(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116522905 B
(45)授权公告日 2024.03.19

(21)申请号 202310801933.X
(22)申请日 2023.07.03
(65)同一申请的已公布的文献号
申请公布号 CN 116522905 A
(43)申请公布日 2023.08.01
(73)专利权人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层
(72)发明人 赵朦
(74)专利代理机构 北京市立方律师事务所 11330
专利代理师 张筱宁
(51)Int.Cl.
G06F 40/232(2020.01)

(56)对比文件
CN 114154487 A,2022.03.08
CN 111723791 A,2020.09.29
CN 113962215 A,2022.01.21
CN 116258137 A,2023.06.13
CN 113627160 A,2021.11.09
US 2021192138 A1,2021.06.24

审查员 王俊杰

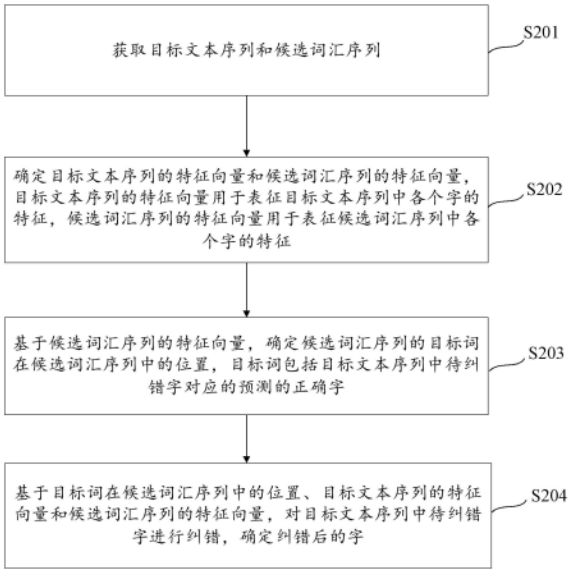
权利要求书3页 说明书16页 附图6页

(54)发明名称

文本纠错方法、装置、设备、可读存储介质及程序产品

(57)摘要

本申请实施例提供了一种文本纠错方法、装置、设备、可读存储介质及程序产品,涉及人工智能、地图等领域,应用场景包括但不限于文本纠错场景。该方法包括:获取目标文本序列和候选词汇序列;确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置;基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。



CN 116522905 B

1.一种文本纠错方法,其特征在于,包括:

获取目标文本序列和候选词汇序列;

确定所述目标文本序列的特征向量和所述候选词汇序列的特征向量,所述目标文本序列的特征向量用于表征所述目标文本序列中各个字的特征,所述候选词汇序列的特征向量用于表征所述候选词汇序列中各个字的特征;

基于所述候选词汇序列的特征向量,确定所述候选词汇序列的目标词在所述候选词汇序列中的位置,所述目标词包括所述目标文本序列中待纠错字对应的预测的正确字;

基于所述目标词在所述候选词汇序列中的位置、所述目标文本序列的特征向量和所述候选词汇序列的特征向量,对所述目标文本序列中待纠错字进行纠错,确定纠错后的字;

所述获取目标文本序列和候选词汇序列,包括:

获取目标文本;

基于所述目标文本,进行短语化处理,得到所述目标文本对应的短语;

将所述目标文本对应的短语,进行拼音转换处理,得到所述短语对应的拼音;

将所述短语对应的拼音和预设知识库中词汇进行匹配,从所述预设知识库中确定候选词汇;

基于所述目标文本和所述候选词汇,进行拼接处理,得到拼接文本,所述拼接文本包括目标文本序列和候选词汇序列;

所述预设知识库用于通过自定义的所述预设知识库的接口提供业务领域知识;

所述候选词汇序列的目标词在所述候选词汇序列中的位置包括所述目标词在所述候选词汇序列中的起始位置和所述目标词在所述候选词汇序列中的终止位置,所述基于所述目标词在所述候选词汇序列中的位置、所述目标文本序列的特征向量和所述候选词汇序列的特征向量,对所述目标文本序列中待纠错字进行纠错,确定纠错后的字,包括:

基于所述起始位置和所述终止位置,从所述候选词汇序列的特征向量中确定所述目标词的特征向量;

基于所述目标词的特征向量和所述目标文本序列的特征向量,进行拼接处理,得到纠错特征向量;

基于所述纠错特征向量,对所述目标文本序列中待纠错字进行纠错,确定纠错后的字;

所述基于所述纠错特征向量,对所述目标文本序列中待纠错字进行纠错,确定纠错后的字,包括:

基于所述纠错特征向量,确定未归一化纠错概率;

基于所述未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量;

基于所述归一化处理后的概率向量,确定纠错后的字的索引号;

基于所述纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字;

所述基于所述归一化处理后的概率向量,确定纠错后的字的索引号,包括:

将所述归一化处理后的概率向量的各元素中最大元素,确定为纠错后的字的索引号。

2.根据权利要求1所述的方法,其特征在于,在所述获取目标文本之前,还包括:

通过预设分词器,对预设的训练数据进行分词处理,得到分词集合;

基于所述分词集合,进行词性过滤处理,得到过滤后的分词集合,所述过滤后的分词集

合中分词的类型包括名称、动词中至少一项；

基于预设拼音库和所述过滤后的分词集合，确定所述过滤后的分词集合对应的拼音；

基于所述过滤后的分词集合对应的拼音，构建所述预设知识库。

3.根据权利要求1所述的方法，其特征在于，所述确定所述目标文本序列的特征向量和所述候选词汇序列的特征向量，包括：

基于所述拼接文本，进行特征提取处理，得到所述拼接文本的特征向量；

基于所述拼接文本的特征向量，进行切分处理，得到所述目标文本序列的特征向量和所述候选词汇序列的特征向量。

4.根据权利要求1所述的方法，其特征在于，所述基于所述候选词汇序列的特征向量，确定所述候选词汇序列的目标词在所述候选词汇序列中的位置，包括：

基于所述候选词汇序列的特征向量，确定所述候选词汇序列的特征向量对应的未归一化概率；

基于所述未归一化概率的候选词汇序列长度维度，进行归一化处理，得到归一化后的概率；

基于所述归一化后的概率，确定所述候选词汇序列的目标词的起始位置概率和终止位置概率；

基于所述目标词的起始位置概率，确定所述目标词在所述候选词汇序列中的起始位置，并基于所述目标词的终止位置概率，确定所述目标词在所述候选词汇序列中的终止位置，所述起始位置和所述终止位置相互匹配。

5.一种文本纠错装置，其特征在于，包括：

第一处理模块，用于获取目标文本序列和候选词汇序列；

第二处理模块，用于确定所述目标文本序列的特征向量和所述候选词汇序列的特征向量，所述目标文本序列的特征向量用于表征所述目标文本序列中各个字的特征，所述候选词汇序列的特征向量用于表征所述候选词汇序列中各个字的特征；

第三处理模块，用于基于所述候选词汇序列的特征向量，确定所述候选词汇序列的目标词在所述候选词汇序列中的位置，所述目标词中包括预测的纠错后的字；

第四处理模块，用于基于所述目标词在所述候选词汇序列中的位置、所述目标文本序列的特征向量和所述候选词汇序列的特征向量，对所述目标文本序列中待纠错字进行纠错，确定纠错后的字；

所述第一处理模块，具体用于：

获取目标文本；

基于所述目标文本，进行短语化处理，得到所述目标文本对应的短语；

将所述目标文本对应的短语，进行拼音转换处理，得到所述短语对应的拼音；

将所述短语对应的拼音和预设知识库中词汇进行匹配，从所述预设知识库中确定候选词汇；

基于所述目标文本和所述候选词汇，进行拼接处理，得到拼接文本，所述拼接文本包括目标文本序列和候选词汇序列；

所述预设知识库用于通过自定义的所述预设知识库的接口提供业务领域知识；

所述候选词汇序列的目标词在所述候选词汇序列中的位置包括所述目标词在所述候

选词汇序列中的起始位置和所述目标词在所述候选词汇序列中的终止位置,所述第四处理模块,具体用于:

基于所述起始位置和所述终止位置,从所述候选词汇序列的特征向量中确定所述目标词的特征向量;

基于所述目标词的特征向量和所述目标文本序列的特征向量,进行拼接处理,得到纠错特征向量;

基于所述纠错特征向量,对所述目标文本序列中待纠错字进行纠错,确定纠错后的字;

所述第四处理模块,具体用于:

基于所述纠错特征向量,确定未归一化纠错概率;

基于所述未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量;

基于所述归一化处理后的概率向量,确定纠错后的字的索引号;

基于所述纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字;

所述第四处理模块,具体用于:

将所述归一化处理后的概率向量的各元素中最大元素,确定为纠错后的字的索引号。

6.一种电子设备,包括存储器、处理器及存储在存储器上的计算机程序,其特征在于,所述处理器执行所述计算机程序以实现权利要求1-4中任一项所述方法的步骤。

7.一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权利要求1-4中任一项所述方法的步骤。

8.一种计算机程序产品,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权利要求1-4中任一项所述方法的步骤。

文本纠错方法、装置、设备、可读存储介质及程序产品

技术领域

[0001] 本申请涉及计算机技术领域,具体而言,本申请涉及一种文本纠错方法、装置、设备、可读存储介质及程序产品。

背景技术

[0002] 现有技术中,CSC(Chinese Spelling Correction,中文拼写纠错)是中文应用系统(中文应用系统例如搜索引擎、媒体AI中台等)的数据处理入口;因此,文本纠错的效率和准确度会极大影响例如意图识别、实体识别、文本检索等下游任务的有效性。但是,业界中文拼写纠错算法的基础方案都是基于原始文本,通过PLM(Pre-trained Language Model,预训练语言模型),对文本中的字进行纠错,往往导致文本纠错的效率和准确度都较低。

发明内容

[0003] 本申请针对现有的方式的缺点,提出一种文本纠错方法、装置、设备、计算机可读存储介质及计算机程序产品,用于解决如何提高文本纠错的效率和准确度的问题。

[0004] 第一方面,本申请提供了一种文本纠错方法,包括:

[0005] 获取目标文本序列和候选词汇序列;

[0006] 确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;

[0007] 基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;

[0008] 基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0009] 在一个实施例中,获取目标文本序列和候选词汇序列,包括:

[0010] 获取目标文本;

[0011] 基于目标文本,进行短语化处理,得到目标文本对应的短语;

[0012] 将目标文本对应的短语,进行拼音转换处理,得到短语对应的拼音;

[0013] 将短语对应的拼音和预设知识库中词汇进行匹配,从预设知识库中确定候选词汇;

[0014] 基于目标文本和候选词汇,进行拼接处理,得到拼接文本,拼接文本包括目标文本序列和候选词汇序列。

[0015] 在一个实施例中,在获取目标文本之前,还包括:

[0016] 通过预设分词器,对预设的训练数据进行分词处理,得到分词集合;

[0017] 基于分词集合,进行词性过滤处理,得到过滤后的分词集合,过滤后的分词集合中分词的类型包括名称、动词中至少一项;

[0018] 基于预设拼音库和过滤后的分词集合,确定过滤后的分词集合对应的拼音;

[0019] 基于过滤后的分词集合对应的拼音,构建预设知识库。

[0020] 在一个实施例中,确定目标文本序列的特征向量和候选词汇序列的特征向量,包括:

[0021] 基于拼接文本,进行特征提取处理,得到拼接文本的特征向量;

[0022] 基于拼接文本的特征向量,进行切分处理,得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0023] 在一个实施例中,基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,包括:

[0024] 基于候选词汇序列的特征向量,确定候选词汇序列的特征向量对应的未归一化概率;

[0025] 基于未归一化概率的候选词汇序列长度维度,进行归一化处理,得到归一化后的概率;

[0026] 基于归一化后的概率,确定候选词汇序列的目标词的起始位置概率和终止位置概率;

[0027] 基于目标词的起始位置概率,确定目标词在候选词汇序列中的起始位置,并基于目标词的终止位置概率,确定目标词在候选词汇序列中的终止位置,起始位置和终止位置相互匹配。

[0028] 在一个实施例中,候选词汇序列的目标词在候选词汇序列中的位置包括目标词在候选词汇序列中的起始位置和目标词在候选词汇序列中的终止位置,基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字,包括:

[0029] 基于起始位置和终止位置,从候选词汇序列的特征向量中确定目标词的特征向量;

[0030] 基于目标词的特征向量和目标文本序列的特征向量,进行拼接处理,得到纠错特征向量;

[0031] 基于纠错特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0032] 在一个实施例中,基于纠错特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字,包括:

[0033] 基于纠错特征向量,确定未归一化纠错概率;

[0034] 基于未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量;

[0035] 基于归一化处理后的概率向量,确定纠错后的字的索引号;

[0036] 基于纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字。

[0037] 在一个实施例中,基于归一化处理后的概率向量,确定纠错后的字的索引号,包括:

[0038] 将归一化处理后的概率向量的各元素中最大元素,确定为纠错后的字的索引号。

[0039] 第二方面,本申请提供了一种文本纠错装置,包括:

[0040] 第一处理模块,用于获取目标文本序列和候选词汇序列;

[0041] 第二处理模块,用于确定目标文本序列的特征向量和候选词汇序列的特征向量,

目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;

[0042] 第三处理模块,用于基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词中包括预测的纠错后的字;

[0043] 第四处理模块,用于基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0044] 第三方面,本申请提供了一种电子设备,包括:处理器、存储器和总线;

[0045] 总线,用于连接处理器和存储器;

[0046] 存储器,用于存储操作指令;

[0047] 处理器,用于通过调用操作指令,执行本申请第一方面的文本纠错方法。

[0048] 第四方面,本申请提供了一种计算机可读存储介质,存储有计算机程序,计算机程序被用于执行本申请第一方面的文本纠错方法。

[0049] 第五方面,本申请提供了一种计算机程序产品,包括计算机程序,计算机程序被处理器执行时实现本申请第一方面中文本纠错方法的步骤。

[0050] 本申请实施例提供的技术方案,至少具有如下有益效果:

[0051] 获取目标文本序列和候选词汇序列;确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字;如此,由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。

附图说明

[0052] 为了更清楚地说明本申请实施例中的技术方案,下面将对本申请实施例描述中所需要使用的附图作简单地介绍。

[0053] 图1为本申请实施例提供的文本纠错系统的架构示意图;

[0054] 图2为本申请实施例提供的一种文本纠错方法的流程示意图;

[0055] 图3为本申请实施例提供的一种文本纠错的架构示意图;

[0056] 图4为本申请实施例提供的一种文本纠错的示意图;

[0057] 图5为本申请实施例提供的一种文本纠错的示意图;

[0058] 图6为本申请实施例提供的一种文本纠错方法的流程示意图;

[0059] 图7为本申请实施例提供的一种文本纠错装置的结构示意图;

[0060] 图8为本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

[0061] 下面结合本申请中的附图描述本申请的实施例。应理解,下面结合附图所阐述的

实施方式,是用于解释本申请实施例的技术方案的示例性描述,对本申请实施例的技术方案不构成限制。

[0062] 本技术领域技术人员可以理解,除非特意声明,这里使用的单数形式“—”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是,本申请实施例所使用的术语“包括”以及“包含”是指相应特征可以实现为所呈现的特征、信息、数据、步骤、操作、元件和/或组件,但不排除实现为本技术领域所支持其他特征、信息、数据、步骤、操作、元件、组件和/或它们的组合等。应该理解,当我们称一个元件被“连接”或“耦接”到另一元件时,该一个元件可以直接连接或耦接到另一元件,也可以指该一个元件和另一元件通过中间元件建立连接关系。此外,这里使用的“连接”或 “耦接”可以包括无线连接或无线耦接。这里使用的术语“和/或”指示该术语所限定的项目中的至少一个,例如“A和/或B”指示实现为“A”,或者实现为“B”,或者实现为“A和B”。

[0063] 可以理解的是,在本申请的具体实施方式中,涉及到文本纠错相关的数据,当本申请以上实施例运用到具体产品或技术中时,需要获得用户许可或者同意,且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0064] 为使本申请的目的、技术方案和优点更加清楚,下面将结合附图对本申请实施方式作进一步地详细描述。

[0065] 本申请实施例是识别系统提供的一种文本纠错方法,该文本纠错方法涉及人工智能、地图等领域。

[0066] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0067] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习、自动驾驶、智慧交通等几大方向。

[0068] 智能交通系统(Intelligent Traffic System,ITS)又称智能运输系统(Intelligent Transportation System),是将先进的科学技术(信息技术、计算机技术、数据通信技术、传感器技术、电子控制技术、自动控制理论、运筹学、人工智能等)有效地综合运用于交通运输、服务控制和车辆制造,加强车辆、道路、使用者三者之间的联系,从而形成一种保障安全、提高效率、改善环境、节约能源的综合运输系统。

[0069] 为了更好的理解及说明本申请实施例的方案,下面对本申请实施例中所涉及到的一些技术用语进行简单说明。

[0070] MRC:机器阅读理解(Machine Reading Comprehension,机器阅读理解)作为自然语言处理领域中的一个基本任务,要求模型就给定的一段文本和与文本相关的问题进行作答。

[0071] 编辑距离:编辑距离是一种标准的方法,编辑距离用来表示经过插入、删除和替换

操作,从一个字符串转换到另外一个字符串的最小操作步数。

[0072] kd树:kd-tree(kd树)是一种对k维空间中的实例点进行存储,以便对其进行快速检索的树形数据结构。

[0073] jieba分词器:jieba分词器的主要功能是做中文分词,jieba分词器可以进行简单分词、并行分词、命令行分词等,jieba分词器还支持关键词提取、词性标注、词位置查询等。

[0074] 概率向量:针对任意一个向量U,若向量U内部的各个元素为非负数,而且各个元素的总和等于1,则该向量U称为概率向量。

[0075] Softmax:归一化指数函数,或称Softmax函数,是逻辑函数的一种推广;Softmax能将一个含任意实数的K维向量z“压缩”到另一个K维实向量σ(z)中,使得每一个元素的范围都在(0,1)之间,并且所有元素的和为1。

[0076] 本申请实施例提供的方案涉及人工智能技术,下面以具体的实施例对本申请的技术方案进行详细说明。下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图,对本申请的实施例进行描述。

[0077] 为了更好的理解本申请实施例提供的方案,下面结合具体的一个应用场景对该方案进行说明。

[0078] 在一个实施例中,图1中示出了本申请实施例所适用的一种文本纠错系统的架构示意图,可以理解的是,本申请实施例所提供的文本纠错方法可以适用于但不限于应用于如图1所示的应用场景中。

[0079] 本示例中,如图1所示,该示例中的文本纠错系统的架构可以包括但不限于服务器10、终端20和数据库30。服务器10、终端20和数据库30之间可以通过网络40进行交互。

[0080] 服务器10获取目标文本序列和候选词汇序列;服务器10确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;服务器10基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;服务器10基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。服务器10将目标文本序列中待纠错字进行纠错,得到纠错文本序列;服务器10将纠错文本序列发送给终端20;服务器10并将纠错文本序列保存在数据库30中。

[0081] 可理解,上述仅为一种示例,本实施例在此不作限定。

[0082] 其中,终端包括但不限于智能手机(如Android手机、iOS手机等)、手机模拟器、平板电脑、笔记本电脑、数字广播接收器、MID(Mobile Internet Devices,移动互联网设备)、PDA(个人数字助理)、智能语音交互设备、智能家电、车载终端等。

[0083] 服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器或服务器集群。

[0084] 云计算(cloud computing)是一种计算模式,它将计算任务分布在大量计算机构成的资源池上,使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源

的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的,并且可以随时获取,按需使用,随时扩展,按使用付费。

[0085] 作为云计算的基础能力提供商,会建立云计算资源池(简称云平台,一般称为IaaS(Infrastructure as a Service,基础设施即服务)平台,在资源池中部署多种类型的虚拟资源,供外部客户选择使用。云计算资源池中主要包括:计算设备(为虚拟化机器,包含操作系统)、存储设备、网络设备。

[0086] 按照逻辑功能划分,在IaaS(Infrastructure as a Service,基础设施即服务)层上可以部署PaaS(Platform as a Service,平台即服务)层,PaaS层之上再部署SaaS(Software as a Service,软件即服务)层,也可以直接将SaaS部署在IaaS上。PaaS为软件运行的平台,如数据库、web容器等。SaaS为各式各样的业务软件,如web门户网站、短信群发器等。一般来说,SaaS和PaaS相对于IaaS是上层。

[0087] 所谓人工智能云服务,一般也被称作是AIaaS(AI as a Service,中文为“AI即服务”)。这是目前主流的一种人工智能平台的服务方式,具体来说AIaaS平台会把几类常见的AI服务进行拆分,并在云端提供独立或者打包的服务。这种服务模式类似于开了一个AI主题商城:所有的开发者都可以通过API接口的方式来接入使用平台提供的一种或者是多种人工智能服务,部分资深的开发者还可以使用平台提供的AI框架和AI基础设施来部署和运维自已专属的云人工智能服务。

[0088] 上述网络可以包括但不限于:有线网络,无线网络,其中,该有线网络包括:局域网、城域网和广域网,该无线网络包括:蓝牙、Wi-Fi及其他实现无线通信的网络。具体也可基于实际应用场景需求确定,在此不作限定。

[0089] 参见图2,图2示出了本申请实施例提供的一种文本纠错方法的流程示意图,其中,该方法可以由任一电子设备执行,如可以是服务器等;作为一可选实施方式,该方法可以由服务器执行,为了描述方便,在下文的一些可选实施例的描述中,将以服务器作为该方法执行主体为例进行说明。如图2所示,本申请实施例提供的文本纠错方法包括如下步骤:

[0090] S201,获取目标文本序列和候选词汇序列。

[0091] 具体地,获取拼接文本,拼接文本包括目标文本序列和候选词汇序列。例如,拼接文本input_texts=[“<CLS>”,“遇”,“到”,“逆”,“竟”,“时”,“,”,“我”,“们”,“必”,“须”,“勇”,“于”,“面”,“对”,“,”,“而”,“且”,“要”,“愈”,“挫”,“愈”,“勇”,“。”,“<SEP>”,“舞”,“蹈”,“误”,“导”,“道”,“理”,“逆”,“境”,“历”,“经”,“离”,“情”,“行”,“使”,“药”,“物”,“无”,“用”,“控”,“制”,“艰”,“辛”,“<SEP>”]。目标文本序列为:“<CLS>”,“遇”,“到”,“逆”,“竟”,“时”,“,”,“我”,“们”,“必”,“须”,“勇”,“于”,“面”,“对”,“,”,“而”,“且”,“要”,“愈”,“挫”,“愈”,“勇”,“。”。候选词汇序列为:“<SEP>”,“舞”,“蹈”,“误”,“导”,“道”,“理”,“逆”,“境”,“历”,“经”,“离”,“情”,“行”,“使”,“药”,“物”,“无”,“用”,“控”,“制”,“艰”,“辛”,“<SEP>”。

[0092] S202,确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征。

[0093] 具体地,可以通过拼接文本的特征向量,进行切分处理,得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0094] S203,基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字。

[0095] 具体地,例如,目标文本序列为:“<CLS>”,“遇”,“到”,“逆”,“竟”,“时”,“,”,“我”,“们”,“必”,“须”,“勇”,“于”,“面”,“对”,“,”,“而”,“且”,“要”,“愈”,“挫”,“愈”,“勇”,“。”。其中,目标文本序列中待纠错字为“竟”,待纠错字“竟”对应的预测的正确字为“境”。

[0096] S204,基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0097] 具体地,例如,目标文本序列中待纠错字为“竟”,纠错后的字为“境”,待纠错字“竟”对应的预测的正确字为“境”,预测的正确字“境”和纠错后的字为“境”相同,即预测成功。

[0098] 本申请实施例中,获取目标文本序列和候选词汇序列;确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字;如此,由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。

[0099] 在一个实施例中,获取目标文本序列和候选词汇序列,包括步骤A1-A5:

[0100] 步骤A1,获取目标文本。

[0101] 具体地,例如,如图3中所示的原始文本为“遇到逆竟时,我们必须勇于面对,而且要愈挫愈勇。”;原始文本为目标文本。

[0102] 步骤A2,基于目标文本,进行短语化处理,得到目标文本对应的短语。

[0103] 具体地,例如,将如图3中所示的原始文本bi-gram化,即进行短语化处理,得到ngram_list:[“遇到”,“到逆”,“逆竟”,“竟时”,“我们”,“们必”,“必须”,“须勇”,“勇于”,“于面”,“面对”,“而且”,“且要”,“要愈”,“愈挫”,“挫愈”,“愈勇”]。原始文本对应的短语例如ngram_list中各个词,各个词例如“遇到”、“到逆”、“逆竟”、“竟时”、“我们”、“们必”、“必须”、“须勇”、“勇于”、“于面”、“面对”、“而且”、“且要”、“要愈”、“愈挫”、“挫愈”和“愈勇”。原始文本为目标文本。

[0104] 步骤A3,将目标文本对应的短语,进行拼音转换处理,得到短语对应的拼音。

[0105] 具体地,例如,将ngram_list进行拼音转换处理,得到短语对应的拼音pinyin_ngram_list:[“yudao”,“daoni”,“nijing”,“jingshi”,“women”,“menbi”,“bixu”,“xuyong”,“yongyu”,“yumian”,“miandui”,“erqie”,“qieyao”,“yaoyu”,“yuchuo”,“yuyong”]。

[0106] 步骤A4,将短语对应的拼音和预设知识库中词汇进行匹配,从预设知识库中确定候选词汇。

[0107] 具体地,预设知识库例如kd树。通过短语对应的拼音pinyin_ngram_list,在kd树中搜索编辑距离为1的拼音对应的多个候选词汇,将多个候选词汇合并成一个列表cands=

[“舞蹈”,“误导”,“道理”,“逆境”,“历经”,“离情”,“行使”,“药物”,“无用”,“控制”,“艰辛”]。

[0108] 保存cands中各cand对应的目标文本序列中位置ori_pos=[(0,2),(0,2),(1,3),(2,4),(2,4),(2,4),(3,5),(17,19),(20,22),(23,25),(30,33)];其中,cand为候选词汇。

[0109] 步骤A5,基于目标文本和候选词汇,进行拼接处理,得到拼接文本,拼接文本包括目标文本序列和候选词汇序列。

[0110] 具体地,例如,拼接文本input_texts=[“<CLS>”,“遇”,“到”,“逆”,“竟”,“时”,“,”,“我”,“们”,“必”,“须”,“勇”,“于”,“面”,“对”,“,”,“而”,“且”,“要”,“愈”,“挫”,“愈”,“勇”,“。”,“<SEP>”,“舞”,“蹈”,“误”,“导”,“道”,“理”,“逆”,“境”,“历”,“经”,“离”,“情”,“行”,“使”,“药”,“物”,“无”,“用”,“控”,“制”,“艰”,“辛”,“<SEP>”]。目标文本序列为:“<CLS>”,“遇”,“到”,“逆”,“竟”,“时”,“,”,“我”,“们”,“必”,“须”,“勇”,“于”,“面”,“对”,“,”,“而”,“且”,“要”,“愈”,“挫”,“愈”,“勇”,“。”。候选词汇序列为:“<SEP>”,“舞”,“蹈”,“误”,“导”,“道”,“理”,“逆”,“境”,“历”,“经”,“离”,“情”,“行”,“使”,“药”,“物”,“无”,“用”,“控”,“制”,“艰”,“辛”,“<SEP>”。

[0111] 在一个实施例中,在获取目标文本之前,还包括步骤B1-B4:

[0112] 步骤B1,通过预设分词器,对预设的训练数据进行分词处理,得到分词集合。

[0113] 具体地,预设分词器例如jieba分词器,通过jieba分词器对训练数据进行分词,得到分词集合cutted_words=[舞蹈,误导,道理,而且,……]。

[0114] 步骤B2,基于分词集合,进行词性过滤处理,得到过滤后的分词集合,过滤后的分词集合中分词的类型包括名称、动词中至少一项。

[0115] 具体地,对分词集合cutted_words=[舞蹈,误导,道理,而且,……],进行词性过滤处理,得到过滤后的分词集合filtered_words=[舞蹈,误导,道理,……],过滤后的分词集合中一般只保留名词和动词。

[0116] 步骤B3,基于预设拼音库和过滤后的分词集合,确定过滤后的分词集合对应的拼音。

[0117] 具体地,预设拼音库例如pypinyin,pypinyin为Python中拼音库。基于预设拼音库pypinyin和过滤后的分词集合filtered_words,将过滤后的分词集合filtered_words转为过滤后的分词集合filtered_words对应的拼音filtered_pinyin2word。

[0118] 步骤B4,基于过滤后的分词集合对应的拼音,构建预设知识库。

[0119] 具体地,预设知识库例如kd树。例如,基于拼音的编辑距离,构建如图3中所示的kd树;其中,拼音为过滤后的分词集合对应的拼音,拼音的编辑距离是差异度,例如shi和si之间的差异度为1,1就是分数,基于这些分数构建kd树。

[0120] 在一个实施例中,确定目标文本序列的特征向量和候选词汇序列的特征向量,包括步骤C1-C2:

[0121] 步骤C1,基于拼接文本,进行特征提取处理,得到拼接文本的特征向量。

[0122] 具体地,例如,如图3所示,将拼接文本输入至BERT预训练语言模型(Pre-trained Language Model,PLM),进行特征提取处理,得到拼接文本的特征向量。

[0123] 步骤C2,基于拼接文本的特征向量,进行切分处理,得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0124] 具体地,例如,如图3所示,通过全连接层,按分隔符[SEP]对拼接文本的特征向量进行切分,分别得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0125] 在一个实施例中,基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,包括步骤D1-D4:

[0126] 步骤D1,基于候选词汇序列的特征向量,确定候选词汇序列的特征向量对应的未归一化概率。

[0127] 具体地,例如,如图4所示,将候选词汇序列的特征向量输入至一个(hidden_size,2)的全连接层,得到候选词汇序列的特征向量对应的未归一化概率(span_logits),span_logits的维度为(batch_size,seq_len,2);其中,hidden_size表示隐藏层的特征维度,batch_size表示有几句话,即句子个数维度,seq_len表示一句话有几个词,即句子长度维度;全连接层可以为MLP(MultiLayer Perceptron,多层感知机)。

[0128] 步骤D2,基于未归一化概率的候选词汇序列长度维度,进行归一化处理,得到归一化后的概率。

[0129] 具体地,例如,如图4所示,将未归一化概率(span_logits)的候选词汇序列长度维度seq_len,通过softmax进行归一化处理,得到归一化后的概率。

[0130] 步骤D3,基于归一化后的概率,确定候选词汇序列的目标词的起始位置概率和终止位置概率。

[0131] 具体地,将归一化后的概率中最后一维拆成2个probs,这2个probs分别为start_probs(目标词的起始位置概率)和end_probs(目标词的终止位置概率),probs的维度为(batch_size,seq_len)。

[0132] 步骤D4,基于目标词的起始位置概率,确定目标词在候选词汇序列中的起始位置,并基于目标词的终止位置概率,确定目标词在候选词汇序列中的终止位置,起始位置和终止位置相互匹配。

[0133] 具体地,将start_probs(目标词的起始位置概率)输入至第一分类器,并将end_probs(目标词的终止位置概率)输入至第二分类器,第一分类器输出为span_start(目标词在候选词汇序列中的起始位置),第二分类器输出为span_end(目标词在候选词汇序列中的终止位置),其中,第一分类器和第二分类器都为图4中所示的分类器。

[0134] 例如,start_probs(目标词的起始位置概率)输入至第一分类器,确定start_probs中最后一维概率最大的idx,idx为索引号,并将该idx作为span_start;end_probs(目标词的终止位置概率)输入至第二分类器,确定end_probs中最后一维概率最大的idx,idx为索引号,并将该idx作为span_end。又例如,若start_probs中最后一维概率最大的idx是3,end_probs中最后一维概率最大的idx是5,则“3-5这个短语”为预测出来的目标词;其中,idx表示句子中的位置,从句子的概率数组中选一个概率最大的idx作为目标词在句子中的起始位置或终止位置,句子为候选词汇序列。

[0135] 例如,可以将候选词汇序列的特征向量对应的未归一化概率span_logits通过全连接层和softmax进行一系列变换,得到位置匹配特征;基于位置匹配特征,通过分类器,预测起始位置和终止位置是否相互匹配。

[0136] 在一个实施例中,候选词汇序列的目标词在候选词汇序列中的位置包括目标词在候选词汇序列中的起始位置和目标词在候选词汇序列中的终止位置,基于目标词在候选词

汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字,包括步骤E1-E3:

[0137] 步骤E1,基于起始位置和终止位置,从候选词汇序列的特征向量中确定目标词的特征向量。

[0138] 具体地,基于起始位置(span_start)和终止位置(span_end),从候选词汇序列的特征向量中抽取目标词的特征向量(span_word_feature)。

[0139] 步骤E2,基于目标词的特征向量和目标文本序列的特征向量,进行拼接处理,得到纠错特征向量。

[0140] 具体地,基于span_start和span_end,通过ori_pos(匹配上的字所对应的目标文本序列中位置)取得对应的word_ori_pos(匹配上的字所处词对应的目标文本序列中位置);如此,可以将目标词的特征向量(span_word_feature)与相应的目标文本序列的特征向量(seq_feature)进行拼接,得到纠错特征向量(corr_feature);其中,纠错特征向量(corr_feature)的维度为(batch_size,seq_len,hidden_size*2)。

[0141] 例如,通过确定word_ori_pos,可以将“逆境”特征向量和目标文本序列的特征向量中“逆竞”特征向量拼接在一起;其中,word_ori_pos为“逆境”对应的目标文本序列中位置。

[0142] 步骤E3,基于纠错特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0143] 具体地,例如,如图5所示,将纠错特征向量(corr_feature)输入至一个(hidden_size,vocab_size)大小的全连接层,得到未归一化纠错概率(corr_logits);其中,未归一化纠错概率(corr_logits)的维度为(batch_size,seq_len,vocab_size),vocab_size表示字表大小维度,可以用字表表示整个汉字字库;对未归一化纠错概率(corr_logits)的字表大小维度(vocab_size)通过softmax,进行归一化处理,得到归一化处理后的概率向量;其中,归一化处理后的概率向量的维度为(batch_size,seq_len,vocab_size),batch_size、seq_len、vocab_size分别表示一段话的单词个数维度、截取的段数维度和字表大小维度;将归一化处理后的概率向量通过分类器,确定归一化处理后的概率向量中最大的idx,该idx为纠错后的字的corr_token_id,即纠错后的字的索引号;将纠错后的字的corr_token_id通过分词器tokenizer分词器,将corr_token_id解码为字表中的字,即得到纠错后的字;待纠错字为如图5中所示的“逆竞”中的“竞”,纠错后的字例如图5中所示的“逆境”中的“境”。

[0144] 在一个实施例中,基于纠错特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字,包括步骤F1-F4:

[0145] 步骤F1,基于纠错特征向量,确定未归一化纠错概率。

[0146] 具体地,将纠错特征向量(corr_feature)输入至一个(hidden_size,vocab_size)大小的全连接层,得到未归一化纠错概率(corr_logits);其中,未归一化纠错概率(corr_logits)的维度为(batch_size,seq_len,vocab_size),vocab_size表示字表大小维度,可以用字表表示整个汉字字库。

[0147] 步骤F2,基于未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量。

[0148] 具体地,对未归一化纠错概率(corr_logits)的字表大小维度(vocab_size)通过softmax,进行归一化处理,得到归一化处理后的概率向量;其中,归一化处理后的概率向量的维度为(batch_size,seq_len,vocab_size),batch_size、seq_len、vocab_size分别表示一段话的单词个数维度、截取的段数维度和字表大小维度。

[0149] 步骤F3,基于归一化处理后的概率向量,确定纠错后的字的索引号。

[0150] 具体地,将归一化处理后的概率向量通过分类器,确定归一化处理后的概率向量中最大的idx,该idx为纠错后的字的corr_token_id,即纠错后的字的索引号。

[0151] 步骤F4,基于纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字。

[0152] 具体地,分词器可以为tokenizer;例如,将纠错后的字的corr_token_id通过tokenizer,将corr_token_id解码为字表中的字,即得到纠错后的字。

[0153] 在一个实施例中,基于归一化处理后的概率向量,确定纠错后的字的索引号,包括:

[0154] 将归一化处理后的概率向量的各元素中最大元素,确定为纠错后的字的索引号。

[0155] 具体地,将归一化处理后的概率向量通过分类器,确定归一化处理后的概率向量的各元素中最大元素,即归一化处理后的概率向量中最大的idx,该idx为纠错后的字的索引号。

[0156] 应用本申请实施例,至少具有如下有益效果:

[0157] 由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。

[0158] 为了更好的理解本申请实施例所提供的方法,下面结合具体应用场景的示例对本申请实施例的方案进行进一步说明。

[0159] 本申请实施例所提供的方法可以应用到多种纯文本及多模态任务中,包括视频摘要、视频文本标签提取、多模态检索、OCR(Optical Character Recognition,文字识别)识别等领域,可以提升算法的效果,进而提升产品的体验;应用到不同的任务时,可以根据对应场景的训练数据生成知识库,提高下游任务的效果,也可以继续添加客户自定义的知识库,提升定制化纠错能力。

[0160] 在一个具体应用场景实施例中,例如文本纠错场景,参见图6,示出了一种文本纠错方法的处理流程,如图6所示,本申请实施例提供的文本纠错方法的处理流程包括如下步骤:

[0161] S601,服务器构建知识库。

[0162] 具体地,知识库例如图3中所示的kd树。例如,通过分词器,对训练数据进行分词处理,得到分词集合;基于分词集合,进行词性过滤处理,得到过滤后的分词集合,过滤后的分词集合中分词的类型包括名称和动词;基于预设拼音库和过滤后的分词集合,确定过滤后的分词集合对应的拼音;基于过滤后的分词集合对应的拼音,构建kd树。

[0163] S602,服务器获取目标文本。

[0164] 具体地,例如,如图3中所示的原始文本为“遇到逆竟时,我们必须勇于面对,而且要愈挫愈勇。”;原始文本为目标文本。

[0165] S603,服务器进行候选词汇匹配。

[0166] 具体地,基于目标文本,进行短语化处理,得到目标文本对应的短语;将目标文本对应的短语,进行拼音转换处理,得到短语对应的拼音;将短语对应的拼音和知识库中词汇进行匹配,从知识库中确定候选词汇。例如,如图3中所示的候选词汇匹配。

[0167] S604,服务器将目标文本和候选词汇进行拼接处理,得到拼接文本。

[0168] 具体地,例如,如图3中所示的候选词汇拼接。拼接文本包括目标文本序列和候选词汇序列。

[0169] S605,服务器基于拼接文本,确定目标文本序列的特征向量和候选词汇序列的特征向量。

[0170] 具体地,例如,如图3所示,将拼接文本输入至序列特征提取模块中的BERT预训练语言模型,进行特征提取,得到拼接文本的特征向量;通过序列特征提取模块中的全连接层,按分隔符[SEP]对拼接文本的特征向量进行切分,分别得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0171] S606,服务器基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的起始位置和终止位置。

[0172] 具体地,例如,如图4所示,将候选词汇序列的特征向量输入至候选词汇答案选择模块中的一个(hidden_size,2)的全连接层,得到候选词汇序列的特征向量对应的未归一化概率(span_logits),span_logits的维度为(batch_size,seq_len,2);其中,候选词汇答案选择模块为如图3所示的候选词汇答案选择模块,hidden_size表示隐藏层的特征维度,batch_size表示有几句话,即句子个数维度,seq_len表示一句话有几个词,即句子长度维度;全连接层可以为MLP(MultiLayer Perceptron,多层感知机)。例如,如图4所示,将未归一化概率(span_logits)的候选词汇序列长度维度seq_len,通过候选词汇答案选择模块中的softmax进行归一化处理,得到归一化后的概率。将归一化后的概率中最后一维拆成2个probs,这2个probs分别为start_probs(目标词的起始位置概率)和end_probs(目标词的终止位置概率),probs的维度为(batch_size,seq_len)。将start_probs(目标词的起始位置概率)输入至第一分类器,并将end_probs(目标词的终止位置概率)输入至第二分类器,第一分类器输出为span_start(目标词在候选词汇序列中的起始位置),第二分类器输出为span_end(目标词在候选词汇序列中的终止位置),其中,第一分类器和第二分类器都为图4中所示的候选词汇答案选择模块中的分类器。例如,如图4所示,将候选词汇序列的特征向量对应的未归一化概率span_logits通过全连接层和softmax进行一系列变换,得到位置匹配特征;基于位置匹配特征,通过候选词汇答案选择模块中分类器,预测起始位置和终止位置是否相互匹配。

[0173] S607,服务器基于起始位置和终止位置,从候选词汇序列的特征向量中确定目标词的特征向量。

[0174] S608,服务器基于目标词的特征向量和目标文本序列的特征向量,进行拼接处理,得到纠错特征向量。

[0175] S609,服务器基于纠错特征向量,确定未归一化纠错概率。

[0176] 具体地,例如,如图5所示,将纠错特征向量(corr_feature)输入至字表分类模块中一个(hidden_size,vocab_size)大小的全连接层,得到未归一化纠错概率(corr_logits);其中,字表分类模块为如图3所示的字表分类模块,未归一化纠错概率(corr_

logits)的维度为(batch_size,seq_len,vocab_size),vocab_size表示字表大小维度,可以用字表表示整个汉字字库。

[0177] S610,服务器基于未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量。

[0178] 具体地,例如,如图5所示,对未归一化纠错概率(corr_logits)的字表大小维度(vocab_size)通过字表分类模块中softmax,进行归一化处理,得到归一化处理后的概率向量;其中,归一化处理后的概率向量的维度为(batch_size,seq_len,vocab_size),batch_size、seq_len、vocab_size分别表示一段话的单词个数维度、截取的段数维度和字表大小维度。

[0179] S611,服务器基于归一化处理后的概率向量,确定纠错后的字的索引号。

[0180] 具体地,将归一化处理后的概率向量通过字表分类模块中分类器,确定归一化处理后的概率向量中最大的idx,该idx为纠错后的字的corr_token_id,即纠错后的字的索引号。

[0181] S612,服务器基于纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字。

[0182] 具体地,分词器可以为tokenizer。将纠错后的字的corr_token_id通过tokenizer,将corr_token_id解码为字表中的字,即得到纠错后的字;待纠错字为如图5中所示的“逆竞”中的“竞”,纠错后的字例如图5中所示的“逆境”中的“境”。

[0183] 应用本申请实施例,至少具有如下有益效果:

[0184] 针对多种应用场景,例如视频摘要、视频文本标签提取、多模态检索、OCR识别等,由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。能够利用外部词汇,例如知识库,引导模型进行纠错,这样既能提升模型在通用场景下的效果,又能根据特定场景的知识库进一步提升特定场景下的效果,其中,模型例如图3中所示的架构。例如,针对云平台通用纠错服务,通过大规模数据训练一个知识库,再提供一个自定义的知识库的接口,使用户能够根据业务需要,自行添加业务领域知识,并从候选词汇序列中确定目标词,并将该目标词参与目标文本序列中字的纠错,最终提升业务领域的纠错效果。能够根据对应业务需要,使用不同的知识库,如此,能快速适应多种应用场景的文本纠错,文本纠错例如中文拼写纠错。

[0185] 本申请实施例还提供了一种文本纠错装置,该文本纠错装置的结构示意图如图7所示,文本纠错装置70,包括第一处理模块701、第二处理模块702、第三处理模块703和第四处理模块704。

[0186] 第一处理模块701,用于获取目标文本序列和候选词汇序列;

[0187] 第二处理模块702,用于确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;

[0188] 第三处理模块703,用于基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词中包括预测的纠错后的字;

[0189] 第四处理模块704,用于基于目标词在候选词汇序列中的位置、目标文本序列的特

征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0190] 在一个实施例中,第一处理模块701,具体用于:

[0191] 获取目标文本;

[0192] 基于目标文本,进行短语化处理,得到目标文本对应的短语;

[0193] 将目标文本对应的短语,进行拼音转换处理,得到短语对应的拼音;

[0194] 将短语对应的拼音和预设知识库中词汇进行匹配,从预设知识库中确定候选词汇;

[0195] 基于目标文本和候选词汇,进行拼接处理,得到拼接文本,拼接文本包括目标文本序列和候选词汇序列。

[0196] 在一个实施例中,第一处理模块701,还用于:

[0197] 通过预设分词器,对预设的训练数据进行分词处理,得到分词集合;

[0198] 基于分词集合,进行词性过滤处理,得到过滤后的分词集合,过滤后的分词集合中分词的类型包括名称、动词中至少一项;

[0199] 基于预设拼音库和过滤后的分词集合,确定过滤后的分词集合对应的拼音;

[0200] 基于过滤后的分词集合对应的拼音,构建预设知识库。

[0201] 在一个实施例中,第二处理模块702,具体用于:

[0202] 基于拼接文本,进行特征提取处理,得到拼接文本的特征向量;

[0203] 基于拼接文本的特征向量,进行切分处理,得到目标文本序列的特征向量和候选词汇序列的特征向量。

[0204] 在一个实施例中,第三处理模块703,具体用于:

[0205] 基于候选词汇序列的特征向量,确定候选词汇序列的特征向量对应的未归一化概率;

[0206] 基于未归一化概率的候选词汇序列长度维度,进行归一化处理,得到归一化后的概率;

[0207] 基于归一化后的概率,确定候选词汇序列的目标词的起始位置概率和终止位置概率;

[0208] 基于目标词的起始位置概率,确定目标词在候选词汇序列中的起始位置,并基于目标词的终止位置概率,确定目标词在候选词汇序列中的终止位置,起始位置和终止位置相互匹配。

[0209] 在一个实施例中,候选词汇序列的目标词在候选词汇序列中的位置包括目标词在候选词汇序列中的起始位置和目标词在候选词汇序列中的终止位置,第四处理模块704,具体用于:

[0210] 基于起始位置和终止位置,从候选词汇序列的特征向量中确定目标词的特征向量;

[0211] 基于目标词的特征向量和目标文本序列的特征向量,进行拼接处理,得到纠错特征向量;

[0212] 基于纠错特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字。

[0213] 在一个实施例中,第四处理模块704,具体用于:

[0214] 基于纠错特征向量,确定未归一化纠错概率;

[0215] 基于未归一化纠错概率的字表大小维度,进行归一化处理,得到归一化处理后的概率向量;

[0216] 基于归一化处理后的概率向量,确定纠错后的字的索引号;

[0217] 基于纠错后的字的索引号,通过分词器进行解码处理,得到纠错后的字。

[0218] 在一个实施例中,第四处理模块704,具体用于:

[0219] 将归一化处理后的概率向量的各元素中最大元素,确定为纠错后的字的索引号。

[0220] 应用本申请实施例,至少具有如下有益效果:

[0221] 获取目标文本序列和候选词汇序列;确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字;如此,由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。

[0222] 本申请实施例还提供了一种电子设备,该电子设备的结构示意图如图8所示,图8所示的电子设备4000包括:处理器4001和存储器4003。其中,处理器4001和存储器4003相连,如通过总线4002相连。可选地,电子设备4000还可以包括收发器4004,收发器4004可以用于该电子设备与其他电子设备之间的数据交互,如数据的发送和/或数据的接收等。需要说明的是,实际应用中收发器4004不限于一个,该电子设备4000的结构并不构成对本申请实施例的限定。

[0223] 处理器4001可以是CPU(Central Processing Unit,中央处理器),通用处理器,DSP(Digital Signal Processor,数据信号处理器),ASIC(Application Specific Integrated Circuit,专用集成电路),FPGA(Field Programmable Gate Array,现场可编程门阵列)或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框,模块和电路。处理器4001也可以是实现计算功能的组合,例如包含一个或多个微处理器组合,DSP和微处理器的组合等。

[0224] 总线4002可包括一通路,在上述组件之间传送信息。总线4002可以是PCI(Peripheral Component Interconnect,外设部件互连标准)总线或EISA(Extended Industry Standard Architecture,扩展工业标准结构)总线等。总线4002可以分为地址总线、数据总线、控制总线等。为便于表示,图8中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0225] 存储器4003可以是ROM(Read Only Memory,只读存储器)或可存储静态信息和指令的其他类型的静态存储设备,RAM(Random Access Memory,随机存取存储器)或者可存储信息和指令的其他类型的动态存储设备,也可以是EEPROM(Electrically Erasable Programmable Read Only Memory,电可擦可编程只读存储器)、CD-ROM(Compact Disc

Read Only Memory,只读光盘)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质、其他磁存储设备、或者能够用于携带或存储计算机程序并能够由计算机读取的任何其他介质,在此不做限定。

[0226] 存储器4003用于存储执行本申请实施例的计算机程序,并由处理器4001来控制执行。处理器4001用于执行存储器4003中存储的计算机程序,以实现前述方法实施例所示的步骤。

[0227] 其中,电子设备包括但不限于:服务器等。

[0228] 应用本申请实施例,至少具有如下有益效果:

[0229] 获取目标文本序列和候选词汇序列;确定目标文本序列的特征向量和候选词汇序列的特征向量,目标文本序列的特征向量用于表征目标文本序列中各个字的特征,候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征;基于候选词汇序列的特征向量,确定候选词汇序列的目标词在候选词汇序列中的位置,目标词包括目标文本序列中待纠错字对应的预测的正确字;基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量,对目标文本序列中待纠错字进行纠错,确定纠错后的字;如此,由于目标词包括目标文本序列中待纠错字对应的预测的正确字,从候选词汇序列中确定该目标词,并将该目标词参与目标文本序列中字的纠错,从而提高了目标文本序列中文本纠错的效率和准确度。

[0230] 本申请实施例提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,计算机程序被处理器执行时可实现前述方法实施例的步骤及相应内容。

[0231] 本申请实施例还提供了一种计算机程序产品,包括计算机程序,计算机程序被处理器执行时可实现前述方法实施例的步骤及相应内容。

[0232] 基于与本申请实施例提供的方法相同的原理,本申请实施例还提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述本申请任一可选实施例中提供的方法。

[0233] 应该理解的是,虽然本申请实施例的流程图中通过箭头指示各个操作步骤,但是这些步骤的实施顺序并不受限于箭头所指示的顺序。除非本文中有明确的说明,否则在本申请实施例的一些实施场景中,各流程图中的实施步骤可以按照需求以其他的顺序执行。此外,各流程图中的部分或全部步骤基于实际的实施场景,可以包括多个子步骤或者多个阶段。这些子步骤或者阶段中的部分或全部可以在同一时刻被执行,这些子步骤或者阶段中的每个子步骤或者阶段也可以分别在不同的时刻被执行。在执行时刻不同的场景下,这些子步骤或者阶段的执行顺序可以根据需求灵活配置,本申请实施例对此不限制。

[0234] 以上所述仅是本申请部分实施场景的可选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本申请的方案技术构思的前提下,采用基于本申请技术思想的其他类似实施手段,同样属于本申请实施例的保护范畴。

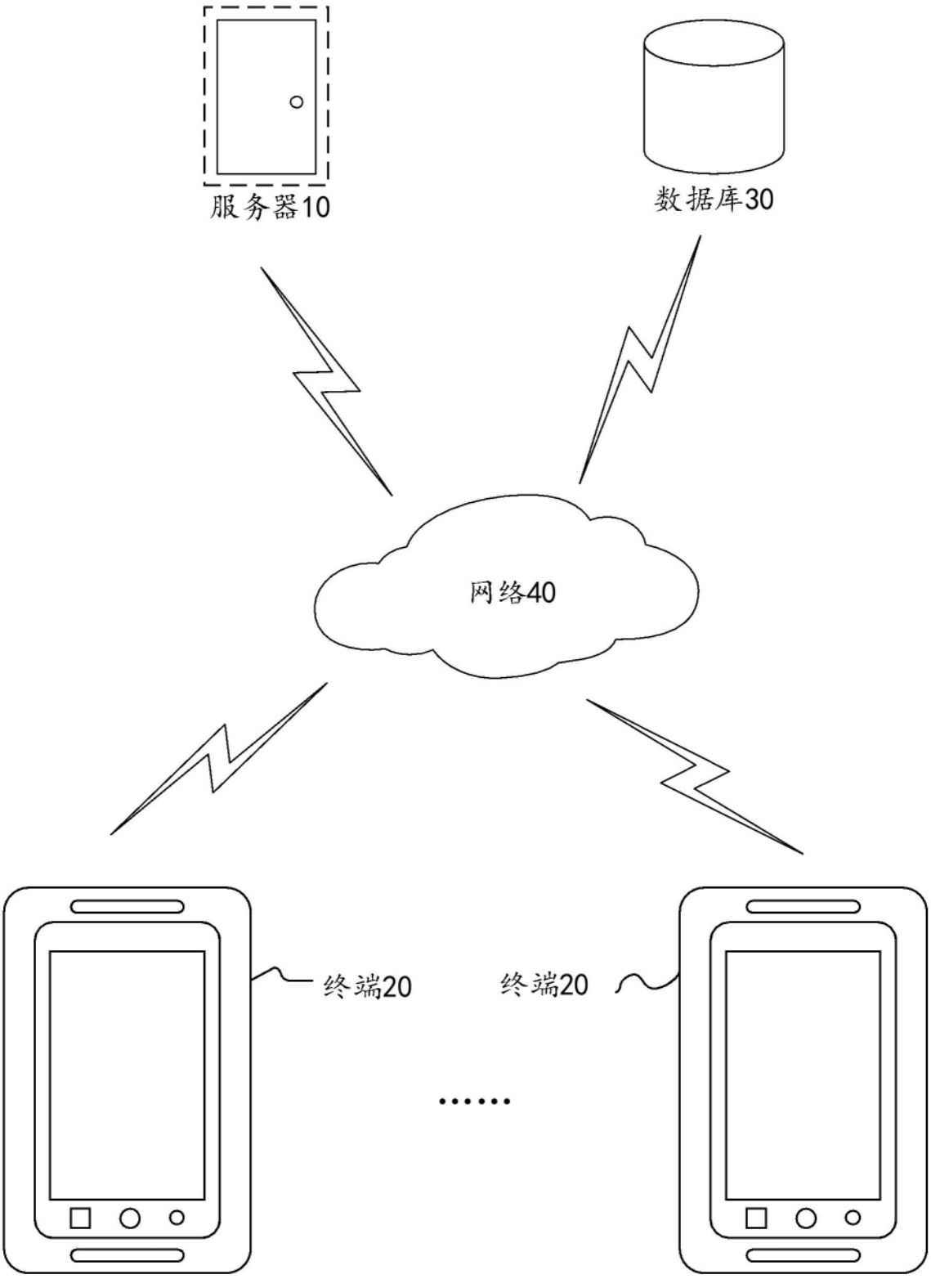


图1

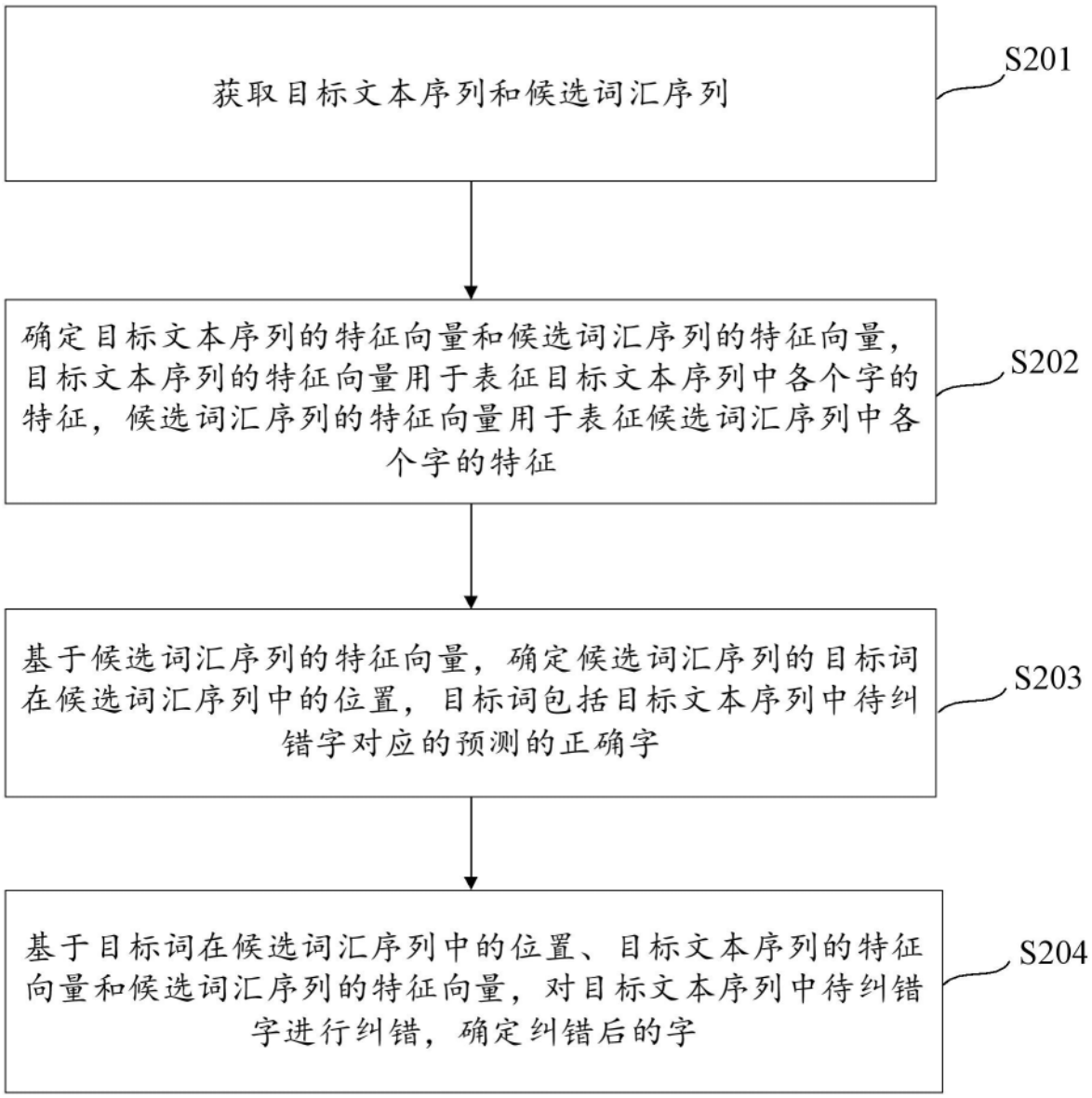

获取目标文本序列和候选词汇序列
S201
确定目标文本序列的特征向量和候选词汇序列的特征向量，目标文本序列的特征向量用于表征目标文本序列中各个字的特征，候选词汇序列的特征向量用于表征候选词汇序列中各个字的特征
S202
基于候选词汇序列的特征向量，确定候选词汇序列的目标词在候选词汇序列中的位置，目标词包括目标文本序列中待纠错字对应的预测的正确字
S203
基于目标词在候选词汇序列中的位置、目标文本序列的特征向量和候选词汇序列的特征向量，对目标文本序列中待纠错字进行纠错，确定纠错后的字
S204

图2

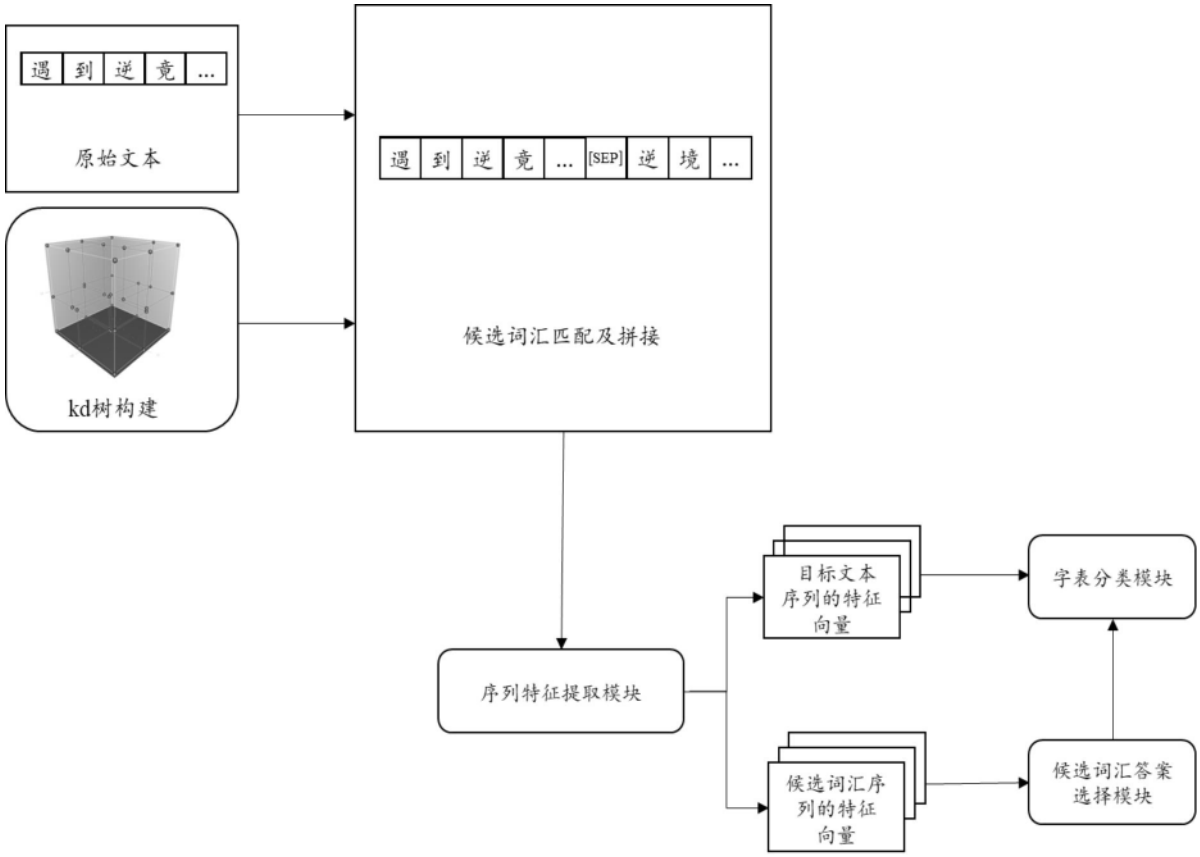

遇
到
逆
竟
...
原始文本
kd树构建
遇
到
逆
竟
...
[SEP]
逆
境
...
候选词汇匹配及拼接
序列特征提取模块
目标文本序列的特征向量
候选词汇序列的特征向量
字表分类模块
候选词汇答案选择模块

图3

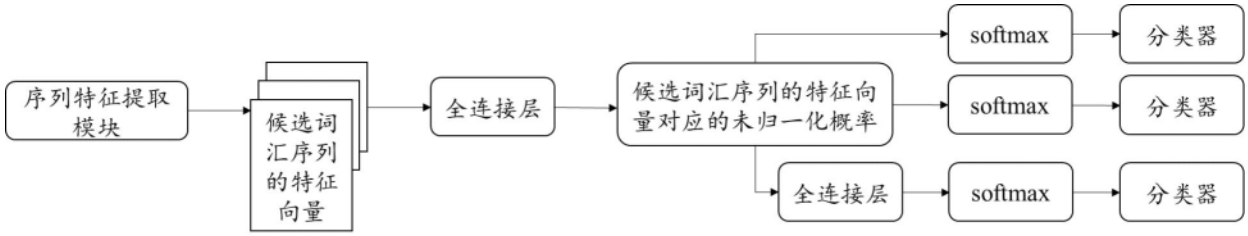

序列特征提取模块
候选词汇序列的特征向量
全连接层
候选词汇序列的特征向量对应的未归一化概率
softmax
softmax
softmax
分类器
分类器
分类器
全连接层

图4

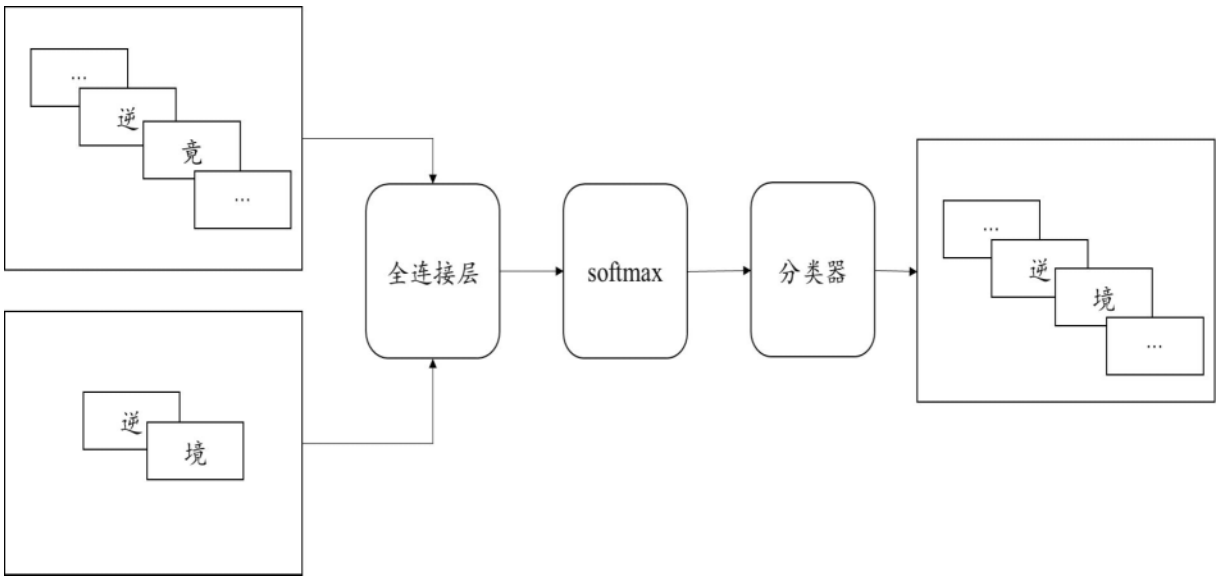


图5

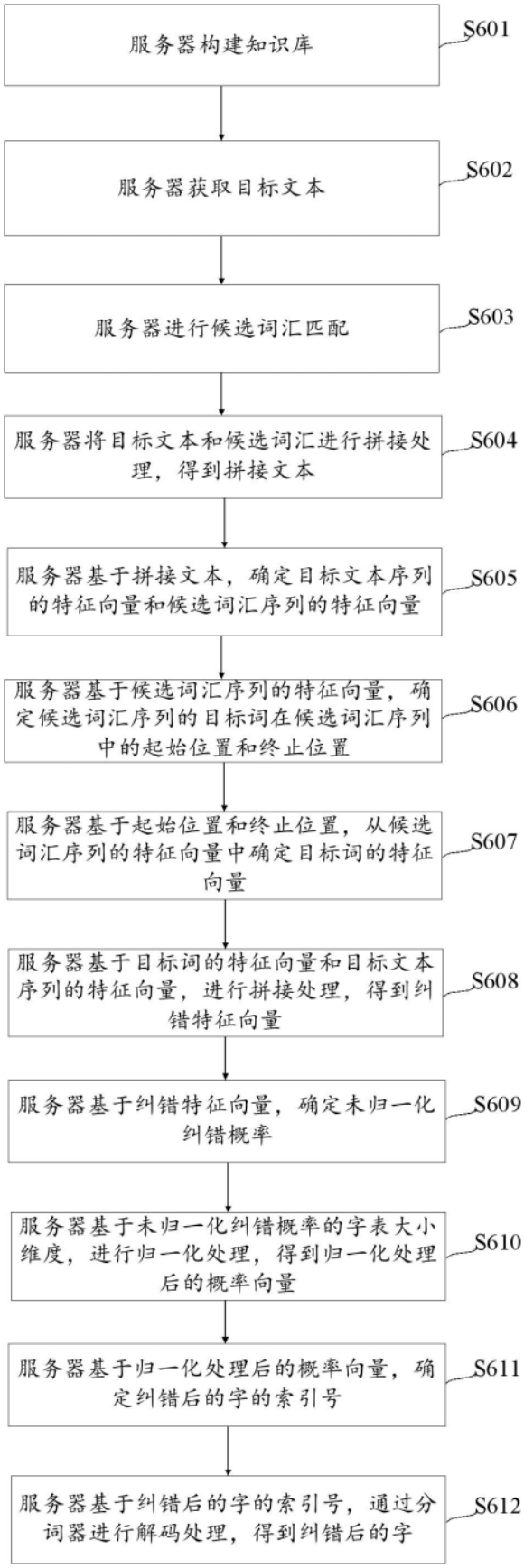

服务器构建知识库
S601
服务器获取目标文本
S602
服务器进行候选词汇匹配
S603
服务器将目标文本和候选词汇进行拼接处理，得到拼接文本
S604
服务器基于拼接文本，确定目标文本序列的特征向量和候选词汇序列的特征向量
S605
服务器基于候选词汇序列的特征向量，确定候选词汇序列的目标词在候选词汇序列中的起始位置和终止位置
S606
服务器基于起始位置和终止位置，从候选词汇序列的特征向量中确定目标词的特征向量
S607
服务器基于目标词的特征向量和目标文本序列的特征向量，进行拼接处理，得到纠错特征向量
S608
服务器基于纠错特征向量，确定未归一化纠错概率
S609
服务器基于未归一化纠错概率的字表大小维度，进行归一化处理，得到归一化处理后的概率向量
S610
服务器基于归一化处理后的概率向量，确定纠错后的字的索引号
S611
服务器基于纠错后的字的索引号，通过分词器进行解码处理，得到纠错后的字
S612

图6

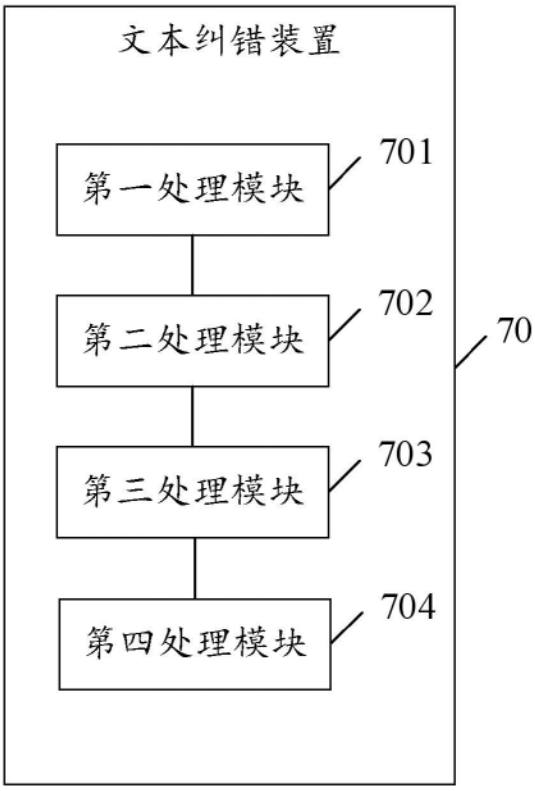

文本纠错装置
第一处理模块
701
第二处理模块
702
70
第三处理模块
703
第四处理模块
704

图7

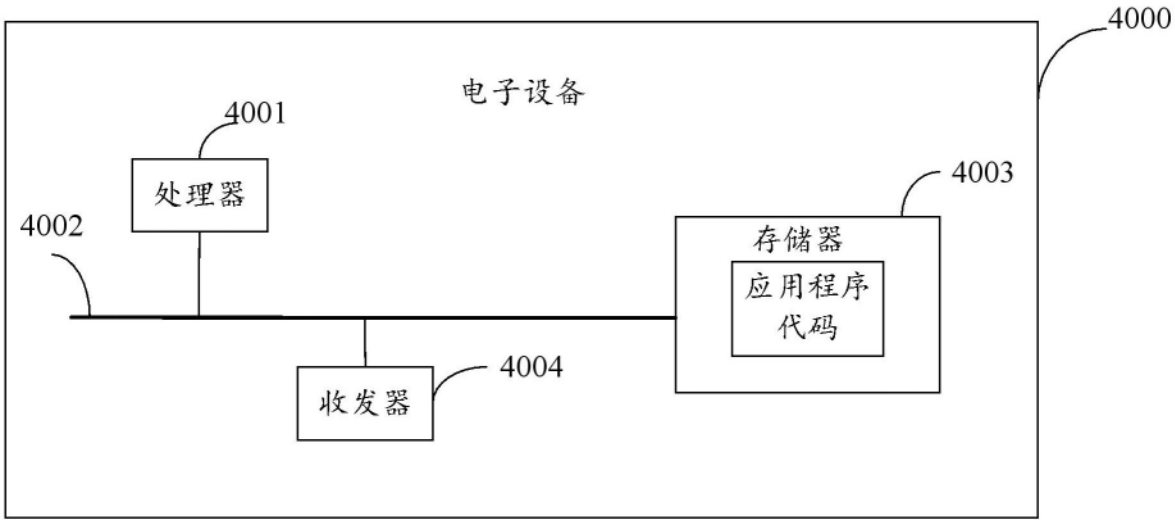

4000
电子设备
4001
处理器
4002
4003
存储器
应用程序
代码
4004
收发器

图8