(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 111738010 A
(43)申请公布日 2020.10.02

(21)申请号 201910211647.1

(22)申请日 2019.03.20

(71)申请人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 黄强 卜建辉 陈林 吴伟佳 谢炜坚

(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
代理人 王达佐 马晓亚

(51)Int.Cl.
*G06F 40/30*(2020.01)
*G06F 16/332*(2019.01)
*G06F 16/33*(2019.01)
*G06K 9/62*(2006.01)

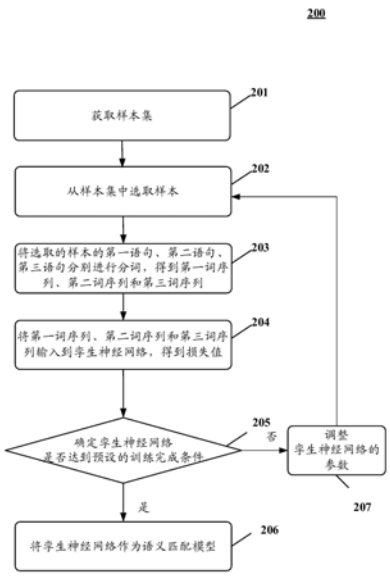
权利要求书2页 说明书11页 附图5页

(54)发明名称

用于生成语义匹配模型的方法和装置

(57)摘要

本申请实施例公开了用于生成语义匹配模型的方法和装置。该方法的一具体实施方式包括:获取样本集,其中,样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句;从样本集中选取样本,以及执行以下训练步骤:将选取的样本的第一语句、第二语句、第三语句分别进行分词,得到第一词序列、第二词序列和第三词序列;将第一词序列、第二词序列和第三词序列输入到孪生神经网络,得到损失值;根据损失值确定孪生神经网络是否训练完成;响应于确定孪生神经网络训练完成,将孪生神经网络作为语义匹配模型。该实施方式能够提高语义匹配的准确性。




CN 111738010 A

1.一种用于生成语义匹配模型的方法,包括:

获取样本集,其中,所述样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句;

从所述样本集中选取样本,以及执行以下训练步骤:将选取的样本的第一语句、第二语句、第三语句分别进行分词,得到第一词序列、第二词序列和第三词序列;将所述第一词序列、所述第二词序列和所述第三词序列输入到孪生神经网络,得到损失值;根据所述损失值确定孪生神经网络是否训练完成;响应于确定所述孪生神经网络训练完成,将所述孪生神经网络作为语义匹配模型。

2.根据权利要求1所述的方法,其中,所述将所述第一词序列、所述第二词序列和所述第三词序列输入到孪生神经网络,得到损失值,包括:

将所述第一词序列、所述第二词序列和所述第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列;

将所述第一词向量序列、所述第二词向量序列和所述第三词向量序列分别转换成第一语句向量、第二语句向量和第三语句向量;

确定所述第一语句向量与所述第二语句向量之间的第一余弦相似度和所述第一语句向量与所述第三语句向量之间的第二余弦相似度;

根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

3.根据权利要求1所述的方法,其中,所述第一语句为用户搜索的语句、所述第二语句为用户点击查看的语句、第三语句为用户未点击查看的语句。

4.根据权利要求1所述的方法,其中,一个样本中第三语句的数量大于等于1。

5.根据权利要求1所述的方法,其中,所述孪生神经网络包括第一子网络和至少一个第二子网络,所述第一子网络的输出结果为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值,所述第二子网络的输出结果为1减去第一语句和第三语句的余弦相似度的sigmoid函数值。

6.根据权利要求1-5之一所述的方法,其中,所述方法还包括:

响应于确定出所述孪生神经网络未训练完成,调整所述孪生神经网络中的相关参数,以及从所述样本集中重新选取样本,使用调整后的孪生神经网络继续执行所述训练步骤。

7.一种用于输出信息的方法,包括:

接收待匹配的第一目标语句和第二目标语句;

将所述第一目标语句和所述第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列;

将所述第一目标词序列、所述第二目标词序列输入如权利要求1-6之一所述的方法生成的语义匹配模型中,生成所述第一目标语句和所述第二目标语句的语义匹配结果。

8.一种用于生成语义匹配模型的装置,包括:

获取单元,被配置成获取样本集,其中,所述样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句;

训练单元,被配置成从所述样本集中选取样本,以及执行以下训练步骤:将选取的样本的第一语句、第二语句、第三语句分别进行分词,得到第一词序列、第二词序列和第三词序列;将所述第一词序列、所述第二词序列和所述第三词序列输入到孪生神经网络,得到损失

值;根据所述损失值确定孪生神经网络是否训练完成;响应于确定所述孪生神经网络训练完成,将所述孪生神经网络作为语义匹配模型。

9.根据权利要求8所述的装置,其中,所述训练单元进一步被配置成:

将所述第一词序列、所述第二词序列和所述第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列;

将所述第一词向量序列、所述第二词向量序列和所述第三词向量序列分别转换成第一语句向量、第二语句向量和第三语句向量;

确定所述第一语句向量与所述第二语句向量之间的第一余弦相似度和所述第一语句向量与所述第三语句向量之间的第二余弦相似度;

根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

10.根据权利要求8所述的装置,其中,所述第一语句为用户搜索的语句、所述第二语句为用户点击查看的语句、第三语句为用户未点击查看的语句。

11.根据权利要求8所述的装置,其中,一个样本中第三语句的数量大于等于1。

12.根据权利要求8所述的装置,其中,所述孪生神经网络包括第一子网络和至少一个第二子网络,所述第一子网络的输出结果为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值,所述第二子网络的输出结果为1减去第一语句和第三语句的余弦相似度的sigmoid函数值。

13.根据权利要求8-12所述的装置,其中,所述装置还包括调整单元,被配置成:

响应于确定出所述孪生神经网络未训练完成,调整所述孪生神经网络中的相关参数,以及从所述样本集中重新选取样本,使用调整后的孪生神经网络继续执行所述训练步骤。

14.一种用于输出信息的装置,包括:

接收单元,被配置成接收待匹配的第一目标语句和第二目标语句;

转换单元,被配置成将所述第一目标语句和所述第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列;

输出单元,被配置成将所述第一目标词序列、所述第二目标词序列输入如权利要求1-6之一所述的方法生成的语义匹配模型中,生成所述第一目标语句和所述第二目标语句的语义匹配结果。

15.一种电子设备,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1-7中任一所述的方法。

16.一种计算机可读介质,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1-7中任一所述的方法。

# 用于生成语义匹配模型的方法和装置

## 技术领域

[0001] 本公开的实施例涉及计算机技术领域，具体涉及用于生成语义匹配模型的方法和装置。

## 背景技术

[0002] 对很多行业来说，建设一个自动问答系统是一个非常热门的主题，在自动问答系统中一个关键的问题是，从自动问答知识库中检索出一个给定问题的最相似问题，这可以被重新定义为一个语义句子匹配问题。

[0003] 现有的技术方案主要分为两种，一种是基于词或者同义词的文本相似度对句子进行相关性计算，这种方法只能计算句子的文本相似度，并不能理解句子的语义。另外一种则是，基于自由文本语料和语言模型去获取近似的句子语义表示，但是这种近似的方式会造成句子语义表示发生偏差，导致效果不好。

## 发明内容

[0004] 本公开的实施例提出了用于生成语义匹配模型的方法和装置。

[0005] 第一方面，本公开的实施例提供了一种用于生成语义匹配模型的方法，包括：获取样本集，其中，样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句；从样本集中选取样本，以及执行以下训练步骤：将选取的样本的第一语句、第二语句、第三语句分别进行分词，得到第一词序列、第二词序列和第三词序列；将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值；根据损失值确定孪生神经网络是否训练完成；响应于确定孪生神经网络训练完成，将孪生神经网络作为语义匹配模型。

[0006] 在一些实施例中，将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值，包括：将第一词序列、第二词序列和第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列；将第一词向量序列、第二词向量序列和第三词向量序列分别转换成第一语句向量、第二语句向量和第三语句向量；确定第一语句向量与第二语句向量之间的第一余弦相似度和第一语句向量与第三语句向量之间的第二余弦相似度；根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

[0007] 在一些实施例中，第一语句为用户搜索的语句、第二语句为用户点击查看的语句、第三语句为用户未点击查看的语句。

[0008] 在一些实施例中，一个样本中第三语句的数量大于等于1。

[0009] 在一些实施例中，孪生神经网络包括第一子网络和至少一个第二子网络，第一子网络的输出结果为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值，第二子网络的输出结果为1减去第一语句和第三语句的余弦相似度的sigmoid函数值。

[0010] 在一些实施例中，该方法还包括：响应于确定出孪生神经网络未训练完成，调整孪生神经网络中的相关参数，以及从样本集中重新选取样本，使用调整后的孪生神经网络继

续执行训练步骤。

[0011] 第二方面，本公开的实施例提供了一种用于输出信息的方法，包括：接收待匹配的第一目标语句和第二目标语句；将第一目标语句和第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列；将第一目标词序列、第二目标词序列输入如第一方面之一的方法生成的语义匹配模型中，生成第一目标语句和第二目标语句的语义匹配结果。

[0012] 第三方面，本公开的实施例提供了一种用于生成语义匹配模型的装置，包括：获取单元，被配置成获取样本集，其中，样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句；训练单元，被配置成从样本集中选取样本，以及执行以下训练步骤：将选取的样本的第一语句、第二语句、第三语句分别进行分词，得到第一词序列、第二词序列和第三词序列；将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值；根据损失值确定孪生神经网络是否训练完成；响应于确定孪生神经网络训练完成，将孪生神经网络作为语义匹配模型。

[0013] 在一些实施例中，训练单元进一步被配置成：将第一词序列、第二词序列和第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列；将第一词向量序列、第二词向量序列和第三词向量序列分别转换成第一语句向量、第二语句向量和第三语句向量；确定第一语句向量与第二语句向量之间的第一余弦相似度和第一语句向量与第三语句向量之间的第二余弦相似度；根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

[0014] 在一些实施例中，第一语句为用户搜索的语句、第二语句为用户点击查看的语句、第三语句为用户未点击查看的语句。

[0015] 在一些实施例中，其中，一个样本中第三语句的数量大于等于1。

[0016] 在一些实施例中，孪生神经网络包括第一子网络和至少一个第二子网络，第一子网络的输出结果为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值，第二子网络的输出结果为1减去第一语句和第三语句的余弦相似度的sigmoid函数值。

[0017] 在一些实施例中，该装置还包括调整单元，被配置成：响应于确定出孪生神经网络未训练完成，调整孪生神经网络中的相关参数，以及从样本集中重新选取样本，使用调整后的孪生神经网络继续执行训练步骤。

[0018] 第四方面，本公开的实施例提供了一种用于输出信息的装置，包括：接收单元，被配置成接收待匹配的第一目标语句和第二目标语句；转换单元，被配置成将第一目标语句和第二目标语句分别进行切词后转换成第一目标词序列、第二目标词序列；输出单元，被配置成将第一目标词序列、第二目标词序列输入如第一方面之一的方法生成的语义匹配模型中，生成第一目标语句和第二目标语句的语义匹配结果。

[0019] 第五方面，本公开的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中任一的方法。

[0020] 第六方面，本公开的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中任一的方法。

[0021] 本公开的实施例提供的用于生成语义匹配模型的方法和装置，通过孪生神经网络和海量的搜索数据生成一个语义匹配模型，用于通用问答场景的文本语义相关性计算。在

FAQ(Frequently Asked Questions,常见问题解答)系统中,在没有任何训练数据的情况下,本公开的语义匹配模型能够较为准确地计算用户提出的问题跟FAQ知识库中的标准问题的语义相关性,从而检索出最佳的答案。除此之外产出的句子向量还能够作为其他自然语言处理的基础特征

## 附图说明

[0022] 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:

[0023] 图1是本公开可以应用于其中的示例性系统架构图;

[0024] 图2是根据本公开用于生成语义匹配模型的方法的一个实施例的流程图;

[0025] 图3是根据本公开用于生成语义匹配模型的方法的一个应用场景的示意图;

[0026] 图4是根据本公开用于生成语义匹配模型的装置的一个实施例的结构示意图;

[0027] 图5是根据本公开用于输出信息的方法的一个实施例的流程图;

[0028] 图6是根据本公开用于输出信息的装置的一个实施例的结构示意图;

[0029] 图7是适于用来实现本公开实施例的电子设备的计算机系统的结构示意图。

## 具体实施方式

[0030] 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。

[0031] 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。

[0032] 图1示出了可以应用本申请实施例的用于生成语义匹配模型的方法、用于生成语义匹配模型的装置、用于输出信息的方法或用于输出信息的装置的示例性系统架构100。

[0033] 如图1所示,系统架构100可以包括终端101、102,网络103、数据库服务器104和服务器105。网络103用以在终端101、102,数据库服务器104与服务器105之间提供通信链路的介质。网络103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。

[0034] 用户110可以使用终端101、102通过网络103与服务器105进行交互,以接收或发送消息等。终端101、102上可以安装有各种客户端应用,例如模型训练类应用、自动问答类应用、购物类应用、支付类应用、网页浏览器和即时通讯工具等。

[0035] 这里的终端101、102可以是硬件,也可以是软件。当终端101、102为硬件时,可以是具有显示屏的各种电子设备,包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture Experts Group Audio Layer III,动态影像专家压缩标准音频层面3)、膝上型便携计算机和台式计算机等等。当终端101、102为软件时,可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块(例如用来提供分布式服务),也可以实现成单个软件或软件模块。在此不做具体限定。

[0036] 当终端101、102为硬件时,其上还可以安装有问答信息采集设备。问答信息采集设备可以是各种能实现采集问答信息功能的设备,如麦克风、键盘等等。用户110可以利用终端101、102上的问答信息采集设备,来采集问题和答案。

[0037] 数据库服务器104可以是提供各种服务的数据库服务器。例如数据库服务器中可以存储有样本集。样本集中包含有大量的样本。其中，样本可以包括第一语句、第二语句、标注信息。这样，用户110也可以通过终端101、102，从数据库服务器104所存储的样本集中选取样本。

[0038] 服务器105也可以是提供各种服务的服务器，例如对终端101、102上显示的各种应用提供支持的后台服务器。后台服务器可以利用终端101、102发送的样本集中的样本，对初始语义匹配模型进行训练，并可以将训练结果(如生成的语义匹配模型)发送给终端101、102。这样，用户可以应用生成的语义匹配模型进行语句匹配。可将用户输入的问题与数据库中预设的问题进行匹配，查找到语义相同的预设问题，再将预设问题对应的答案反馈给用户。

[0039] 这里的数据库服务器104和服务器105同样可以是硬件，也可以是软件。当它们为硬件时，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。当它们为软件时，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)，也可以实现成单个软件或软件模块。在此不做具体限定。

[0040] 需要说明的是，本申请实施例所提供的用于生成语义匹配模型的方法或用于输出信息的方法一般由服务器105执行。相应地，用于生成语义匹配模型的装置或用于输出信息的装置一般也设置于服务器105中。

[0041] 需要指出的是，在服务器105可以实现数据库服务器104的相关功能的情况下，系统架构100中可以不设置数据库服务器104。

[0042] 应该理解，图1中的终端、网络、数据库服务器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、网络、数据库服务器和服务器。

[0043] 继续参考图2，示出了根据本公开的用于生成语义匹配模型的方法的一个实施例的流程200。该用于生成语义匹配模型的方法，包括以下步骤：

[0044] 步骤201，获取样本集。

[0045] 在本实施例中，用于生成语义匹配模型的方法的执行主体(例如图1所示的服务器105)可以通过多种方式来获取样本集。例如，执行主体可以通过有线连接方式或无线连接方式，从数据库服务器(例如图1所示的数据库服务器104)中获取存储于其中的现有的样本集。再例如，用户可以通过终端(例如图1所示的终端101、102)来收集样本。这样，执行主体可以接收终端所收集的样本，并将这些样本存储在本地，从而生成样本集。

[0046] 在这里，样本集中可以包括至少一个样本。其中，样本可以包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句。

[0047] 在本实施例的一些可选的实现方式中，样本可以包括多个与第一语句语义不匹配的第三语句。例如，使用5个与第一语句语义不匹配的第三语句。

[0048] 在本实施例的一些可选的实现方式中，第一语句为用户搜索的语句、第二语句为用户点击查看的语句、第三语句为用户未点击查看的语句。搜集海量的搜索用户行为数据，对数据进行清洗和过滤，产出基本的训练数据，作为句子语义模型的数据支撑。例如，将用户搜索的语句(query)作为第一语句，将用户点击查看的语句(title)作为第二语句。随机负采样的方法，模拟真实的点击分布得到第三语句。

[0049] 步骤202，从样本集中选取样本。

[0050] 在本实施例中，执行主体可以从步骤201中获取的样本集中选取样本，以及执行步骤203至步骤208的训练步骤。其中，样本的选取方式和选取数量在本申请中并不限制。例如可以是随机选取至少一个样本，也可以是从中选取语句长度较长(例如，超过20个字)的样本。

[0051] 步骤203，将选取的样本的第一语句、第二语句、第三语句分别进行分词，得到第一词序列、第二词序列和第三词序列。

[0052] 在本实施例中，分词又称切词，指的是将一个汉字序列切分成一个一个单独的词。中文分词是文本挖掘的基础，对于输入的一段中文，成功的进行中文分词，可以达到电脑自动识别语句含义的效果。现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。可基于有现的分词算法将三个语句分别分词得到三种词序列。

[0053] 步骤204，将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值。

[0054] 在本实施例中，孪生神经网络是一类包含两个或更多个相同子网络的神经网络架构。这里相同是指它们具有相同的配置即具有相同的参数和权重。参数更新在两个子网上共同进行。本公开的孪生神经网络包括两种类型的子网络，每个子网络可包括投影层、余弦相似度函数和二分类的分类器(例如，sigmoid分类器等常规二分类的分类器)。投影层用于将一个语句的词向量序列转换成语句向量。现有技术中有多种语句向量的转换方法，例如将词向量序列的平均值作为语句向量。第一种类型的子网络(简称“第一子网络”)将语义匹配的第一词序列和第二词序列作为输入，输出为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值。假设第一语句向量为Xq，第二语句向量为Xd。第一语句向量和第二语句向量的余弦相似度S＝cos(Xq，Xd)。第一子网络的输出为L1＝sigmoid(S)。第二种类型的子网络(简称“第二子网络”)将语义不匹配的第一词序列和第三词序列作为输入，输出为1减去第一语句向量和第三语句向量的余弦相似度的sigmoid函数值。假设第一语句向量为Xq，第三语句向量为Xd1(可有多个第二子网络，第三语句向量可为Xd1…Xdn)。第一语句向量和第三语句向量的余弦相似度S1＝cos(Xq，Xd1)。第二子网络的输出为L2＝1-sigmoid(S1)。可有多个第二子网络，则其它第二子网络的输出为L3＝1-sigmoid(S3)..Ln＝1-sigmoid(Sn)。孪生神经网络的输出为各子网络的输出结果之积即输出L＝L1*L2*L3…*Ln。将通过模型输出的L与理论计算的结果(例如1)相对比，确定出损失值。损失函数可采用二分类交叉熵，或者.Contrastive Loss(对比损失)等常见损失函数。

[0055] 在本实施例的一些可选地实现方式中，将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值，包括：

[0056] 步骤2041，将第一词序列、第二词序列和第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列。

[0057] 在本实施例中，对于词序列中的每个词，可通过词袋模型将该词转换成词向量，那么一个词序列可转换成多个词向量，这些词向量就组成了词向量序列。第一词向量序列、第二词向量序列和第三词向量序列的长度是相等的。词向量是可根据权重调整的。通过调整词向量的权重使得语句向量变化，从而使得孪生神经网络的损失值趋近目标值。

[0058] 步骤2042，将第一词向量序列、第二词向量序列和第三词向量序列分别转换成第

一语句向量、第二语句向量和第三语句向量。

[0059] 在本实施例中，可将每个词向量序列中的多个词向量进行平均得到语句向量。也可通过累加等其它方式得到语句向量。

[0060] 步骤2043，确定第一语句向量与第二语句向量之间的第一余弦相似度和第一语句向量与第三语句向量之间的第二余弦相似度。

[0061] 在本实施例中，可由孪生神经网络的第一子网络通过出第一语句向量与第二语句向量之间的第一余弦相似度。由孪生神经网络的第二子网络通过出第一语句向量与第三语句向量之间的第二余弦相似度。

[0062] 步骤2044，根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

[0063] 在本实施例中，可将第一余弦相似度的sigmoid函数值、(1-第二余弦相似度的的sigmoid函数值)之积作为孪生神经网络的输出结果。通过模型输出的结果与理论计算的结果(例如1)相对比，确定出损失值。

[0064] 步骤205，根据损失值确定孪生神经网络是否训练完成。

[0065] 在本实施例中，可以将选取的样本的损失值与目标值进行比较。目标值一般可以用于表示预测值(如孪生神经网络各子网络输出结果之积)与真实值(如1)之间的不一致程度的理想情况。也就是说，当损失值达到目标值时，可以认为预测值接近或近似真值。目标值可以根据实际需求来设置。

[0066] 需要说明的是，若步骤202中选取有多个(至少两个)样本，则执行主体可以将每个样本的损失值分别与目标值进行比较。从而可以确定每个样本的损失值是否达到目标值。

[0067] 如果损失值小于目标值，则认为达到预设的训练完成条件。根据步骤205中的比较结果，执行主体可以确定孪生神经网络是否训练完成。作为示例，如果步骤202中选取有多个样本，那么在每个样本的损失值均达到目标值的情况下，执行主体可以确定孪生神经网络训练完成。再例如，执行主体可以统计损失值达到目标值的样本占选取的样本的比例。且在该比例达到预设样本比例(如95％)，可以确定孪生神经网络训练完成。

[0068] 步骤206，响应于确定孪生神经网络训练完成，将孪生神经网络作为语义匹配模型。

[0069] 在本实施例中，若执行主体确定孪生神经网络已训练完成，则可以将训练完成的孪生神经网络作为语义匹配模型。

[0070] 可选地，执行主体可以将生成的生成语义匹配模型存储在本地，也可以将其发送给终端或数据库服务器。

[0071] 步骤207，响应于确定出孪生神经网络未训练完成，调整孪生神经网络中的相关参数，以及从样本集中重新选取样本，使用调整后的孪生神经网络继续执行训练步骤。

[0072] 在本实施例中，若执行主体确定孪生神经网络未训练完成，则可以调整孪生神经网络中的相关参数。例如采用反向传播技术修改孪生神经网络中各层中的权重，例如更新词向量的权重，即调整词向量，从而更新语句向量。以及可以返回步骤202，从样本集中重新选取样本。从而可以继续执行上述训练步骤。训练完成之后，可记录样本语句的语句向量，以便于将语句向量作为基础的特征用于其它自然语言处理任务。

[0073] 需要说明的是，这里的选取方式在本申请中也不限制。例如在样本集中有大量样本的情况下，执行主体可以从中选取未被选取过的样本。

[0074] 进一步参见图3,图3是根据本实施例的用于生成语义匹配模型的方法的一个应用场景的示意图。在图3的应用场景中,用户所使用的终端上可以安装有模型训练类应用。当用户打开该应用,并上传样本集或样本集的存储路径后,对该应用提供后台支持的服务器可以运行用于生成语义匹配模型的方法,包括:

[0075] 首先,可以获取样本集。其中,样本集中的样本可以包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的多个第三语句。之后,可以从样本集中选取样本,以及执行以下训练步骤:将选取的样本的第一语句、第二语句、第三语句分别进行分词得到第一词序列、第二词序列和第三词序列。对于任意一个词序列,将该词序列中每个词转换成词向量,用W1、W2、W3表示。再通过平均的方法将该词序列的词向量转换成语句向量。其中,第一语句转换成第一语句向量Xq,第二语句转换成第二语句向量Xd,第三语句1转换成第三语句向量Xd1,第三语句2转换成语句向量Xd2…。将第一语句向量Xq和第二语句向量Xd输入孪生神经网络的第一子网络,计算第一语句向量Xq和第二语句向量Xd的余弦相似度S=cos(Xq,Xd),然再求出sigmoid(S)。将第一语句向量Xq和第三语句向量Xd1输入孪生神经网络的第二子网络,计算第一语句向量Xq和第三语句向量Xd1的余弦相似度S1=cos(Xq,Xd1),然再求出1-sigmoid(S1)。将第一语句向量Xq和第三语句向量Xd2输入孪生神经网络的第二子网络,计算第一语句向量Xq和第三语句向量Xd2的余弦相似度S2=cos(Xq,Xd2),然再求出1-sigmoid(S2)…。最后计算出孪生神经网络的输出L=sigmoid(S)*(1-sigmoid(S1))*(1-sigmoid(S2))*…。比较L的值与期望值(例如1)的差距,如果差距大于目标值,则采用梯度下降算法反向传播调整各词向量。再继续选取样本执行上述训练过程,直到L与期望值的差距小于目标值,则将孪生神经网络作为语义匹配模型。

[0076] 本公开的上述实施例提供的方法,能够提高通用句子语义相关性计算的准确率。并为其他自然语言处理任务提供句子表示,作为基础的特征。

[0077] 继续参见图4,作为对上述各图所示方法的实现,本申请提供了一种用于生成语义匹配模型的装置的一个实施例。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。

[0078] 如图4所示,本实施例的用于生成语义匹配模型的装置400可以包括:获取单元401,被配置成获取样本集,其中,样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句;训练单元402,被配置成从样本集中选取样本,以及执行以下训练步骤:将选取的样本的第一语句、第二语句、第三语句分别进行分词,得到第一词序列、第二词序列和第三词序列;将第一词序列、第二词序列和第三词序列输入到孪生神经网络,得到损失值;根据损失值确定孪生神经网络是否训练完成;响应于确定孪生神经网络训练完成,将孪生神经网络作为语义匹配模型。

[0079] 在本实施例的一些可选的实现方式中,训练单元402进一步被配置成:将第一词序列、第二词序列和第三词序列分别转换成第一词向量序列、第二词向量序列和第三词向量序列;将第一词向量序列、第二词向量序列和第三词向量序列分别转换成第一语句向量、第二语句向量和第三语句向量;确定第一语句向量与第二语句向量之间的第一余弦相似度和第一语句向量与第三语句向量之间的第二余弦相似度;根据第一余弦相似度、第二余弦相似度和损失函数确定损失值。

[0080] 在本实施例的一些可选的实现方式中,第一语句为用户搜索的语句、第二语句为

用户点击查看的语句、第三语句为用户未点击查看的语句。

[0081] 在本实施例的一些可选的实现方式中，一个样本中第三语句的数量大于等于1。

[0082] 在本实施例的一些可选的实现方式中，孪生神经网络包括第一子网络和至少一个第二子网络，第一子网络的输出结果为第一语句向量和第二语句向量的余弦相似度的sigmoid函数值，第二子网络的输出结果为1减去第一语句和第三语句的余弦相似度的sigmoid函数值。

[0083] 在本实施例的一些可选的实现方式中，装置400还包括调整单元403，被配置成：响应于确定出孪生神经网络未训练完成，调整孪生神经网络中的相关参数，以及从样本集中重新选取样本，使用调整后的孪生神经网络继续执行训练步骤。

[0084] 可以理解的是，该装置400中记载的诸单元与参考图2描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作、特征以及产生的有益效果同样适用于装置400及其中包含的单元，在此不再赘述。

[0085] 请参见图5，其示出了本申请提供的用于输出信息的方法的一个实施例的流程500。该用于输出信息的方法可以包括以下步骤：

[0086] 步骤501，接收待匹配的第一目标句和第二目标句。

[0087] 在本实施例中，用于输出信息的方法的执行主体(例如图1所示的服务器105)可以通过多种方式来获取待匹配的第一目标句和第二目标句。例如，执行主体可以通过有线连接方式或无线连接方式，从数据库服务器(例如图1所示的数据库服务器104)中获取存储于其中的问答信息集合，每条问答信息包括问题和答案，并将问题集合确定为第二目标句集合。再例如，执行主体也可以接收终端(例如图1所示的终端101、102)或其他设备采集的用户想要询问的问题作为第一目标句。将第一目标句依次与第二目标句集合中的第二目标句进行匹配，如果找到语义匹配的第二目标句，则将第二目标句对应的答案确定为第一目标句的答案。

[0088] 步骤502，将第一目标语句和第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列；

[0089] 在本实施例中，先将每个目标语进行切词生成词序列。可采用常见的自然语言切词方法，例如最大逆向匹配法等。

[0090] 步骤503，将第一目标词序列、第二目标词序列输入语义匹配模型中，输出第一目标句和第二目标句的语义匹配结果。

[0091] 在本实施例中，执行主体可以将步骤502生成的第一目标词序列、第二目标词序列输入语义匹配模型中，从而生成第一目标句和第二目标句的语义匹配结果。语义匹配结果可以是用于描述第一目标句和第二目标句是否语义匹配的信息，例如0表示不匹配或1表示匹配。语义匹配结果还可以是两个目标句的余弦相似度。可将任意两个语句输入第一子网络来判断是否语义匹配。

[0092] 在本实施例中，语义匹配模型可以是采用如上述图2实施例所描述的方法而生成的。具体生成过程可以参见图2实施例的相关描述，在此不再赘述。

[0093] 需要说明的是，本实施例用于输出信息的方法可以用于测试上述各实施例所生成的语义匹配模型。进而根据测试结果可以不断地优化语义匹配模型。该方法也可以是上述各实施例所生成的语义匹配模型的实际应用方法。采用上述各实施例所生成的语义匹配模

型，来进行自动问答，有助于提高自动问答的性能。如找到的答案较多，找到的答案比较准确等。

[0094] 继续参见图6，作为对上述图5所示方法的实现，本申请提供了一种用于输出信息的装置的一个实施例。该装置实施例与图5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。

[0095] 如图6所示，本实施例的用于输出信息的装置600可以包括：接收单元601，被配置成接收待匹配的第一目标语句和第二目标语句；转换单元602，被配置成将第一目标语句和第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列；输出单元603，被配置成将将第一目标词序列、第二目标词序列输入如图5的方法生成的语义匹配模型中，生成第一目标语句和第二目标语句的语义匹配结果。

[0096] 可以理解的是，该装置600中记载的诸单元与参考图5描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作、特征以及产生的有益效果同样适用于装置600及其中包含的单元，在此不再赘述。

[0097] 下面参考图7，其示出了适于用来实现本公开的实施例的电子设备(例如图1中的服务器)700的结构示意图。图7示出的服务器仅仅是一个示例，不应对本公开的实施例的功能和使用范围带来任何限制。

[0098] 如图7所示，电子设备700可以包括处理装置(例如中央处理器、图形处理器等)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装置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RAM 703中，还存储有电子设备700操作所需的各种程序和数据。处理装置701、ROM 702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。

[0099] 通常，以下装置可以连接至I/O接口705：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置706；包括例如液晶显示器(LCD)、扬声器、振动器等的输出装置707；包括例如磁带、硬盘等的存储装置708；以及通信装置709。通信装置709可以允许电子设备700与其他设备进行无线或有线通信以交换数据。虽然图7示出了具有各种装置的电子设备700，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图7中示出的每个方框可以代表一个装置，也可以根据需要代表多个装置。

[0100] 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装置709从网络上被下载和安装，或者从存储装置708被安装，或者从ROM 702被安装。在该计算机程序被处理装置701执行时，执行本公开的实施例的方法中限定的上述功能。需要说明的是，本公开的实施例所述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。在本公开的实施例中，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。而在本公开的实施例中，计算机可读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，该计算机可读信号介质可以发送、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：电线、光缆、RF(射频)等等，或者上述的任意合适的组合。

[0101] 上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。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电子设备执行时，使得该电子设备：获取样本集，其中，样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句；从样本集中选取样本，以及执行以下训练步骤：获取样本集，其中，样本集中的样本包括第一语句、与第一语句语义匹配的第二语句、与第一语句语义不匹配的第三语句；从样本集中选取样本，以及执行以下训练步骤：将选取的样本的第一语句、第二语句、第三语句分别进行分词，得到第一词序列、第二词序列和第三词序列；将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值；根据损失值确定孪生神经网络是否训练完成；响应于确定孪生神经网络训练完成，将孪生神经网络作为语义匹配模型。

[0102] 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公开的实施例的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用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)—连接到用户计算机，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)。

[0103] 附图中的流程图和框图，图示了按照本公开各种实施例的系统、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，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。也要注意的是，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。

[0104] 描述于本公开的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：一种处理器包括获取单元、训练单元。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，获取单元还可以被描述为“获取样本集的单元”。

[0105] 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所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。

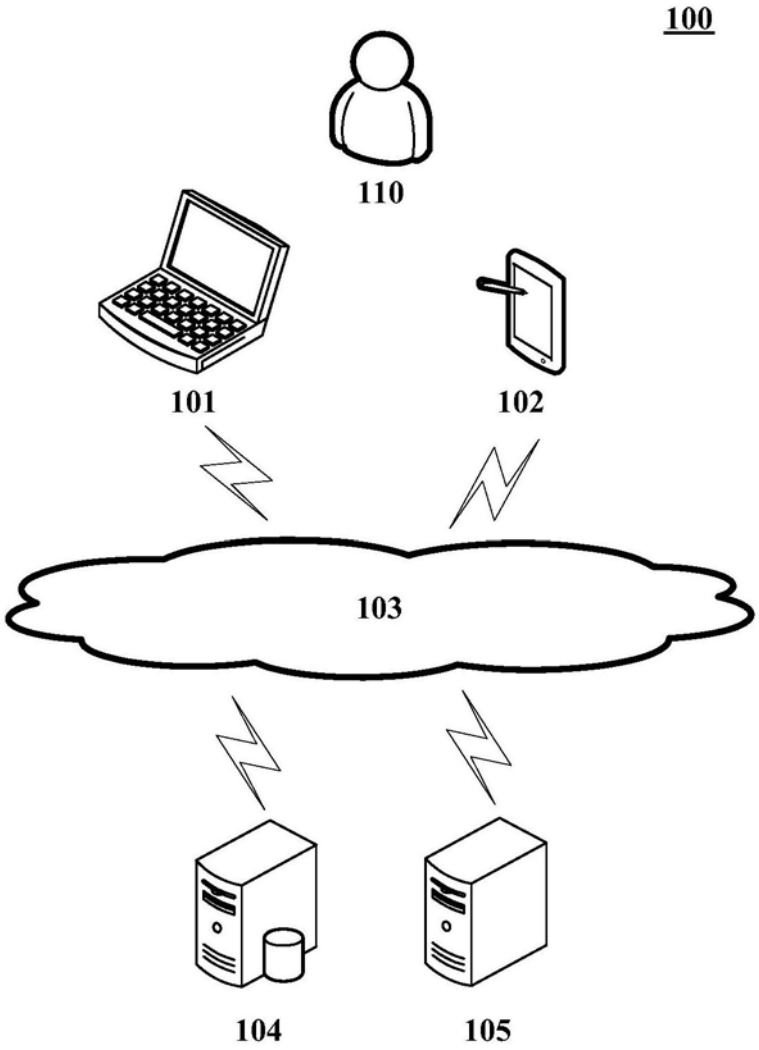



图1

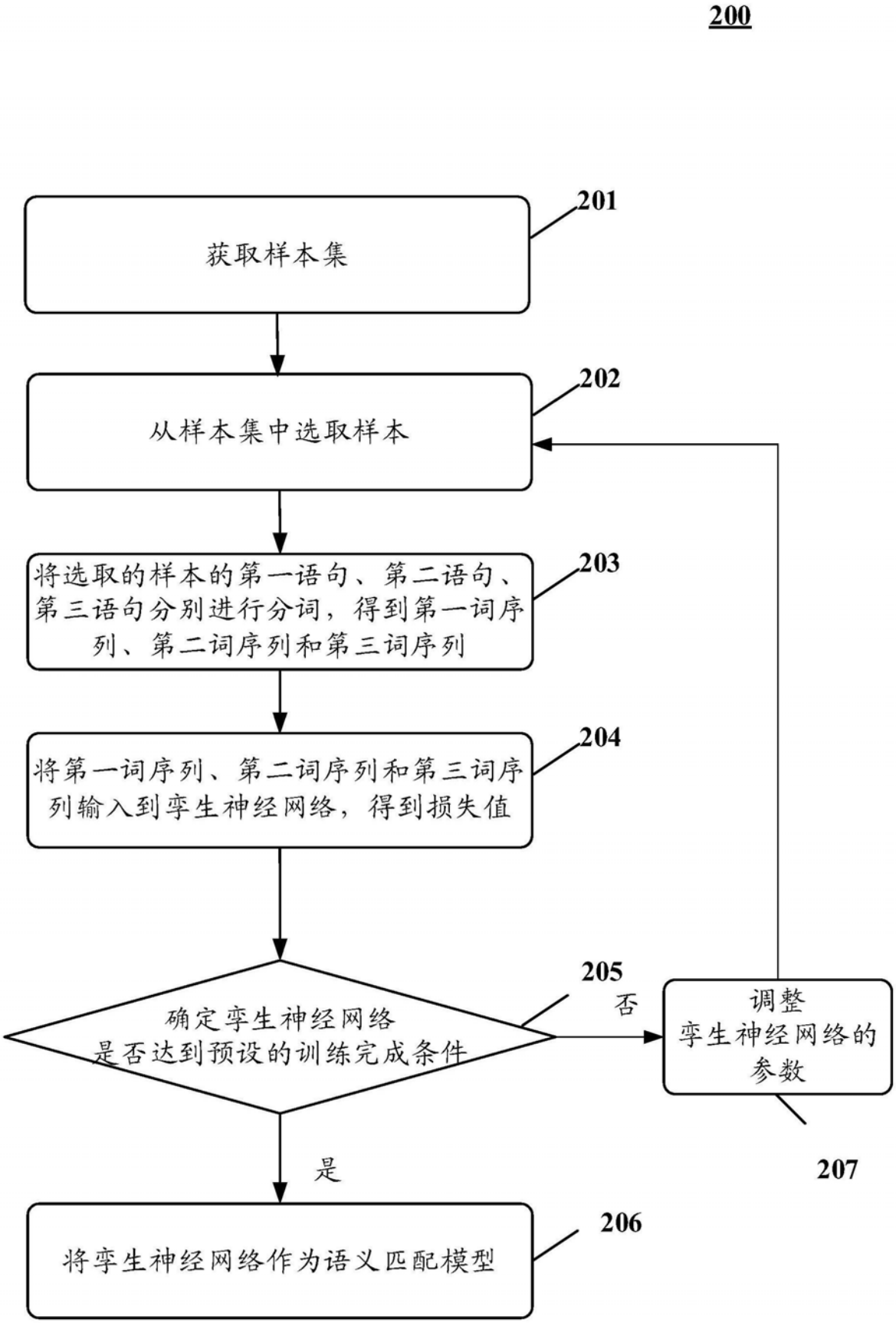

200
201
获取样本集
202
从样本集中选取样本
203
将选取的样本的第一语句、第二语句、第三语句分别进行分词，得到第一词序列、第二词序列和第三词序列
204
将第一词序列、第二词序列和第三词序列输入到孪生神经网络，得到损失值
205
确定孪生神经网络是否达到预设的训练完成条件
否
调整孪生神经网络的参数
207
是
206
将孪生神经网络作为语义匹配模型

图2

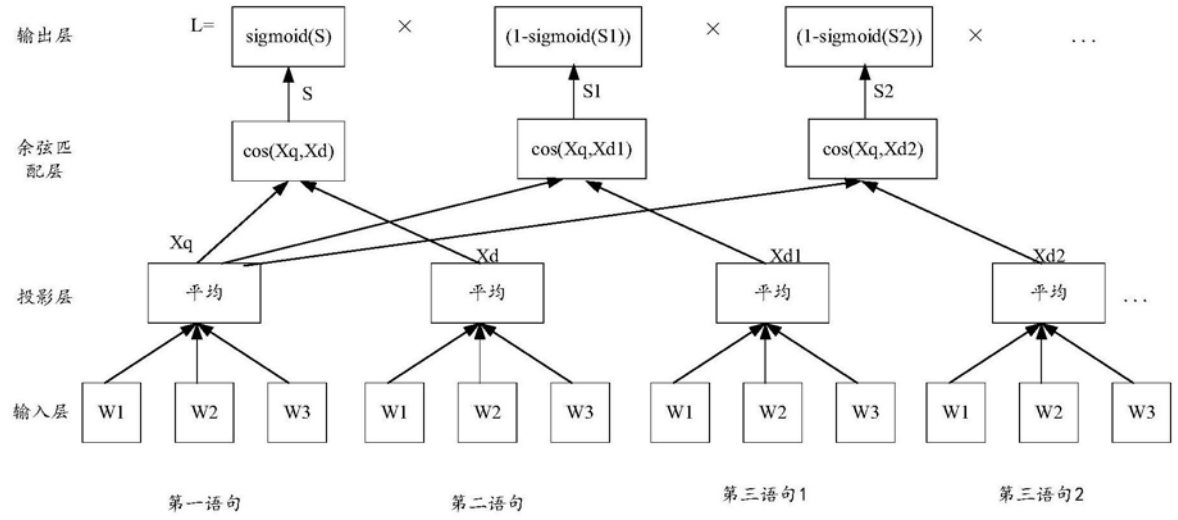

输出层
L=
sigmoid(S)
×
(1-sigmoid(S1))
×
(1-sigmoid(S2))
×
...
S
S1
S2
余弦匹配层
cos(Xq,Xd)
cos(Xq,Xd1)
cos(Xq,Xd2)
Xq
Xd
Xd1
Xd2
投影层
平均
平均
平均
平均
...
输入层
W1
W2
W3
W1
W2
W3
W1
W2
W3
W1
W2
W3
第一语句
第二语句
第三语句1
第三语句2

图3

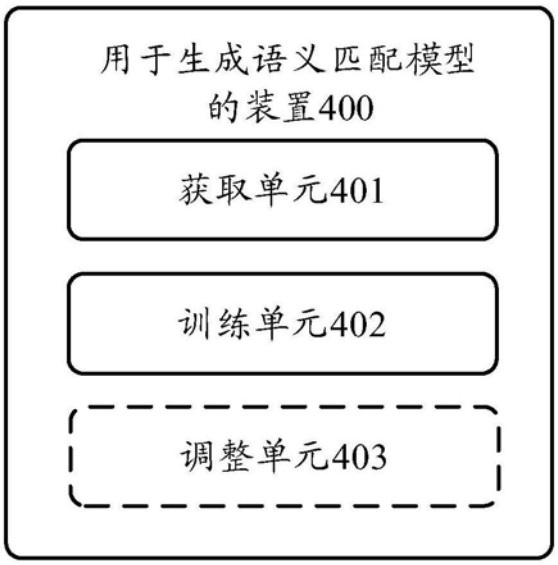

用于生成语义匹配模型的装置400
获取单元401
训练单元402
调整单元403

图4

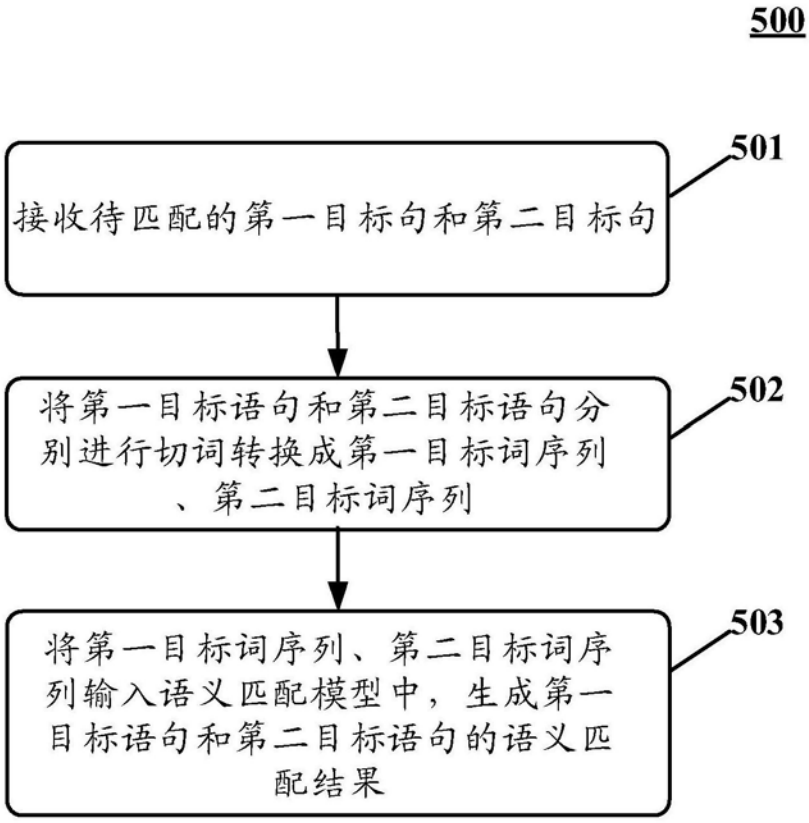

500
501
接收待匹配的第一目标句和第二目标句
502
将第一目标语句和第二目标语句分别进行切词转换成第一目标词序列、第二目标词序列
503
将第一目标词序列、第二目标词序列输入语义匹配模型中，生成第一目标语句和第二目标语句的语义匹配结果

图5


用于输出信息的装置600
接收单元601
转换单元602
输出单元603

图6

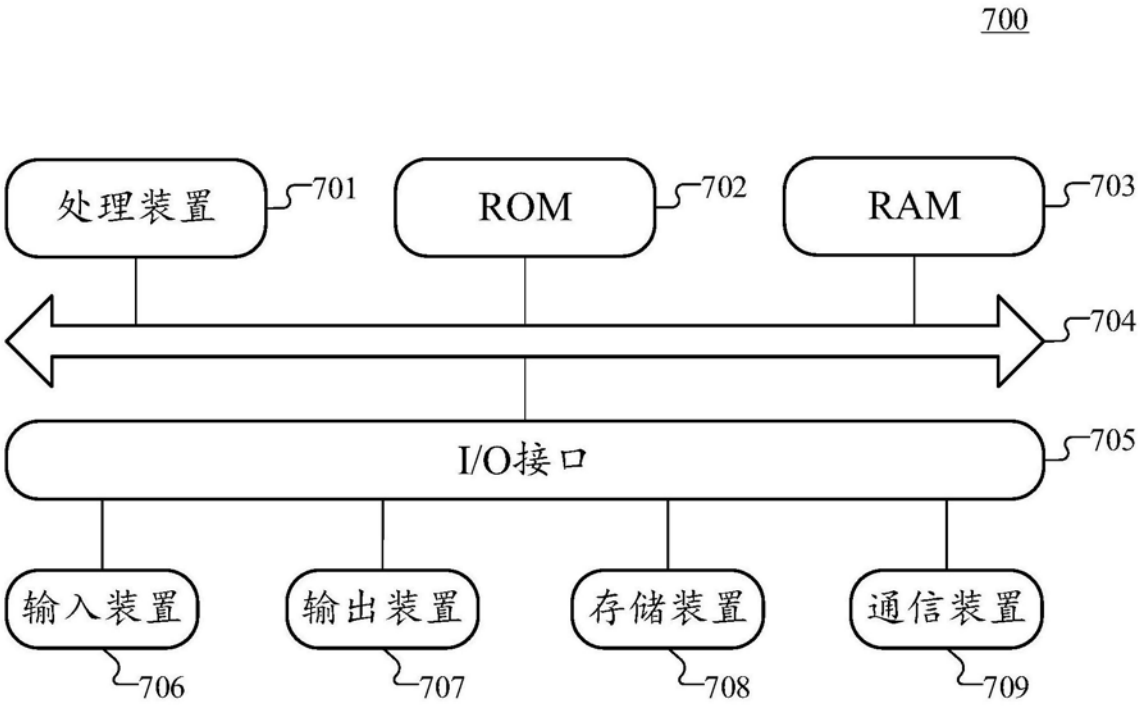

700
处理装置
701
ROM
702
RAM
703
704
I/O接口
705
输入装置
706
输出装置
707
存储装置
708
通信装置
709

图7