(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116894091 A
(43)申请公布日 2023.10.17

(21)申请号 202310911838.5

(22)申请日 2023.07.24

(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 王培英 段超群 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
专利代理师 张效荣 张一军

(51)Int.Cl.
*G06F 16/35*(2019.01)
*G06F 40/216*(2020.01)
*G06F 40/284*(2020.01)
*G06F 40/289*(2020.01)
*G06F 18/2415*(2023.01)
*G06N 3/045*(2023.01)
*G06N 3/0464*(2023.01)
*G06N 3/08*(2023.01)

权利要求书2页 说明书16页 附图9页

(54)发明名称

一种文本顺滑检测的方法和装置

(57)摘要

本发明公开了一种文本顺滑检测的方法和装置，涉及计算机技术领域。文本顺滑检测的方法的一具体实施方式包括：响应于接收到目标文本，根据预先设置的预训练模型，对所述目标文本进行编码，生成对应的隐向量；对所述隐向量进行多尺度特征提取，得到特征向量；对所述隐向量和所述特征向量进行组合，得到组合向量，根据预先设置的激活函数对所述组合向量进行分类，生成所述目标文本对应的分类标签。该实施方式能够捕获不同长度的词语之间的相关性，提高冗余口语词的检出率，使文本更顺滑、便于理解。

CN 116894091 A

1.一种文本顺滑检测的方法,其特征在于,包括:

响应于接收到目标文本,根据预先设置的预训练模型,对所述目标文本进行编码,生成对应的隐向量;

对所述隐向量进行多尺度特征提取,得到特征向量;

对所述隐向量和所述特征向量进行组合,得到组合向量,根据预先设置的激活函数对所述组合向量进行分类,生成所述目标文本对应的分类标签。

2.根据权利要求1所述的方法,其特征在于,对所述隐向量进行多尺度特征提取,得到特征向量,包括:

使用不同窗口尺寸的卷积核对所述隐向量进行卷积运算,得到多个短语向量表征;

根据所述隐向量与所述多个短语向量表征之间的相关性,生成多个相关性矩阵;

对所述多个相关性矩阵进行池化操作,得到多个池化向量;

对所述多个池化向量进行拼接,得到所述特征向量。

3.根据权利要求2所述的方法,其特征在于,根据所述隐向量与所述多个短语向量表征之间的相关性,生成多个相关性矩阵,包括:

根据每个短语向量表征与所述隐向量之间的相关性,生成对应的第一相关性矩阵;

根据不同短语向量表征之间的相关性,生成对应的第二相关性矩阵。

4.根据权利要求3所述的方法,其特征在于,对所述多个池化向量进行拼接,包括:

根据所述卷积核的窗口尺寸确定所述多个池化向量的拼接顺序;

根据所述拼接顺序对所述多个池化向量进行拼接。

5.根据权利要求1所述的方法,其特征在于,在接收到目标文本之前,所述方法还包括:

基于预先设置的训练文本集对预先设置的神经网络进行迭代训练,得到所述训练文本集对应的预测结果;

根据所述预测结果和预先设置的真实顺滑文本确定对应的分类损失;

根据所述分类损失对所述神经网络包括的参数进行优化。

6.根据权利要求5所述的方法,其特征在于,所述方法还包括:

确定所述训练文本集包括的训练文本;

识别所述真实顺滑文本包括的关键词,将所述关键词从所述真实顺滑文本中去除,得到后处理文本;

根据所述训练文本、所述真实顺滑文本和所述后处理文本之间的相似性,确定所述推理结果对应的对比损失;

根据所述分类损失和所述对比损失,对所述神经网络进行训练。

7.根据权利要求6所述的方法,其特征在于,将所述关键词从所述真实顺滑文本中去除,包括:

根据所述关键词的词频和所述关键词的逆文档频率,对所述关键词进行排序,得到关键词排序结果;

根据预先设置的关键词个数和所述关键词排序结果确定目标关键词;

将所述目标关键词从所述真实顺滑文本中去除。

8.一种文本顺滑检测的装置,其特征在于,包括:

编码模块,用于响应于接收到目标文本,根据预先设置的预训练模型,对所述目标文本

进行编码,生成对应的隐向量;

提取模块,用于对所述隐向量进行多尺度特征提取,得到特征向量;

分类模块,用于对所述隐向量和所述特征向量进行组合,得到组合向量,根据预先设置的激活函数对所述组合向量进行分类,生成所述目标文本对应的分类标签。

9.一种电子设备,其特征在于,包括:

一个或者多个处理器;

存储装置,用于存储一个或者多个程序,

当所述一个或者多个程序被所述一个或者多个处理器执行时,所述一个或者多个处理器实现如权利要求1-7中任一所述的方法。

10.一种计算机可读介质,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1-7中任一所述的方法。

# 一种文本顺滑检测的方法和装置

## 技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种文本顺滑检测的方法和装置。

## 背景技术

[0002] 语音对话系统能够接收用户的语音，将语音转换为文本并进行语义理解。但是在对话过程中，口语化的表达通常存在词语冗余、语气词较多等问题，因此需要对转换得到的文本进行检测，使文本更加顺滑，便于系统理解。目前，在进行文本顺滑检测时，通常采用的方法是单独预测各个字符的标签，根据每个字符与其他字符的相似性、每个字符与该字符周围的字符是否重复来识别冗余词。

[0003] 在实现本发明的过程中，发明人发现现有技术至少存在如下问题：

[0004] 仅对单个字符进行标签预测难以捕获不同长度的词语之间的相似性，使得文本顺滑检测准确率不高，文本顺滑效果较差。

## 发明内容

[0005] 有鉴于此，本发明实施例提供一种文本顺滑检测的方法和装置，能够捕获不同长度的词语之间的相关性，提高冗余口语词的检出率。

[0006] 为实现上述目的，根据本发明实施例的第一方面，提供一种文本顺滑检测的方法，包括：

[0007] 响应于接收到目标文本，根据预先设置的预训练模型，对所述目标文本进行编码，生成对应的隐向量；

[0008] 对所述隐向量进行多尺度特征提取，得到特征向量；

[0009] 对所述隐向量和所述特征向量进行组合，得到组合向量，根据预先设置的激活函数对所述组合向量进行分类，生成所述目标文本对应的分类标签。

[0010] 可选地，对所述隐向量进行多尺度特征提取，得到特征向量，包括：

[0011] 使用不同窗口尺寸的卷积核对所述隐向量进行卷积运算，得到多个短语向量表征；

[0012] 根据所述隐向量与所述多个短语向量表征之间的相关性，生成多个相关性矩阵；

[0013] 对所述多个相关性矩阵进行池化操作，得到多个池化向量；

[0014] 对所述多个池化向量进行拼接，得到所述特征向量。

[0015] 可选地，根据所述隐向量与所述多个短语向量表征之间的相关性，生成多个相关性矩阵，包括：

[0016] 根据每个短语向量表征与所述隐向量之间的相关性，生成对应的第一相关性矩阵；

[0017] 根据不同短语向量表征之间的相关性，生成对应的第二相关性矩阵。

[0018] 可选地，对所述多个池化向量进行拼接，包括：

[0019] 根据所述卷积核的窗口尺寸确定所述多个池化向量的拼接顺序；

[0020] 根据所述拼接顺序对所述多个池化向量进行拼接。

[0021] 可选地，在接收到目标文本之前，所述方法还包括：

[0022] 基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到所述训练文本集对应的预测结果；

[0023] 根据所述预测结果和预先设置的真实顺滑文本确定对应的分类损失；

[0024] 根据所述分类损失对所述神经网络包括的参数进行优化。

[0025] 可选地，所述方法还包括：

[0026] 确定所述训练文本集包括的训练文本；

[0027] 识别所述真实顺滑文本包括的关键词，将所述关键词从所述真实顺滑文本中去除，得到后处理文本；

[0028] 根据所述训练文本、所述真实顺滑文本和所述后处理文本之间的相似性，确定所述推理结果对应的对比损失；

[0029] 根据所述分类损失和所述对比损失，对所述神经网络进行训练。

[0030] 可选地，将所述关键词从所述真实顺滑文本中去除，包括：

[0031] 根据所述关键词的词频和所述关键词的逆文档频率，对所述关键词进行排序，得到关键词排序结果；

[0032] 根据预先设置的关键词个数和所述关键词排序结果确定目标关键词；

[0033] 将所述目标关键词从所述真实顺滑文本中去除。

[0034] 根据本发明实施例的第二方面，提供一种文本顺滑检测的装置，包括：

[0035] 编码模块，用于响应于接收到目标文本，根据预先设置的预训练模型，对所述目标文本进行编码，生成对应的隐向量；

[0036] 提取模块，用于对所述隐向量进行多尺度特征提取，得到特征向量；

[0037] 分类模块，用于对所述隐向量和所述特征向量进行组合，得到组合向量，根据预先设置的激活函数对所述组合向量进行分类，生成所述目标文本对应的分类标签。

[0038] 可选地，对所述隐向量进行多尺度特征提取，得到特征向量，包括：

[0039] 使用不同窗口尺寸的卷积核对所述隐向量进行卷积运算，得到多个短语向量表征；

[0040] 根据所述隐向量与所述多个短语向量表征之间的相关性，生成多个相关性矩阵；

[0041] 对所述多个相关性矩阵进行池化操作，得到多个池化向量；

[0042] 对所述多个池化向量进行拼接，得到所述特征向量。

[0043] 可选地，根据所述隐向量与所述多个短语向量表征之间的相关性，生成多个相关性矩阵，包括：

[0044] 根据每个短语向量表征与所述隐向量之间的相关性，生成对应的第一相关性矩阵；

[0045] 根据不同短语向量表征之间的相关性，生成对应的第二相关性矩阵。

[0046] 可选地，对所述多个池化向量进行拼接，包括：

[0047] 根据所述卷积核的窗口尺寸确定所述多个池化向量的拼接顺序；

[0048] 根据所述拼接顺序对所述多个池化向量进行拼接。

[0049] 可选地，所述装置还包括：

[0050] 第一训练模块，用于基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到所述训练文本集对应的预测结果；

[0051] 第一确定模块，用于根据所述预测结果和预先设置的真实顺滑文本确定对应的分类损失；

[0052] 优化模块，用于根据所述分类损失对所述神经网络包括的参数进行优化。

[0053] 可选地，所述装置还包括：

[0054] 第二确定模块，用于确定所述训练文本集包括的训练文本；

[0055] 去除模块，用于识别所述真实顺滑文本包括的关键词，将所述关键词从所述真实顺滑文本中去除，得到后处理文本；

[0056] 对比模块，用于根据所述训练文本、所述真实顺滑文本和所述后处理文本之间的相似性，确定所述推理结果对应的对比损失；

[0057] 第二训练模块，用于根据所述分类损失和所述对比损失，对所述神经网络进行训练。

[0058] 可选地，将所述关键词从所述真实顺滑文本中去除，包括：

[0059] 根据所述关键词的词频和所述关键词的逆文档频率，对所述关键词进行排序，得到关键词排序结果；

[0060] 根据预先设置的关键词个数和所述关键词排序结果确定目标关键词；

[0061] 将所述目标关键词从所述真实顺滑文本中去除。

[0062] 根据本发明实施例的第三方面，提供一种电子设备，包括：

[0063] 一个或多个处理器；

[0064] 存储装置，用于存储一个或多个程序，

[0065] 当所述一个或多个程序被所述一个或多个处理器执行时，所述一个或多个处理器实现上述任一实施例所述的方法。

[0066] 根据本发明实施例的第四方面，提供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上述任一实施例所述的方法。

[0067] 上述发明中的一个实施例具有如下优点或有益效果：对目标文本进行预训练，得到对应的隐向量，对隐向量进行多尺度特征提取，能够准确地获取到不同长度的词语之间的相关性，便于根据多种相关性信息进行文本顺滑检测；多尺度特征提取包括：基于不同尺寸卷积核的卷积运算、根据相关性生成相关性矩阵、最大池化操作、向量拼接等步骤，使得得到的特征向量具有良好的特征表示；根据短语向量表征与隐向量之间的相关性生成第一相关性矩阵，根据不同短语向量表征之间的相关性，生成第二相关矩阵，能够进一步获取不同长度的词语之间的多种相关性；根据卷积核的窗口尺寸确定池化向量的拼接顺序，根据拼接顺序对池化向量进行拼接，能够提高池化向量拼接的灵活性；根据分类损失对神经网络包括的参数进行优化，能够提高神经网络的优化效果和优化效率；根据多种文本之间的相似性确定推理结果对应的对比损失，能够引入对目标文本的语义一致性约束，能够提高神经网络对关键词的识别能力，避免将关键词从目标文本中去除；使用词频和逆文档频率获取关键词，能够提高关键词的获取效率，提高多种后处理文本的生成效率和对比损失的计算效率。

[0068] 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

[0069] 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：

[0070] 图1是根据本发明实施例的文本顺滑检测的方法的主要流程的示意图；

[0071] 图2是根据本发明一个可参考实施例的文本顺滑检测的方法的整体流程的示意图；

[0072] 图3是根据本发明一个可参考实施例的文本顺滑检测的方法的多尺度特征提取的示意图；

[0073] 图4是根据本发明一个可参考实施例的文本顺滑检测的方法的神经网络模型框架的示意图；

[0074] 图5是根据本发明一个可参考实施例的文本顺滑检测的方法的主要流程的示意图；

[0075] 图6是根据本发明另一个可参考实施例的文本顺滑检测的方法的主要流程的示意图；

[0076] 图7是根据本发明又一个可参考实施例的文本顺滑检测的方法的主要流程的示意图；

[0077] 图8是根据本发明实施例的文本顺滑检测的装置的主要模块的示意图；

[0078] 图9是本发明实施例可以应用于其中的示例性系统架构图；

[0079] 图10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

[0080] 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。

[0081] 需要说明的是，本发明的技术方案中，所涉及的用户个人信息的收集、使用、保存、共享和转移等处理，均符合相关法律法规的规定，且需要告知用户并获得用户的同意或授权，当适用时，对用户个人信息进行了去标识化和/或匿名化和/或加密的技术处理。

[0082] 语音对话系统能够接收用户的语音，将语音转换为文本并进行语义理解。但是在对话过程中，口语化的表达通常存在词语冗余、语气词较多等问题，因此需要对转换得到的文本进行检测，使文本更加顺滑，便于系统理解。目前，在进行文本顺滑检测时，通常采用的方法是单独预测各个字符的标签，根据每个字符与其他字符的相似性、每个字符与该字符周围的字符是否重复来识别冗余词，然后对识别出的字符设置对应的标签，根据标签对文本中的冗余成本、语气词成分进行删除，从而实现文本顺滑。

[0083] 单独预测各个字符是否冗余或者是否为语气词，忽略了不同长度的词语之间的相似性，导致对不同长度的词语之间的词语冗余问题不能准确识别；并且难以识别出文本中的关键词，导致在进行文本顺滑时容易将关键词从文本中去除，改变文本的含义，造成后续

的语义理解不准确。

[0084] 有鉴于此，根据本发明实施例的第一方面，提供一种文本顺滑检测的方法。

[0085] 图1是根据本发明实施例的文本顺滑检测的方法的主要流程的示意图。如图1所示，根据本发明实施例的文本顺滑检测的方法主要包括如下的步骤S101至步骤S103。

[0086] 步骤S101，响应于接收到目标文本，根据预先设置的预训练模型，对所述目标文本进行编码，生成对应的隐向量。

[0087] 目标文本是需要进行文本顺滑检测的文本，例如，在智能手机、平板电脑或者计算机中内置的语音对话系统接收到用户发来的语音之后，语音对话系统将用户的语音转换为目标文本并发送至本发明实施例的执行主体，或者本发明实施例的执行主体直接接收用户输入的目标文本。

[0088] 在得到目标文本之后，在目标文本的起始位置添加分类特殊标记，分类特殊标记可以用于后续对目标文本的分类任务，在目标文本的结束位置添加句子分隔特殊标记，句子分隔特殊标记用于对目标文本进行分隔，在目标文本包括多个文本段的情况下，在每个文本段的结束位置添加句子分隔特殊标记；例如，分类特殊标记为“[CLS]”，句子分隔特殊标记为“[SEP]”，在目标文本包括：A1、A2、A3等3个文本段，对目标文本添加标记后得到“[CLS]A1[SEP]A2[SEP]A3[SEP]”。

[0089] 预先设置预训练模型，预训练模型是一种编码器，预训练模型包括多层网络，将每层网络作为一个子编码器，每层网络包括自注意力模块和前馈网络模块，其中，自注意力模块对每个字符进行编码，前馈网络模块通过以全连接层的形式对目标文本进行非线性变换，具体地，预训练模型可以是BERT模型。将添加标记之后的目标文本输入预先设置的预训练模型中，是预训练模型对目标文本进行编码，得到目标文本包括的各个字符对应的隐向量。隐向量是目标文本包括的每个字符的向量表示，隐向量的维度与目标文本的维度相同。

[0090] 图2是根据本发明一个可参考实施例的文本顺滑检测的方法的整体流程的示意图。示例性地，如图2所示，本发明实施例的执行主体包括预训练模型、多尺度特征提取、全连接层、激活函数等多个模块，在得到目标文本之后，将目标文本输入预训练模型，使预训练模型对目标文本进行编码，得到隐向量。

[0091] 再示例性地，目标文本为“$\{x_0, x_1, x_2, \ldots, x_s\}$”，其中，“$x_0$”表示分类特殊标记，“$x_s$”表示句子分隔特殊标记，将目标文本输入BERT模型之后，得到隐向量为“$\{h_0, h_1, h_2, \ldots, h_s\}$”，其中，“$h_1$”是“$x_1$”的向量表示、“$h_2$”是“$x_2$”的向量表示、“$h_s$”是“$x_s$”的向量表示。

[0092] 将目标文本输入预处理模型能够得到目标文本的向量表示，便于之后对隐向量进行多尺度特征提取。

[0093] 步骤S102，对所述隐向量进行多尺度特征提取，得到特征向量。

[0094] 在得到隐向量之后，根据多种预先设置的卷积核对隐向量进行多次卷积运算，得到不同长度的短语向量表征；然后计算不同长度的短语向量表征之间的相关性，计算不同长度的短语向量表征之间的注意力得分，获取不同长度的短语向量表征之间的相关性矩阵，对得到的相关性矩阵进行降维并保留相关性矩阵中的部分相关性信息，得到降维矩阵，每个相关性矩阵对应一个降维矩阵，对多个降维矩阵进行拼接，得到特征向量。

[0095] 示例性地，预先设置二维卷积核、三维卷积核和四维卷积核，使用上述三种卷积核分别对隐向量进行卷积运算，得到短语向量表征B2、短语向量表征B3、短语向量表征B4等三

个不同长度的短语向量表征，其中，“B2”表示二维卷积核对应的短语向量表征、“B3”表示三维卷积核对应的短语向量表征、“B4”表示四维卷积核对应的短语向量表征，计算多个短语向量表征之间的注意力得分，得到多个相关性矩阵，包括：短语向量表征B2和短语向量表征B3之间的相关性矩阵、短语向量表征B2和短语向量表征B4之间的相关性矩阵、短语向量表征B3和短语向量表征B4之间的相关性矩阵，对每个相关性矩阵进行降维，例如，将每个相关性矩阵划分为多个尺寸相等区域，确定每个区域内的数据对应的中位数或者众数，将多个区域的中位数或者众数按区域位置进行组合，得到降维矩阵，按随机顺序对多个降维矩阵进行拼接，得到特征矩阵。

[0096] 使用不同维度的卷积核对隐向量进行卷积运算，能够得到不同长度的词语对应的向量表示，从而能够计算不同长度的词语之间的相关性，便于对不同长度的词语进行分类、预测；对相关性矩阵进行降维，能够有效控制相关性矩阵的维度，提高特征提取的效率。

[0097] 根据本发明一个可参考的实施例，对所述隐向量进行多尺度特征提取，得到特征向量，包括：使用不同窗口尺寸的卷积核对所述隐向量进行卷积运算，得到多个不同长度的短语向量表征；根据所述隐向量与所述多个不同长度的短语向量表征之间的相关性，生成多个相关性矩阵；对所述多个相关性矩阵进行池化操作，得到多个池化向量；对所述多个池化向量进行拼接，得到所述特征向量。

[0098] 预先设置卷积神经网络，卷积神经网络包括多个不同窗口尺寸的卷积核，使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到具有不同长度、不同特征信息的多个短语向量表征，例如，对隐向量进行卷积运算的公式为：$H_n = Conv_n(H)$，其中，$Conv_n$表示卷积操作，H表示隐向量，$H_n$表示长度为n的短语向量表征，n表示卷积窗口大小(即卷积核的维度)，n的取值可以是2、3或者目标文本的长度值等等。

[0099] 在得到多个短语向量表征之后，计算隐向量与多个短语向量表征之间的注意力得分，得到隐向量与每个短语向量表征之间的相关性矩阵。在得到多个相关性矩阵之后，对每个相关性矩阵进行池化操作，例如平均池化操作，优选地，对每个相关性矩阵进行最大池化操作。在得到多个池化向量之后，根据卷积核的窗口尺寸确定多个池化向量之间的拼接顺序，根据拼接顺序对多个池化向量进行拼接。示例性地，相关性矩阵C1为隐向量与二维卷积核对应的短语向量表征之间相关性矩阵，相关性矩阵C2为隐向量与三维卷积核对应的短语向量表征之间相关性矩阵，相关性矩阵C3为隐向量与四维卷积核对应的短语向量表征之间相关性矩阵，对相关性矩阵进行最大池化操作，得到相关性矩阵C1对应的池化向量D1，相关性矩阵C2对应的池化向量D2，相关性矩阵C3对应的池化向量D3；按照卷积核的维度从小到大的顺序确定拼接顺序，得到“D1,D2,D3”，按上述拼接顺序对池化向量进行拼接得到特征向量“[D1:D2:D3]”。

[0100] 再示例性地，如图2所示，在得到隐向量之后，将隐向量输入多尺度特征提取模块，具体地，基于多个窗口尺寸的卷积核对隐向量进行卷积运算得到多个短语向量表征，然后分别对多个短语向量表征和隐向量进行相关性分析得到多个第一相关性矩阵，对不同短语向量表征进行相关性分析得到多个第二相关性矩阵，对多个第一相关性矩阵和多个第二相关性矩阵进行最大池化操作，得到池化向量，对多个池化向量进行拼接得到特征向量。

[0101] 使用不同长度的短语向量表征对隐向量进行卷积运算，能够得到不同长度的词语对应的向量表示，从而能够计算不同长度的词语之间的相关性，便于对不同长度的词语进

行分类、预测;对相关性矩阵进行池化操作,能够提高特征提取的效率。

[0102] 根据本发明另一个可参考的实施例,根据所述隐向量与所述多个短语向量表征之间的相关性,生成多个相关性矩阵,包括:根据每个短语向量表征与所述隐向量之间的相关性,生成对应的第一相关性矩阵;根据不同短语向量表征之间的相关性,生成对应的第二相关性矩阵。

[0103] 使用预先设置的激活函数计算每个短语向量表征与隐向量之间的注意力得分,得到每个短语向量表征与隐向量之间的第一相关性矩阵,每个短语向量表征与隐向量之间的注意力得分的计算公式为:

[0104] $$M_n^{t=>p} = softmax(\frac{(HW_q)(H_nW_k)^T}{\sqrt{d}}),$$

[0105] 其中,$M_n^{t=>p}$表示长度为n的短语向量表征与隐向量之间的第一相关性矩阵,H表示隐向量,$H_n$表示长度为n的短语向量表征,$W_q$和$W_k$为参数,d表示隐向量的向量维度,例如,在使用BRET预训练模型的情况下,d表示向量维度768,使用的激活函数为softmax。

[0106] 使用预先设置的激活函数计算不同短语向量表征之间的注意力得分,得到不同短语向量表征之间的第二相关性矩阵,不同短语向量表征之间的注意力得分的计算公式为:

[0107] $$M_n^{p=>p} = softmax(\frac{(H_nW_q)(H_nW_k)^T}{\sqrt{d}}),$$

[0108] 其中,$M_n^{p=>p}$表示不同长度的短语向量表征之间的第二相关性矩阵,$H_n$表示长度为n的短语向量表征,$W_q$和$W_k$为参数,d表示隐向量的向量维度,使用的激活函数为softmax。

[0109] 在得到不同短语向量表征与隐向量之间的第一相关性矩阵和不同短语向量表征之间的第二相关性矩阵之后,对第一相关性矩阵和第二相关性矩阵进行最大池化操作,对相关性矩阵进行最大池化操作的计算公式为:

[0110] $$V_n^{t=>p} = \max_pooling(M_n^{t=>p}),$$

[0111] 其中,$V_n^{t=>p}$表示对长度为n的短语向量表征与隐向量之间的第一相关性矩阵进行最大池化操作得到的池化向量,max_pooling表示最大池化操作。对第二相关性矩阵进行最大池化操作的计算公式为:

[0112] $$V_n^{p=>p} = \max_pooling(M_n^{p=>p}),$$

[0113] 其中,$V_n^{p=>p}$表示对不同长度的短语向量表征之间的第二相关性矩阵进行最大池化操作得到的池化向量,max_pooling表示最大池化操作。

[0114] 计算每个短语向量表征与隐向量之间的第一相关性矩阵,以及每不同短语向量表征之间的第二相关性矩阵,能够获取到不同长度的词语之间的相关性,便于提取目标文本包括的多种特征标表示;对第一相关性矩阵和第二相关性矩阵进行最大池化操作,能够从不同视角对目标文本进行特征提取,便于获取目标文本的分类标签。

[0115] 根据本发明又一个可参考的实施例,对所述多个池化向量进行拼接,包括:根据所述卷积核的窗口尺寸确定所述多个池化向量的拼接顺序;根据所述拼接顺序对所述多个池化向量进行拼接。

[0116] 在得到多个第一相关性矩阵对应的多个池化向量和多个第二相关性矩阵对应的

多个池化向量之后，根据池化向量对应的卷积核的维度确定拼接顺序，根据拼接顺序对池化向量进行拼接。示例性地，预先设置二维卷积核、三维卷积核和四维卷积核，得到二维卷积核对应的短语向量表征E1、得到三维卷积核对应的短语向量表征E2、得到四维卷积核对应的短语向量表征E3，按卷积核从小到大的顺序对短语向量表征进行排序，得到“E1、E2、E3”，由短语向量表征E1和隐向量得到第一相关性矩阵E1-1，由短语向量表征E2和隐向量得到第一相关性矩阵E2-2，由短语向量表征E3和隐向量得到第一相关性矩阵E3-3；由短语向量表征E1和短语向量表征E2得到第二相关性矩阵E1-2，由短语向量表征E2和短语向量表征E3得到第二相关性矩阵E2-3，由短语向量表征E1和短语向量表征E3得到第二相关性矩阵E1-3；对第一相关性矩阵E1-1进行最大池化操作得到池化向量F1-1，对第二相关性矩阵E1-2进行最大池化操作得到池化向量F1-2，对第一相关性矩阵E2-2进行最大池化操作得到池化向量F2-2，对第二相关性矩阵E2-3进行最大池化操作得到池化向量F2-3，对第一相关性矩阵E3-3进行最大池化操作得到池化向量F3-3，对第二相关性矩阵E1-3进行最大池化操作得到池化向量F1-3；根据上述排序结果对池化向量进行排序，得到特征向量“[F1-1:F1-2:F2-2:F2-3:F3-3:F1-3]”。

[0117] 图3是根据本发明一个可参考实施例的文本顺滑检测的方法的多尺度特征提取的示意图。示例性地，如图3所示，将隐向量输入多尺度特征提取模块301，多尺度特征提取模块301包括多个窗口尺寸的卷积核，例如，2*d卷积核、3*d卷积核等等，其中d表示向量维度768，根据每个卷积核得到对应的第一相关性矩阵和第二相关性矩阵，对每个第一相关性矩阵和第二相关性矩阵进行池化操作，得到对应的池化向量，将多个池化向量进行拼接，得到特征向量。

[0118] 基于不同尺寸卷积核的卷积运算、根据相关性生成相关性矩阵、最大池化操作、向量拼接等步骤，使得得到的特征向量具有良好的特征表示。

[0119] 步骤S103，对所述隐向量和所述特征向量进行组合，得到组合向量，根据预先设置的激活函数对所述组合向量进行分类，生成所述目标文本对应的分类标签。

[0120] 在得到特征向量之后，将特征向量经过一个全连接层，得到特征向量的全连接层输出向量，特征向量的全连接层输出向量的计算公式为：$L_p=MLP(V)$，其中，$L_p$表示特征向量的全连接层输出向量，V表示特征向量，MLP(V)表示将特征向量输入全连接层。将隐向量也经过一个全连接层，得到隐向量的全连接层输出向量，隐向量的全连接层输出向量的计算公式为：$L_h=MLP(H)$，其中，$L_h$表示隐向量的全连接层输出向量，H表示特征向量，MLP(H)表示将隐向量输入全连接层。

[0121] 对隐向量和特征向量进行组合，例如，将隐向量和特征向量相加，得到组合向量，使用预先设置的激活函数对组合向量进行处理，得到目标序列文本包括的每个字符的分类标签，处理公式为：$P(y_i|x_r)=softmax(L_p+L_h)$，其中，$P(y_i|x_r)$表示第r个字符的分类标签为$y_i$的概率，分类标签包括保留标签和去除标签，每个字符对应一个保留标签的概率和一个去除标签的概率，例如，在保留标签的概率大于去除标签的概率的情况下，判定字符的分类标签为保留标签，否则判定字符的分类标签为去除标签，在根据分类标签对目标文本进行文本顺滑操作时，留下分类标签为保留标签的字符，删除分类标签为去除标签的字符。

[0122] 示例性地，如图2所示，在得到特征向量之后，将特征向量和隐向量分别输入全连接层，再将两个全连接层的输出向量输入至激活函数，获取目标文本包括的各个字符的分

类标签。

[0123] 使用激活函数对组合向量进行处理，得到目标文本包括的每个字符的分类标签，能够提高分类标签的准确率，优化文本顺滑效果。

[0124] 根据本发明一个可参考的实施例，在接收到目标文本之前，所述方法还包括：基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到所述训练文本集对应的预测结果；根据所述预测结果和预先设置的真实顺滑文本确定对应的分类损失；根据所述分类损失对所述神经网络包括的参数进行优化。

[0125] 预先设置的训练文本集是训练文本的集合，使用训练文本集对预先设置的神经网络进行迭代训练，每次训练结果后得到训练文本集对应的预测结果，预测结果是文本包括的每个字符的预测标签的概率。预先设置真实顺滑文本，真实顺滑文本包括训练文本的每个字符对应的真实标签，使用预先设置的损失函数计算预测结果和真实顺滑文本(即Ground Truth)包括的真实标签之间的损失差值，损失差值为神经网络每个时刻的损失差值之和，优选地，使用交叉熵损失函数计算损失差值，将获得的损失差值作为分类损失。

[0126] 图4是根据本发明一个可参考实施例的文本顺滑检测的方法的神经网络模型框架的示意图。示例性地，如图4所示，首先将训练文本输入BERT预训练模型，得到隐向量，将隐向量输入多尺度特征提取模块，得到特征向量(图4中401内的带横线的矩形)，基于特征向量与隐向量(图4中401内的未带横线的矩形)对训练文本向量的每个字符进行标签预测，得到推理结果，其中，标签“0”表示该字符为保留字符，标签“D”表示该字符为去除字符。在真实顺滑文本中确定关键词，将关键词从真实顺滑文本中去除得到后处理文本，分别对真实顺滑文本和后处理文本进行编码得到对应的隐向量，对多个隐向量进行对比学习，从而拉近训练文本和真实顺滑文本之间的距离，拉远训练文本与后处理文本之间的距离，提高标签的准确性、文本顺滑检测的准确性。

[0127] 根据分类损失对神经网络包括的参数进行优化，能够提高神经网络的优化效果和优化效率。

[0128] 根据本发明另一个可参考的实施例，所述方法还包括：确定所述训练文本集包括的训练文本；识别所述真实顺滑文本包括的关键词，将所述关键词从所述真实顺滑文本中去除，得到后处理文本，；根据所述训练文本、所述真实顺滑文本和所述后处理文本之间的相似性，确定所述推理结果对应的对比损失；根据所述分类损失和所述对比损失，对所述神经网络进行训练。

[0129] 将训练文本集中输入神经网络的文本作为训练文本，根据预先设置的关键词模板，确定真实顺滑文本包括的关键词，将关键词从真实顺滑文本中去除，得到后处理文本。

[0130] 对真实顺滑文本、训练文本和后处理文本进行对比学习，计算真实顺滑文本、训练文本和后处理文本之间的对比损失，对比损失的计算公式为：

[0131] $$L_{cl} = -\sum_{x} log \frac{e^{sim(x,x^{+})/\tau}}{e^{sim(x,x^{+})/\tau} + e^{sim(x,x^{-})/\tau}},$$

[0132] 其中，$L_{cl}$表示对比损失，x表示训练文本，$x^{+}$表示真实顺滑文本，$x^{-}$表示后处理文本，sim表示相似度函数，τ为温度系数。通过计算对比损失，能够缩小训练文本与真实顺滑文本之间的距离，扩大训练文本与后处理文本之间的距离。

[0133] 在得到分类损失和对比损失之后，根据上述两种损失计算神经网络的联合损失，联合损失的计算公式为：$Loss=\lambda L_{ce}+(1-\lambda)L_{cl}$，其中，Loss为联合损失，$L_{ce}$表示分类损失，$L_{cl}$表示对比损失，λ为系数，两种损失函数的系数不同。根据联合损失对神经网络进行训练。需要说明的是，经过参数优化的神经网络能够用于目标文本的文本顺滑检测方法，具体地，用于输出目标文本的分类标签。

[0134] 根据多种文本之间的相似性确定推理结果对应的对比损失，将分类损失和对比损失作为总的损失函数，对神经网络进行训练，能够引入对目标文本的语义一致性约束，能够提高神经网络对关键词的识别能力，避免将关键词从目标文本中去除。

[0135] 根据本发明又一个可参考的实施例，将所述关键词从所述真实顺滑文本中去除，包括：根据所述关键词的词频和所述关键词的逆文档频率，对所述关键词进行排序，得到关键词排序结果；根据预先设置的关键词个数和所述关键词排序结果确定目标关键词；将所述目标关键词从所述真实顺滑文本中去除。

[0136] 在确定真实顺滑文本包括的关键词之后，计算每个关键词的词频和逆文档频率，关键词的词频表示关键词在真实顺滑文本中出现的频率，即关键词的词频＝g1/g2，其中，g1表示关键词在真实顺滑文本中出现的次数，g2表示真实顺滑文本的总字数；关键词的逆文档频率表示文本总数与含有关键词的文本的比值，例如，关键词的逆文档频率＝log(g3/(g4+1))，其中，g3表示文本总数，g4表示含义关键词的文本的数量。将每个关键词的词频和逆文档频率相乘，得到排序值，对排序值进行降序排列。获取预先设置的关键词个数，关键词个数表示从真实顺滑文本中删除的关键词的个数，例如，关键词个数为s，则将前s个排序值对应的关键词作为目标关键词，将目标关键词从真实顺滑文本中去除，得到后处理文本。

[0137] 示例性地，关键词J1的词频为0.4，逆文档频率为4，排序值为1.6；关键词J2的词频为0.5，逆文档频率为3，排序值为1.5；关键词J3的词频为0.6，逆文档频率为3，排序值为1.8；预先设置的关键词个数为2，则按排序值从大到小排序，确定需要去除的目标关键词为关键词J3和关键词J1。

[0138] 使用词频和逆文档频率获取关键词，能够提高关键词的获取效率，提高多种文本的生成效率和对比损失的计算效率。

[0139] 图5是根据本发明一个可参考实施例的文本顺滑检测的方法的主要流程的示意图。如图5所示，所述文本顺滑检测的方法可以包括：

[0140] 步骤S501，响应于接收到目标文本，根据预先设置的预训练模型，对目标文本进行编码，生成对应的隐向量；

[0141] 步骤S502，使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到多个短语向量表征；

[0142] 步骤S503，根据所述隐向量与多个短语向量表征之间的相关性，生成多个相关性矩阵；

[0143] 步骤S504，对多个相关性矩阵进行池化操作，得到多个池化向量；

[0144] 步骤S505，对多个池化向量进行拼接，得到特征向量；

[0145] 步骤S506，将特征向量和隐向量分别输入全连接层；

[0146] 步骤S507，根据预先设置的激活函数对全连接层的输出结果进行标签预测，生成目标文本对应的分类标签。

[0147] 上述本发明一个可参考实施例的文本顺滑检测的方法的具体实施内容，在上面所述文本顺滑检测的方法中已经详细说明了，故在此重复内容不再说明。

[0148] 图6是根据本发明另一个可参考实施例的文本顺滑检测的方法的主要流程的示意图。如图6所示，所述文本顺滑检测的方法可以包括：

[0149] 步骤S601，基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到预测结果；

[0150] 步骤S602，根据预测结果确定对应的分类损失；

[0151] 步骤S603，根据分类损失对神经网络包括的参数进行优化；

[0152] 步骤S604，响应于接收到目标文本，根据预先设置的预训练模型，对目标文本进行编码，生成对应的隐向量；

[0153] 步骤S605，使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到多个短语向量表征；

[0154] 步骤S606，根据每个短语向量表征与隐向量之间的相关性，生成每个短语向量表征与所述隐向量之间的第一相关性矩阵；

[0155] 步骤S607，根据不同短语向量表征之间的相关性，生成对应的第二相关性矩阵；

[0156] 步骤S608，对多个相关性矩阵进行池化操作，得到多个池化向量；

[0157] 步骤S609，对多个池化向量进行拼接，得到特征向量；

[0158] 步骤S610，将特征向量和隐向量分别输入全连接层；

[0159] 步骤S611，根据预先设置的激活函数对全连接层的输出结果进行标签预测，生成对应的分类标签。

[0160] 上述本发明另一个可参考实施例的文本顺滑检测的方法的具体实施内容，在上面所述文本顺滑检测的方法中已经详细说明了，故在此重复内容不再说明。

[0161] 图7是根据本发明又一个可参考实施例的文本顺滑检测的方法的主要流程的示意图。如图7所示，所述文本顺滑检测的方法可以包括：

[0162] 步骤S701，基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到预测结果；

[0163] 步骤S702，根据预测结果确定对应的交叉熵损失；

[0164] 步骤S703，确定训练文本集所包括的训练文本；

[0165] 步骤S704，识别真实顺滑文本包括的关键词，将关键词从真实顺滑文本中去除，得到后处理文本；

[0166] 步骤S705，根据训练文本、真实顺滑文本和后处理文本之间的相似性，确定对应的对比损失；

[0167] 步骤S706，根据对比损失对神经网络进行训练；

[0168] 步骤S707，响应于接收到目标文本，根据预先设置的预训练模型，对目标文本进行编码，生成对应的隐向量；

[0169] 步骤S708，使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到多个短语向量表征；

[0170] 步骤S709，根据每个短语向量表征与隐向量之间的相关性，生成每个短语向量表征与隐向量之间的第一相关性矩阵；

[0171] 步骤S710,根据不同短语向量表征之间的相关性,生成对应的第二相关性矩阵;

[0172] 步骤S711,根据卷积核的窗口尺寸确定拼接顺序;

[0173] 步骤S712,根据拼接顺序对得到的多个池化向量进行拼接,得到特征向量;

[0174] 步骤S713,将特征向量和隐向量分别输入全连接层;

[0175] 步骤S714,根据预先设置的激活函数对全连接层的输出结果进行标签预测,生成对应的分类标签。

[0176] 上述本发明又一个可参考实施例的文本顺滑检测的方法的具体实施内容,在上面所述文本顺滑检测的方法中已经详细说明了,故在此重复内容不再说明。

[0177] 根据本发明实施例的第二方面,提供一种文本顺滑检测的装置。

[0178] 图8是根据本发明实施例的文本顺滑检测的装置的主要模块的示意图,如图8所示,所述文本顺滑检测的装置800主要包括:

[0179] 编码模块801,用于响应于接收到目标文本,根据预先设置的预训练模型,对所述目标文本进行编码,生成对应的隐向量;

[0180] 提取模块802,用于对所述隐向量进行多尺度特征提取,得到特征向量;

[0181] 分类模块803,用于对所述隐向量和所述特征向量进行组合,得到组合向量,根据预先设置的激活函数对所述组合向量进行分类,生成所述目标文本对应的分类标签。

[0182] 根据本发明一个可参考的实施例,对所述隐向量进行多尺度特征提取,得到特征向量,包括:

[0183] 使用不同窗口尺寸的卷积核对所述隐向量进行卷积运算,得到多个短语向量表征;

[0184] 根据所述隐向量与所述多个短语向量表征之间的相关性,生成多个相关性矩阵;

[0185] 对所述多个相关性矩阵进行池化操作,得到多个池化向量;

[0186] 对所述多个池化向量进行拼接,得到所述特征向量。

[0187] 根据本发明另一个可参考的实施例,根据所述隐向量与所述多个短语向量表征之间的相关性,生成多个相关性矩阵,包括:

[0188] 根据每个短语向量表征与所述隐向量之间的相关性,生成对应的第一相关性矩阵;

[0189] 根据不同卷积向量之间的相关性,生成对应的第二相关性矩阵。

[0190] 根据本发明又一个可参考的实施例,对所述多个池化向量进行拼接,包括:

[0191] 根据所述卷积核的窗口尺寸确定所述多个池化向量的拼接顺序;

[0192] 根据所述拼接顺序对所述多个池化向量进行拼接。

[0193] 根据本发明还一个可参考的实施例,所述文本顺滑检测的装置800还包括:

[0194] 第一训练模块,用于基于预先设置的训练文本集对预先设置的神经网络进行迭代训练,得到所述训练文本集对应的预测结果;

[0195] 第一确定模块,用于根据所述预测结果和预先设置的真实顺滑文本确定对应的分类损失;

[0196] 优化模块,用于根据所述分类损失对所述神经网络包括的参数进行优化。

[0197] 根据本发明再一个可参考的实施例,所述文本顺滑检测的装置800还包括:

[0198] 第二确定模块,用于确定所述训练文本集包括的训练文本;

[0199] 去除模块，用于识别所述真实顺滑文本包括的关键词，将所述关键词从所述真实顺滑文本中去除，得到后处理文本；

[0200] 对比模块，用于根据所述训练文本、所述真实顺滑文本和所述后处理文本之间的相似性，确定所述推理结果对应的对比损失；

[0201] 第二训练模块，用于根据所述分类损失和所述对比损失，对所述神经网络进行训练。

[0202] 根据本发明一个可参考的实施例，将所述关键词从所述真实顺滑文本中去除，包括：

[0203] 根据所述关键词的词频和所述关键词的逆文档频率，对所述关键词进行排序，得到关键词排序结果；

[0204] 根据预先设置的关键词个数和所述关键词排序结果确定目标关键词；

[0205] 将所述目标关键词从所述真实顺滑文本中去除。

[0206] 需要说明的是，在本发明实施例中所述文本顺滑检测的装置的具体实施内容，在上面所述文本顺滑检测的方法中已经详细说明了，故在此重复内容不再说明。

[0207] 根据本发明实施例的技术方案，对目标文本进行预训练，得到对应的隐向量，对隐向量进行多尺度特征提取，能够准确地获取到不同长度的词语之间的相关性，便于根据多种相关性信息进行文本顺滑检测；多尺度特征提取包括：基于不同尺寸卷积核的卷积运算、根据相关性生成相关性矩阵、最大池化操作、向量拼接等步骤，使得得到的特征向量具有良好的特征表示；根据短语向量表征与隐向量之间的相关性生成第一相关性矩阵，根据不同短语向量表征之间的相关性，生成第二相关性矩阵，能够进一步获取不同长度的词语之间的多种相关性；根据卷积核的窗口尺寸确定拼接顺序，根据拼接顺序对多个池化向量进行拼接，能够提高池化向量拼接的灵活性；根据分类损失对神经网络包括的参数进行优化，能够提高神经网络的优化效果和优化效率；根据多种文本之间的相似性确定推理结果对应的对比损失，能够引入对目标文本的语义一致性约束，能够提高神经网络对关键词的识别能力，避免将关键词从目标文本中去除；使用词频和逆文档频率获取关键词，能够提高关键词的获取效率，提高后处理文本的生成效率和对比损失的计算效率。

[0208] 根据本发明实施例的第三方面，提供一种文本顺滑检测的电子设备，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本发明实施例第一方面提供的方法。

[0209] 根据本发明实施例的第四方面，提供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本发明实施例第一方面提供的方法。

[0210] 图9示出了可以应用本发明实施例的文本顺滑检测的方法或者文本顺滑检测的装置的示例性系统架构900。

[0211] 如图9所示，系统架构900可以包括终端设备901、902、903，网络904和服务器905。网络904用以在终端设备901、902、903和服务器905之间提供通信链路的介质。网络904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。

[0212] 用户可以使用终端设备901、902、903通过网络904与服务器905交互，以接收或发送消息等。终端设备901、902、903上可以安装有各种通讯客户端应用，例如文本顺滑检测类应用、文本识别类应用、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)。

[0213] 终端设备901、902、903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。

[0214] 服务器905可以是提供各种服务的服务器,例如对上游利用终端设备901、902、903所发来的文本顺滑检测的请求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响应于接收到目标文本,根据预先设置的预训练模型,对所述目标文本进行编码,生成对应的隐向量;对所述隐向量进行多尺度特征提取,得到特征向量;对所述隐向量和所述特征向量进行组合,得到组合向量,根据预先设置的激活函数对所述组合向量进行分类,生成所述目标文本对应的分类标签;并将文本顺滑检测的情况(仅为示例)反馈给终端设备。

[0215] 需要说明的是,本发明实施例所提供的文本顺滑检测的方法一般由服务器905执行,相应地,所述文本顺滑检测的装置一般设置在服务器905中。本发明实施例所提供的文本顺滑检测的方法也可以由终端设备901、902、903执行,相应地,所述文本顺滑检测的装置可以设置在终端设备901、902、903中。

[0216] 应该理解,图9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。

[0217] 下面参考图10,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1000的结构示意图。图10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。

[0218] 如图10所示,计算机系统1000包括中央处理单元(CPU)1001,其可以根据存储在只读存储器(ROM)1002中的程序或者从存储部分908加载到随机访问存储器(RAM)1003中的程序而执行各种适当的动作和处理。在RAM 1003中,还存储有系统1000操作所需的各种程序和数据。CPU 1001、ROM 1002以及RAM 1003通过总线1004彼此相连。输入/输出(I/O)接口1005也连接至总线1004。

[0219] 以下部件连接至I/O接口1005:包括键盘、鼠标等的输入部分1006;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1007;包括硬盘等的存储部分1008;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1009。通信部分1009经由诸如因特网的网络执行通信处理。驱动器1010也根据需要连接至I/O接口1005。可拆卸介质1011,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1010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1008。

[0220] 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,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1009从网络上被下载和安装,和/或从可拆卸介质1011被安装。在该计算机程序被中央处理单元(CPU)1001执行时,执行本发明实施例的系统中限定的上述功能。

[0221] 需要说明的是,本发明实施例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。而在本发明实施例中,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。

[0222] 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。也要注意的是,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。

[0223] 描述于本发明实施例中所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:一种处理器包括编码模块、提取模块、分类模块,其中,这些模块的名称在某种情况下并不构成对该模块本身的限定,例如,编码模块还可以被描述为“对目标文本进行编码的模块”。

[0224] 作为另一方面,本发明实施例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,该设备实现如下方法:响应于接收到目标文本,根据预先设置的预训练模型,对所述目标文本进行编码,生成对应的隐向量;对所述隐向量进行多尺度特征提取,得到特征向量;对所述隐向量和所述特征向量进行组合,得到组合向量,根据预先设置的激活函数对所述组合向量进行分类,生成所述目标文本对应的分类标签。

[0225] 根据本发明实施例的技术方案,对目标文本进行预训练,得到对应的隐向量,对隐向量进行多尺度特征提取,能够准确地获取到不同长度的词语之间的相关性,便于根据多种相关性信息进行文本顺滑检测;多尺度特征提取包括:基于不同尺寸卷积核的卷积运算、根据相关性生成相关性矩阵、最大池化操作、向量拼接等步骤,使得得到的特征向量具有良好的特征表示;根据短语向量表征与隐向量之间的相关性生成第一相关性矩阵,根据不同短语向量表征之间的相关性,生成第二相关矩阵,能够进一步获取不同长度的词语之间的多种相关性;根据卷积核的窗口尺寸确定拼接顺序,根据拼接顺序对多个池化向量进行拼接,能够提高池化向量拼接的灵活性;根据分类损失和对神经网络包括的参数进行优化,能

够提高神经网络的优化效果和优化效率;根据多种文本之间的相似性确定推理结果对应的对比损失,能够引入对目标文本的语义一致性约束,能够提高神经网络对关键词的识别能力,避免将关键词从目标文本中去除;使用词频和逆文档频率获取关键词,能够提高关键词的获取效率,提高后处理文本的生成效率和对比损失的计算效率。

[0226] 上述具体实施方式,并不构成对本发明实施例保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和其他因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明实施例的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明实施例保护范围之内。

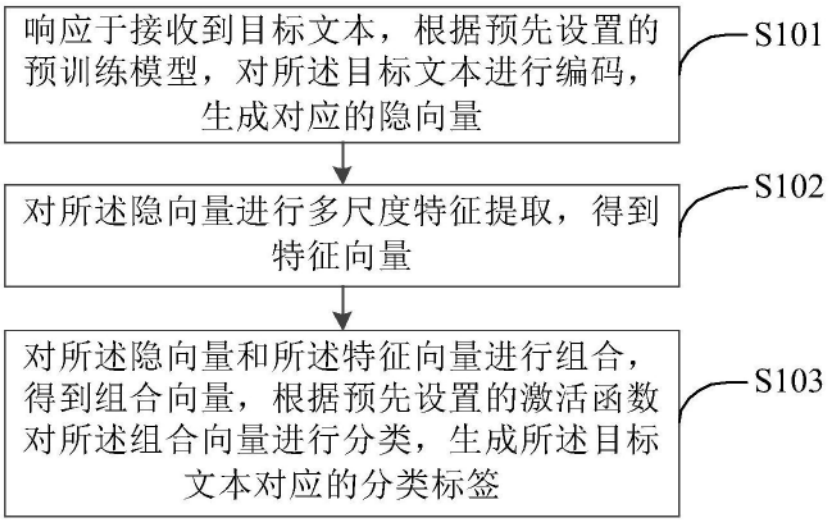
响应于接收到目标文本，根据预先设置的预训练模型，对所述目标文本进行编码，生成对应的隐向量
S101
对所述隐向量进行多尺度特征提取，得到特征向量
S102
对所述隐向量和所述特征向量进行组合，得到组合向量，根据预先设置的激活函数对所述组合向量进行分类，生成所述目标文本对应的分类标签
S103

图1

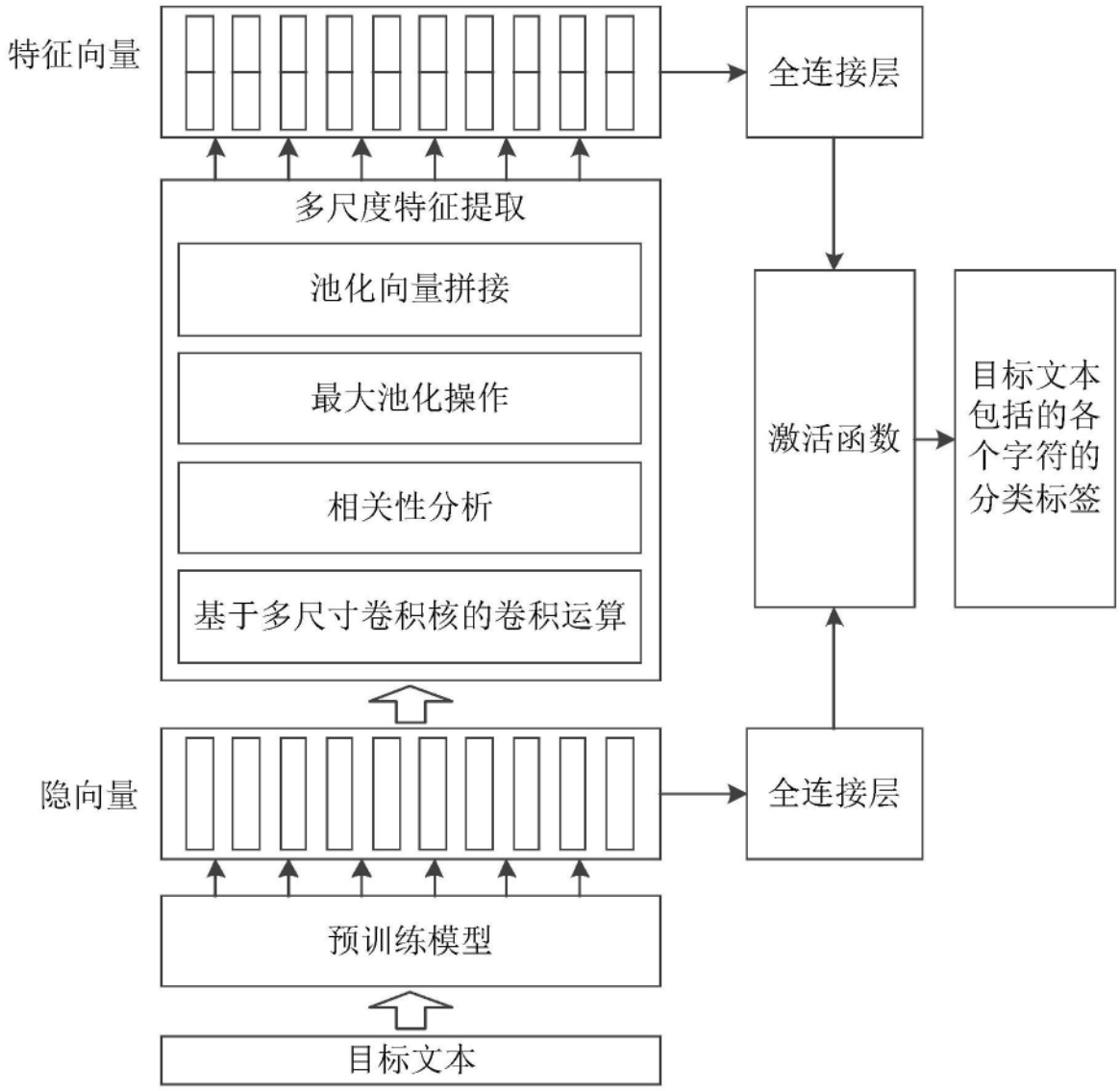
特征向量
全连接层
多尺度特征提取
池化向量拼接
最大池化操作
相关性分析
基于多尺寸卷积核的卷积运算
激活函数
目标文本包括的各个字符的分类标签
隐向量
全连接层
预训练模型
目标文本

图2

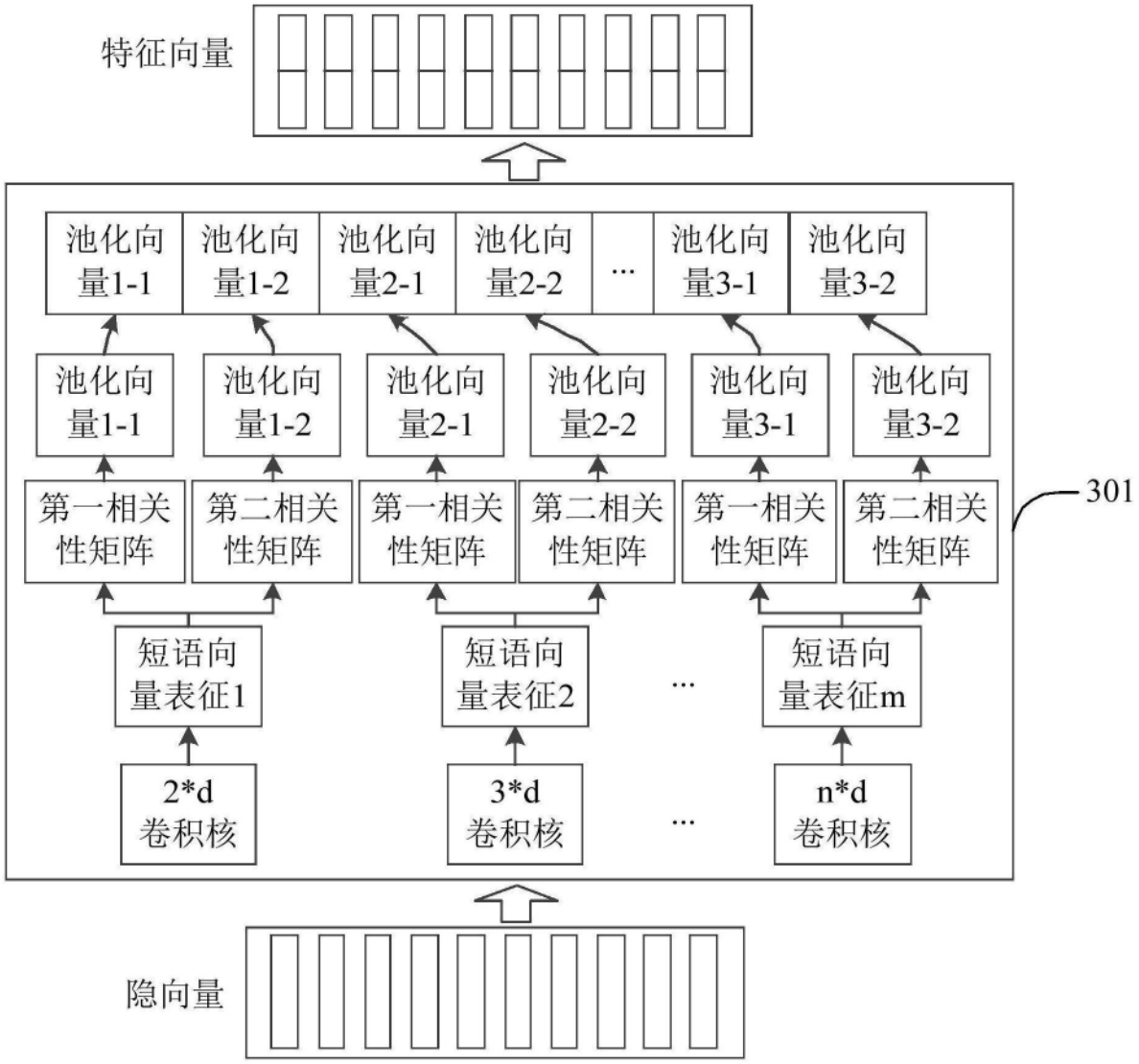
特征向量
池化向量1-1
池化向量1-2
池化向量2-1
池化向量2-2
…
池化向量3-1
池化向量3-2
池化向量1-1
池化向量1-2
池化向量2-1
池化向量2-2
池化向量3-1
池化向量3-2
第一相关性矩阵
第二相关性矩阵
第一相关性矩阵
第二相关性矩阵
第一相关性矩阵
第二相关性矩阵
301
短语向量表征1
短语向量表征2
…
短语向量表征m
2*d卷积核
3*d卷积核
…
n*d卷积核
隐向量

图3

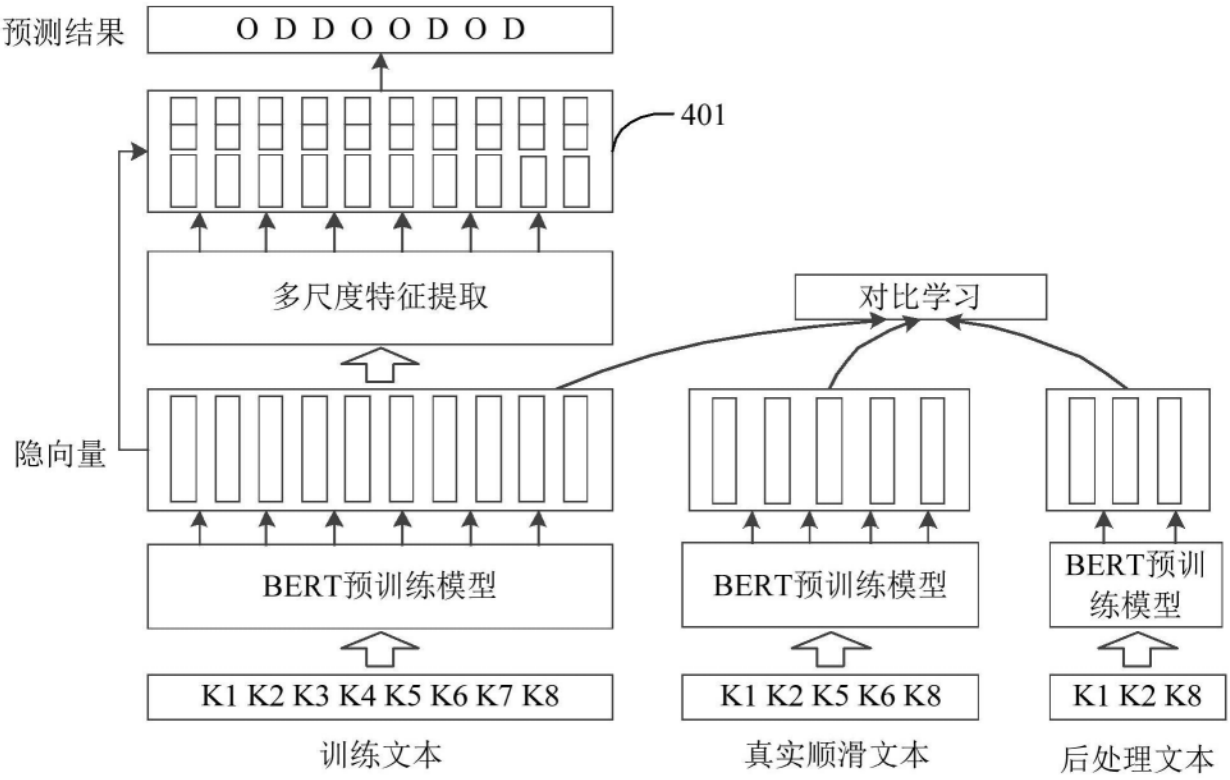
预测结果
O D D O O D O D
401
多尺度特征提取
对比学习
隐向量
BERT预训练模型
BERT预训练模型
BERT预训练模型
K1 K2 K3 K4 K5 K6 K7 K8
K1 K2 K5 K6 K8
K1 K2 K8
训练文本
真实顺滑文本
后处理文本

图4

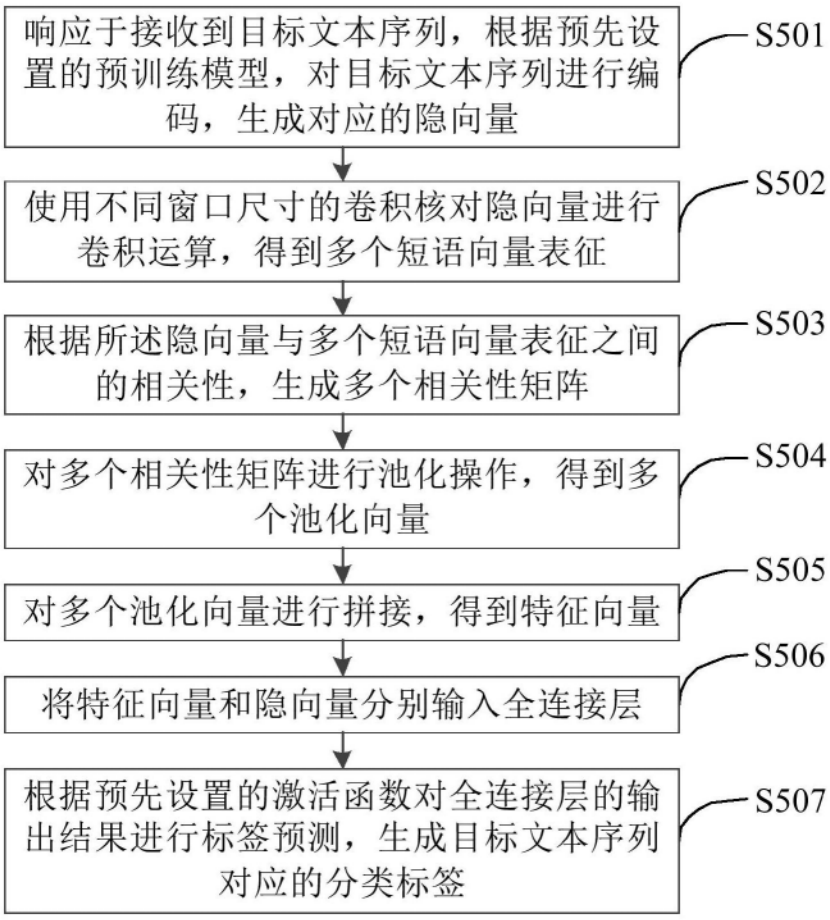
响应于接收到目标文本序列，根据预先设置的预训练模型，对目标文本序列进行编码，生成对应的隐向量
S501
使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到多个短语向量表征
S502
根据所述隐向量与多个短语向量表征之间的相关性，生成多个相关性矩阵
S503
对多个相关性矩阵进行池化操作，得到多个池化向量
S504
对多个池化向量进行拼接，得到特征向量
S505
将特征向量和隐向量分别输入全连接层
S506
根据预先设置的激活函数对全连接层的输出结果进行标签预测，生成目标文本序列对应的分类标签
S507

图5

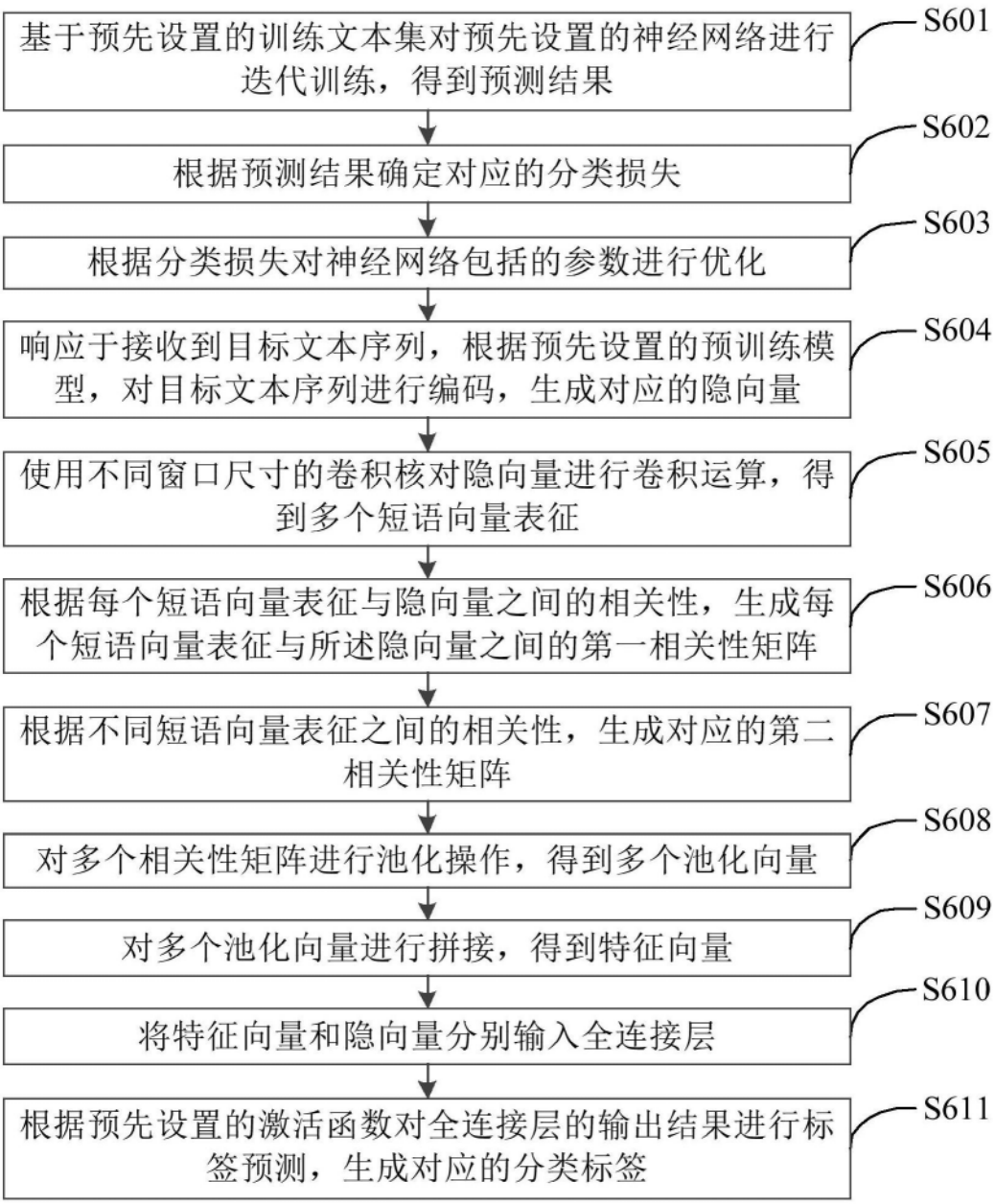
基于预先设置的训练文本集对预先设置的神经网络进行迭代训练，得到预测结果
S601
根据预测结果确定对应的分类损失
S602
根据分类损失对神经网络包括的参数进行优化
S603
响应于接收到目标文本序列，根据预先设置的预训练模型，对目标文本序列进行编码，生成对应的隐向量
S604
使用不同窗口尺寸的卷积核对隐向量进行卷积运算，得到多个短语向量表征
S605
根据每个短语向量表征与隐向量之间的相关性，生成每个短语向量表征与所述隐向量之间的第一相关性矩阵
S606
根据不同短语向量表征之间的相关性，生成对应的第二相关性矩阵
S607
对多个相关性矩阵进行池化操作，得到多个池化向量
S608
对多个池化向量进行拼接，得到特征向量
S609
将特征向量和隐向量分别输入全连接层
S610
根据预先设置的激活函数对全连接层的输出结果进行标签预测，生成对应的分类标签
S611

图6

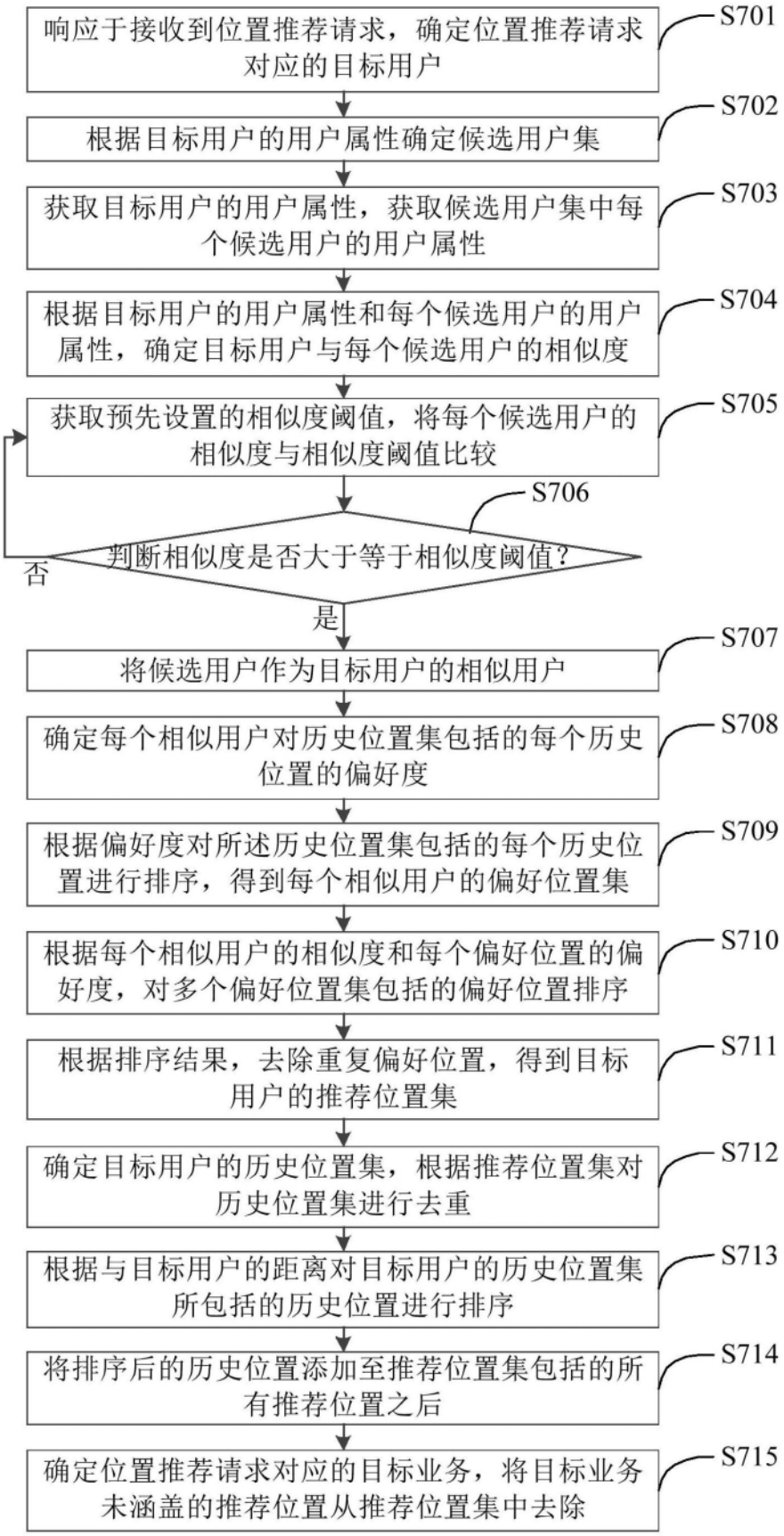
响应于接收到位置推荐请求，确定位置推荐请求对应的目标用户
S701
根据目标用户的用户属性确定候选用户集
S702
获取目标用户的用户属性，获取候选用户集中每个候选用户的用户属性
S703
根据目标用户的用户属性和每个候选用户的用户属性，确定目标用户与每个候选用户的相似度
S704
获取预先设置的相似度阈值，将每个候选用户的相似度与相似度阈值比较
S705
S706
判断相似度是否大于等于相似度阈值？
否
是
将候选用户作为目标用户的相似用户
S707
确定每个相似用户对历史位置集包括的每个历史位置的偏好度
S708
根据偏好度对所述历史位置集包括的每个历史位置进行排序，得到每个相似用户的偏好位置集
S709
根据每个相似用户的相似度和每个偏好位置的偏好度，对多个偏好位置集包括的偏好位置排序
S710
根据排序结果，去除重复偏好位置，得到目标用户的推荐位置集
S711
确定目标用户的历史位置集，根据推荐位置集对历史位置集进行去重
S712
根据与目标用户的距离对目标用户的历史位置集所包括的历史位置进行排序
S713
将排序后的历史位置添加至推荐位置集包括的所有推荐位置之后
S714
确定位置推荐请求对应的目标业务，将目标业务未涵盖的推荐位置从推荐位置集中去除
S715

图7

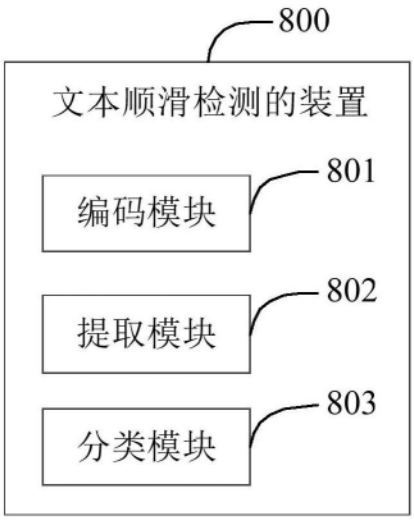
800
文本顺滑检测的装置
801
编码模块
802
提取模块
803
分类模块

图8

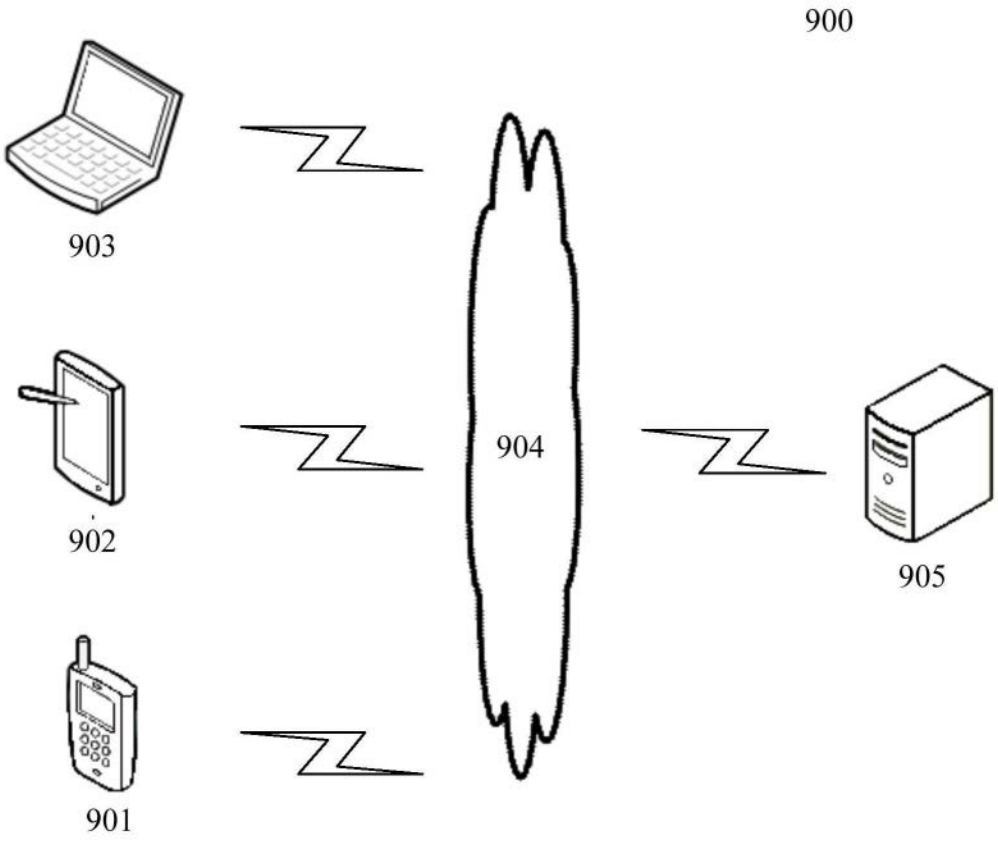
900
903
902
901
904
905

图9

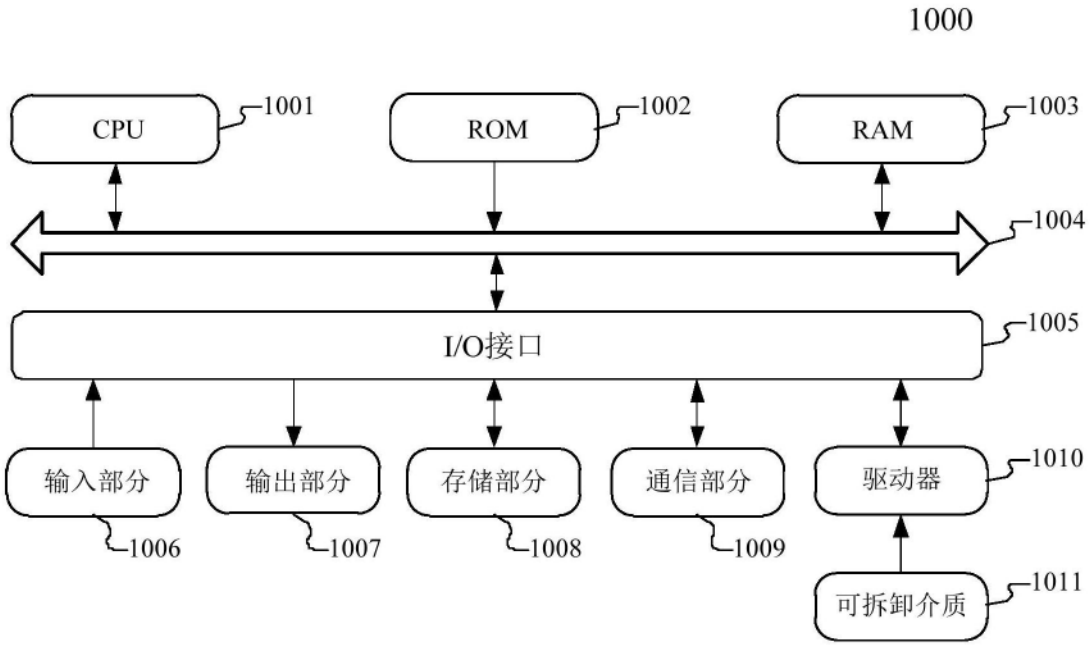
1000
CPU
1001
ROM
1002
RAM
1003
1004
I/O接口
1005
输入部分
输出部分
存储部分
通信部分
驱动器
1010
1006
1007
1008
1009
可拆卸介质
1011

图10