



(12) 发明专利申请

(10) 申请公布号 CN 111767737 A

(43) 申请公布日 2020. 10. 13

(21) 申请号 201910465585.7

(22) 申请日 2019.05.30

(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 潘博 陈蒙

(74) 专利代理机构 北京品源专利代理有限公司
11332
代理人 孟金喆

(51) Int. Cl .
G06F 40/30 (2020.01)
G06N 3/04 (2006.01)

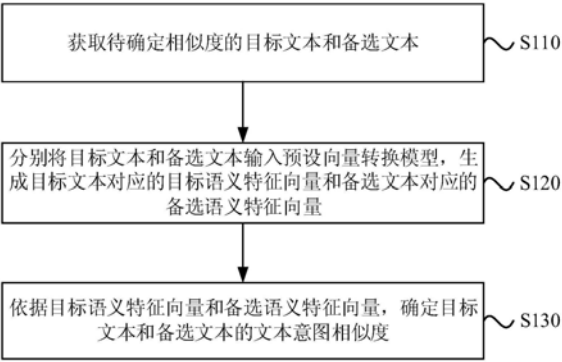
权利要求书2页 说明书10页 附图5页

(54) 发明名称

文本意图相似度确定方法、装置、电子设备和存储介质

(57) 摘要

本发明实施例公开了一种文本相似度确定方法、装置、电子设备和存储介质。该方法包括：获取待确定相似度的目标文本和备选文本；分别将所述目标文本和所述备选文本输入预设向量转换模型，生成所述目标文本对应的目标语义特征向量和所述备选文本对应的备选语义特征向量，其中，所述预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得，语义特征向量反映文本在各预设意图类别下的语义信息，所述预设意图类别基于业务场景而预先设定；依据所述目标语义特征向量和所述备选语义特征向量，确定所述目标文本和所述备选文本的文本意图相似度。通过上述技术方案，实现了更加准确地确定文本之间的相似度。



1. 一种文本相似度确定方法,其特征在于,包括:

获取待确定相似度的目标文本和备选文本;

分别将所述目标文本和所述备选文本输入预设向量转换模型,生成所述目标文本对应的目标语义特征向量和所述备选文本对应的备选语义特征向量,其中,所述预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,所述预设意图类别基于业务场景而预先设定;

依据所述目标语义特征向量和所述备选语义特征向量,确定所述目标文本和所述备选文本的文本意图相似度。

2. 根据权利要求1所述的方法,其特征在于,所述预设向量转换模型通过如下方式基于文本意图分类的卷积神经网络模型预先训练获得:

获得与各所述预设意图类别对应的训练样本,所述训练样本包含待训练文本和所述待训练文本对应的预设意图类别;

获得目标模型参数利用所述训练样本训练所述文本意图分类的卷积神经网络模型,获得目标模型参数;

依据所述目标模型参数、所述文本意图分类的卷积神经网络模型中的卷积层和池化层,生成所述预设向量转换模型。

3. 根据权利要求2所述的方法,其特征在于,所述文本意图分类的卷积神经网络模型为文本分类卷积神经网络模型,且在所述基于文本意图分类的卷积神经网络模型的训练过程中采用注意力机制。

4. 根据权利要求1所述的方法,其特征在于,所述业务场景为电商智能问答,所述目标文本为用户问题,所述备选文本为答案数据库中的备选答案,各所述预设意图类别包含陈述事实类别、物品属性类别、物品价格类别、物流类别、物品安装类别、促销信息类别、信息推荐类别和其他类别。

5. 根据权利要求4所述的方法,其特征在于,获取待确定相似度的备选文本包括:

依据所述用户问题和所述答案数据库,基于预设相似度确定算法,生成所述用户问题对应的初始参照答案,作为所述备选文本。

6. 根据权利要求4所述的方法,其特征在于,在依据所述目标语义特征向量和所述备选语义特征向量,确定所述目标文本和所述备选文本的文本意图相似度之后,还包括:

依据各所述文本意图相似度对各所述备选文本进行排序,生成备选文本相似度排序结果,作为所述用户问题的目标参照答案。

7. 根据权利要求6所述的方法,其特征在于,依据各所述文本意图相似度对各所述备选文本进行排序,生成备选文本相似度排序结果包括:

将各所述文本意图相似度及除所述文本意图相似度之外的至少一个其余特征维度的各相似度输入快速梯度提升排序模型,生成所述备选文本相似度排序结果。

8. 一种文本相似度确定装置,其特征在于,包括:

文本确定模块,用于获取待确定相似度的目标文本和备选文本;

特征向量生成模块,用于分别将所述目标文本和所述备选文本输入预设向量转换模型,生成所述目标文本对应的目标语义特征向量和所述备选文本对应的备选语义特征向量,其中,所述预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语

义特征向量反映文本在各预设意图类别下的语义信息,所述预设意图类别基于业务场景而预先设定;

相似度确定模块,用于依据所述目标语义特征向量和所述备选语义特征向量,确定所述目标文本和所述备选文本的文本意图相似度。

9.一种电子设备,其特征在于,所述电子设备包括:

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1-7中任一所述的文本相似度确定方法。

10.一种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1-7中任一所述的文本相似度确定方法。

文本意图相似度确定方法、装置、电子设备和存储介质

技术领域

[0001] 本发明实施例涉及计算机技术,尤其涉及一种文本意图相似度确定方法、装置、电子设备和存储介质。

背景技术

[0002] 随着科学技术的不断发展,文本意图相似度在诸如信息检索、智能化问答及机器翻译等领域的应用越来越多。例如,智能化问答系统便是以文本意图相似度为基础,从答案库中查找并返回与用户问题相似度较高的答案,作为反馈用户问题的候选答案。可见,在智能问答系统中,文本意图相似度的准确性很大程度上直接影响智能回答的有效性。

[0003] 现有的文本意图相似度确定方案可以大致归为2类:一类是基于词法特征的相似度确定方案,该方案是基于整个文本的全局性建模,其将用户问题和答案库中的备选答案输入神经网络模型,模型内部计算出用户问题和备选答案的向量表示,然后计算两个向量的相似度,作为用户问题与备选答案之间的相似度。该方案中神经网络模型的设计原理便是基于文本的词法特征,例如句型特征或词对文本重要性的词频特征等。另一类是基于语义特征的相似度确定方案,该方案是基于文本中的词的局部性建模,其首先计算用户问题与备选答案的词之间匹配的特征,获得各个词的匹配特征矩阵,然后再将特征矩阵输入神经网络模型中,计算两个文本之间的相似度。

[0004] 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:无论是基于词法特征的相似度确定方案还是基于语义特征的相似度确定方案,均没有考虑用户问题的意图信息。如果单纯地只考虑文本的语义特征或词法特征,对于用户一些带有明确主题意图的问题,上述文本意图相似度确定方案往往难以准确地找到答案进行匹配,尤其是对于出现率较低的非常见问题(称为长尾问题)的答案匹配准确度更低。

发明内容

[0005] 本发明实施例提供一种文本意图相似度确定方法、装置、电子设备和存储介质,以实现更加准确地确定文本之间的相似度。

[0006] 第一方面,本发明实施例提供了一种文本相似度确定方法,包括:

[0007] 获取待确定相似度的目标文本和备选文本;

[0008] 分别将所述目标文本和所述备选文本输入预设向量转换模型,生成所述目标文本对应的目标语义特征向量和所述备选文本对应的备选语义特征向量,其中,所述预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,所述预设意图类别基于业务场景而预先设定;

[0009] 依据所述目标语义特征向量和所述备选语义特征向量,确定所述目标文本和所述备选文本的文本意图相似度。

[0010] 第二方面,本发明实施例还提供了一种文本相似度确定装置,该装置包括:

[0011] 文本确定模块,用于获取待确定相似度的目标文本和备选文本;

[0012] 特征向量生成模块,用于分别将所述目标文本和所述备选文本输入预设向量转换模型,生成所述目标文本对应的目标语义特征向量和所述备选文本对应的备选语义特征向量,其中,所述预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,所述预设意图类别基于业务场景而预先设定;

[0013] 相似度确定模块,用于依据所述目标语义特征向量和所述备选语义特征向量,确定所述目标文本和所述备选文本的文本意图相似度。

[0014] 第三方面,本发明实施例还提供了一种电子设备,该电子设备包括:

[0015] 一个或多个处理器;

[0016] 存储装置,用于存储一个或多个程序,

[0017] 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本发明任意实施例所提供的文本相似度确定方法。

[0018] 第四方面,本发明实施例还提供了一种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本发明任意实施例所提供的文本相似度确定方法。

[0019] 本发明实施例通过基于文本意图分类的卷积神经网络模型预先训练获得预设向量转换模型,实现了将文本转换为带有用户意图的向量的模型训练,为后续基于意图的文本相似度计算提供模型基础。通过分别将获得的目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量;依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。实现了将目标文本和备选文本均转换为能够反映文本在各预设意图类别下的语义信息的特征向量,引入了文本对应的用户意图信息,并基于转换而来的两个特征向量计算文本相似度,能够获得基于文本意图的文本意图相似度,提高了文本相似度的准确性。

附图说明

[0020] 图1是本发明实施例一中的一种文本相似度确定方法的流程图;

[0021] 图2是本发明实施例二中的一种文本相似度确定方法中的预设向量转换模型的训练方法的流程图;

[0022] 图3是本发明实施例二中的文本意图分类的卷积神经网络模型的模型结构示意图;

[0023] 图4是本发明实施例二中的预设向量转换模型将文本转换为语义特征向量的过程示意图;

[0024] 图5是本发明实施例三中的一种文本相似度确定方法的流程图;

[0025] 图6是本发明实施例三中的应用预设向量转换模型确定文本意图相似度的过程示意图;

[0026] 图7是本发明实施例四中的一种文本相似度确定装置的结构示意图;

[0027] 图8是本发明实施例五中的一种电子设备的结构示意图。

具体实施方式

[0028] 下面结合附图和实施例对本发明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本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部结构。

[0029] 实施例一

[0030] 本实施例提供的文本相似度确定方法可适用于确定任意两个文本之间的文本相似度。该方法可以由文本相似度确定装置来执行,该装置可以由软件和/或硬件的方式实现,该装置可以集成在具有数据运算功能的电子设备中,例如平板电脑、台式电脑、服务器或网络设备等。参见图1,本实施例的方法具体包括如下步骤:

[0031] S110、获取待确定相似度的目标文本和备选文本。

[0032] 其中,目标文本是需要与其他文本计算相似度的文本。备选文本是配合目标文本进行文本相似度计算的文本,即上述其他文本。

[0033] 在计算文本相似度之前,需要先确定出目标文本和备选文本。目标文本可以是外部输入内容所得的文本,如用户通过输入设备输入的文本或输入的语音转换的文本,也可以是文本数据库中的文本。这里文本数据库是指由大量文本构成的数据库,其数据来源与业务需求有关。例如业务需求仅为计算文本间的相似度,那么文本数据库的数据来源可以是任何网络平台中的文本,而且是来源越多越好;如果业务需求为给用户问题匹配相似答案,那么由于不同智能问答系统中答案的侧重点有所不同,故文本数据库的数据来源便只能是目标文本对应的智能问答系统中的答案。备选文本可以是文本数据库中的任一文本,也可以是文本数据库中的文本经过初步筛选后所得的筛选文本中的任一文本。

[0034] S120、分别将目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量。

[0035] 其中,预设向量转换模型是预先训练的模型,用于将文本表达为向量。示例性地,预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得。卷积神经网络(Convolutional Neural Network,CNN)模型是一类包含卷积计算且具有深度结构的前馈神经网络(Feedforward Neural Networks),是深度学习(deep learning)的代表算法之一。由于CNN在文本分类上效果显著,因此经常被研究者使用进行用户意图分类。文本意图分类的卷积神经网络模型便是根据文本意图进行文本分类的卷积神经网络。本发明实施例正是利用了文本意图分类的卷积神经网络模型的这个特点来引入文本意图信息,以达到提高文本相似度精度的目的。又由于文本意图分类的卷积神经网络模型的输出是输入文本所属的意图类别,而本发明实施例中需要输出文本对应的表征向量,所以预设向量转换模型需要基于文本意图分类的卷积神经网络模型进行进一步的改进和训练。语义特征向量是隐含了文本意图信息和文本语义特征的向量。示例性地,语义特征向量反映文本在各预设意图类别下的语义信息,预设意图类别基于业务场景而预先设定。这里预设意图类别是指文本所属的意图的类别,例如物品属性类别或物品价格类别等,其是模型训练过程中人为的根据业务场景而设定的,或者说其是由模型训练时所采用的训练数据来客观确定的,而训练数据是人为预先准备的。

[0036] 将上述确定的目标文本输入预设向量转换模型,输出该目标文本的向量表征,即生成目标文本对应的目标语义特征向量。同样地,将确定的备选文本输入预设向量转换模

型,输出该备选文本的向量表征,即生成备选文本对应的备选语义特征向量。

[0037] S130、依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。

[0038] 其中,文本意图相似度是指考虑了文本意图信息的文本相似度。

[0039] 根据目标语义特征向量和备选语义特征向量进行向量余弦计算或欧式距离计算等向量相似度计算,便可确定目标语义特征向量和备选语义特征向量之间的相似度,该相似度便是目标文本和备选文本之间的相似度。由于目标语义特征向量和备选语义特征向量中均隐含了文本意图信息,故上述确定的相似度便为目标文本和备选文本之间的文本意图相似度。

[0040] 如果需要与目标文本进行相似度计算的备选文本数量为至少两个,那么针对每个备选文本均重复上述整个流程,即可获得目标文本与每个备选文本之间的文本意图相似度。为了提高实际操作中的运算速度,可以预先将每个备选文本输入预设向量转换模型,生成对应的备选语义特征向量或由向量构成的备选语义特征矩阵,而后将实时生成的目标语义特征向量与预先生成的每个备选语义特征向量或备选语义特征矩阵中的每个行向量计算相应的文本意图相似度。

[0041] 本实施例的技术方案,通过基于文本意图分类的卷积神经网络模型预先训练获得预设向量转换模型,实现了将文本转换为带有用户意图的向量的模型训练,为后续基于意图的文本相似度计算提供模型基础。通过分别将获得的目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量;依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。实现了将目标文本和备选文本均转换为能够反映文本在各预设意图类别下的语义信息的特征向量,引入了文本对应的用户意图信息,并基于转换而来的两个特征向量计算文本相似度,能够获得基于文本意图的文本意图相似度,提高了文本相似度的准确性。

[0042] 实施例二

[0043] 本实施例在上述实施例一的基础上,进一步说明了预设向量转换模型的训练过程。其中与上述各实施例相同或相应的术语的解释在此不再赘述。参见图 2,本实施例提供的文本相似度确定方法中的预设向量转换模型的训练方法包括:

[0044] S210、获得与各预设意图类别对应的训练样本,训练样本包含待训练文本和待训练文本对应的预设意图类别。

[0045] 在模型训练之前,需要准备训练样本。本实施例训练样本中的待训练文本从文本数据库中选取,如果业务场景为智能问答系统中为用户问题匹配相似答案,那么训练样本则从包含问题和答案的问答数据库中选取。确定了待训练文本之后,需要为每个选取的待训练文本标注其所属的预设意图类别,待训练文本及其所属的预设意图类别便为一组训练样本。所选取的训练样本应该包含所有的预设意图类别,这样训练样本的数量至少为预设意图类别的数量。为了取得更好的训练效果,应该增加训练样本的数量,且每个预设意图类别对应的训练样本数量应当大致相等,以使后续训练的模型能够很好地兼容每种预设意图类别。

[0046] S220、利用训练样本训练文本意图分类的卷积神经网络模型,获得目标模型参数。

[0047] 参见图3,文本意图分类的卷积神经网络模型包含卷积层、池化层和全连接层,其

输入为一个文本,输出为该文本对应的意图类别。在模型训练初期,文本意图分类的卷积神经网络模型的模型参数为初始模型参数,该初始模型参数可以是基于零的较小随机数,也可以是人为设定的初始值。模型训练的过程便是不断修正模型参数,直至获得满意的模型输出结果,此时的模型参数便为目标模型参数,其可以直接用于后续模型预测(或称模型使用)。

[0048] 在进行模型训练时,将训练样本中的每个待训练文本输入文本意图分类的卷积神经网络模型,经过卷积层、池化层和全连接层softmax的处理,便可输出相应待训练文本对应的模型输出意图类别。该过程可以表征为公式(1):

$$[0049] \quad \hat{T} = \operatorname{argmax}_T \{\log P(T|S)\} \quad (1)$$

[0050] 其中, $\hat{T}$ 表示待训练文本S的模型输出意图类别,其为所有预设意图类别中最大 $\log P(T|S)$ 值对应的预设意图类别, $P(T|S)$ 表示待训练文本S属于预设意图类别T的后验概率。

[0051] 将模型输出意图类别 $\hat{T}$ 与待训练文本对应的预设意图类别进行比较,并结合损失函数进行误差反传来修正模型参数。不断循环上述整个模型训练过程,直至模型输出意图类别与待训练文本对应的预设意图类别满足一定的误差条件,此时的模型参数便为目标模型参数。

[0052] 示例性地,文本意图分类的卷积神经网络模型为文本分类卷积神经网络模型,且在基于文本意图分类的卷积神经网络模型的训练过程中采用注意力机制。文本分类卷积神经网络模型模型(Convolutional Neural Networks for Sentence Classification, textCNN)是常用的文本意图分类的卷积神经网络模型中的一种。注意力机制即attention机制,其会根据文本中词的词性而为每个词分配不同的权重,例如实词权重重大,而虚词权重小。注意力机制的采用能够进一步提高textCNN模型的训练效果。

[0053] S230、依据目标模型参数、文本意图分类的卷积神经网络模型中的卷积层和池化层,生成预设向量转换模型。

[0054] 根据上述说明,如果直接使用传统的textCNN模型,输入目标文本(或备选文本),那么会直接输出目标文本(或备选文本)所属的类别,无法进一步判定目标文本与备选文本之间相关度,所以本发明实施例中并未完全采用textCNN模型,而是采用textCNN模型中池化层的输出向量作为输入文本的表征向量。由于全连接层的作用是将池化层输出向量再经过逻辑回归计算后得到输入文本属于每个预设意图类别下的概率,而本发明实施例不需要判断上述概率,故可以直接选择将textCNN模型的卷积层或池化层的输出向量作为文本的表征向量。又考虑到卷积层的输出向量维度过高,数据量过大,故本实施例中选择采用池化层的输出向量,以更小的数据量来更加全面的表征模型输入文本的隐含文本意图的向量。

[0055] 那么,生成预设向量转换模型的过程便为:利用目标模型参数替代文本意图分类的卷积神经网络模型的初始模型参数,获得完善的文本意图分类的卷积神经网络模型,并截取完善后的文本意图分类的卷积神经网络模型的输入层、卷积层和池化层,便可获得预设向量转换模型。

[0056] 参见图4,预设向量转换模型将文本转换为向量的过程为:

[0057] 首先通过公式(2)切割输入的文本,获得输入文本的矩阵表示(简称输入文本矩

阵)。然后,通过公式(3)对输入文本矩阵中的相应向量进行卷积计算,获得每个卷积核对应的输出向量,例如采用了6个卷积核,便可获得6个卷积层向量,如公式(4),每个卷积层向量的列维数和行维数与卷积核的大小及输入文本矩阵的大小相关。最后,通过公式(5)对每个卷积层向量进行最大池化(1-max)操作,获得池化层向量,该池化层向量便是输入文本的语义特征向量,其为输入文本的最终向量表示,池化层向量中每个元素表示某个特定的语义信息。

$$[0058] \quad X_{1:n} = X_1 \oplus X_2 \oplus \cdots \oplus X_n \quad (2)$$

$$[0059] \quad C_i = f(W \cdot X_{i:i+N-1} + b) \quad (3)$$

$$[0060] \quad C = [C_1, C_2, \dots, C_{n-N+1}] \quad (4)$$

$$[0061] \quad \hat{C} = \max[C] \quad (5)$$

[0062] 其中,公式(2)的X表示被N-gram形式切割成的文本片段,例如2-gram的话,文本“货什么时候到达”会被切割成 $\{X_{1:2} = \text{“货什”}; X_{2:3} = \text{“什么”}; X_{3:4} = \text{“么时”}; X_{4:5} = \text{“时候”}; X_{5:6} = \text{“候到”}; X_{6:7} = \text{“到达”}\}$ 。公式(3)的f表示卷积函数,W为卷积矩阵, $X_{i:i+N-1}$ 表示公式(3)中的文本片段的表征向量, C_i 表示每个卷积核输出的卷积层向量。式(5)的max表示对所有卷积层向量中的每个卷积层向量进行最大池化,得出一个最大值。最终所有卷积层向量的最大池化值可以拼接成一个向量,即池化层向量 $\hat{C}$ 。

[0063] 本实施例的技术方案,通过对文本意图分类的卷积神经网络模型进行训练获得目标模型参数,及依据目标模型参数、文本意图分类的卷积神经网络模型中的卷积层和池化层,生成预设向量转换模型,能够更加快速的获得更加稳定的预设向量转换模型,为后续文本意图相似度确定提供模型基础,从而进一步提高文本意图相似度的确定效率。

[0064] 实施例三

[0065] 本实施例在上述实施例一的基础上,进一步优化了“获取待确定相似度的备选文本”。在此基础上,增加了备选文本排序的步骤。其中与上述各实施例相同或相应的术语的解释在此不再赘述。

[0066] 本实施例中的业务场景为电商智能问答,目标文本为用户问题,备选文本为答案数据库中的备选答案,各预设意图类别包含陈述事实类别、物品属性类别、物品价格类别、物流类别、物品安装类别、促销信息类别、信息推荐类别和其他类别。

[0067] 在电商智能问答场景下,需要为用户提出的问题(简称用户问题)匹配合适的参考答案,以供人工客服或智能客服来快速回答用户问题。此时,目标文本便为用户问题,备选文本便是答案数据库中的备选答案。这里答案数据库来源为客服与用户之间的历史对话日志或其他用户回答用户问题的留言答案等。用户问题的预设意图类别设置为陈述事实类别、物品属性类别(如物品尺寸和物品组成等)、物品价格类别、物流类别(如物流承运商和配送周期等)、物品安装类别(如是否免费安装及安装时间等)、促销信息类别(如促销力度和促销时间等)、信息推荐类别(如根据用户需求推荐合适的物品等)和其他类别。

[0068] 参见图5,本实施例提供的文本相似度确定方法包括:

[0069] S310、获取待确定相似度的目标文本。

[0070] 获取用户问题,作为目标文本。该用户问题可以是用户输入的文本,也可以是用户输入的语音或图片等转换而来的文本。

[0071] S320、依据用户问题和答案数据库,基于预设相似度确定算法,生成用户问题对应的初始参照答案,作为备选文本。

[0072] 其中,预设相似度确定算法是指预先选定的用于初步计算文本与文本之间相似度的算法,其特点为运算效率高,返回相似文本的文本召回率高,但是对相似度的确定精度要求较低,例如可以是最长公共子序列或基于关键词贡献的相似度确定算法等。

[0073] 为了减少后续相似度精算的数据量,先采用预设相似度确定算法,对用户问题和答案数据库中的每个备选答案进行初步相似度计算,从所有备选答案中筛选出与用户问题相似度高的部分备选答案,作为用户问题的初始参照答案,也即用户问题对应的各备选文本。

[0074] S330、分别将目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量。

[0075] S340、依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。

[0076] 参见图6,将用户问题和一个备选文本分别输入预设向量转换模型,经过模型卷积层和池化层的处理,便可分别输出用户问题对应的目标语义特征向量和该备选文本对应的备选语义特征向量,最后计算目标语义特征向量和备选语义特征向量之间的余弦距离,便可获得用户问题与该备选文本之间的文本意图相似度。针对每个备选文本均采用上述过程,便可获得用户问题与每个备选文本之间的文本意图相似度,该文本意图相似度的数量与备选文本(即初始参照答案)的数量一致。

[0077] S350、依据各文本意图相似度对各备选文本进行排序,生成备选文本相似度排序结果,作为用户问题的目标参照答案。

[0078] 每个备选文本与用户问题之间的文本意图相似度有大小差异,相似度大的备选文本的参照意义更大,所以根据每个备选文本对应的文本意图相似度的大小,将所有的备选文本进行降序排列,获得备选文本相似度排序结果,作为用户问题的目标参照答案,这样可以进一步提高客服选用备选文本的效率,也提高用户问题的回答效率。

[0079] 示例性地,依据各文本意图相似度对各备选文本进行排序,生成备选文本相似度排序结果包括:将各文本意图相似度及除文本意图相似度之外的至少一个其余特征维度的各相似度输入快速梯度提升排序模型,生成备选文本相似度排序结果。

[0080] 其中,快速梯度提升排序模型(eXtreme Gradient Boosting,XGBoost)是一种提升树模型,是在梯度提升Gradient Boosting框架下实现的机器学习算法。它属于boosting算法的一种,其将许多树模型集成在一起,提供了并行树提升(也称为基于梯度提高的决策树(Gradient Boosting Decision Tree,GBDT)),从而形成一个很强的分类器。它可以快速准确地、高效地及灵活地解决许多数据科学问题,例如分类(classification)和排序(rank)。本发明实施例中利用了XGBoost for rank模型。

[0081] 具体实施时,将基于预设向量转换模型确定的文本意图相似度作为一维特征输入到XGBoost模型中,同时将除文本意图相似度之外的至少一个其余特征维度的各相似度,如S320中确定的各相似度,即基于关键词贡献算法确定的用户问题与各备选文本之间的相似度,及采用最长公共子序列确定的用户问题与各备选文本之间的相似度中的至少一种,也作为一维特征输入到XGBoost模型中,实现从不同维度的相似度对各备选文本排序,获得备

选文本相似度排序结果。这样设置的好处在于提升相似度排序的效率和精度,从而进一步提高客服回答用户问题的效率和准确性。实验表明,相比于仅利用S320中各维度的相似度进行基于XGBoost模型排序的备选文本相似度排序结果,加入了文本意图相似度维度的基于XGBoost模型排序的备选文本相似度排序结果的准确率提升了 5%。

[0082] 本实施例的技术方案,通过依据用户问题和答案数据库,基于预设相似度确定算法,生成用户问题对应的初始参照答案,作为备选文本,实现了答案数据库中所有备选答案的初步筛选,减少了后续相似度精算的数据量,进一步提高了文本意图相似度的确定效率。通过依据各文本意图相似度对各备选文本进行排序,生成备选文本相似度排序结果,作为用户问题的目标参照答案。实现了各备选文本的精确排序,进一步提高了电商智能问答业务场景下用户问题的答复针对性。通过在电商智能问答业务场景下,引入问答对的意图类别信息来确定问答对的相关度,解决了长尾问题的答案匹配率低的问题,提高了长尾问题的备选答案的匹配准确率,也进一步提高了问答对匹配的语义泛化性。

[0083] 实施例四

[0084] 本实施例提供一种文本相似度确定装置,参见图7,该装置具体包括:

[0085] 文本确定模块710,用于获取待确定相似度的目标文本和备选文本;

[0086] 特征向量生成模块720,用于分别将目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量,其中,预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,预设意图类别基于业务场景而预先设定;

[0087] 相似度确定模块730,用于依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。

[0088] 可选地,在上述装置的基础上,该装置还包括模型训练模块,用于通过如下方式基于文本意图分类的卷积神经网络模型预先训练获得预设向量转换模型:

[0089] 获得与各预设意图类别对应的训练样本,训练样本包含待训练文本和待训练文本对应的预设意图类别;

[0090] 获得目标模型参数利用训练样本训练文本意图分类的卷积神经网络模型,获得目标模型参数;

[0091] 依据目标模型参数、文本意图分类的卷积神经网络模型中的卷积层和池化层,生成预设向量转换模型。

[0092] 进一步地,文本意图分类的卷积神经网络模型为文本分类卷积神经网络模型,且在基于文本意图分类的卷积神经网络模型的训练过程中采用注意力机制。

[0093] 可选地,业务场景为电商智能问答,目标文本为用户问题,备选文本为答案数据库中的备选答案,各预设意图类别包含陈述事实类别、物品属性类别、物品价格类别、物流类别、物品安装类别、促销信息类别、信息推荐类别和其他类别。

[0094] 可选地,文本确定模块710具体用于:

[0095] 依据用户问题和答案数据库,基于预设相似度确定算法,生成用户问题对应的初始参照答案,作为备选文本。

[0096] 可选地,在上述装置的基础上,该装置还包括文本排序模块,用于:

[0097] 在依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本

意图相似度之后,依据各文本意图相似度对各备选文本进行排序,生成备选文本相似度排序结果,作为用户问题的目标参照答案。

[0098] 进一步地,文本排序模块具体用于:将各文本意图相似度及除文本意图相似度之外的至少一个其余特征维度的各相似度输入快速梯度提升排序模型,生成备选文本相似度排序结果。

[0099] 通过本发明实施例四的一种文本相似度确定装置,实现了将目标文本和备选文本均转换为能够反映文本在各预设意图类别下的语义信息的特征向量,引入了文本对应的用户意图信息,并基于转换而来的两个特征向量计算文本相似度,能够获得基于文本意图的文本意图相似度,提高了文本相似度的准确性。

[0100] 本发明实施例所提供的文本相似度确定装置可执行本发明任意实施例所提供的文本相似度确定方法,具备执行方法相应的功能模块和有益效果。

[0101] 值得注意的是,上述文本相似度确定装置的实施例中,所包括的各个单元和模块只是按照功能逻辑进行划分的,但并不局限于上述的划分,只要能够实现相应的功能即可;另外,各功能单元的具体名称也只是为了便于相互区分,并不用于限制本发明的保护范围。

[0102] 实施例五

[0103] 参见图8,本实施例提供了一种电子设备,其包括:一个或多个处理器820;存储装置810,用于存储一个或多个程序,当一个或多个程序被一个或多个处理器820执行,使得一个或多个处理器820实现本发明实施例所提供的文本相似度确定方法,包括:

[0104] 获取待确定相似度的目标文本和备选文本;

[0105] 分别将目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量,其中,预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,预设意图类别基于业务场景而预先设定;

[0106] 依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。

[0107] 当然,本领域技术人员可以理解,处理器820还可以实现本发明任意实施例所提供的文本相似度确定方法的技术方案。

[0108] 图8显示的电子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。

[0109] 如图8所示,该电子设备包括处理器820、存储装置810、输入装置830和输出装置840;电子设备中处理器820的数量可以是一个或多个,图8中以一个处理器820为例;电子设备中的处理器820、存储装置810、输入装置830和输出装置840可以通过总线或其他方式连接,图8中以通过总线850连接为例。

[0110] 存储装置810作为一种计算机可读存储介质,可用于存储软件程序、计算机可执行程序以及模块,如本发明实施例中的文本相似度确定方法对应的程序指令/模块(例如,文本相似度确定装置中的文本确定模块、特征向量生成模块和相似度确定模块)。

[0111] 存储装置810可主要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、至少一个功能所需的应用程序;存储数据区可存储根据终端的使用所创建的数据等。此外,存储装置810可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少

一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中,存储装置810可进一步包括相对于处理器820远程设置的存储器,这些远程存储器可以通过网络连接至电子设备。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。

[0112] 输入装置830可用于接收输入的数字或字符信息,以及产生与电子设备的用户设置以及功能控制有关的键信号输入。输出装置840可包括显示屏等显示设备。

[0113] 实施例六

[0114] 本实施例提供一种包含计算机可执行指令的存储介质,计算机可执行指令在由计算机处理器执行时用于执行一种文本相似度确定方法,该方法包括:

[0115] 获取待确定相似度的目标文本和备选文本;

[0116] 分别将目标文本和备选文本输入预设向量转换模型,生成目标文本对应的目标语义特征向量和备选文本对应的备选语义特征向量,其中,预设向量转换模型基于文本意图分类的卷积神经网络模型预先训练获得,语义特征向量反映文本在各预设意图类别下的语义信息,预设意图类别基于业务场景而预先设定;

[0117] 依据目标语义特征向量和备选语义特征向量,确定目标文本和备选文本的文本意图相似度。

[0118] 当然,本发明实施例所提供的一种包含计算机可执行指令的存储介质,其计算机可执行指令不限于如上的方法操作,还可以执行本发明任意实施例所提供的文本相似度确定方法中的相关操作。

[0119] 通过以上关于实施方式的描述,所属领域的技术人员可以清楚地了解到,本发明可借助软件及必需的通用硬件来实现,当然也可以通过硬件实现,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的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,如计算机的软盘、只读存储器 (Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、闪存(FLASH)、硬盘或光盘等,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)执行本发明各个实施例所提供的文本相似度确定方法。

[0120] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

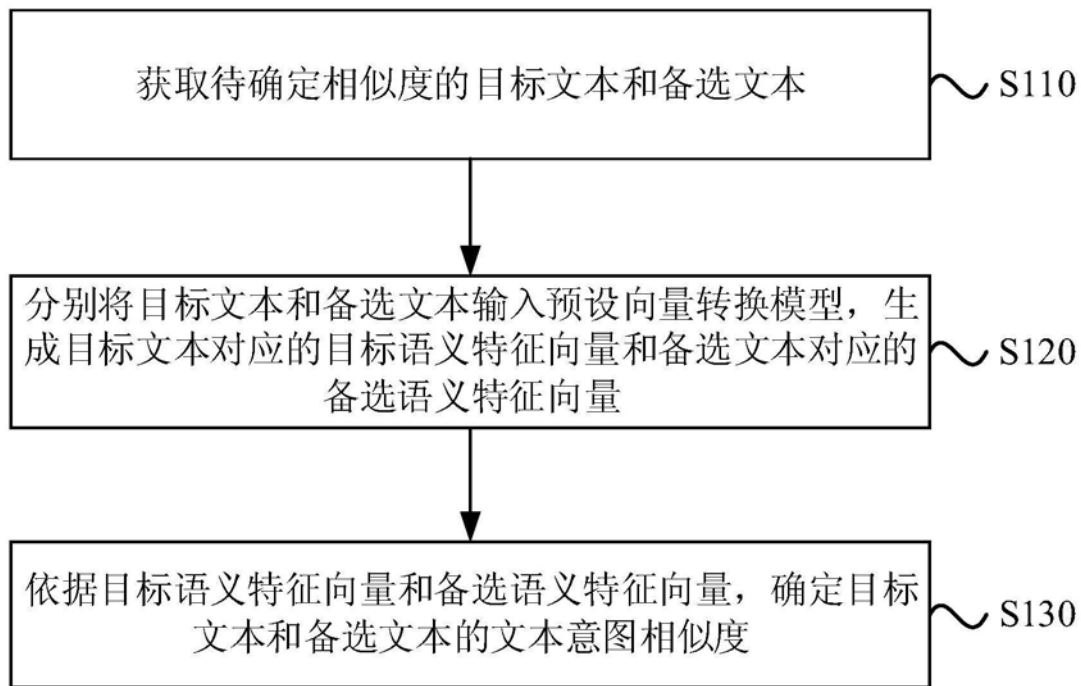


图1

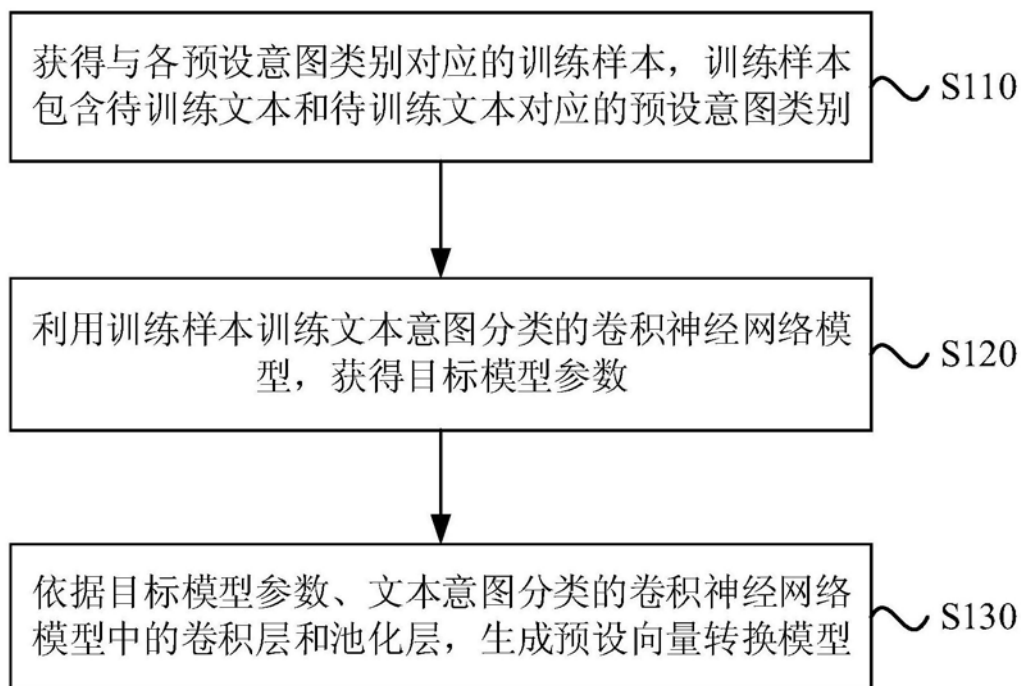


图2

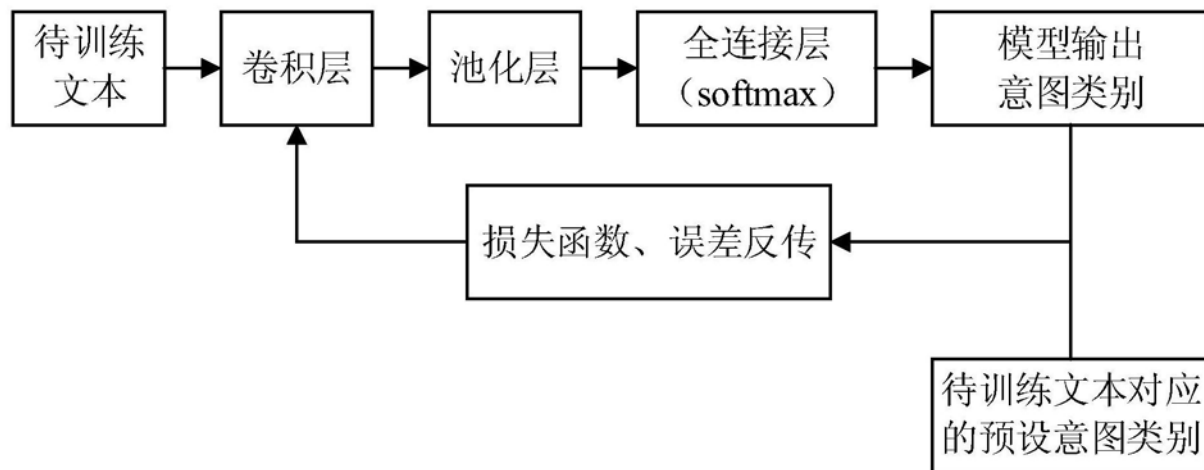


图3

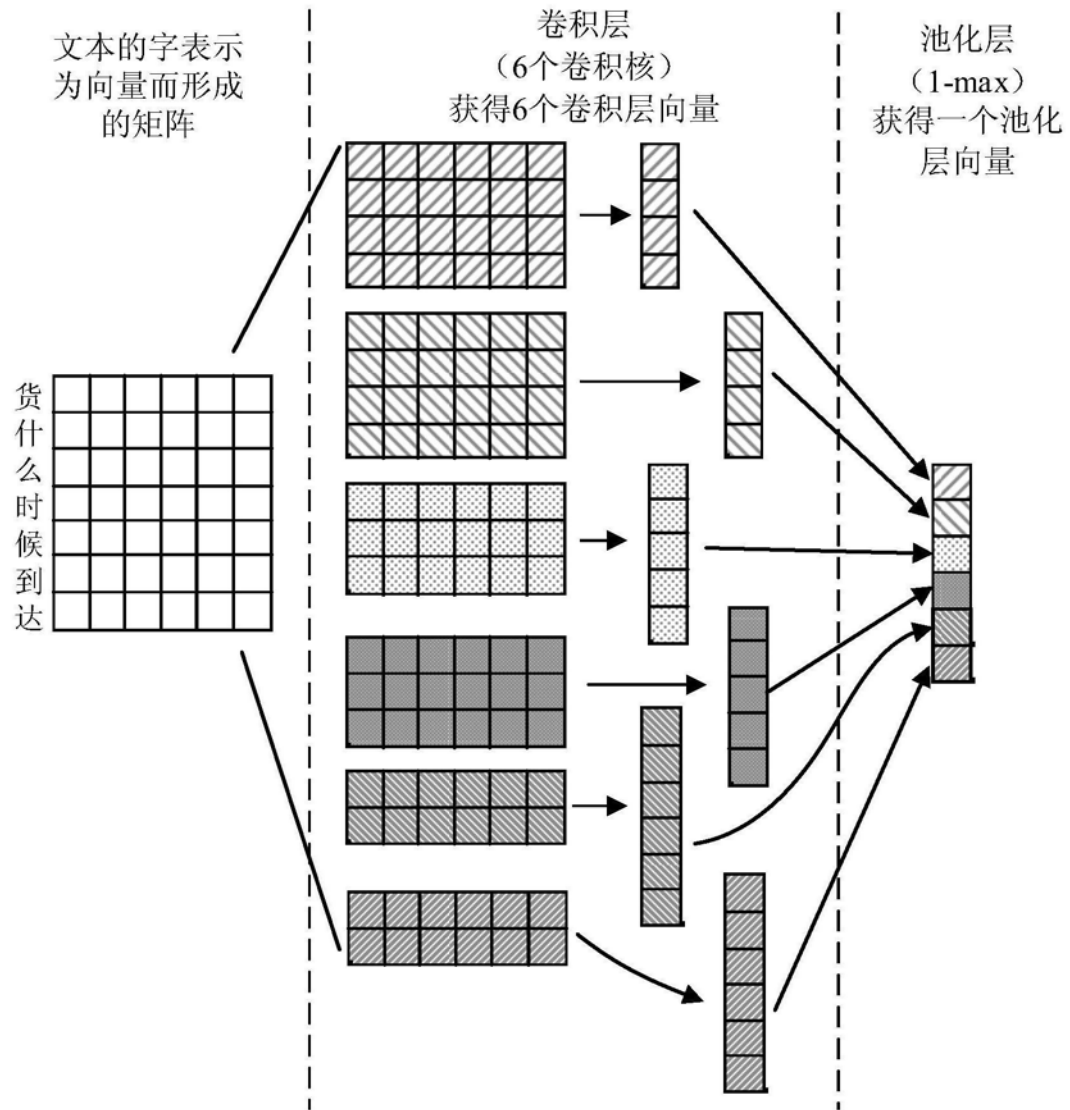


图4

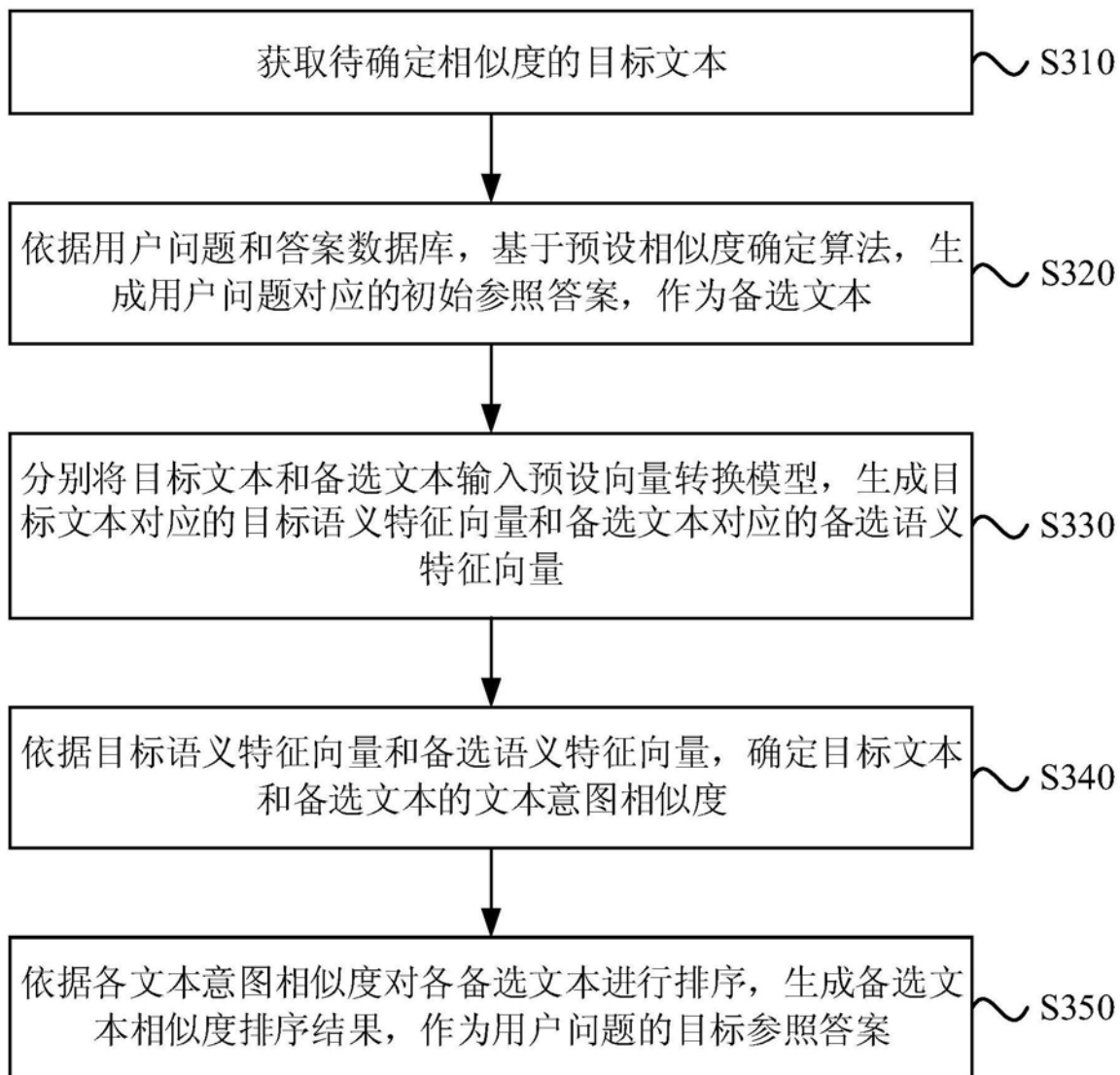


图5

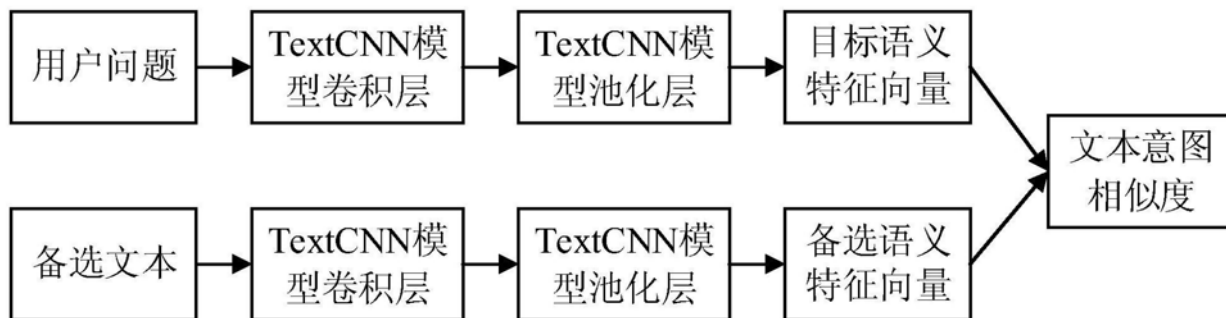


图6

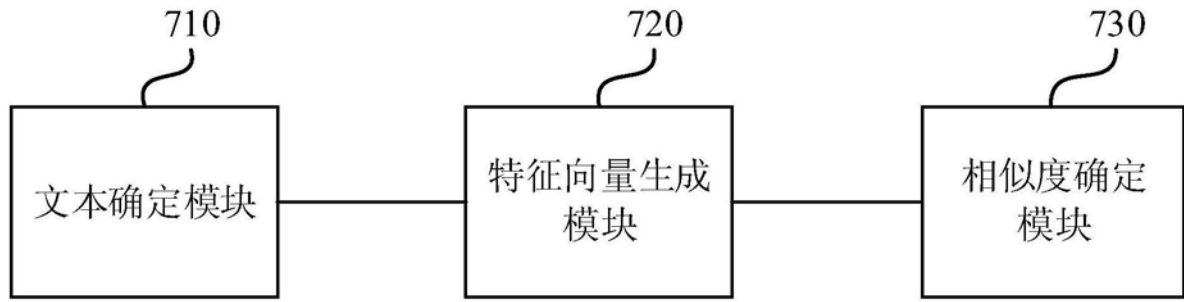


图7

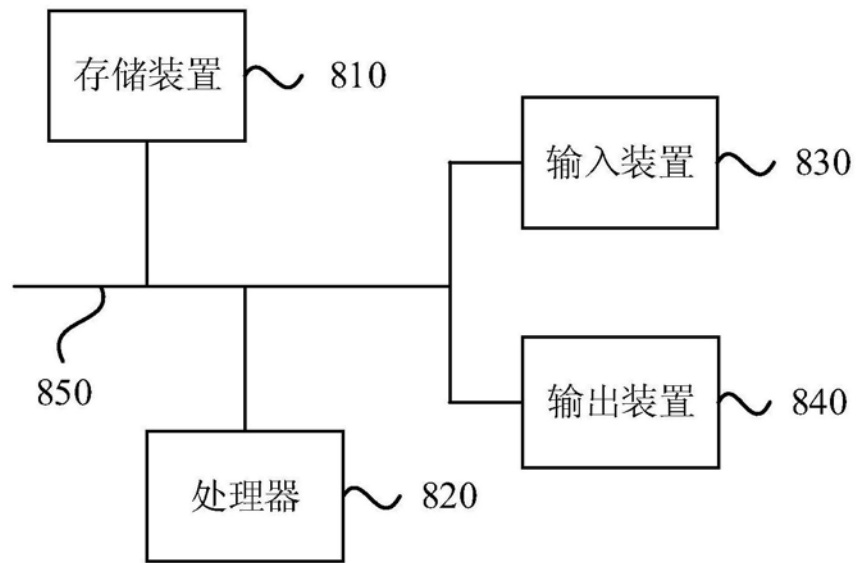


图8