(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108304387 B
(45)授权公告日 2021.06.15

(21)申请号 201810195233.X

(22)申请日 2018.03.09

(65)同一申请的已公布的文献号
申请公布号 CN 108304387 A

(43)申请公布日 2018.07.20

(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地信息产业基地创业路6号

(72)发明人 金宝宝 杨帆 张成松

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王云晓 王宝筠

(51)Int.Cl.
G06F 40/284(2020.01)
G06K 9/62(2006.01)

(56)对比文件
CN 104462378 A,2015.03.25
CN 106815192 A,2017.06.09
CN 107122416 A,2017.09.01
US 2003135356 A1,2003.07.17

审查员 王悦

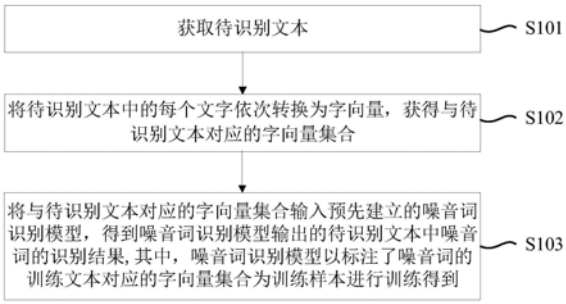
权利要求书2页 说明书8页 附图3页

(54)发明名称

文本中噪音词的识别方法、装置、服务器组及存储介质

(57)摘要

本申请提供了一种文本中噪音词的识别方法、装置、服务器组及存储介质，方法包括：获取待识别文本，将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合，将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的待识别文本中噪音词的识别结果，其中，噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。本申请提供的文本中噪音词的识别方法可通过预先建立的噪音词识别模型对待识别文本进行识别，由于噪音词识别模型基于标注了噪音词的训练文本训练得到，因此，通过该噪音词识别模型可从待识别文本中识别噪音词。



CN 108304387 B

1.一种文本中噪音词的识别方法,其特征在于,包括:

获取待识别文本,所述待识别文本中包括目标词;

将所述待识别文本中的每个文字依次转换为字向量,获得与所述待识别文本对应的字向量集合;

将与所述待识别文本对应的字向量集合输入预先建立的,并经过训练的噪音词识别模型,得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果,所述待识别文本中噪音词的识别结果用于指示所述目标词是否为噪音词;其中,所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到,包括:所述训练文本中包括所述目标词;所述噪音识别模型以标注了所述目标词为噪音词的训练文本对应的字向量集合,以及标注了所述目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到,能够根据目标词所处的上下文环境判断目标词是否为噪音词;

预先建立所述噪音词识别模型的过程,包括:

获取多个标注了噪音词的文本,组成训练文本集合;

将所述训练文本集合中的训练文本中的每个文字依次转换为字向量,得到与所述训练文本对应的字向量集合,其中,不同字向量之间的距离表征其对应的文字之间的关联性;

将所述训练文本对应的字向量集合作为输入,训练循环神经网络,将训练得到的循环神经网络作为所述噪音词识别模型。

2.根据权利要求1所述的文本中噪音词的识别方法,其特征在于,所述将所述训练文本集合中的训练文本中的每个文字依次转换为字向量,包括:

将所述训练文本集合中的训练文本中的每个文字依次处理成矢量数据,并将所述矢量数据转换为字向量,得到与所述训练文本对应的字向量集合。

3.根据权利要求2所述的文本中噪音词的识别方法,其特征在于,所述方法还包括:

获取所述训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系;

所述将所述待识别文本中的每个文字依次转换为字向量,包括:

将所述待识别文本中的每个文字依次转换为矢量数据作为目标矢量数据,并基于所述训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系将所述目标矢量数据转换为字向量。

4.一种文本中噪音词的识别装置,其特征在于,包括:待识别文本获取模块、待识别文本转换模块和噪音识别模块;

所述待识别文本获取模块,用于获取待识别文本,所述待识别文本中包括目标词;

所述待识别文本转换模块,用于将所述待识别文本中的每个文字依次转换为字向量,获得与所述待识别文本对应的字向量集合;

所述噪音识别模块,用于将与所述待识别文本对应的字向量集合输入预先建立的,并经过训练的噪音词识别模型,得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果,所述待识别文本中噪音词的识别结果用于指示所述目标词是否为噪音词;其中,所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到,包括:所述训练文本中包括所述目标词;所述噪音识别模型以标注了所述目标词为噪音词的训练文本对应的字向量集合,以及标注了所述目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到,能够根据目标词所处的上下文环境判断目标词是否

为噪音词,能够根据目标词所处的上下文环境判断目标词是否为噪音词;

还包括:训练文本获取模块、训练文本转换模块和训练模块;

所述训练文本获取模块,用于获取多个标注了噪音词的文本,组成训练文本集合;

所述训练文本转换模块,用于将所述训练文本集合中的训练文本中的每个文字依次转换为字向量,得到与所述训练文本对应的字向量集合,其中,不同字向量之间的距离表征其对应的文字之间的关联性;

所述训练模块,用于将所述训练文本对应的字向量集合作为输入,训练循环神经网络,将训练得到的循环神经网络作为所述噪音词识别模型。

5.一种服务器组,其特征在于,包括:存储器和处理器;

所述存储器,用于存储程序;

所述处理器,用于执行所述程序,以进行以下操作:

获取待识别文本,所述待识别文本中包括目标词;

将所述待识别文本中的每个文字依次转换为字向量,获得与所述待识别文本对应的字向量集合;

将与所述待识别文本对应的字向量集合输入预先建立的,并经过训练的噪音词识别模型,得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果,所述待识别文本中噪音词的识别结果用于指示所述目标词是否为噪音词;其中,所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到,包括:所述训练文本中包括所述目标词;所述噪音识别模型以标注了所述目标词为噪音词的训练文本对应的字向量集合,以及标注了所述目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到,能够根据目标词所处的上下文环境判断目标词是否为噪音词,能够根据目标词所处的上下文环境判断目标词是否为噪音词;

预先建立所述噪音词识别模型的过程,包括:

获取多个标注了噪音词的文本,组成训练文本集合;

将所述训练文本集合中的训练文本中的每个文字依次转换为字向量,得到与所述训练文本对应的字向量集合,其中,不同字向量之间的距离表征其对应的文字之间的关联性;

将所述训练文本对应的字向量集合作为输入,训练循环神经网络,将训练得到的循环神经网络作为所述噪音词识别模型。

6.一种可读存储介质,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3任一项所述的文本中噪音词的识别方法的各个步骤。

文本中噪音词的识别方法、装置、服务器组及存储介质

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种文本中噪音词的识别方法、装置、服务器组及存储介质。

背景技术

[0002] 自然语言处理是人工智能领域最为重要的子领域之一，是当前热门的翻译系统、人机对话系统、问答系统的技术核心。现实世界中产生的文本的不规范性是影响自然语言处理性能的最主要因素之一，而噪音词引起的不规范性尤其显著。

[0003] 其中，噪音词指的是不在停用词范围，但在当前语境下无意义的词。噪音词与相对固定的停用词不同，其并不固定，某些文本中的噪音词在其他文本中有可能不是噪音词，比如“12第5中学”中的数字12这里是无意义的噪音词，但放在“12月中旬”中就不是噪音词，这导致噪音词难以识别。

发明内容

[0004] 有鉴于此，本发明提供了一种文本中噪音词的识别方法、装置、服务器组及存储介质，用以解决现有技术中噪音词难以识别的问题，其技术方案如下：

[0005] 一种文本中噪音词的识别方法，包括：

[0006] 获取待识别文本；

[0007] 将所述待识别文本中的每个文字依次转换为字向量，获得与所述待识别文本对应的字向量集合；

[0008] 将与所述待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果，其中，所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0009] 其中，所述待识别文本中包括目标词；

[0010] 所述训练文本中包括所述目标词；

[0011] 所述噪音识别模型以标注了所述目标词为噪音词的训练文本对应的字向量集合，以及标注了所述目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到；

[0012] 所述待识别文本中噪音词的识别结果用于指示所述目标词是否为噪音词。

[0013] 其中，预先建立所述噪音词识别模型的过程，包括：

[0014] 获取多个标注了噪音词的文本，组成训练文本集合；

[0015] 将所述训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与所述训练文本对应的字向量集合，其中，不同字向量之间的距离表征其对应的文字之间的关联性；

[0016] 将所述训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为所述噪音词识别模型。

[0017] 其中，所述将所述训练文本集合中的训练文本中的每个文字依次转换为字向量，包括：

[0018] 将所述训练文本集合中的训练文本中的每个文字依次处理成矢量数据，并将所述矢量数据转换为字向量，得到与所述训练文本对应的字向量集合。

[0019] 其中，所述文本中噪音词的识别方法还包括：

[0020] 获取所述训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系；

[0021] 所述将所述待识别文本中的每个文字依次转换为字向量，包括：

[0022] 将所述待识别文本中的每个文字依次转换为矢量数据作为目标矢量数据，并基于所述训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系将所述目标矢量数据转换为字向量。

[0023] 一种文本中噪音词的识别装置，包括：待识别文本获取模块、待识别文本转换模块和噪音识别模块；

[0024] 所述待识别文本获取模块，用于获取待识别文本；

[0025] 所述待识别文本转换模块，用于将所述待识别文本中的每个文字依次转换为字向量，获得与所述待识别文本对应的字向量集合；

[0026] 所述噪音识别模块，用于将与所述待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果，其中，所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0027] 其中，所述待识别文本中包括目标词；

[0028] 所述训练文本中包括所述目标词；

[0029] 所述噪音识别模型以标注了所述目标词为噪音词的训练文本对应的字向量集合，以及标注了所述目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到；

[0030] 所述待识别文本中噪音词的识别结果用于指示所述目标词是否为噪音词。

[0031] 所述文本中噪音词的识别装置，还包括：训练文本获取模块、训练文本转换模块和训练模块；

[0032] 所述训练文本获取模块，用于获取多个标注了噪音词的文本，组成训练文本集合；

[0033] 所述训练文本转换模块，用于将所述训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与所述训练文本对应的字向量集合，其中，不同字向量之间的距离表征其对应的文字之间的关联性；

[0034] 所述训练模块，用于将所述训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为所述噪音词识别模型。

[0035] 一种服务器组，包括：存储器和处理器；

[0036] 所述存储器，用于存储程序；

[0037] 所述处理器，用于执行所述程序，以进行以下操作：

[0038] 获取待识别文本；

[0039] 将所述待识别文本中的每个文字依次转换为字向量，获得与所述待识别文本对应

的字向量集合；

[0040] 将与所述待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果，其中，所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0041] 一种可读存储介质，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，实现如所述的文本中噪音词的识别方法的各个步骤。

[0042] 上述技术方案具有如下有益效果：

[0043] 本发明提供的文本中噪音词的识别方法、装置、服务器组及存储介质，首先获取待识别文本，然后将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合，最后将将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的所述待识别文本中噪音词的识别结果，由于噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到，因此，通过噪音词识别模型可从待识别文本中识别噪音词。本发明提供的文本中噪音词的识别方法使得用户不需要较强的行业知识，只需要在训练模型的初期对训练文本进行标注，实现简单，且识别准确率较高。

附图说明

[0044] 为了更清楚地说明本发明实施例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简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的实施例，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，还可以根据提供的附图获得其他的附图。

[0045] 图1为本发明实施例提供的文本中噪音词的识别方法的一流程示意图；

[0046] 图2为本发明实施例提供的预先建立噪音词识别模型的实现方式的流程示意图；

[0047] 图3为本发明实施例提供的文本中噪音词的识别方法的另一流程示意图；

[0048] 图4为本发明实施例提供的预先建立噪音词识别模型的实现方式的流程示意图；

[0049] 图5为本发明实施例提供的文本中噪音词的识别装置的结构示意图；

[0050] 图6为本发明实施例提供的服务器组的结构示意图。

具体实施方式

[0051] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0052] 本发明实施例提供了一种文本中噪音词的识别方法，请参阅图1，示出了该识别方法的流程示意图，可以包括：

[0053] 步骤S101：获取待识别文本。

[0054] 步骤S102：将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合。

[0055] 步骤S103：将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，

得到噪音词识别模型输出的待识别文本中噪音词的识别结果。

[0056] 其中，噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0057] 请参阅图2，示出了本实施例中预先建立噪音词识别模型的一种可能的实现过程的流程示意图，可以包括：

[0058] 步骤S201：获取多个标注了噪音词的文本，组成训练文本集合。

[0059] 具体地，首先获取多个文本，获取途径可以但不限定为从已有的语料库中选取、通过网络爬虫从网络上爬取等，然后分别对每个文本中的噪音词进行标注，从而获得多个标注了噪音词的文本，每个标注了噪音词的文本为一个训练文本，将这些标注了噪音词的文本组成训练文本集合。优选地，可获取不同领域的文本，以使建立的噪音词识别模型适应不同的应用领域。

[0060] 步骤S202：将训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与训练文本对应的字向量集合。

[0061] 其中，不同字向量之间的距离表征其对应的文字之间的关联性。例如，训练文本集合中有大量的“第一人民医院”、“第二中心医院”等与“医院”相关的训练文本，则进行向量转换后，“医”对应的字向量与“院”对应的字向量之间的距离较近，即“医”与“院”之间的关联性较强，而“人”与“中”两个字并没有大量同时出现，因此，“人”对应的字向量与“中”对应的字向量之间的距离较远，即“人”与“中”之间的关联性较弱。

[0062] 具体地，将训练文本集合中的训练文本中的每个文字依次转换为字向量的过程可以包括：将训练文本集合中的训练文本中的每个文字依次处理成矢量数据，并将矢量数据转换为字向量，得到与训练文本对应的字向量集合。

[0063] 步骤S203：将训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为噪音词识别模型。

[0064] 其中，循环神经网络可以但不限定为RNN、LSTM、GRU等带有记忆功能的神经网络模型。

[0065] 本发明实施例提供的文本中噪音词的识别方法，首先获取待识别文本，然后将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合，最后将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的待识别文本中噪音词的识别结果，由于噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到，因此，通过噪音词识别模型可从待识别文本中识别噪音词。本发明实施例提供的文本中噪音词的识别方法，可直接对待识别文本进行全文文本分析，确定待识别文本中是否包含噪音词。本发明实施例提供的识别方法使得用户不需要较强的行业知识，只需要在训练模型的初期对训练文本进行标注，因此实现简单，且识别准确率较高，另外，由于训练文本选自多个不同领域，因此，该方法可适用于多个不同的领域，即适用范围较广。

[0066] 请参阅图3，示出了本发明实施例提供的文本中噪音词的识别方法的另一流程示意图，该识别方法可以包括：

[0067] 步骤S301：获取包含目标词的待识别文本。

[0068] 步骤S302：将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对

应的字向量集合。

[0069] 步骤S303:将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型,得到噪音词识别模型输出的、指示待识别文本中的目标词是否为噪音词的识别结果。

[0070] 其中,噪音词识别模型通过训练文本对应的字向量集合训练得到,其中,训练文本中包括目标词,具体地,噪音识别模型以标注了目标词为噪音词的训练文本对应的字向量集合,以及标注了目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0071] 请参阅图4,示出了本实施例中预先建立噪音词识别模型的一种可能的实现过程的流程示意图,可以包括:

[0072] 步骤S401:获取对包含目标词的文本中的目标词进行标注的文本,组成训练文本集合。

[0073] 具体地,首先获取包括目标词的多个文本,获取途径可以但不限定为从已有的语料库中选取、通过网络爬虫从网络上爬取等,然后分别对每个文本中的目标词进行标注,标注该目标词为噪音词还是非噪音词,从而获得多个对目标词进行标注的文本,将这些对目标词进行标注的文本组成训练文本集合。优选地,获取包括目标词、且属于不同领域的多个文本,以使建立的噪音词识别模型能够适应不同的应用领域。

[0074] 步骤S402:将训练文本集合中的训练文本中的每个文字依次转换为字向量,得到与训练文本对应的字向量集合。

[0075] 其中,不同字向量之间的距离表征其对应的文字之间的关联性。

[0076] 具体地,将训练文本集合中的训练文本中的每个文字依次转换为字向量的过程可以包括:将训练文本集合中的训练文本中的每个文字依次处理成矢量数据,并将矢量数据转换为字向量,得到与训练文本对应的字向量集合。

[0077] 在一种可能的实现方式中,可将训练文本集合中出现的所有字进行one-hot编码,以完成文本数据向计算机可处理的矢量数据的转换。需要说明的是,one-hot编码也叫单热点编码,即给训练文本集合中出现的每一个字一个唯一的编码。

[0078] 具体地,若训练文本集合中总共有N种文字,则每一种文字可用一个N-1维的矢量进行表示,第一种文字的N-1维矢量的所有位均为0,第二种文字的第一位置为1,第三种文字的第二位置为1,以此类推。

[0079] 示例性地,训练文本集合中有两条语句:“123第一人民医院”和“解放一路”,则训练文本集合中共有12种文字,分别为:“1”、“2”、“3”、“第”、“一”、“人”、“民”、“医”、“院”、“解”、“放”、“路”12种文字,则上述12中文字对应的编码依次为:

[0080] “1”对应的编码为:[0,0,0,0,0,0,0,0,0,0,0,0]

[0081] “2”对应的编码为:[1,0,0,0,0,0,0,0,0,0,0,0]

[0082] “3”对应的编码为:[0,1,0,0,0,0,0,0,0,0,0,0]

[0083] “第”对应的编码为:[0,0,1,0,0,0,0,0,0,0,0,0]

[0084] “一”对应的编码为:[0,0,0,1,0,0,0,0,0,0,0,0]

[0085] ……

[0086] “放”对应的编码为:[0,0,0,0,0,0,0,0,0,0,1,0]

[0087] “路”对应的编码为:[0,0,0,0,0,0,0,0,0,0,0,1]

[0088] 在将每种文字处理成矢量数据后，可采用word2vec等方法将每个矢量数据转换为字向量。通过上述过程可获得训练文本集合中出现的所有文字对应的矢量数据及字向量，基于此，对于每个训练样本而言，其包含的所有文字对应的字向量可确定，其包含的所有文字对应的字向量组成字向量集合，如此便可获得训练样本对应的字向量集合。

[0089] 另外，为了实现后续待识别文本中的文字向字向量的转换，可存储训练文本集合中出现的所有文字对应的矢量数据与对应字向量的映射关系。具体的，步骤S102中将待识别文本中的每个文字依次转换为字向量的实现过程可以包括：将待识别文本中的每个文字依次转换为矢量数据作为目标矢量数据，基于上述矢量数据与字向量的对应关系，将目标矢量数据转换为字向量。具体地，在矢量数据与字向量的对应关系中查找与目标矢量数据对应的字向量。

[0090] 步骤S403：将训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为噪音词识别模型。

[0091] 其中，循环神经网络可以但不限定为RNN、LSTM、GRU等带有记忆功能的神经网络模型。

[0092] 本发明实施例提供的文本中噪音词的识别方法，首先获取包含目标词的待识别文本，然后将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合，最后将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的识别结果，由于噪音词识别模型以标注了目标词为噪音词的训练文本对应的字向量集合，以及标注了目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到，因此，通过噪音词识别模型可识别出待识别文本中的目标词是否为噪音词。本发明实施例提供的文本中噪音词的识别方法，可对待识别文本中的目标词进行分析，确定待识别文本中的目标词是否为噪音词。本发明实施例提供的方法不需要较强的行业知识，只需要在训练模型的初期对训练文本进行标注，因此实现简单，并且，由于噪音词识别模型根据目标词所处的上下文环境判断目标词是否为噪音词，因此，识别准确率较高，另外，由于训练文本选自多个不同领域，因此，该方法可适用于多个不同的领域，即适用范围较广。

[0093] 本发明实施例还提供了一种文本中噪音词的识别装置，请参阅图5，示出了该识别装置的结构示意图，可以包括：待识别文本获取模块501、待识别文本转换模块502和噪音识别模块503。其中：

[0094] 待识别文本获取模块501，用于获取待识别文本。

[0095] 待识别文本转换模块502，用于将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合。

[0096] 噪音识别模块503，用于将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的待识别文本中噪音词的识别结果。

[0097] 其中，噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0098] 本发明实施例提供的文本中噪音词的识别装置，可利用预先建立的噪音词识别模型对待识别文本进行分析，确定待识别文本中是否包含噪音词。本发明实施例提供的识别装置使得用户不需要较强的行业知识，只需要在训练模型的初期对训练文本进行标注，因

此实现简单，且识别准确率较高，另外，由于训练文本选自多个不同领域，因此，该方法可适用于多个不同的领域，即适用范围较广。

[0099] 在一种可能的实现方式中，上述实施例中待识别文本获取模块501获取的待识别文本中包括目标词，相应地，训练文本中也包括目标词。噪音识别模型以标注了目标词为噪音词的训练文本对应的字向量集合，以及标注了目标词为非噪音词的训练文本对应的字向量集合为训练样本进行训练得到。噪音识别模块503输出的待识别文本中噪音词的识别结果用于指示目标词是否为噪音词。

[0100] 在一种可能的实现方式中，上述实施例提供的文本中噪音词的识别装置，还可以包括：训练文本获取模块、训练文本转换模块和训练模块。其中：

[0101] 训练文本获取模块，用于获取多个标注了噪音词的文本，组成训练文本集合。

[0102] 训练文本转换模块，用于将训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与训练文本对应的字向量集合。

[0103] 其中，不同字向量之间的距离表征其对应的文字之间的关联性。

[0104] 训练模块，用于将训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为噪音词识别模型。

[0105] 其中，训练文本转换模块，具体用于将训练文本集合中的训练文本中的每个文字依次处理成矢量数据，并将矢量数据转换为字向量，得到与训练文本对应的字向量集合。

[0106] 上述实施例提供的文本中噪音词的识别装置，还可以包括：映射关系获取模块。

[0107] 映射关系获取模块，用于获取训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系。

[0108] 待识别文本转换模块502，具体用于将待识别文本中的每个文字依次转换为矢量数据作为目标矢量数据，并基于训练文本集合中出现的每种文字对应的矢量数据与对应字向量的映射关系将所述目标矢量数据转换为字向量。

[0109] 本发明实施例还提供了一种服务器组，该服务器组可以包括：存储器601和处理器602。

[0110] 存储器601，用于存储程序；

[0111] 处理器602，用于执行所述程序，以进行以下操作：

[0112] 获取待识别文本；

[0113] 将所述待识别文本中的每个文字依次转换为字向量，获得与所述待识别文本对应的字向量集合；

[0114] 将与所述待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到所述噪音词识别模型输出的所述待识别文本中噪音词的识别结果，其中，所述噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到。

[0115] 本发明实施例还提供了一种可读存储介质，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，实现上述任一实施例提供的文本中噪音词的识别方法的各个步骤。

[0116] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。

[0117] 在本申请所提供的几个实施例中，应该理解到，所揭露的方法、装置和设备，可以

通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0118] 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单元来实现本实施例方案的目的。另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0119] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，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)执行本发明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,RandomAccess Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0120] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

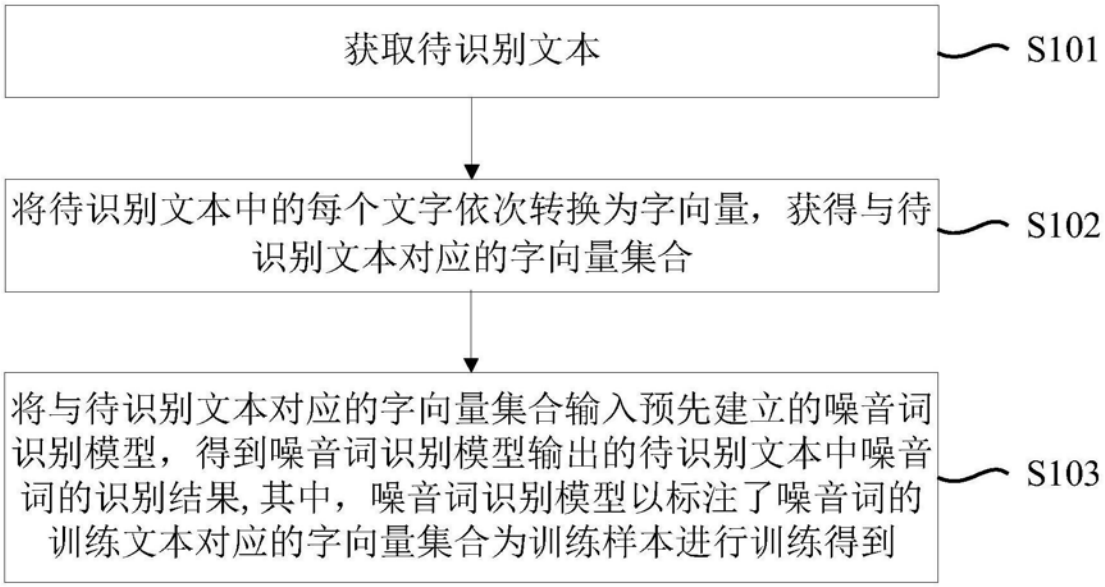

获取待识别文本
S101
将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合
S102
将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的待识别文本中噪音词的识别结果,其中，噪音词识别模型以标注了噪音词的训练文本对应的字向量集合为训练样本进行训练得到
S103

图1

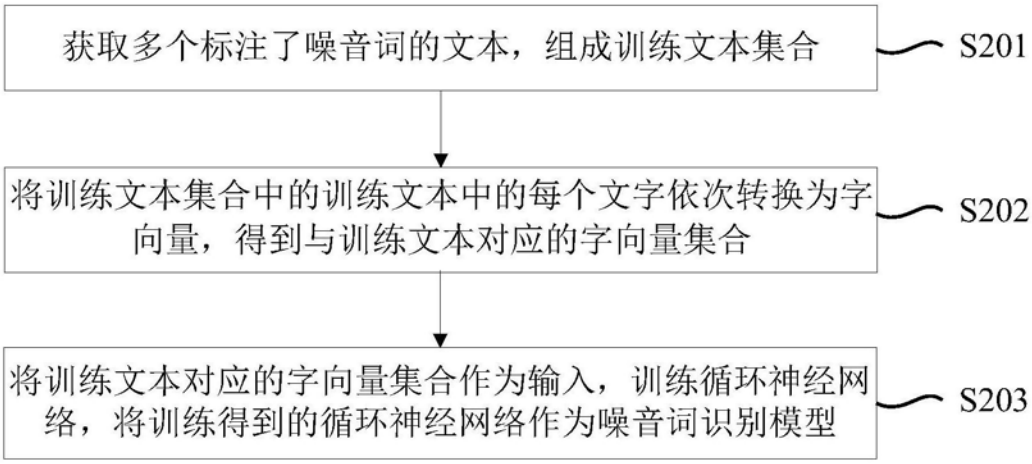

获取多个标注了噪音词的文本，组成训练文本集合
S201
将训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与训练文本对应的字向量集合
S202
将训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为噪音词识别模型
S203

图2

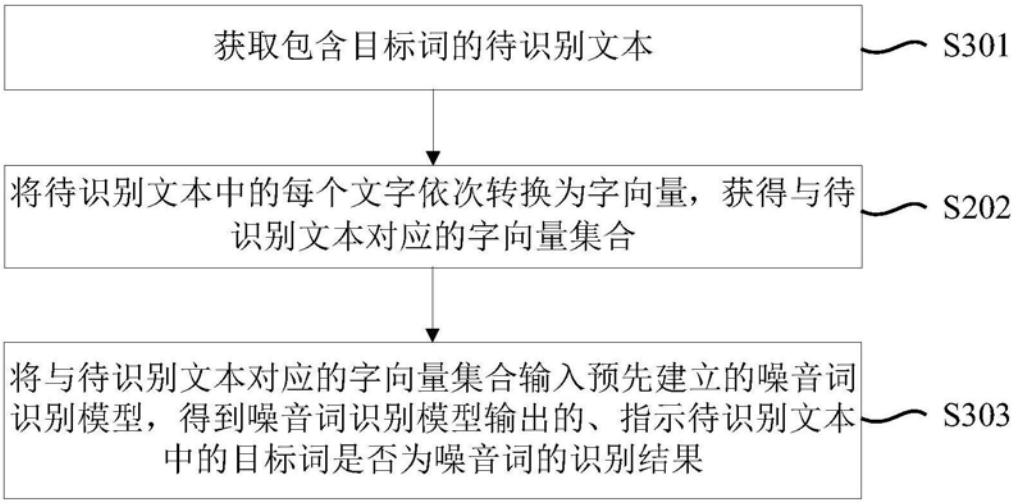

获取包含目标词的待识别文本
S301
将待识别文本中的每个文字依次转换为字向量，获得与待识别文本对应的字向量集合
S202
将与待识别文本对应的字向量集合输入预先建立的噪音词识别模型，得到噪音词识别模型输出的、指示待识别文本中的目标词是否为噪音词的识别结果
S303

图3

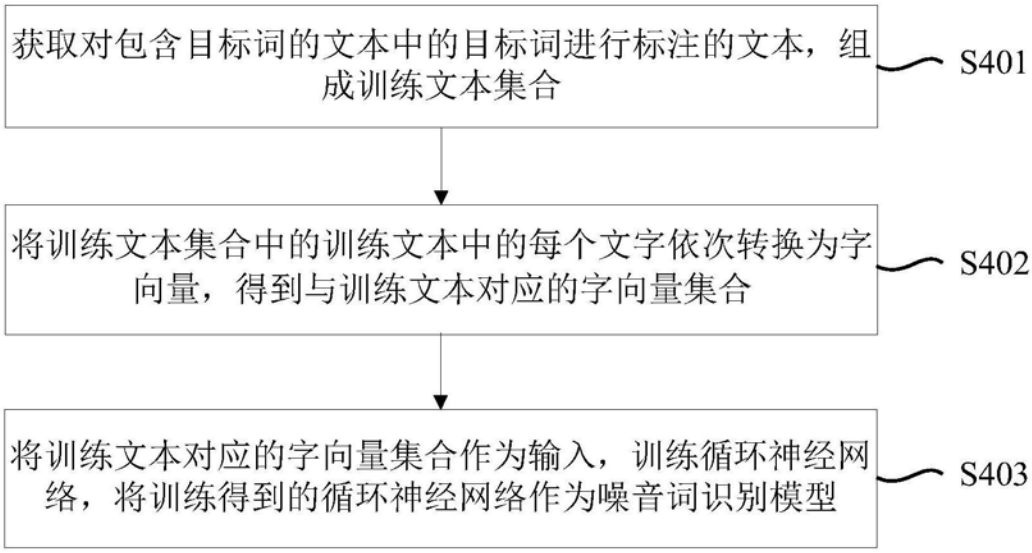

获取对包含目标词的文本中的目标词进行标注的文本，组成训练文本集合
S401
将训练文本集合中的训练文本中的每个文字依次转换为字向量，得到与训练文本对应的字向量集合
S402
将训练文本对应的字向量集合作为输入，训练循环神经网络，将训练得到的循环神经网络作为噪音词识别模型
S403

图4

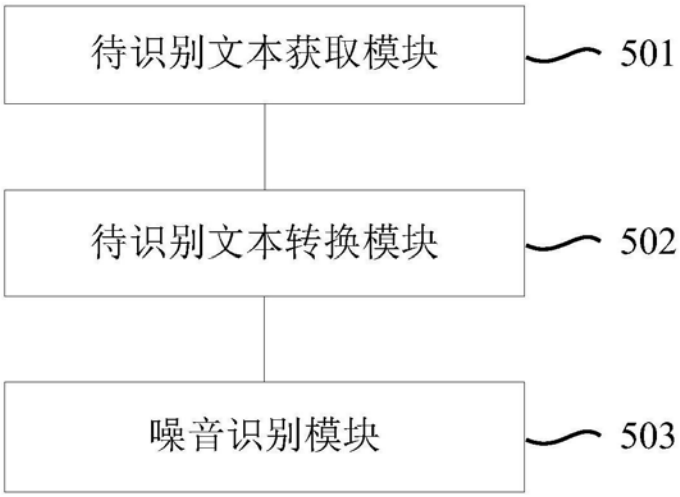


图5

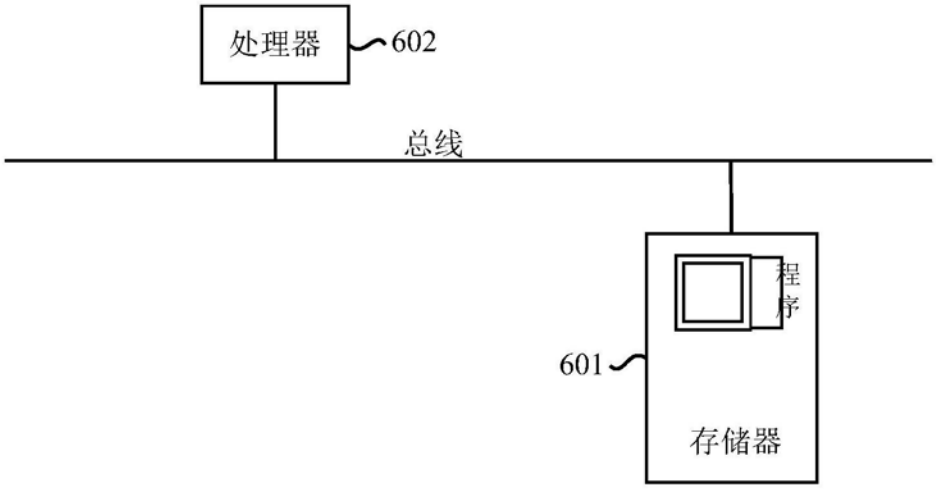


图6