中尽管在附图中以特定的顺序描述操作,但是不应将其理解为要求按照所示的特定顺序或是串行顺序来执行这些操作,或是要求执行全部所示的操作以得到期望的结果。在特定环境中,多任务和并行处理可能是有利的。

[0125] 最后应说明的是:以上实施例仅用以说明本公开的技术方案,而非对其限制;尽管参照前述实施例对本公开进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本公开各实施例技术方案的精神和范围。

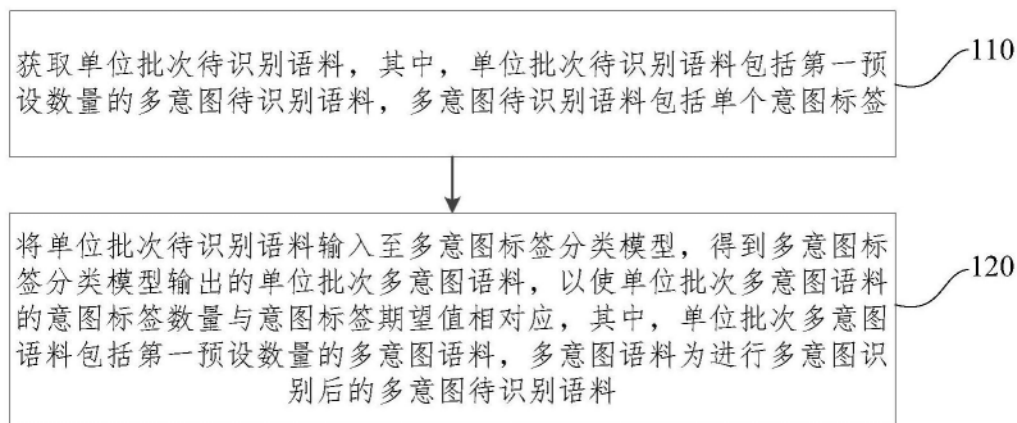


图1

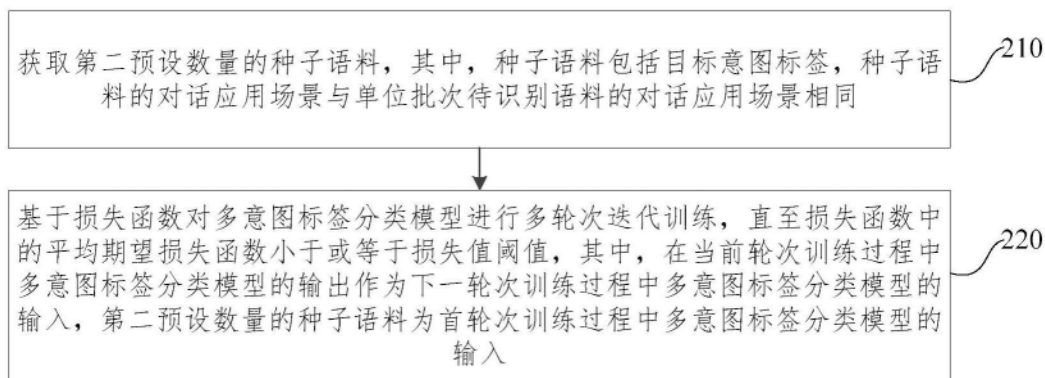


图2

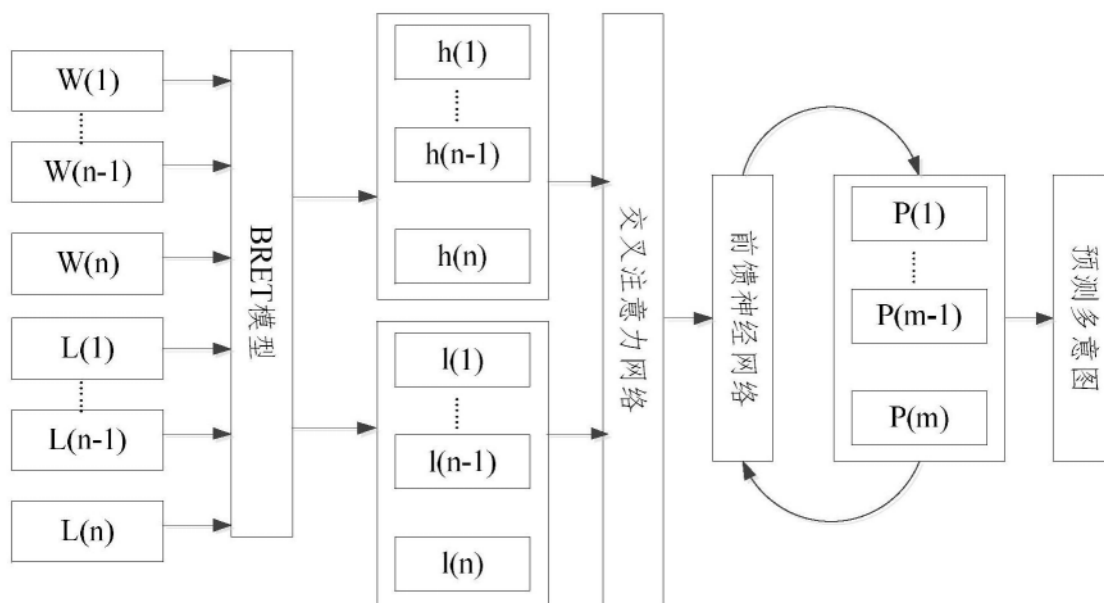


图3

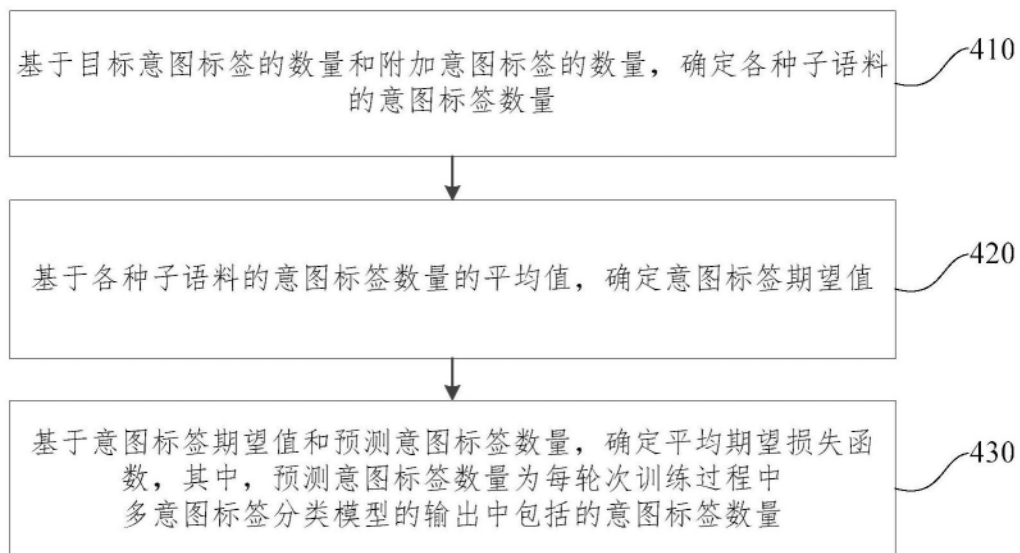


图4

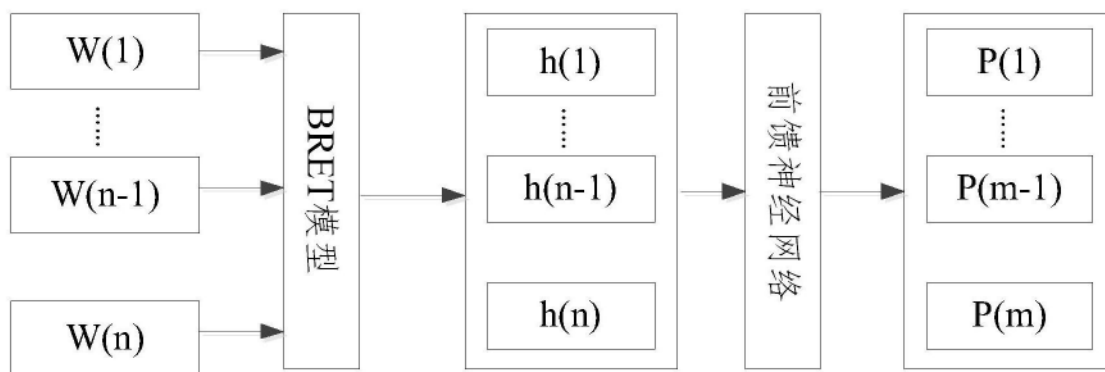


图5

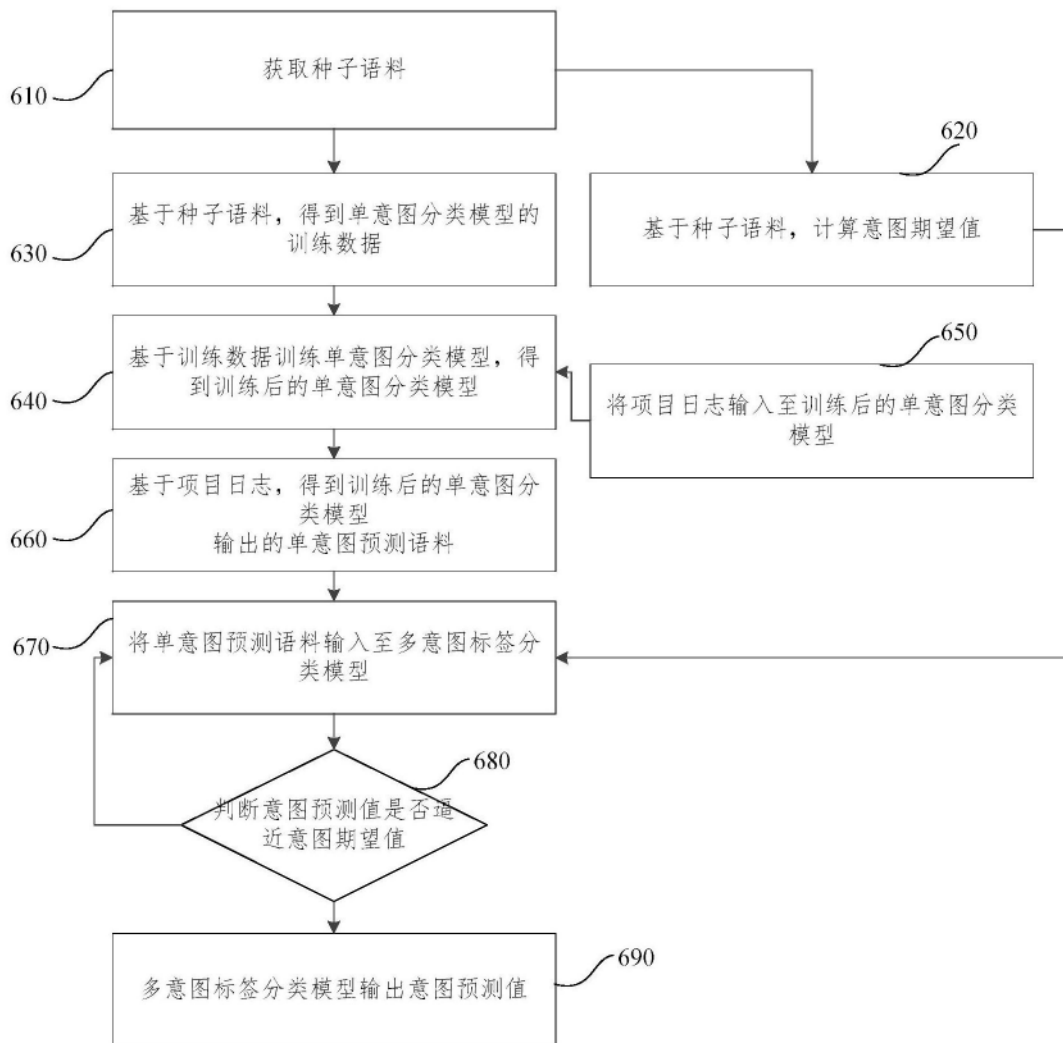


图6



图7

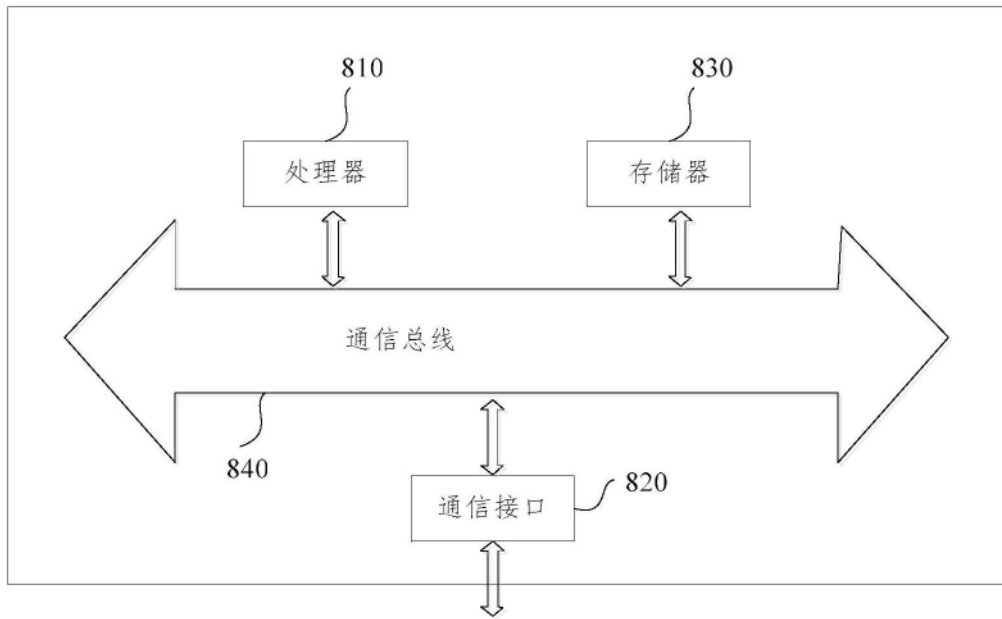


图8