

(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
2020 年 3 月 19 日 (19.03.2020)



(10) 国际公布号
WO 2020/052060 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2018/115953

(22) 国际申请日: 2018 年 11 月 16 日 (16.11.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811074035.4 2018 年 9 月 14 日 (14.09.2018) CN

(71) 申请人: 北京字节跳动网络技术有限公司 (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市石景山区实兴大街 30 号院 3 号楼 2 层 B-0035 房间, Beijing 100041 (CN)。

(72) 发明人: 邓江东 (DENG, Jiangdong); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100190 (CN)。

(74) 代理人: 北京英赛嘉华知识产权代理有限公司 (INSIGHT INTELLECTUAL PROPERTY LIMITED); 中国北京市海淀区知春路甲 48 号盈都大厦 A 座 19A, Beijing 100098 (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, 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: METHOD AND APPARATUS FOR GENERATING CORRECTION STATEMENT

(54) 发明名称: 用于生成修正语句的方法和装置

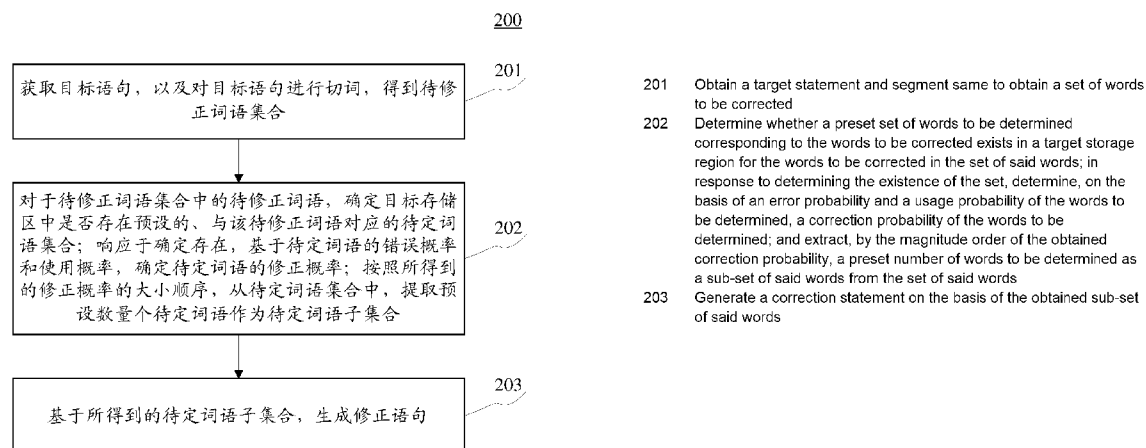


图 2

(57) Abstract: Disclosed are a method and apparatus for generating a correction statement. One specific embodiment of the method comprises: obtaining a target statement and segmenting same to obtain a set of words to be corrected; determining whether a preset set of words to be determined corresponding to the words to be corrected exists in a target storage region for the words to be corrected in the set of said words; in response to determining the existence of the set, determining, on the basis of an error probability and a usage probability of the words to be determined, a correction probability of the words to be determined; extracting, by the magnitude order of the obtained correction probability, a preset number of words to be determined as a sub-set of said words from the set of said words; and generating a correction statement on the basis of the obtained sub-set of said words. The embodiment improves the accuracy of correcting the statement.

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, 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))。

(57) 摘要: 本申请实施例公开了用于生成修正语句的方法和装置。该方法的一具体实施方式包括: 获取目标语句, 以及对目标语句进行切词, 得到待修正词语集合; 对于待修正词语集合中的待修正词语, 确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合; 响应于确定存在, 基于待定词语的错误概率和使用概率, 确定待定词语的修正概率; 按照所得到的修正概率的大小顺序, 从待定词语集合中, 提取预设数量个待定词语作为待定词语子集合; 基于所得到的待定词语子集合, 生成修正语句。该实施方式提高了对语句进行修正的准确性。

用于生成修正语句的方法和装置

5 本专利申请要求于 2018 年 9 月 14 日提交的、申请号为 201811074035.4、申请人为北京字节跳动网络技术有限公司、发明名称为“用于生成修正语句的方法和装置”的中国专利申请的优先权，该申请的全文以引用的方式并入本申请中。

10 技术领域

 本申请实施例涉及计算机技术领域，具体涉及用于生成修正语句的方法和装置。

背景技术

15 当人们利用手机、电脑等电子设备输入文字（例如向搜索引擎输入搜索语句、输入评论语句等）时，通常会产生输入文字错误的情况。这时需要将错误的文字修正。现有技术中，一般是由搜索引擎对用户输入的搜索语句进行纠错，并将纠正后的语句展示出来供用户选择。

20 发明内容

 本申请实施例提出了用于生成修正语句的方法和装置。

 第一方面，本申请实施例提供了一种用于生成修正语句的方法，该方法包括：获取目标语句，以及对目标语句进行切词，得到待修正词语集合；对于待修正词语集合中的待修正词语，确定目标存储区中
25 是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为
30 待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集

合中，提取预设数量个待定词语作为待定词语子集合；基于所得到的待定词语子集合，生成修正语句。

在一些实施例中，目标存储区为预先建立的对应关系表，对应关系表包括待修正词语和待定词语集合，对应关系表用于表征待修正词语与待定词语集合的对应关系，对应关系表还包括待定词语的错误概率和使用概率。

在一些实施例中，待定词语的错误概率和使用概率预先按照如下步骤得到：对于待定词语集合中的待定词语，确定预设的、该待定词语对应的文本子集中的文本包括的、该待定词语的数量为第一数量，其中，文本子集是预设的文本集的子集；确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量；将所确定的第二数量除以第一数量与第二数量之和，得到该待定词语的错误概率；确定该待定词语在文本集中的数量为第三数量，以及将所确定的第三数量与文本集中的文本包括的词语的总数的比值确定为该待定词语的使用概率。

在一些实施例中，基于所得到的待定词语子集合，生成修正语句，包括：利用所得到的待定词语子集合中的待定词语，生成至少一个候选修正语句，以及得到至少一个候选修正语句中的候选修正语句的置信度；基于所得到的置信度的大小顺序，从至少一个候选修正语句中提取候选修正语句作为修正语句。

在一些实施例中，在确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合之后，该方法还包括：响应于确定目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音；基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语；从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

在一些实施例中，所得到的至少一个匹配词语中的匹配词语具有对应的热度数值，热度数值用于表征匹配词语被使用的频率；以及从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合，包括：按照热度数值的大小顺序，从所获取的至少一个匹

配词语中选择预设数量个匹配词语作为待定词语子集合。

第二方面，本申请实施例提供了一种用于生成修正语句的装置，该装置包括：获取单元，被配置成获取目标语句，以及对目标语句进行切词，得到待修正词语集合；提取单元，被配置成对于待修正词语集合中的待修正词语，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集合中，提取预设数量个待定词语作为待定词语子集合；生成单元，被配置成基于所得到的待定词语子集合，生成修正语句。

在一些实施例中，目标存储区为预先建立的对应关系表，对应关系表包括待修正词语和待定词语集合，对应关系表用于表征待修正词语与待定词语集合的对应关系，对应关系表还包括待定词语的错误概率和使用概率。

在一些实施例中，待定词语的错误概率和使用概率预先按照如下步骤得到：对于待定词语集合中的待定词语，确定预设的、该待定词语对应的文本子集中的文本包括的、该待定词语的数量为第一数量，其中，文本子集是预设的文本集的子集；确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量；将所确定的第二数量除以第一数量与第二数量之和，得到该待定词语的错误概率；确定该待定词语在文本集中的数量为第三数量，以及将所确定的第三数量与文本集中的文本包括的词语的总数的比值确定为该待定词语的使用概率。

在一些实施例中，生成单元包括：生成模块，被配置成利用所得到的待定词语子集合中的待定词语，生成至少一个候选修正语句，以及得到至少一个候选修正语句中的候选修正语句的置信度；提取模块，被配置成基于所得到的置信度的大小顺序，从至少一个候选修正语句

中提取候选修正语句作为修正语句。

在一些实施例中，提取单元包括：确定模块，被配置成响应于确定目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音；获取模块，被配置成基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语；选择模块，被配置成从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

在一些实施例中，所得到的至少一个匹配词语中的匹配词语具有对应的热度数值，热度数值用于表征匹配词语被使用的频率；以及选择模块进一步被配置成：按照热度数值的大小顺序，从所得到的至少一个匹配词语中选择预设数量个匹配词语作为待定词语子集合。

第三方面，本申请实施例提供了一种服务器，该服务器包括：一个或多个处理器；存储装置，其上存储有一个或多个程序；当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面中任一实现方式描述的方法。

第四方面，本申请实施例提供了一种计算机可读介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如第一方面中任一实现方式描述的方法。

本申请实施例提供的用于生成修正语句的方法和装置，通过对目标语句进行切词，得到待修正词语集合，然后对于待修正词语集合中的待修正词语，确定该待修正词语对应的待定词语集合，再基于待定词语的错误概率和使用概率，确定待定词语的修正概率，以及按照修正概率的大小顺序从待定词语集合中，提取预设数量个待定词语作为该待修正词语对应的待定词语子集合，最后基于所得到的各个待定词语子集合，生成修正语句，从而提高了对语句进行修正的准确性。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 是本申请的一个实施例可以应用于其中的示例性系统架构图；

图 2 是根据本申请实施例的用于生成修正语句的方法的一个实施例的流程图；

图 3 是根据本申请实施例的用于生成修正语句的方法的一个应用场景的示意图；

5 图 4 是根据本申请实施例的用于生成修正语句的方法的又一个实施例的流程图；

图 5 是根据本申请实施例的用于生成修正语句的装置的一个实施例的结构示意图；

10 图 6 是适于用来实现本申请实施例的服务器的计算机系统的结构示意图。

具体实施方式

下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明
15 的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

20 图 1 示出了可以应用本申请实施例的用于生成修正语句的方法或用于生成修正语句的装置的示例性系统架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，
25 例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 101、102、103 通过网络 104 与服务器 105 交互，以接收或发送消息等。终端设备 101、102、103 上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用等。

30 终端设备 101、102、103 可以是硬件，也可以是软件。当终端设

备 101、102、103 为硬件时，可以是各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3 播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面 3)、MP4 (Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面 4) 播放器、膝上型便携计算机和台式计算机等等。当终端设备 101、102、103 为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块(例如用来提供分布式服务的软件或软件模块)，也可以实现成单个软件或软件模块。在此不做具体限定。

服务器 105 可以是提供各种服务的服务器，例如对终端设备 101、102、103 上传的目标语句进行处理的后台信息处理服务器。后台信息处理服务器可以对获取的目标语句进行处理，并得到处理结果(例如修正语句)。

需要说明的是，本申请实施例所提供的用于生成修正语句的方法一般由服务器 105 执行，相应地，用于生成修正语句的装置一般设置于服务器 105 中。

需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块(例如用来提供分布式服务的软件或软件模块)，也可以实现成单个软件或软件模块。在此不做具体限定。

应该理解，图 1 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。

继续参考图 2，示出了根据本申请的用于生成修正语句的方法的一个实施例的流程 200。该用于生成修正语句的方法，包括以下步骤：

步骤 201，获取目标语句，以及对目标语句进行切词，得到待修正词语集合。

在本实施例中，用于生成修正语句的方法的执行主体(例如图 1 所示的服务器)可以通过有线连接方式或者无线连接方式从远程或从本地获取目标语句。其中，目标语句可以是待对其进行切词的语句，

目标语句可以是用户利用终端设备(例如图 1 所示的终端设备)输入,并发送至上​​述执行主体的语句,或者是预先存储在上述执行主体中的语句。作为示例,目标语句可以是用于搜索信息的搜索语句,或者是预先从诸如文章、文字评论等信息中提取的语句。

5 上述执行主体可以进一步对所获取的目标语句进行切词,得到待修正词语集合。具体地,上述执行主体可以按照现有的切词方法(例如最大正向匹配法、N-gram 模型方法、隐马尔科夫模型方法等)对目标语句进行切词,得到待修正词语集合。

10 需要说明的是,待修正词语集合可以包括对上述目标语句进行切词后所得到的全部词语,也可以包括部分符合预设条件的词语。例如,可以对上述目标语句进行切词后所得到的每个词语进行词性标注,从各个词语中提取预设词性(例如名词和/或动词)的词语作为待修正词语。其中,对词语进行词性标注的方法是​​目前广泛研究和应用的公知技术,在此不再赘述。

15 步骤 202,对于待修正词语集合中的待修正词语,确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合;响应于确定存在,基于待定词语的错误概率和使用概率,确定待定词语的修正概率;按照所得到的修正概率的大小顺序,从待定词语集合中,提取预设数量个待定词语作为待定词语子集合。

20 在本实施例中,对于待修正词语集合中的待修正词语,上述执行主体可以执行如下步骤:

 步骤 2021,确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合。

25 其中,待定词语具有对应的错误概率和使用概率。待定词语所对应的错误概率用于表征待定词语被错换为该待修正词语的概率。作为示例,假设某待定词语 A 对应的错误概率为 10%,表示该待定词语被错用为该修正词语的概率为 10%。使用概率用于表征待定词语被使用的概率。通常,使用概率可以是预先对预设的文本进行词频统计所得到的、待定词语出现的概率。错误概率可以是预先对将待定词语错换为待修正词语的次数进行统计所得到的概率。

30

在本实施例的一些可选的实现方式中，待定词语的错误概率和使用概率可以预先按照如下步骤得到：

对于待定词语集合中的待定词语，执行如下步骤：

5 步骤一，确定预设的、该待定词语对应的文本子集中的文本包括的、该待定词语的数量为第一数量。其中，文本子集是预设的文本集的子集。

步骤二，确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量。

10 作为示例，假设某待定词语为“唢呐”，对应的文本子集中的所有文本包括的词语“唢呐”的数量为第一数量；词语“唢呐”对应的待修正词语为“所纳”，文本子集中的所有文本包括的词语“所纳”的数量为第二数量。需要说明的是，待定词语对应的文本子集可以是技术人员预先设置的文本集包括的文本子集。例如，文本集是预先获取的、用户输入的搜索语句的集合。待定词语对应的文本子集可以是文本集
15 中的、包括待定词语的多条搜索语句和包括待定词语对应的待修正词语的多条搜索语句的集合。

步骤三，将所确定的第二数量除以第一数量与第二数量之和，得到该待定词语的错误概率。

20 继续上述示例，假设文本子集中的文本包括的词语“所纳”（即待修正词语）的数量为 10，包括的词语“唢呐”（即待定词语）的数量为 90，则词语“唢呐”对应的错误概率为 $10/(10+90)$ 。

步骤四，确定该待定词语在文本集中的数量为第三数量，以及将所确定的第三数量与文本集中的文本包括的词语的总数的比值确定为该待定词语的使用概率。

25 在本实施例中，上述目标存储区可以是在上述执行主体中预先分配的存储区，也可以是在与上述执行主体通信连接的电子设备中预先分配的存储区。需要说明的是，目标存储区可以是实际的物理存储区（例如磁盘、芯片等内部的存储区），也可以是虚拟存储区（例如表格等）。

30 作为示例，目标存储区中可以存储有多个待修正词语和多个待定

词语集合，每个待修正词语可以对应一个待定词语集合。

在本实施例的一些可选的实现方式中，目标存储区为预先建立的对应关系表，对应关系表包括待修正词语和待定词语集合，对应关系表用于表征待修正词语与待定词语集合的对应关系，对应关系表还包括待定词语的错误概率和使用概率。作为示例，对应关系表可以是二维表格，该表格可以存储有多个待修正词语，每个待修正词语对应二维表中的三行（或三列），各行（或各列）分别用于存储待修正词语对应的待定词语、待定词语的错误概率和待定词语的使用概率。通过使用对应关系表，可以准确地确定待修正词语对应的待定词语集合，以及有助于直观地呈现待修正词语和对应的待定词语集合。

步骤 2022，响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率。

其中，修正概率用于表征将该待修正词语修改为待定词语的概率。通常，修正概率越大，表示该待修正词语发生错误的可能性越大，即应当将该待修正词语修改为待定词语的可能性越大。作为示例，上述执行主体可以基于贝叶斯公式，计算得到待定词语的修正概率。具体地，贝叶斯公式为：

$$P(c|w)=P(w|c) \times P(c)/P(w),$$

其中，c 用于表征待定词语；w 用于表征待修正词语；P(c)为待定词语的使用概率；P(w|c)为待定词语 c 对应的错误概率，即在待定词语 c 正确的前提下，待定词语 c 被错误地写成待修正词语 w 的概率；P(w)为待修正词语的使用概率；P(c|w)为待定词语 c 的修正概率，即待修正词语 w 应当被修改为待定词语 c 的概率。其中，P(w)可以是对预设的文本集合中的文本包括的待修正词语进行统计得到概率。例如，将待修正词语的数量与文本集中的文本包括的词语的总数的比值确定为待修正词语的使用概率。

作为另一示例，对于某个待修正词语的使用概率 P(w)，由于该待修正词语的使用概率 P(w)是固定的，因此，上述公式中的 P(c|w)正比于 P(w|c) × P(c)，确定待定词语的修正概率的公式可以近似为如下公式：

$$P(c|w)=P(w|c) \times P(c)。$$

步骤 2023, 按照所得到的修正概率的大小顺序, 从待定词语集合中, 提取预设数量个待定词语作为待定词语子集合。

5 作为示例, 上述执行主体可以按照各个修正概率由大到小的顺序, 从待定词语集合中, 提取预设数量个待定词语。需要说明的是, 当待定词语集合包括的待定词语的数量小于等于上述预设数量时, 可以将待定词语集合作为待修正词语对应的待定词语子集合。

步骤 203, 基于所得到的待定词语子集合, 生成修正语句。

在本实施例中, 上述执行主体可以基于所得到的待定词语子集合, 生成修正语句。

10 作为示例, 上述执行主体可以按照修正概率由大到小的顺序, 从各个待定词语子集合中分别提取待定词语, 将提取的待定词语与待修正词语集合中的、不具有对应的待定词语集合的待修正词语按照待修正词语集合中的待修正词语的排列顺序, 组合为修正语句。

15 在本实施例的一些可选的实现方式中, 上述执行主体可以按照如下步骤生成修正语句:

首先, 利用所得到的待定词语子集合中的待定词语, 生成至少一个候选修正语句, 以及得到至少一个候选修正语句中的候选修正语句的置信度。具体地, 上述执行主体可以利用现有的语句生成算法, 生成至少一个候选修正语句。作为示例, 假设对目标语句切词后得到的待修正词语集合为 “A B C D E”, 其中, 待修正词语 B 对应的待定词语子集合包括待定词语 “B1” 和 “B2”, 待修正词语 E 对应的待定词语子集合包括待定词语 “E1” 和 “E2”, 则生成的至少一个候选修正语句可以包括: “A B1 C D E1”、“A B2 C D E1”、“A B1 C D E2”、“A B1 C D E2”。

25 其中, 语句生成算法可以包括但不限于以下至少一种: 集束搜索算法、贪心搜索算法、维特比算法。

通常, 利用上述语句生成算法得到的至少一个候选修正语句中的每个候选修正语句, 可以具有对应的置信度。具体地, 置信度可以是候选修正语句包括的词语中, 各相邻词语对之间的概率之和, 其中, 30 相邻词语之间的概率用于表征两个相邻词语是正确的词语的概率。置

信度用于表征候选修正语句为正确语句的概率。置信度越大，表示步骤 201 中获取的目标语句被修改为该置信度对应的候选修正语句的概率越大。需要说明的是，上述语句生成算法是目前广泛研究和应用的公知技术，在此不再赘述。

5 然后，基于所得到的置信度的大小顺序，从至少一个候选修正语句中提取候选修正语句作为修正语句。具体地，修正语句的数量可以是至少一个。

 可选地，上述执行主体可以进一步将修正语句以各种方式输出。例如，将修正语句在与上述执行主体连接的显示器上显示，或者将修正语句发送到与上述执行主体通信连接的终端设备上显示。

10 继续参见图 3，图 3 是根据本实施例的用于生成修正语句的方法的应用场景的一个示意图。在图 3 的应用场景中，服务器 301 首先获取到用户利用终端设备 302 输入的搜索语句作为目标语句 303（例如“我喜欢吃流连”）。然后，服务器 301 对目标语句 303 进行切词，得到待修正词语集合 304（例如包括“我”，“喜欢”，“吃”，“流连”）。接着，服务器 301 确定目标存储区 305 中，词语“流连”具有对应的待
15 定词语集合 306（例如包括“榴莲”，“留恋”等）。而其他词语没有对应的待修正词语集合。再然后，服务器 301 利用贝叶斯公式，以及利用待
 定词语的错误概率和使用概率，计算出各个待
 定词语的修正概率
20 （例如“榴莲”的修正概率 3061 为 8%，
 “留恋”的修正概率 3062 为 6%）。再按照所得到的修正概率的大小顺序，从待
 定词语集合 306 中，提取预设数量个（例如两个）待
 定词语作为待
 定词语子集合 307。最后，服务器 301 利用集束搜索算法，生成修正语句 308（例如“我喜欢吃榴莲”），并将修正语句 308 发送给终端设备 302，以使终端设备 302
25 上显示该修正语句 308。

 本申请的上述实施例提供的方法，通过对目标语句进行切词，得到待修正词语集合，然后对于待修正词语集合中的待修正词语，确定该待修正词语对应的待
30 定词语集合，再基于待
 定词语的错误概率和使用概率，确定待
 定词语的修正概率，以及按照修正概率的大小顺序从待
 定词语集合中，提取预设数量个待
 定词语作为该待修正词语对应的

待定词语子集合，最后基于所得到的各个待定词语子集合，生成修正语句，从而提高了对语句进行修正的准确性。

进一步参考图 4，其示出了用于生成修正语句的方法的又一个实施例的流程 400。该用于生成修正语句的方法的流程 400，包括以下步骤：

步骤 401，获取目标语句，以及对目标语句进行切词，得到待修正词语集合。

在本实施例中，步骤 401 与图 2 对应实施例中的步骤 201 基本一致，这里不再赘述。

步骤 402，对于待修正词语集合中的待修正词语，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合；响应于确定不存在，确定该待修正词语包含的汉字的拼音；基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语；从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

在本实施例中，对于待修正词语集合中的待修正词语，上述执行主体可以执行如下步骤：

步骤 4021，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合。

在本实施例中，步骤 4021 与图 2 对应实施例中的步骤 2021 基本一致，这里不再赘述。

步骤 4022，响应于确定目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音。

在本实施例中，用于生成修正语句的方法的执行主体（例如图 1 所示的服务器）可以响应于确定目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音。例如，待修正词语为“正骨膏”，目标存储区中不存在与“正骨膏”对应的待定词语集合，则确定该待修正词语的拼音“zheng gu gao”。

步骤 4023，基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语。

在本实施例中，上述执行主体可以从预设的词库中，获取至少一个与所确定的拼音相匹配的匹配词语。其中，预设的词库可以预先设置在上述执行主体中，或者预先设置在与上述执行主体通信连接的电子设备中。作为示例，词库可以包括多个词语，以及每个词语的拼音。

5 所提取的至少一个匹配词语中的匹配词语可以是包括的汉字的拼音与所确定的拼音相同或相似的词语。具体地，待修正词语的拼音可以对应于预设的拼音集合，例如，该待修正词语的拼音为“zheng gu gao”，对应的拼音集合可以包括：“zhuang gu gao”、“zheng gu gao”，
10 则可以从预设的词库中，提取拼音为“zhuang gu gao”或“zheng gu gao”的词语作为匹配词语（例如至少一个匹配词语包括“壮骨膏”、“壮骨高”、“正骨膏”等）。

步骤 4024，从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

15 在本实施例中，上述执行主体可以按照各种方式从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。例如，可以从至少一个匹配词语中，随机选择预设数量个匹配词语作为待定词语子集合。

20 在本实施例的一些可选的实现方式中，所得到的至少一个匹配词语中的匹配词语具有对应的热度数值，热度数值用于表征词语被使用的频率。作为示例，热度数值可以是预设时间段（例如过去一个月）内，用户使用词语（例如使用词语进行搜索）的次数；或者可以是预设时间段内，用户使用词语的次数与使用预设的词语集合（例如预设时间段内某些用户使用的所有词语）中的词语的总次数的比值。

25 上述执行主体可以基于匹配词语对应的热度数值，按照如下步骤从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合：

30 按照热度数值的大小顺序，从所获取的至少一个匹配词语中选择预设数量个匹配词语作为待定词语子集合。作为示例，上述执行主体可以按照热度数值由大到小的顺序，从所获取的至少一个匹配词语中选择预设数量个匹配词语。通常，热度数值越大，表示匹配词语被使

用的频率越高，即匹配词语是待修正词语应当修改为的词语的可能性越大。

作为示例，假设预设数量为 2，至少一个匹配词语中，匹配词语“壮骨膏”的热度数值最大，匹配词语“正骨膏”的热度数值居其次。

5 则确定的待定词语子集合可以包括“壮骨膏”和“正骨膏”。

步骤 403，基于所得到的待定词语子集合，生成修正语句。

在本实施例中，步骤 403 与图 2 对应实施例中的步骤 203 基本一致，这里不再赘述。

10 从图 4 中可以看出，与图 2 对应的实施例相比，本实施例中的用于生成修正语句的方法的流程 400 突出了在目标存储区中不存在与待修正词语对应的待定词语集合的情况下，确定待定词语子集合的步骤。由此，本实施例描述的方案可以使待定词语子集合更加全面，从而可以进一步提高对语句进行修正的准确性。

15 进一步参考图 5，作为对上述各图所示方法的实现，本申请提供了一种用于生成修正语句的装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 5 所示，本实施例的用于生成修正语句的装置 500 包括：获取单元 501，被配置成获取目标语句，以及对目标语句进行切词，得到
20 待修正词语集合；提取单元 502，被配置成对于待修正词语集合中的待修正词语，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集合中，提取预设数量个待定词语作为待定词语子集合；生成单元 503，被配置成基于所得到的待定词语子集合，生成修正语句。

30 在本实施例中，获取单元 501 可以通过有线连接方式或者无线连

接方式从远程或从本地获取目标语句。其中，目标语句可以是用户利用其使用的终端设备（例如图 1 所示的终端设备）输入，并发送至上
述装置 500 的语句，或者是预先存储在上述装置 500 中的语句。上述
获取单元 501 可以进一步对所获取的目标语句进行切词，得到待修正
5 词语集合。具体地，上述获取单元 501 可以按照现有的切词方法（例
如最大正向匹配法、N-gram 模型方法、隐马尔科夫模型方法等）对目
标语句进行切词，得到待修正词语集合。

需要说明的是，待修正词语集合可以包括对上述目标语句进行切
词后所得到的全部词语，也可以包括部分符合预设条件的词语。例如，
10 可以对上述目标语句进行切词后所得到的每个词语进行词性标注，从
各个词语中提取预设词性（例如名词和/或动词）的词语作为待修正词
语。其中，对词语进行词性标注的方法是目前广泛研究和应用的公知
技术，在此不再赘述。

在本实施例中，对于待修正词语集合中的待修正词语，上述提取
15 单元 502 可以执行如下步骤：

步骤 5021，确定目标存储区中是否存在预设的、与该待修正词语
对应的待定词语集合。

其中，待定词语具有对应的错误概率和使用概率，错误概率用于
表征待定词语被错换为该待修正词语的概率。作为示例，假设某待定
20 词语 A 对应的错误概率为 10%，表示该待修正词语有 10% 的概率被错
用，且正确的词语为该待定词语 A。使用概率用于表征待定词语被使
用的概率。通常，使用概率可以是预先对预设的文本进行词频统计所
得到的、待定词语出现的频率。错误概率可以是预先对将待定词语错
换为待修正词语的次数进行统计所得到的概率。

25 步骤 5022，响应于确定存在，基于待定词语的错误概率和使用概
率，确定待定词语的修正概率。

其中，修正概率用于表征将该待修正词语修改为待定词语的概率。
通常，修正概率越大，表示该待修正词语发生错误的可能性越大，即
应当将该待修正词语修改为待定词语的可能性越大。作为示例，上述
30 提取单元 502 可以基于贝叶斯公式，计算得到待定词语的修正概率。

步骤 5023, 按照所得到的修正概率的大小顺序, 从待定词语集合中, 提取预设数量个待定词语作为待定词语子集合。

作为示例, 上述提取单元 502 可以按照各个修正概率由大到小的顺序, 从待定词语集合中, 提取预设数量个待定词语。需要说明的是, 5 当待定词语集合包括的待定词语的数量小于等于上述预设数量时, 可以将待定词语集合作为待修正词语对应的待定词语子集合。

在本实施例中, 生成单元 503 可以基于所得到的待定词语子集合, 生成修正语句。

作为示例, 上述生成单元 503 可以按照修正概率由大到小的顺序, 10 从各个待定词语子集合中分别提取待定词语, 将提取的待定词语与待修正词语集合中的、不具有对应的待定词语集合的待修正词语组合为修正语句。

在本实施例的一些可选的实现方式中, 目标存储区为预先建立的对应关系表, 对应关系表包括待修正词语和待定词语集合, 对应关系表用于表征待修正词语与待定词语集合的对应关系, 对应关系表还包 15 括待定词语的错误概率和使用概率。

在本实施例的一些可选的实现方式中, 待定词语的错误概率和使用概率预先按照如下步骤得到: 对于待定词语集合中的待定词语, 确定预设的、该待定词语对应的文本子集中的文本包括的、该待定词语 20 的数量为第一数量, 其中, 文本子集是预设的文本集的子集; 确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量; 将所确定的第二数量除以第一数量与第二数量之和, 得到该待定词语的错误概率; 确定该待定词语在文本集中的数量为第三数量, 以及将所确定的第三数量与文本集中的文本包括的 25 词语的总数的比值确定为该待定词语的使用概率。

在本实施例的一些可选的实现方式中, 生成单元 503 可以包括: 生成模块(图中未示出), 被配置成利用所得到的待定词语子集合中的待定词语, 生成至少一个候选修正语句, 以及得到至少一个候选修正语句中的候选修正语句的置信度; 提取模块(图中未示出), 被配置成 30 基于所得到的置信度的大小顺序, 从至少一个候选修正语句中提取候

选修正语句作为修正语句。

在本实施例的一些可选的实现方式中，提取单元 502 可以包括：
确定模块（图中未示出），被配置成响应于确定目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含
5 的汉字的拼音；获取模块（图中未示出），被配置成基于所确定的拼音，
获取至少一个与拼音相匹配的匹配词语；选择模块（图中未示出），被
配置成从所获取的至少一个匹配词语中，选择预设数量个匹配词语作
为待定词语子集合。

在本实施例的一些可选的实现方式中，所得到的至少一个匹配词
10 语中的匹配词语具有对应的热度数值，热度数值用于表征匹配词语被
使用的频率；以及选择模块可以进一步被配置成：按照热度数值的大
小顺序，从所得到的至少一个匹配词语中选择预设数量个匹配词语作
为待定词语子集合。

本申请的上述实施例提供的装置，通过对目标语句进行切词，得
15 到待修正词语集合，然后对于待修正词语集合中的待修正词语，确定
该待修正词语对应的待定词语集合，再基于待定词语的错误概率和使
用概率，确定待定词语的修正概率，以及按照修正概率的大小顺序从
待定词语集合中，提取预设数量个待定词语作为该待修正词语对应的
待定词语子集合，最后基于所得到的各个待定词语子集合，生成修正
20 语句，从而提高了对语句进行修正的准确性。

下面参考图 6，其示出了适于用来实现本申请实施例的服务器的
计算机系统 600 的结构示意图。图 6 示出的服务器仅仅是一个示例，
不应对本申请实施例的功能和使用范围带来任何限制。

如图 6 所示，计算机系统 600 包括中央处理单元（CPU）601，其
25 可以根据存储在只读存储器（ROM）602 中的程序或者从存储部分 608
加载到随机访问存储器（RAM）603 中的程序而执行各种适当的动作
和处理。在 RAM 603 中，还存储有系统 600 操作所需的各种程序和数
据。CPU 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入
30 /输出（I/O）接口 605 也连接至总线 604。

以下部件连接至 I/O 接口 605: 包括键盘、鼠标等的输入部分 606; 包括诸如液晶显示器 (LCD) 等以及扬声器等的输出部分 607; 包括硬盘等的存储部分 608; 以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 609。通信部分 609 经由诸如因特网的网络执行通信处理。驱动器 610 也根据需要连接至 I/O 接口 605。可拆卸介质 611, 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 610 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 608。

特别地, 根据本公开的实施例, 上文参考流程图描述的过程可以被实现为计算机软件程序。例如, 本公开的实施例包括一种计算机程序产品, 其包括承载在计算机可读介质上的计算机程序, 该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中, 该计算机程序可以通过通信部分 609 从网络上被下载和安装, 和/或从可拆卸介质 611 被安装。在该计算机程序被中央处理单元 (CPU) 601 执行时, 执行本申请的方法中限定的上述功能。

需要说明的是, 本申请所述的计算机可读介质可以是计算机可读信号介质或者计算机可读介质或者是上述两者的任意组合。计算机可读介质例如可以是一——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。在本申请中, 计算机可读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。而在本申请中, 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。

5 可以以一种或多种程序设计语言或其组合来编写用于执行本申请的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在
10 远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

15 附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。
20 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，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者
25 可以用专用硬件与计算机指令的组合来实现。

描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器包括获取单元、提取单元、生成单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，
30 例如，获取单元还可以被描述为“获取目标语句，以及对目标语句进

行切词，得到待修正词语集合的单元”。

作为另一方面，本申请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服务器中所包含的；也可以是单独存在，而未装配入该服务器中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该服务器执行时，使得该服务器：获取目标语句，以及对目标语句进行切词，得到待修正词语集合；对于待修正词语集合中的待修正词语，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响

5 应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集合中，提取预设数量个待定词语作为待定词语子集合；基于所得到的待定词语子集合，生成修正语句。

10 15

以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

20

权 利 要 求 书

1、一种用于生成修正语句的方法，包括：

5 获取目标语句，以及对所述目标语句进行切词，得到待修正词语集合；

 对于所述待修正词语集合中的待修正词语，确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错
10 换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集合中，提取预设数量个待定词语作为待定词语子集合；

15 基于所得到的待定词语子集合，生成修正语句。

 2、根据权利要求1所述的方法，其中，所述目标存储区为预先建立的对对应关系表，所述对应关系表包括待修正词语和待定词语集合，所述对应关系表用于表征待修正词语与待定词语集合的对对应关系，所
20 述对应关系表还包括待定词语的错误概率和使用概率。

 3、根据权利要求1所述的方法，其中，待定词语的错误概率和使用概率预先按照如下步骤得到：

 对于待定词语集合中的待定词语，确定预设的、该待定词语对应的
25 的文本子集中的文本包括的、该待定词语的数量为第一数量，其中，文本子集是预设的文本集的子集；确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量；将所确定的第二数量除以第一数量与第二数量之和，得到该待定词语的错误概率；确定该待定词语在所述文本集中的数量为第三数量，以及
30 将所确定的第三数量与所述文本集中的文本包括的词语的总数的比值确定为该待定词语的使用概率。

4、根据权利要求 1 所述的方法，其中，所述基于所得到的待定词语子集合，生成修正语句，包括：

5 利用所得到的待定词语子集合中的待定词语，生成至少一个候选修正语句，以及得到所述至少一个候选修正语句中的候选修正语句的置信度；

基于所得到的置信度的大小顺序，从所述至少一个候选修正语句中提取候选修正语句作为修正语句。

10 5、根据权利要求 1-4 之一所述的方法，其中，在所述确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合之后，所述方法还包括：

响应于确定所述目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音；

15 基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语；

从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

20 6、根据权利要求 5 所述的方法，其中，所得到的至少一个匹配词语中的匹配词语具有对应的热度数值，热度数值用于表征匹配词语被使用的频率；以及

所述从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合，包括：

25 按照热度数值的大小顺序，从所获取的至少一个匹配词语中选择预设数量个匹配词语作为待定词语子集合。

7、一种用于生成修正语句的装置，包括：

获取单元，被配置成获取目标语句，以及对所述目标语句进行切词，得到待修正词语集合；

30 提取单元，被配置成对于所述待修正词语集合中的待修正词语，

确定目标存储区中是否存在预设的、与该待修正词语对应的待定词语集合，其中，待定词语具有对应的错误概率和使用概率，错误概率用于表征待定词语被错换为该待修正词语的概率，使用概率用于表征待定词语被使用的概率；响应于确定存在，基于待定词语的错误概率和使用概率，确定待定词语的修正概率，其中，修正概率用于表征将该待修正词语修改为待定词语的概率；按照所得到的修正概率的大小顺序，从待定词语集合中，提取预设数量个待定词语作为待定词语子集合；

生成单元，被配置成基于所得到的待定词语子集合，生成修正语句。

8、根据权利要求7所述的装置，其中，所述目标存储区为预先建立的对应关系表，所述对应关系表包括待修正词语和待定词语集合，所述对应关系表用于表征待修正词语与待定词语集合的对应关系，所述对应关系表还包括待定词语的错误概率和使用概率。

9、根据权利要求7所述的装置，其中，待定词语的错误概率和使用概率预先按照如下步骤得到：

对于待定词语集合中的待定词语，确定预设的、该待定词语对应的文本子集中的文本包括的、该待定词语的数量为第一数量，其中，文本子集是预设的文本集的子集；确定该待定词语对应的文本子集中的文本包括的、该待定词语对应的待修正词语的数量为第二数量；将所确定的第二数量除以第一数量与第二数量之和，得到该待定词语的错误概率；确定该待定词语在所述文本集中的数量为第三数量，以及将所确定的第三数量与所述文本集中的文本包括的词语的总数的比值确定为该待定词语的使用概率。

10、根据权利要求7所述的装置，其中，所述生成单元包括：

生成模块，被配置成利用所得到的待定词语子集合中的待定词语，生成至少一个候选修正语句，以及得到所述至少一个候选修正语句中

的候选修正语句的置信度；

提取模块，被配置成基于所得到的置信度的大小顺序，从所述至少一个候选修正语句中提取候选修正语句作为修正语句。

5 11、根据权利要求 7-10 之一所述的装置，其中，所述提取单元包括：

确定模块，被配置成响应于确定所述目标存储区中不存在预设的、与该待修正词语对应的待定词语集合，确定该待修正词语包含的汉字的拼音；

10 获取模块，被配置成基于所确定的拼音，获取至少一个与拼音相匹配的匹配词语；

选择模块，被配置成从所获取的至少一个匹配词语中，选择预设数量个匹配词语作为待定词语子集合。

15 12、根据权利要求 11 所述的装置，其中，所得到的至少一个匹配词语中的匹配词语具有对应的热度数值，热度数值用于表征匹配词语被使用的频率；以及

所述选择模块进一步被配置成：

20 按照热度数值的大小顺序，从所得到的至少一个匹配词语中选择预设数量个匹配词语作为待定词语子集合。

13、一种服务器，包括：

一个或多个处理器；

存储装置，其上存储有一个或多个程序，

25 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-6 中任一所述的方法。

14、一种计算机可读介质，其上存储有计算机程序，其中，该程序被处理器执行时实现如权利要求 1-6 中任一所述的方法。

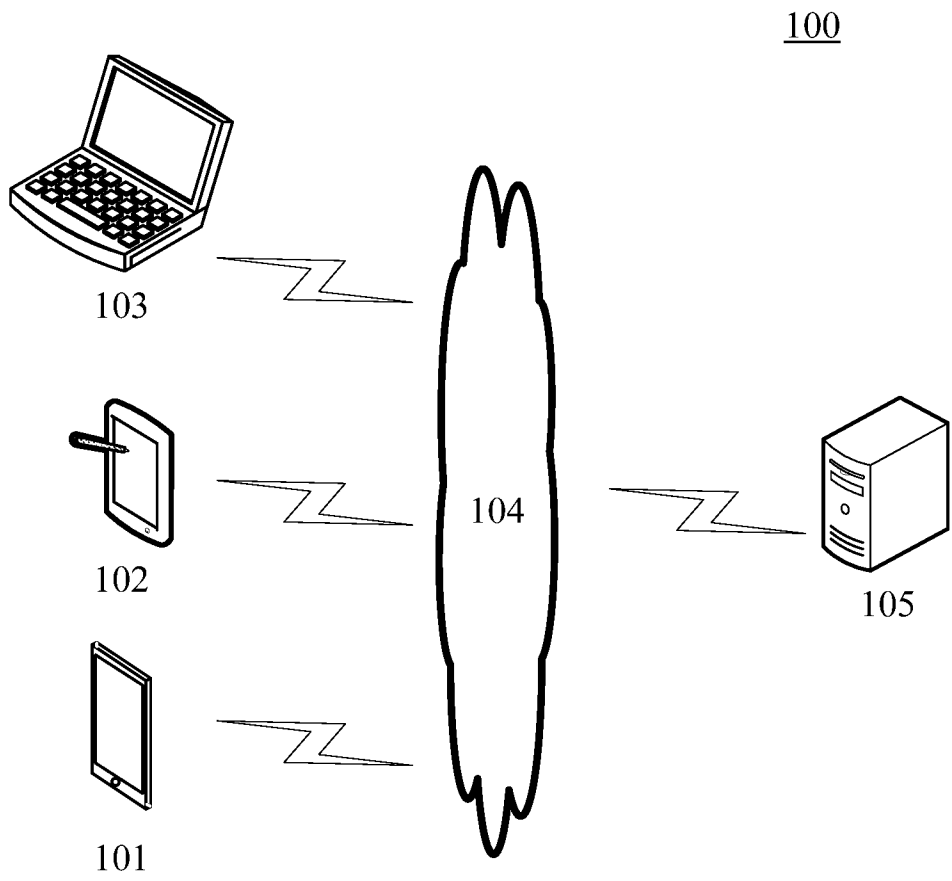


图 1

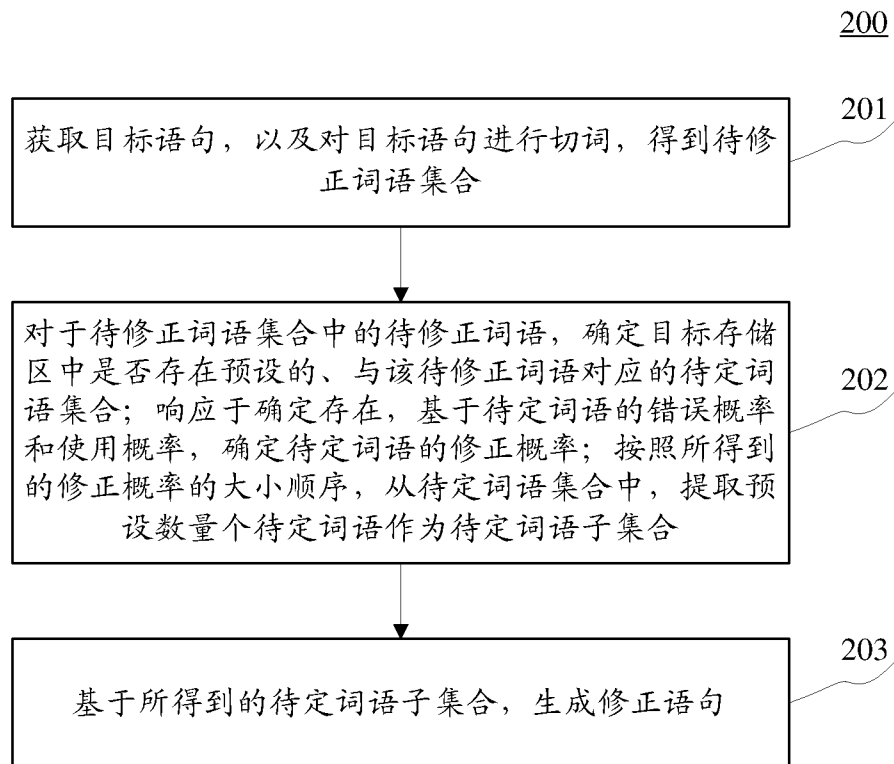


图 2

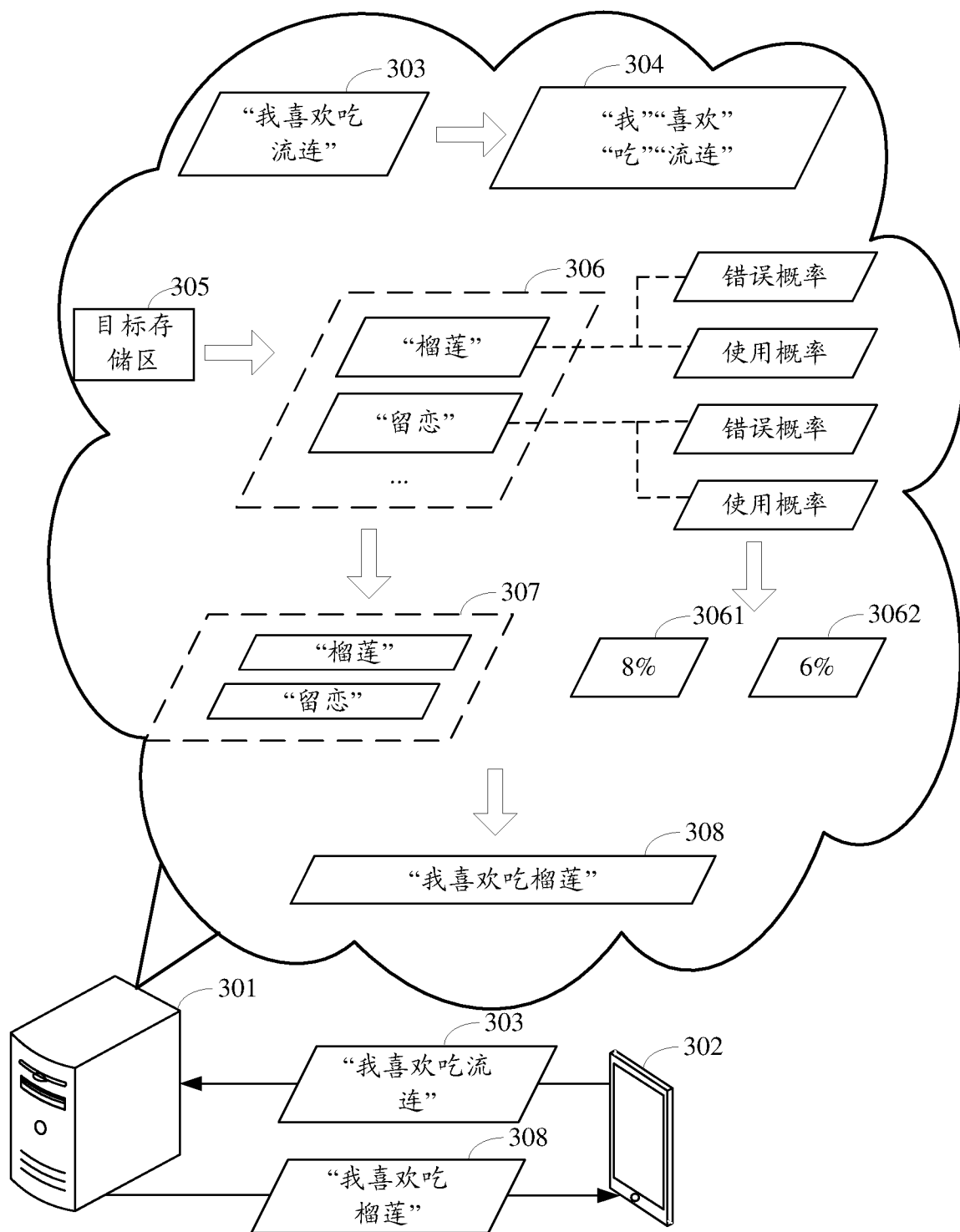


图 3

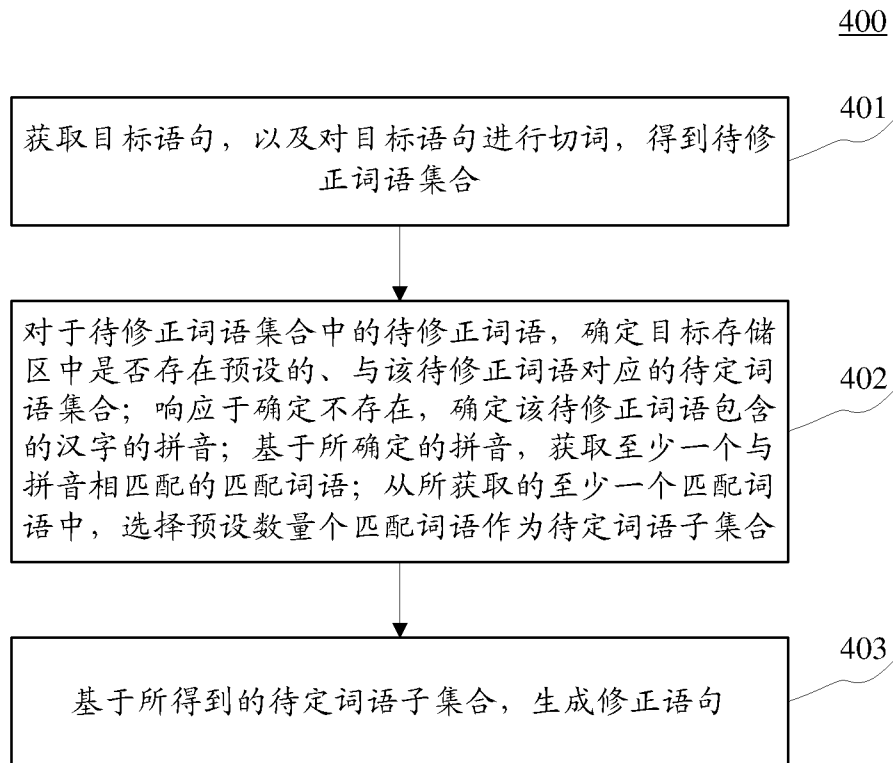


图 4

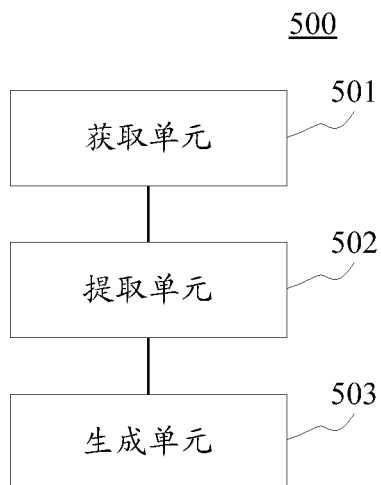


图 5

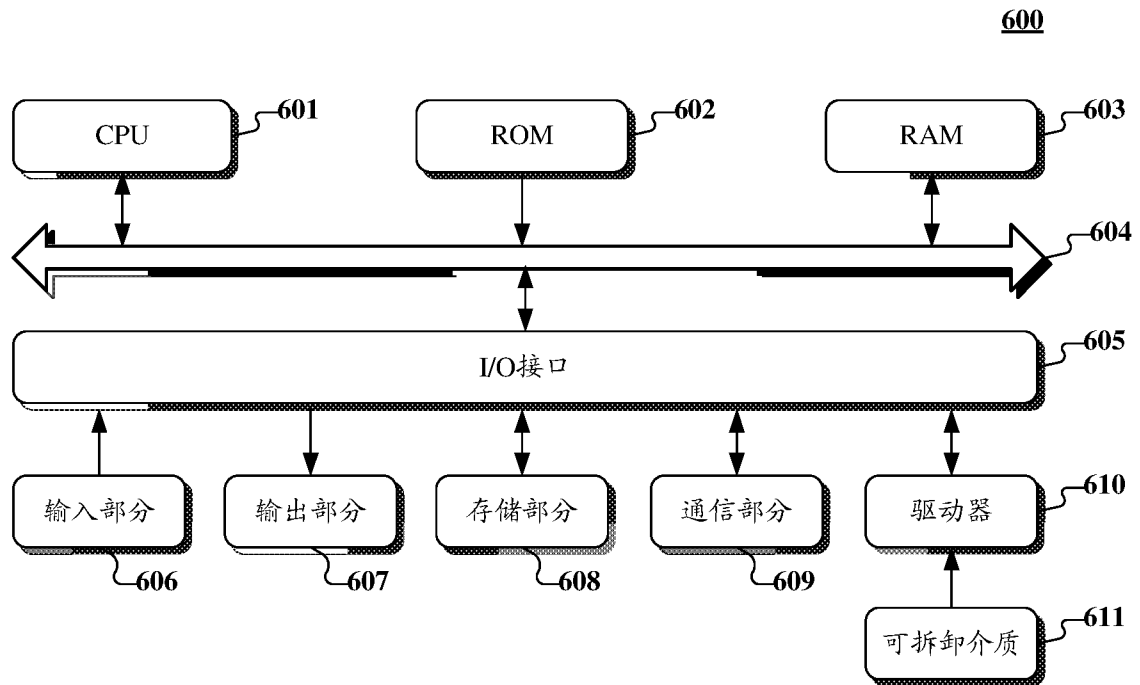


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/115953

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)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)

CNTXT; CNABS; CNKI; VEN; USTXT; WOTXT; EPTXT: 错误, 更正, 纠错, 修正, 拼写, 输入, 概率, 目标, 候选, 词语, 置信度, error, correct, spell, input, probability, target, candidate, word, confidence, degree

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 1387650 A (MICROSOFT CORPORATION) 25 December 2002 (2002-12-25) claim 27, description, p. 8, paragraphs 1-4, and p. 15, paragraph 4 to p. 18, paragraph 4, and figures 3-5	1-3, 5-9, 11-14
Y	CN 1387650 A (MICROSOFT CORPORATION) 25 December 2002 (2002-12-25) claim 27, description, p. 8, paragraphs 1-4, and p. 15, paragraph 4 to p. 18, paragraph 4, and figures 3-5	4, 10
Y	CN 106708893 A (HUAWEI TECHNOLOGIES CO., LTD.) 24 May 2017 (2017-05-24) description, paragraphs 102-123, and figures 3 and 4	4, 10
X	US 9471566 B1 (ORACLE AMERICA INC.) 18 October 2016 (2016-10-18) description, columns 15-17, and figures 7 and 8	1-3, 5-9, 11-14
A	CN 106250364 A (IFLYTEK CO., LTD.) 21 December 2016 (2016-12-21) entire document	1-14



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

03 June 2019

Date of mailing of the international search report

13 June 2019

Name and mailing address of the ISA/CN

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

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/115953

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	1387650	A	25 December 2002	WO	0135250	A3	06 June 2002
				AU	1086801	A	06 June 2001
				WO	0135250	A2	17 May 2001
				US	7302640	B2	27 November 2007
				CN	1205572	C	08 June 2005
				US	6848080	B1	25 January 2005
				HK	1050411	A1	09 December 2005
				US	2005044495	A1	24 February 2005
				US	2005086590	A1	21 April 2005
				JP	5535417	B2	02 July 2014
				JP	2003514304	A	15 April 2003
				US	7424675	B2	09 September 2008
CN	106708893	A	24 May 2017	CN	106708893	B	28 September 2018
				WO	2017084506	A1	26 May 2017
US	9471566	B1	18 October 2016	None			
CN	106250364	A	21 December 2016	None			

国际检索报告

国际申请号

PCT/CN2018/115953

A. 主题的分类 G06F 17/27 (2006.01) 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)) CNTXT; CNABS; CNKI; VEN; USTXT; WOTXT; EPTXT; 错误, 更正, 纠错, 修正, 拼写, 输入, 概率, 目标, 候选, 词语, 置信度, error, correct, spell, input, probability, target, candidate, word, confidence, degree																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5</td> <td>1-3、5-9、11-14</td> </tr> <tr> <td>Y</td> <td>CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5</td> <td>4、10</td> </tr> <tr> <td>Y</td> <td>CN 106708893 A (华为技术有限公司) 2017年 5月 24日 (2017 - 05 - 24) 说明书第102-123段, 图3-4</td> <td>4、10</td> </tr> <tr> <td>X</td> <td>US 9471566 B1 (ORACLE AMERICA INC) 2016年 10月 18日 (2016 - 10 - 18) 说明书第15-17栏, 图7-8</td> <td>1-3、5-9、11-14</td> </tr> <tr> <td>A</td> <td>CN 106250364 A (科大讯飞股份有限公司) 2016年 12月 21日 (2016 - 12 - 21) 全文</td> <td>1-14</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5	1-3、5-9、11-14	Y	CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5	4、10	Y	CN 106708893 A (华为技术有限公司) 2017年 5月 24日 (2017 - 05 - 24) 说明书第102-123段, 图3-4	4、10	X	US 9471566 B1 (ORACLE AMERICA INC) 2016年 10月 18日 (2016 - 10 - 18) 说明书第15-17栏, 图7-8	1-3、5-9、11-14	A	CN 106250364 A (科大讯飞股份有限公司) 2016年 12月 21日 (2016 - 12 - 21) 全文	1-14
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5	1-3、5-9、11-14																		
Y	CN 1387650 A (微软公司) 2002年 12月 25日 (2002 - 12 - 25) 权利要求27、说明书第8页第1-4段、第15页第4段-第18页第4段, 图3-5	4、10																		
Y	CN 106708893 A (华为技术有限公司) 2017年 5月 24日 (2017 - 05 - 24) 说明书第102-123段, 图3-4	4、10																		
X	US 9471566 B1 (ORACLE AMERICA INC) 2016年 10月 18日 (2016 - 10 - 18) 说明书第15-17栏, 图7-8	1-3、5-9、11-14																		
A	CN 106250364 A (科大讯飞股份有限公司) 2016年 12月 21日 (2016 - 12 - 21) 全文	1-14																		
☐ 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 2019年 6月 3日		国际检索报告邮寄日期 2019年 6月 13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 刘细金 电话号码 86-(20)-28950395																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/115953

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	1387650	A	2002年 12月 25日	WO	0135250	A3	2002年 6月 6日
				AU	1086801	A	2001年 6月 6日
				WO	0135250	A2	2001年 5月 17日
				US	7302640	B2	2007年 11月 27日
				CN	1205572	C	2005年 6月 8日
				US	6848080	B1	2005年 1月 25日
				HK	1050411	A1	2005年 12月 9日
				US	2005044495	A1	2005年 2月 24日
				US	2005086590	A1	2005年 4月 21日
				JP	5535417	B2	2014年 7月 2日
				JP	2003514304	A	2003年 4月 15日
				US	7424675	B2	2008年 9月 9日
CN	106708893	A	2017年 5月 24日	CN	106708893	B	2018年 9月 28日
				WO	2017084506	A1	2017年 5月 26日
US	9471566	B1	2016年 10月 18日	无			
CN	106250364	A	2016年 12月 21日	无			