

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 8 月 11 日 (11.08.2022)



(10) 国际公布号
WO 2022/166808 A1

(51) 国际专利分类号:
G06F 40/166 (2020.01)

(21) 国际申请号: PCT/CN2022/074583

(22) 国际申请日: 2022 年 1 月 28 日 (28.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110158872.0 2021年2月4日 (04.02.2021) CN

(71) 申请人: 维沃移动通信有限公司 (VIVO
MOBILE COMMUNICATION CO., LTD.) [CN/CN];
中国广东省东莞市长安镇维沃路 1 号,
Guangdong 523863 (CN)。

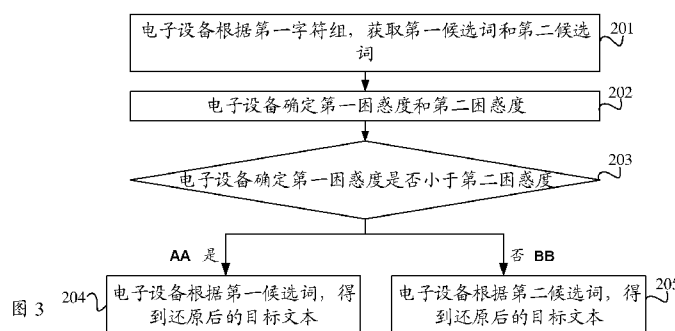
(72) 发明人: 佟禹 (TONG, Yu); 中国广东省东莞市长
安镇维沃路 1 号, Guangdong 523863 (CN)。

(74) 代理人: 北京远志博慧知识产权代理事务所(普通
合伙) (BOHUI INTELLECTUAL PROPERTY); 中
国北京市海淀区交大东路 31 号东区 10 号楼等
17 幢 31 幢 108, Beijing 100044 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: TEXT RESTORATION METHOD AND APPARATUS, AND ELECTRONIC DEVICE

(54) 发明名称: 文本还原方法、装置及电子设备



201 An electronic device obtains first candidate words and second candidate words according to a first character group

202 The electronic device determines a first confusion degree and a second confusion degree

203 The electronic device determines whether the first confusion degree is smaller than the second confusion degree

204 The electronic device obtains restored target text according to the first candidate words

205 The electronic device obtains restored target text according to the second candidate words

AA Yes

BB No

(57) Abstract: The present application relates to the technical field of language recognition and discloses a text restoration method and apparatus, and an electronic device. The text restoration method comprises: obtaining first candidate words and second candidate words according to a first character group; determining a first confusion degree and a second confusion degree, the first confusion degree being the confusion degree corresponding to a first sentence obtained by replacing the first character group and a second character group in a target sentence with the first candidate words, and the second confusion degree being the confusion degree corresponding to

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

a second sentence obtained by replacing the first character group and the second character group in the target sentence with the second candidate words; if the first confusion degree is smaller than the second confusion degree, obtaining restored target text according to the first candidate words; or if the second confusion degree is smaller than the first confusion degree, obtaining restored target text according to the second candidate words. The present method is applied in text restoration scenarios.

(57) 摘要: 本申请公开了一种文本还原方法、装置及电子设备, 属于语言识别技术领域。该文本还原方法包括: 根据第一字符组, 获取第一候选词和第二候选词; 确定第一困惑度和第二困惑度, 第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度, 第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度; 在第一困惑度小于第二困惑度的情况下, 根据第一候选词, 得到还原后的目标文本; 或在第二困惑度小于第一困惑度的情况下, 根据第二候选词, 得到还原后的目标文本。该方法应用于还原文本的场景中。

文本还原方法、装置及电子设备

相关申请的交叉引用

本申请主张在 2021 年 02 月 04 日在中国提交的中国专利申请号为
5 202110158872.0 的优先权，其全部内容通过引用包含于此。

技术领域

本申请属于语言识别技术领域，具体涉及一种文本还原方法、装置及电子设备。

背景技术

10 在编辑文本的过程中，若位于该文本的某一行行末的西文字符组（例如英文字符组）不能全部显示在该行时，则可以将该西文字符组从自动换行的位置断开，并在断行的位置添加一个分隔符，例如图 1 中的标记处 1，标记处 2，标记处 3，标记处 4，标记处 5，标记处 6。

目前，如果将上述文本复制到另外一个文件中，那么可以根据这些分隔符
15 自动还原字符组。具体的，可以直接去掉位于该文本行末的分隔符，使得分隔符前后的字符组组成一个字符组，并显示在复制得到的文本中，例如图 2 所示的文本即为图 1 所示的文本复制之后得到的文本。

然而，在上述过程中，由于有的字符组是复合词，即该字符组本身是包括
20 分隔符的，因此通过直接去掉分隔符的方式，可能会导致还原后的文本中的字符组有误，例如图 2 中的标记处 3，标记处 5，标记处 6 标记的字符组。因此，如何准确还原文本成为一个亟待解决的问题。

发明内容

本申请实施例的目的是提供一种文本还原方法、装置及电子设备，能够解决现有的电子设备还原文本不准确的问题。

25 为了解决上述技术问题，本申请是这样实现的：

第一方面，本申请实施例提供了一种文本还原方法，该方法包括：根据第一字符组，获取第一候选词和第二候选词，第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，第一候选词为第一字符组与第二字符组合得到的词，第二候选词为第三字符组与第二字符组合得到的词，第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组，第三字符组为第一字符组除去分隔符后得到的字符组；确定第一困惑度和第二困惑度，第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度；在第一困惑度小于第二困惑度的情况下，根据第一候选词，得到还原后的目标文本；或在第二困惑度小于第一困惑度的情况下，根据第二候选词，得到还原后的目标文本。

第二方面，本申请实施例提供了一种文本还原装置，该文本还原装置包括获取模块，确定模块和还原模块。获取模块，用于根据第一字符组，获取第一候选词和第二候选词，第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，第一候选词为第一字符组与第二字符组合得到的词，第二候选词为第三字符组与第二字符组合得到的词，第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组，第三字符组为第一字符组除去分隔符后得到的字符组；确定模块，用于确定第一困惑度和第二困惑度，第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度；还原模块，用于在第一困惑度小于第二困惑度的情况下，根据第一候选词，得到还原后的目标文本；或在第二困惑度小于第一困惑度的情况下，根据第二候选词，得到还原后的目标文本。

第三方面，本申请实施例提供了一种电子设备，该电子设备包括处理器、存储器及存储在存储器上并可在处理器上运行的程序或指令，该程序或指令被处理器执行时，实现如上述第一方面中的文本还原方法的步骤。

第四方面，本申请实施例提供了一种可读存储介质，该可读存储介质上存

储程序或指令，该程序或指令被处理器执行时，实现如上述第一方面中的文本还原方法的步骤。

第五方面，本申请实施例提供了一种芯片，该芯片包括处理器和通信接口，通信接口和处理器耦合，处理器用于运行程序或指令，实现如上述第一方面中的文本还原方法的步骤。

在本申请实施例中，可以根据第一字符组，获取第一候选词和第二候选词，第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，第一候选词为第一字符组与第二字符组合得到的词，第二候选词为第三字符组与第二字符组合得到的词，第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组，第三字符组为第一字符组除去分隔符后得到的字符组；确定第一困惑度和第二困惑度，第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度；在第一困惑度小于第二困惑度的情况下，根据第一候选词，得到还原后的目标文本；或在第二困惑度小于第一困惑度的情况下，根据第二候选词，得到还原后的目标文本。通过该方案，由于语句对应的困惑度越小，表示语句越流畅，即语句对应的困惑度越小，语句越准确，因此通过比较根据第一候选词得到的第一语句对应的困惑度和根据第二候选词得到的第二语句对应的困惑度，可以确定第一候选词和第二候选词哪个是正确地，即可以确定目标文本中的第一字符组和第二字符组组成的正确的词，从而可以准确地还原文本。

附图说明

图 1 为本申请实施例提供的一种待还原的文本示意图；

图 2 为本申请实施例提供的一种还原后的文本示意图；

图 3 为本申请实施例提供的文本还原方法的流程示意图；

图 4 为本申请实施例提供的文本还原装置的结构示意图；

图 5 为本申请实施例提供的电子设备的结构示意图；

图 6 为本申请实施例提供的电子设备的硬件示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不用来描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”等所区分的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”，一般表示前后关联对象是一种“或”的关系。

在本申请实施例中，“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请实施例中被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

下面结合附图，通过具体的实施例及其应用场景，对本申请实施例提供的文本还原方法进行详细地说明。

如图3所示，本申请实施例提供一种文本还原方法，该方法包括下述的步骤201-步骤204，或步骤201-步骤203和步骤205。

需要说明的是，本申请实施例提供的文本还原方法的执行主体可以为文本还原装置，或者该文本还原装置中的用于执行文本还原方法的控制模块，还可以为电子设备。下面将以文本还原装置为例，对本申请实施例提供的文本还原方法进行示例性的说明。

可选地，本申请实施例中，当本申请实施例提供的文本还原方法的执行主体为电子设备时，该电子设备可以包括本申请实施例提供的文本还原装置，或外接该文本还原装置。具体可以根据实际使用需求确定，本申请实施例不作限

定。

步骤 201, 电子设备根据第一字符组, 获取第一候选词和第二候选词。

其中, 上述第一字符组可以为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组, 第一候选词为第一字符组与第二字符组合得到的词, 第二候选词为第三字符组与第二字符组合得到的词, 第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组, 第三字符组为第一字符组除去分隔符后得到的字符组, N 为正整数。

本申请实施例中, 在电子设备获取上述待还原的目标文本之后, 电子设备可以根据上述第一字符组, 获取上述第一候选词和第二候选词, 从而可以根据该第一候选词和第二候选词中正确地词, 还原目标文本。

可选地, 本申请实施例提供的文本还原方法可以应用于下述两种可能的场景中:

场景一: 电子设备将目标文本从一个位置复制至另一个位置, 例如将目标文本从一个文档复制至另一个文档。

场景二: 目标文本为目标图像中的文本, 电子设备通过光学字符(optical character recognition, OCR)技术识别该目标图像中的文本。

可选地, 在上述场景二中, 目标图像中的文本可以是横向排版, 也可以是竖向排版。当目标图像中的文本是竖向排版时, 上述第一字符组可以为待还原的目标文本中的第 M 列的列末、且以分隔符结尾的字符组, 第二字符组为待还原的目标文本中的第 M+1 列的第一个字符组, M 为正整数。

当然, 实际实现时, 本申请实施例提供的文本还原方法还可以应用于其他任意可能的场景中, 具体可以根据实际使用需求确定, 本申请实施例不作限定。

可选地, 本申请实施例涉及的字符组可以为西文字符组, 例如英文字符组、法文字符组、德文字符组、俄文字符组或葡萄牙文字符组等, 具体可以根据实际使用需求确定, 本申请实施例不作限定。其中, 本申请实施例是以英文字符组为例进行示例性说明的。

本申请实施例中, 在电子设备获得上述待还原的目标文本之后, 电子设备

可以逐行检测该目标文本中的每行文本的行末是否以分隔符或特定分割符（如“-”）结束，如果是，那么电子设备可以将包括该分隔符的字符组作为上述第一字符组。如果不是，那么电子设备可以继续检测下一行的文本。

可选地，本申请实施例中，电子设备获取上述第一候选词和第二候选词的方式可以为：

步骤 1、电子设备将当前行的行末的分隔符之前的字符组（即上述第三字符组），分隔符（如“-”），以及当前行的下一行的第一个字符组组成一个单元（以下称为处理候选集）。

示例性地，以图 1 所示的文本为例，电子设备可以从第一行得到待处理候选集 {representa, -, tion}，从第四行得到待处理候选集 {repre, -, sentation}，从第六行得到待处理候选集 {pre, -, train}，从第九行得到待处理候选集 {re, -, sult}，从第十行得到待处理候选集 {fine, -, tuned}，从第十四行得到待处理候选集 {task, -, specific}。

步骤 2、对每个待处理候选集中的所有候选项，合并分隔符前后的字符组，得到候选词，例如 {representation}，{representation}，{pretrain}，{result}，{finetuned} 和 {taskspecific}，即可以得到第一候选词。

步骤 3、] 对每个待处理候选集中的所有候选项，生成保留分割符的复合词候，例如 {representa-tion}，{repre-sentation}，{pre-train}，{re-sult}，{fine-tuned} 和 {task-specific}，即可以得到上述第二候选词。

可以理解，本申请实施例中，第二候选词为复合词，如此电子设备可以采用分隔符前后字符组合并后的词和由该分隔符形成的复合词，分别进行词的合法性检测以及句子流畅度检测，从而可以保证还原后的目标文本的准确度。

步骤 202，电子设备确定第一困惑度和第二困惑度。

其中，上述第一困惑度可以为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度。

本申请实施例中，在电子设备获取上述第一候选词和第二候选词之后，电

子设备可以确定上述第一困惑度和第二困惑度，从而可以从上述第一候选词和第二候选词中确定正确的词，以还原上述目标文本。

可选地，本申请实施例中，对于上述步骤 202，电子设备可以对上述第一候选词和第二候选词分别执行下述的步骤 202a 和步骤 202b，从而可以确定上述第一困惑度和第二困惑度。

可以理解，下述的步骤 202a 和步骤 202b 是以上述第一候选词和第二候选词中的一个候选词（例如本申请实施例中的目标候选词）进行示例性说明的。

步骤 202a，电子设备基于目标候选词中的每个字符在目标文本中出现的概率，确定目标参数。

其中，上述目标候选词可以为上述第一候选词或上述第二候选词。

步骤 202b，电子设备根据目标参数，确定目标候选词对应的困惑度。

其中，上述目标参数可以包括：目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值。目标词组可以包括目标候选词、第四字符组和第五字符组，第四字符组可以为目标文本中位于第一字符组之前的字符组，第五字符组为目标文本中位于第二字符组之后的字符组。

本申请实施例中，电子设备可以基于目标候选词（第一候选词或第二候选词）中的每个字符在目标文本中出现的概率，确定目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值，从而可以得到上述目标参数，然后电子设备可以根据该目标参数，确定目标候选词对应的困惑度（例如上述第一困惑度或第二困惑度）。

本申请实施例中，电子设备可以将目标候选词（第一候选词或第二候选词）和目标文本输入语言模型，然后语言模型可以进行目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值的计算，从而可以得到上述目标参数。

本申请实施例中，上述目标候选词的合法性值可以为目标文本中，目标候选词在目标文本中出现的概率（记为 Score₁）。

可选地，本申请实施例中，目标候选词的合法性值可以为目标候选词中的

每个字符在目标文本中出现的概率之间的乘积。

其中，目标候选词中的第 K 个字符在目标文本中出现的概率是指：在目标文本中出现第六字符组的情况下出现第 K 个字符的概率，第六字符组由目标候选词中的第 1 个字符至第 $(K-1)$ 个字符组成， K 为大于 1 的整数。

5 需要说明的是，本申请实施例中涉及的“在出现某一字符组或字符（记为 A ）的情况下，出现另一字符（记为 B ）”是指：在文本中， B 位于 A 之后，且 A 与 B 之间没有分隔符。

具体地，目标候选词的合法性值可以表示为：

$$P(W)=p(C_1) \times p(C_2|C_1) \times \dots \times p(C_K|C_1,C_2,\dots C_{K-1});$$

10 其中， $P(W)$ 表示目标候选词的合法性值， $p(C_1)$ 表示目标候选词中的第一个字符在目标文本中出现的概率， $p(C_K|C_1,C_2,\dots C_{K-1})$ 表示在目标文本中出现第六字符组的情况下出现第 K 个字符的概率的乘积，第六字符组由目标候选词中的第 1 个字符至第 $(K-1)$ 个字符组成。

15 示例性地，通过语言模型来判断，语言模型见下述公式（1）中， W 表示一个候选词， C_1 表示候选词中的第一个字符， C_k 表示候选词中的最后一个字符，通过计算由 C_1 至 C_k 的字符组成候选词 W 的概率，判断 W 是否为一个合法的单词。其中，计算单词的概率公式见下述公式（2），其中 $p(C_1)$ 表示字符 C_1 在目标文本中出现的概率，计算公式见下述公式（3）。示例性地，如果 C_1 表示字符 r ，目标文本中的字符总数为 100，字符 r 出现了 10 次，那么 r 出现的概率为 $10/100=0.1$ ，即 $p(C_1)=0.1$ 。

公式(4)中， $p(C_2|C_1)$ 表示 C_2 的出现与 C_1 是相关的，即在出现 C_1 的条件下出现 C_2 的概率。示例性地， C_1 如果表示字符“w”， C_2 表示字符“e”，那么在出现字符“w”的条件下出现字符“e”的概率出现为： $P(e|w)=P(we)/P(w)$ 。

$$W=C_1,C_2,C_3,\dots C_K \quad (1)$$

$$25 \quad P(W)=P(C_1,C_2,C_3,\dots C_K)=p(C_1) \times p(C_2|C_1) \times \dots \times p(C_K|C_1,C_2,\dots C_{K-1}) \quad (2)$$

$$p(C_k)=\text{字符 } k \text{ 出现的次数/文档总字符数} \quad (3)$$

$$P(C_2|C_1)=P(C_1C_2)/P(C_1) \quad (4)$$

本申请实施例中，上述目标词组的流畅度值可以为目标候选词、第四字符组和第五字符组组成的词组在目标文本中出现的概率（记为 Score₂）。

本申请实施例中，上述目标词组的流畅度值可以根据下述的公式（5）计算得到。

$$5 \quad \text{PPL}(S)=P(W_1 W_2 \dots W_N)^{\frac{1}{N}} \quad (5)$$

其中，S 代表一个句子或词组，由单词 $W_1 \dots W_N$ 组成。一般困惑度越小，句子或词组越流畅。

示例性地，如图 1 中的标记 1 所示，假设上述目标候选词为“representation”，位于“representa-”之前的单词为“language”，位于“tion”之后的单词为“model”，
10 那么通过公式（5）可以得到 $\text{Score}_2=P(\text{language}, \text{representation}, \text{model})^{\frac{1}{3}}$ 。

本申请实施例中，上述目标语句的流畅度值可以为目标语句在目标文本中出现的概率（记为 Score₃）。

本申请实施例中，上述目标语句的流畅度值可以根据上述公式（5）计算得到。

15 示例性地，如图 1 中的标记 1 所示，假设上述目标候选词为“representation”，“representa-”所在的语句为“We introduce a new language representation model called BERT”，那么通过公式（5）可以得到：

$$\text{Score}_2=P(\text{We introduce a new language representation model called BERT})^{\frac{1}{9}}$$

20 可选地，本申请实施例中，上述步骤 202b 具体可以通过下述的步骤 202b1 实现。

步骤 202b1，电子设备根据目标候选词的合法性值与第一系数的乘积、目标词组的流畅度值与第二系数的乘积、目标语句的流畅度值与第三系数的乘积之和，得到目标候选词对应的困惑度。

其中，上述第一系数、第二系数和第三系数之和等于 1。

25 本申请实施例中，在电子设备确定上述目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值之后，电子设备可以法计算目标候选词的合法性值与第一系数（记为 α ）的乘积、目标词组的流畅度值与第二系数（记为

β) 的乘积、目标语句的流畅度值与第三系数(记为 γ) 的乘积之和,从而可以得到目标候选词对应的困惑度(记为 Score)。

即, $\text{Score} = \alpha \times \text{Score}_1 + \beta \times \text{Score}_2 + \gamma \times \text{Score}_3$ 。

可选地,本申请实施例中,上述第一系数、第二系数和第三系数的取值可以
5 以为任意可能的正数,且第一系数、第二系数和第三系数之和等于 1。

步骤 203,电子设备确定第一困惑度是否小于第二困惑度。

本申请实施例中,在电子设备确定第一困惑度和第二困惑度之后,电子设备可以比较第一困惑度和第二困惑度的大小。从而确定上述第一候选词和第二候选词中的哪个候选词是正确的。

10 本申请实施例中,若第一困惑度小于第二困惑度,则电子设备可以根据第一候选词,得到还原后的目标文本,即在第一困惑度小于第二困惑度的情况下,电子设备可以执行下述的步骤 204。若第二困惑度小于第一困惑度,则电子设备可以根据第二候选词,得到还原后的目标文本,即在第二困惑度小于第一困惑度的情况下,电子设备可以执行下述的步骤 205。

15 可以理解,本申请实施例中,下述步骤 204 和步骤 205 是择一执行的。

步骤 204,电子设备根据第一候选词,得到还原后的目标文本。

本申请实施例中,在第一困惑度小于第二困惑度的情况下,电子设备可以根据第一候选词,还原目标文本,从而可以得到还原后的目标文本。

20 一种可能的实现方式,电子设备可以直接采用第一候选词,替换目标文本中的第一字符组和第二字符组,从而可以得到还原后的目标文本。

另一种可能的实现方式,电子设备可以采用上述第一语句(第一语句包括第一候选词),替换目标文本中的目标语句,从而可以得到还原后的目标文本。

步骤 205,电子设备根据第二候选词,得到还原后的目标文本。

25 本申请实施例中,在第二困惑度小于第一困惑度的情况下,电子设备可以根据第二候选词,还原目标文本,从而可以得到还原后的目标文本。

一种可能的实现方式,电子设备可以直接采用第二候选词,替换目标文本中的第一字符组和第二字符组,从而可以得到还原后的目标文本。

另一种可能的实现方式，电子设备可以采用上述第二语句（第二语句包括第二候选词），替换目标文本中的目标语句，从而可以得到还原后的目标文本。

本申请实施例提供的文本还原方法，由于语句对应的困惑度越小，表示语句越流畅，即语句对应的困惑度越小，语句越准确，因此通过比较根据第一候选词得到的第一语句对应的困惑度和根据第二候选词得到的第二语句对应的困惑度，可以确定第一候选词和第二候选词哪个是正确地，即可以确定目标文本中的第一字符组和第二字符组组成的正确的词，从而可以准确地还原文本。

可选地，本申请实施例中，在电子设备得到还原后的目标文本之后，本申请实施例提供的文本还原方法还可以包括下述的步骤 206。

10 步骤 206，电子设备基于关键词识别模型，获取还原后的目标文本的关键词。

其中，上述关键词的内容类型可以与关键词识别模型中预设的内容类型相同。

15 本申请实施例中，在电子设备得到还原后的目标文本之后，电子设备可以将还原后的目标文本输入关键词识别模型，从而可以基于该关键词识别模型，获取还原后的目标文本中的关键词，如此可以得到准确的关键词，进而可以提升关键词识别的准确率。

20 可选地，本申请实施例中，在关键词识别模型识别还原后的目标文本中的关键词之后，关键词识别模型可以向电子设备输出关键词列表。其中，该关键词列表中包括还原后的目标文本中的所有关键词。

示例性地，假设关键词识别模型中预设的关键词的内容类型为“地名”，那么在电子设备将还原后的目标文本输入关键词识别模型之后，关键词识别模型可以从还原后的目标文本中提取并输出所有与“地名”相关的词，从而得到上述关键词。

25 本申请实施例中，在电子设备将还原后的目标文本输入关键词识别模型之后，关键词识别模型可以对还原后的目标文本进行关键词识别，从而可以得到还原后的目标文本中的关键词，并向电子设备输出这些关键词的列表，从而可

以准确得到目标文本中的关键词。

下面将以本申请实施例中以文本还原装置执行文本还原方法为例，说明本申请实施例提供的文本还原装置。

如图 4 所示，本申请实施例提供一种文本还原装置 300，文本还原装置 300
5 包括获取模块 301，确定模块 302 和还原模块 303。获取模块 301，用于根据第一字符组，获取第一候选词和第二候选词，第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，第一候选词为第一字符组与第二字符组合得到的词，第二候选词为第三字符组与第二字符组合得到的词，第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组，第三字符组为第一字符组除去分隔符后得到的字符组；确定模块 302，用于确定第一困惑度和第二困惑度，第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度；还原模块 303，用于在第一困惑度小于第二困惑度的情况下，根据第一候选词，得到还
10 原后的目标文本；或在第二困惑度小于第一困惑度的情况下，根据第二候选词，得到还原后的目标文本。

可选地，确定模块，具体用于对第一候选词和第二候选词分别执行以下步骤：基于目标候选词中的每个字符在目标文本中出现的概率，确定目标参数，目标候选词为第一候选词或第二候选词；根据目标参数，确定目标候选词对应的困惑度；其中，目标参数包括：目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值；目标词组包括目标候选词、第四字符组和第五字符组，第四字符组为目标文本中位于第一字符组之前的字符组，第五字符组为目标文本中位于第二字符组之后的字符组。

可选地，确定模块，具体用于根据目标候选词的合法性值与第一系数的乘积、目标词组的流畅度值与第二系数的乘积、目标语句的流畅度值与第三系数的乘积之和，得到目标候选词对应的困惑度；其中，第一系数、第二系数和第三系数之和等于 1。

可选地，目标候选词的合法性值为目标文本中，目标候选词在目标文本中出现的概率；目标词组的流畅度值为目标候选词、第四字符组和第五字符组组成的词组在目标文本中出现的概率；目标语句的流畅度值为目标语句在目标文本中出现的概率。

5 可选地，目标候选词的合法性值为目标候选词中的每个字符在目标文本中出现的概率之间的乘积；其中，目标候选词中的第 K 个字符在目标文本中出现的概率是指：在目标文本中出现第六字符组的情况下出现第 K 个字符的概率，第六字符组由目标候选词中的第 1 个字符至第 (K-1) 个字符组成，K 为大于 1 的整数。

10 可选地，确定模块，还用于基于关键词识别模型，获取还原后的目标文本的关键词，该关键词的内容类型与关键词识别模型中预设的内容类型相同。

 本申请实施例提供一种文本还原装置，由于语句对应的困惑度越小，表示语句越流畅，即语句对应的困惑度越小，语句越准确，因此通过比较根据第一候选词得到的第一语句对应的困惑度和根据第二候选词得到的第二语句对应的困惑度，可以确定第一候选词和第二候选词哪个是正确地，即可以确定目标文本中的第一字符组和第二字符组组成的正确的词，从而可以准确地还原文本。

 本申请实施例中的文本还原装置可以是装置，也可以是电子设备中的部件、集成电路、或芯片。该装置可以是移动电子设备，也可以为非移动电子设备。示例性的，移动电子设备可以为手机、平板电脑、笔记本电脑、掌上电脑、车载电子设备、可穿戴设备、超级移动个人计算机（ultra-mobile personal computer, UMPC）、上网本或者个人数字助理（personal digital assistant, PDA）等，非移动电子设备可以为个人计算机（personal computer, PC）、电视机（television, TV）、柜员机或者自助机等，本申请实施例不作具体限定。

25 本申请实施例中的文本还原装置可以为具有操作系统的装置。该操作系统可以为安卓（Android）操作系统，可以为 ios 操作系统，还可以为其他可能的操作系统，本申请实施例不作具体限定。

本申请实施例提供的文本还原装置能够实现上述方法实施例实现的各个过程，为避免重复，这里不再赘述。

可选地，如图 5 所示，本申请实施例还提供一种电子设备 500，包括处理器 501，存储器 502，存储在存储器 502 上并可在处理器 501 上运行的程序或指令，该程序或指令被处理器 501 执行时实现上述文本还原方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

需要说明的是，本申请实施例中的电子设备包括上述的移动电子设备和非移动电子设备。

图 6 为实现本申请实施例的一种电子设备的硬件结构示意图。

10 该电子设备 100 包括但不限于：射频单元 101、网络模块 102、音频输出单元 103、输入单元 104、传感器 105、显示单元 106、用户输入单元 107、接口单元 108、存储器 109、以及处理器 110 等部件。

本领域技术人员可以理解，电子设备 100 还可以包括给各个部件供电的电源（比如电池），电源可以通过电源管理系统与处理器 110 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。图 6 中示出的电子设备结构并不构成对电子设备的限定，电子设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置，在此不再赘述。

其中，处理器 110，可以用于根据第一字符组，获取第一候选词和第二候选词，第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，第一候选词为第一字符组与第二字符组合得到的词，第二候选词为第三字符组与第二字符组合得到的词，第二字符组为待还原的目标文本中的第 N+1 行的第一个字符组，第三字符组为第一字符组除去分隔符后得到的字符组；并确定第一困惑度和第二困惑度，第一困惑度为第一候选词替换目标语句中的第一字符组和第二字符组得到的第一语句对应的困惑度，第二困惑度为第二候选词替换目标语句中的第一字符组和第二字符组得到的第二语句对应的困惑度；以及在第一困惑度小于第二困惑度的情况下，根据第一候选词，得到还原后的目标文本；或在第二困惑度小于第一困惑度的情况下，根据第二

候选词，得到还原后的目标文本。

可选地，处理器 110，具体用于对第一候选词和第二候选词分别执行以下步骤：基于目标候选词中的每个字符在目标文本中出现的概率，确定目标参数，目标候选词为第一候选词或第二候选词；根据目标参数，确定目标候选词对应的困惑度；其中，目标参数包括：目标候选词的合法性值、目标词组的流畅度值和目标语句的流畅度值；目标词组包括目标候选词、第四字符组和第五字符组，第四字符组为目标文本中位于第一字符组之前的字符组，第五字符组为目标文本中位于第二字符组之后的字符组。

可选地，处理器 110，具体用于根据目标候选词的合法性值与第一系数的乘积、目标词组的流畅度值与第二系数的乘积、目标语句的流畅度值与第三系数的乘积之和，得到目标候选词对应的困惑度；其中，第一系数、第二系数和第三系数之和等于 1。

可选地，目标候选词的合法性值为目标文本中，目标候选词在目标文本中出现的概率；目标词组的流畅度值为目标候选词、第四字符组和第五字符组组成的词组在目标文本中出现的概率；目标语句的流畅度值为目标语句在目标文本中出现的概率。

可选地，目标候选词的合法性值为目标候选词中的每个字符在目标文本中出现的概率之间的乘积；其中，目标候选词中的第 K 个字符在目标文本中出现的概率是指：在目标文本中出现第六字符组的情况下出现第 K 个字符的概率，第六字符组由目标候选词中的第 1 个字符至第 (K-1) 个字符组成，K 为大于 1 的整数。

可选地，处理器 110，还用于基于关键词识别模型，获取还原后的目标文本的关键词，该关键词的内容类型与关键词识别模型中预设的内容类型相同。

本申请实施例提供一种电子设备，由于语句对应的困惑度越小，表示语句越流畅，即语句对应的困惑度越小，语句越准确，因此通过比较根据第一候选词得到的第一语句对应的困惑度和根据第二候选词得到的第二语句对应的困惑度，可以确定第一候选词和第二候选词哪个是正确地，即可以确定目标文本

中的第一字符组和第二字符组组成的正确的词，从而可以准确地还原文本。

需要说明的是，本申请实施例中，上述文本还原装置中的获取模块、确定模块、还原模块和输入模块均可以通过上述处理器 110 实现。

应理解的是，本申请实施例中，射频单元 101 包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器、双工器等。电子设备通过网络模块 102 为用户提供无线的宽带互联网访问，如帮助用户收发电子邮件、浏览网页和访问流式媒体等。音频输出单元 103 可以包括扬声器、蜂鸣器以及受话器等。输入单元 104 可以包括图形处理器 (Graphics Processing Unit, GPU) 1041 和麦克风 1042，图形处理器 1041 对在视频捕获模式或图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。显示单元 106 可包括显示面板 1061，可以采用液晶显示器、有机发光二极管等形式来配置显示面板 1061。用户输入单元 107 包括触控面板 1071 以及其他输入设备 1072。触控面板 1071，也称为触摸屏。触控面板 1071 可包括触摸检测装置和触摸控制器两个部分。其他输入设备 1072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。存储器 109 可用于存储软件程序以及各种数据，包括但不限于应用程序和操作系统。处理器 110 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器 110 中。

本申请实施例还提供一种可读存储介质，该可读存储介质上存储有程序或指令，该程序或指令被处理器执行时实现上述文本还原方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

其中，上述处理器为上述实施例中的电子设备中的处理器。可读存储介质可以包括计算机可读存储介质，如计算机只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等。

本申请实施例另提供了一种芯片，该芯片包括处理器和通信接口，通信接口和处理器耦合，处理器用于运行程序或指令，实现上述文本还原方法实施例

的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

应理解，本申请实施例提到的芯片还可以称为系统级芯片、系统芯片、芯片系统或片上系统芯片等。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。此外，需要指出的是，本申请实施方式中的方法和装置的范围不限按示出或讨论的顺序来执行功能，还可包括根据所涉及的功能按基本同时的方式或按相反的顺序来执行功能，例如，可以按不同于所描述的次序来执行所描述的方法，并且还可以添加、省去、或组合各种步骤。另外，参照某些示例所描述的特征可在其他示例中被组合。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台电子设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

权 利 要 求 书

1. 一种文本还原方法，包括：

根据第一字符组，获取第一候选词和第二候选词，所述第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符组，所述第一候选词为所述第一字符组与第二字符组合得到的词，所述第二候选词为第三字符组与所述第二字符组合得到的词，所述第二字符组为待还原的所述目标文本中的第 N+1 行的第一个字符组，所述第三字符组为所述第一字符组除去所述分隔符后得到的字符组；

确定第一困惑度和第二困惑度，所述第一困惑度为所述第一候选词替换目标语句中的所述第一字符组和所述第二字符组得到的第一语句对应的困惑度，所述第二困惑度为所述第二候选词替换目标语句中的所述第一字符组和所述第二字符组得到的第二语句对应的困惑度；

在所述第一困惑度小于所述第二困惑度的情况下，根据所述第一候选词，得到还原后的所述目标文本；或在所述第二困惑度小于所述第一困惑度的情况下，根据所述第二候选词，得到还原后的所述目标文本。

2. 根据权利要求 1 所述的方法，其中，所述确定第一困惑度和第二困惑度，包括：

对所述第一候选词和所述第二候选词分别执行以下步骤：

基于目标候选词中的每个字符在所述目标文本中出现的概率，确定目标参数，所述目标候选词为所述第一候选词或所述第二候选词；

根据所述目标参数，确定所述目标候选词对应的困惑度；

其中，所述目标参数包括：目标候选词的合法性值、目标词组的流畅度值和所述目标语句的流畅度值；所述目标词组包括所述目标候选词、第四字符组和第五字符组，所述第四字符组为所述目标文本中位于所述第一字符组之前的字符组，所述第五字符组为所述目标文本中位于所述第二字符组之后的字符组。

3. 根据权利要求 2 所述的方法，其中，所述根据所述目标参数，确定所述

目标候选词对应的困惑度，包括：

根据所述目标候选词的合法性值与第一系数的乘积、所述目标词组的流畅度值与第二系数的乘积、所述目标语句的流畅度值与第三系数的乘积之和，得到所述目标候选词对应的困惑度；

5 其中，所述第一系数、第二系数和第三系数之和等于 1。

4. 根据权利要求 2 或 3 所述的方法，其中，所述目标候选词的合法性值为所述目标文本中，所述目标候选词在所述目标文本中出现的概率；

所述目标词组的流畅度值为所述目标候选词、所述第四字符组和所述第五字符组组成的词组在所述目标文本中出现的概率；

10 所述目标语句的流畅度值为所述目标语句在所述目标文本中出现的概率。

5. 根据权利要求 4 所述的方法，其中，所述目标候选词的合法性值为目标候选词中的每个字符在所述目标文本中出现的概率之间的乘积；

其中，所述目标候选词中的第 K 个字符在所述目标文本中出现的概率是指：在所述目标文本中出现第六字符组的情况下出现第 K 个字符的概率，所述
15 第六字符组由所述目标候选词中的第 1 个字符至第 (K-1) 个字符组成，K 为大于 1 的整数。

6. 一种文本还原装置，包括获取模块，确定模块和还原模块；

获取模块，用于根据第一字符组，获取第一候选词和第二候选词，所述第一字符组为处于待还原的目标文本中的第 N 行的行末、且以分隔符结尾的字符
20 组，所述第一候选词为所述第一字符组与第二字符组合得到的词，所述第二候选词为第三字符组与所述第二字符组合得到的词，所述第二字符组为待还原的所述目标文本中的第 N+1 行的第一个字符组，所述第三字符组为所述第一字符组除去所述分隔符后得到的字符组；

确定模块，用于确定第一困惑度和第二困惑度，所述第一困惑度为所述第一候选词替换目标语句中的所述第一字符组和所述第二字符组得到的第一语句对应的困惑度，所述第二困惑度为所述第二候选词替换目标语句中的所述第一字符组和所述第二字符组得到的第二语句对应的困惑度；

还原模块，用于在所述第一困惑度小于所述第二困惑度的情况下，根据所述第一候选词，得到还原后的所述目标文本；或在所述第二困惑度小于所述第一困惑度的情况下，根据所述第二候选词，得到还原后的所述目标文本。

7. 根据权利要求 6 所述的装置，其中，所述确定模块，具体用于对所述第一候选词和所述第二候选词分别执行以下步骤：

基于目标候选词中的每个字符在所述目标文本中出现的概率，确定目标参数，所述目标候选词为所述第一候选词或所述第二候选词；

根据所述目标参数，确定所述目标候选词对应的困惑度；

其中，所述目标参数包括：目标候选词的合法性值、目标词组的流畅度值和所述目标语句的流畅度值；所述目标词组包括所述目标候选词、第四字符组和第五字符组，所述第四字符组为所述目标文本中位于所述第一字符组之前的字符组，所述第五字符组为所述目标文本中位于所述第二字符组之后的字符组。

8. 根据权利要求 7 所述的装置，其中，所述确定模块，具体用于根据所述目标候选词的合法性值与第一系数的乘积、所述目标词组的流畅度值与第二系数的乘积、所述目标语句的流畅度值与第三系数的乘积之和，得到所述目标候选词对应的困惑度；

其中，所述第一系数、第二系数和第三系数之和等于 1。

9. 根据权利要求 7 或 8 所述的装置，其中，所述目标候选词的合法性值为所述目标文本中，所述目标候选词在所述目标文本中出现的概率；

所述目标词组的流畅度值为所述目标候选词、所述第四字符组和所述第五字符组组成的词组在所述目标文本中出现的概率；

所述目标语句的流畅度值为所述目标语句在所述目标文本中出现的概率。

10. 根据权利要求 9 所述的装置，其中，所述目标候选词的合法性值为目标候选词中的每个字符在所述目标文本中出现的概率之间的乘积；

其中，所述目标候选词中的第 K 个字符在所述目标文本中出现的概率是指：在所述目标文本中出现第六字符组的情况下出现第 K 个字符的概率，所述

第六字符组由所述目标候选词中的第 1 个字符至第 (K-1) 个字符组成, K 为大于 1 的整数。

11. 一种电子设备, 包括处理器, 存储器及存储在所述存储器上并可在所述处理器上运行的程序或指令, 所述程序或所述指令被所述处理器执行时实现
5 如权利要求 1-5 中任一项所述的文本还原方法的步骤。

12. 一种可读存储介质, 所述可读存储介质上存储程序或指令, 所述程序或所述指令被处理器执行时实现如权利要求 1-5 中任一项所述的文本还原方法的步骤。

13. 一种计算机程序产品, 所述计算机程序产品被至少一个处理器执行以实现如权利要求 1-5 中任一项所述的文本还原方法的步骤。
10

14. 一种芯片, 所述芯片包括处理器和通信接口, 所述通信接口和所述处理器耦合, 所述处理器用于运行程序或指令, 实现如权利要求 1-5 任一项所述的文本还原方法的步骤。

15. 一种电子设备, 所述电子设备被配置成用于执行如权利要求 1-5 任一项
15 所述的文本还原方法的步骤。

We introduce a new language representation¹ model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation² models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train³ deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a re-⁴ sult, the pre-trained BERT model can be fine-⁵ tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-⁶ specific architecture modifications.

图 1

We introduce a new language representation¹ model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation² models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain³ deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result⁴, the pre-trained BERT model can be finetuned with⁵ just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific⁶ architecture modifications.

图 2

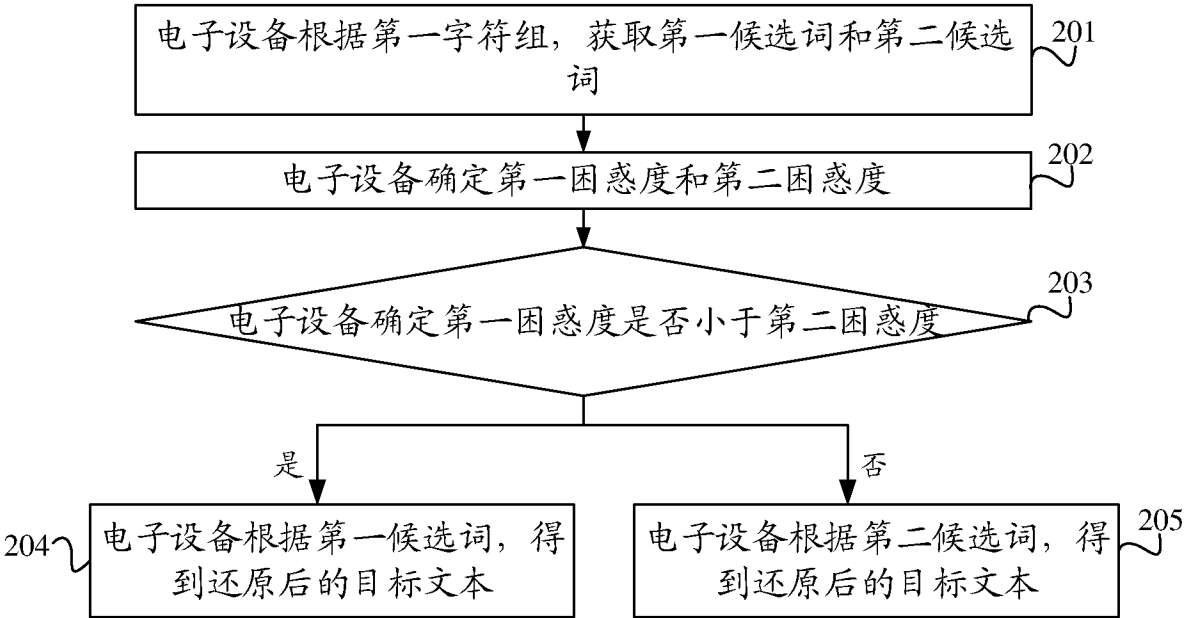


图 3

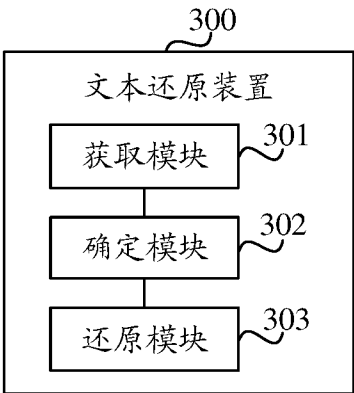


图 4

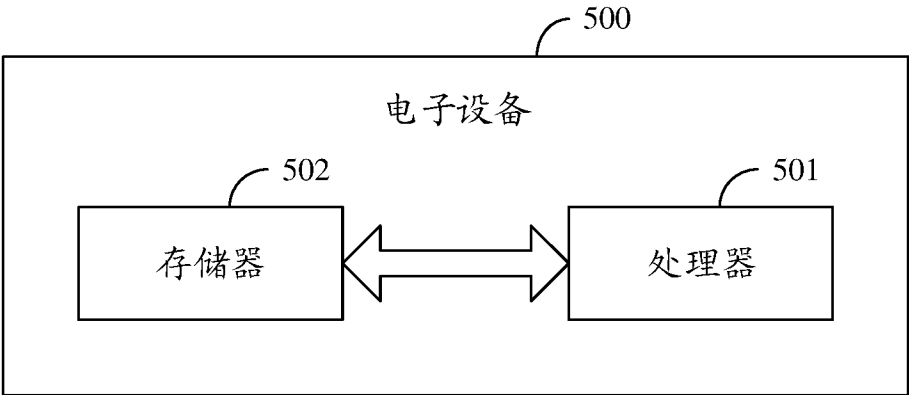


图 5

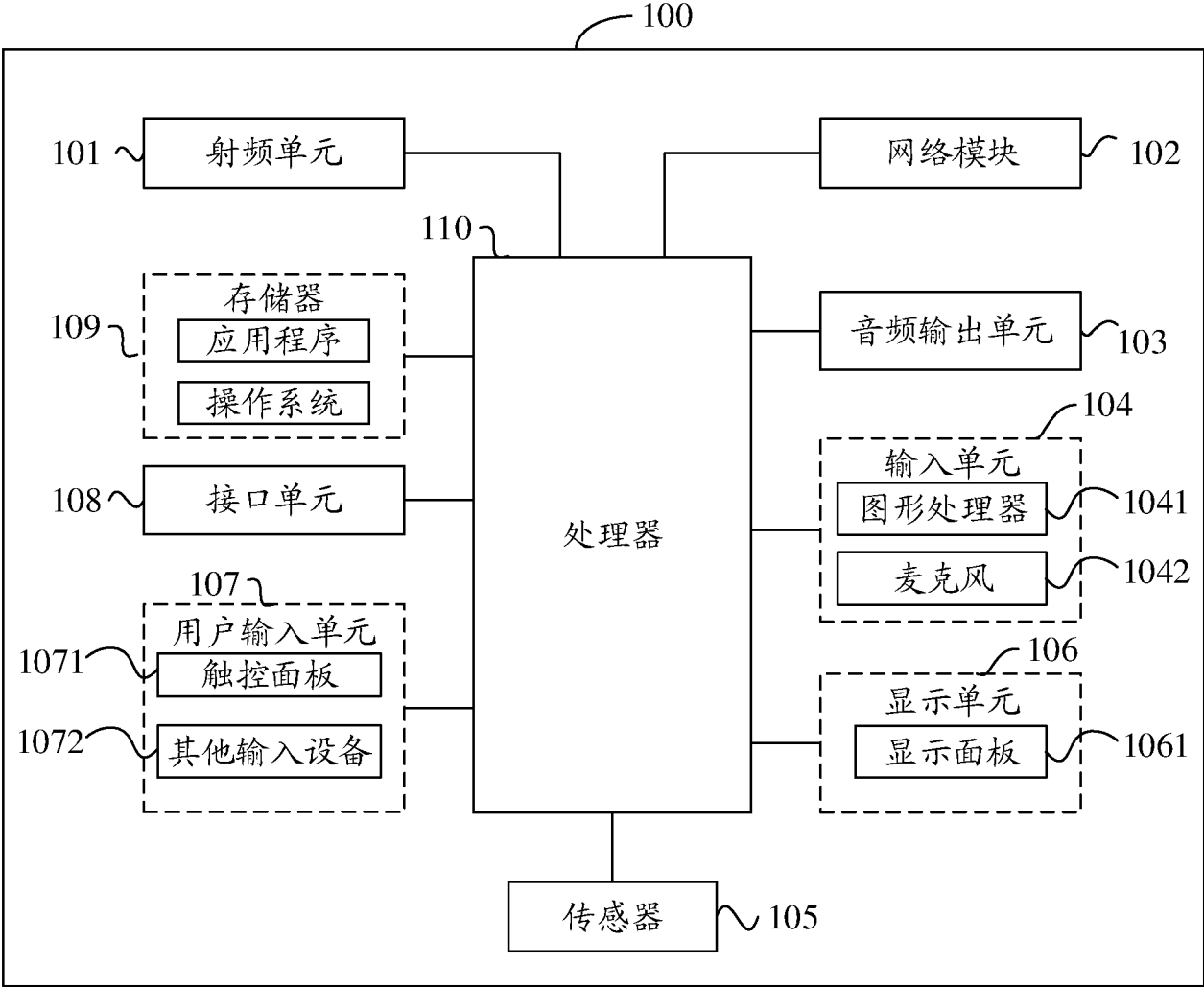


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/074583

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/166(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, EPODOC, WPI: 分割符, 字符, 英文, 文本, 连字符, 还原, 概率, 困惑度, text, restoration, separator, identification, confusion degree, probability

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112949261 A (VIVO COMMUNICATION TECHNOLOGY CO., LTD.) 11 June 2021 (2021-06-11) claims 1-12, description paragraph [0109]	1-15208715994
X	CN 108899016 A (IFLYTEK CO., LTD.) 27 November 2018 (2018-11-27) description paragraphs [0056]-[0138]	1-15208715994
A	CN 111401004 A (SUZHOU JISHU XINWEI TECHNOLOGY CO., LTD.) 10 July 2020 (2020-07-10) entire document	1-15
A	US 10402490 B1 (SHUTTERSTOCK, INC.) 03 September 2019 (2019-09-03) entire document	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

11 April 2022

Date of mailing of the international search report

26 April 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/074583

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	112949261	A	11 June 2021	None	
CN	108899016	A	27 November 2018	None	
CN	111401004	A	10 July 2020	None	
US	10402490	B1	03 September 2019	None	

国际检索报告

国际申请号

PCT/CN2022/074583

A. 主题的分类

G06F 40/166(2020.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNKI, CNPAT, EPDOC, WPI: 分割符, 字符, 英文, 文本, 连字符, 还原, 概率, 困惑度, text, restoration, separator, identification, confusion degree, probability

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 112949261 A (维沃移动通信有限公司) 2021年6月11日 (2021 - 06 - 11) 权利要求1-12, 说明书第[0109]段	1-15
X	CN 108899016 A (科大讯飞股份有限公司) 2018年11月27日 (2018 - 11 - 27) 说明书第[0056]-[0138]段	1-15
A	CN 111401004 A (苏州机数芯微科技有限公司) 2020年7月10日 (2020 - 07 - 10) 全文	1-15
A	US 10402490 B1 (SHUTTERSTOCK, INC.) 2019年9月3日 (2019 - 09 - 03) 全文	1-15

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年4月11日

国际检索报告邮寄日期

2022年4月26日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

赵晓春

电话号码 86-(10)-53961295

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/074583

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	112949261	A	2021年6月11日	无	
CN	108899016	A	2018年11月27日	无	
CN	111401004	A	2020年7月10日	无	
US	10402490	B1	2019年9月3日	无	