(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 102193639 B
(45)授权公告日 2014.03.12

(21)申请号 201010120044.X

(22)申请日 2010.03.04

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛资本大厦一座四层 847 号邮箱

(72)发明人 薛永刚 陈培军 秦吉胜 侯磊

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘

(51)Int.Cl.
G06F 3/023(2006.01)

审查员 聂鹏

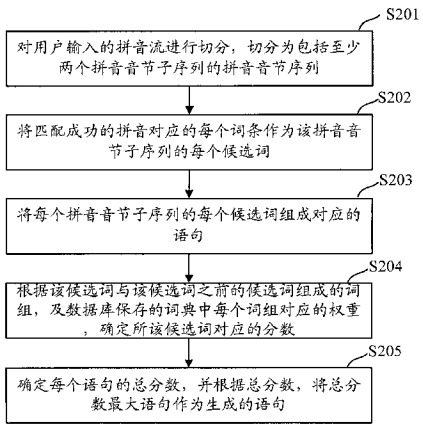
权利要求书3页 说明书11页 附图5页

(54)发明名称

一种语句生成方法及装置

(57)摘要

本申请公开了一种语句生成方法及装置，用以解决现有技术中拼音输入法生成的语句准确性低的问题。该方法将拼音流切分后的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，确定该拼音音节子序列的每个候选词，将每个候选词组成对应的语句，针对每个语句的候选词与该候选词之前或之后的候选词组成的词组，及词典中每个词组对应的权重，确定该候选词对应的分数，根据所述每个语句中每个候选词的分数，确定每个语句的总分数，将总分数最大的语句作为生成的语句。由于只有经常出现的词组对应的权重才会比较高，即经常出现的词组一定是用户经常使用，或满足语言规则的词组，因此采用该方法可以使生成的语句更加的准确。



CN 102193639 B

1. 一种语句生成方法，其特征在于，包括：

将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；其中，所述词典包括一元词典和二元词典，其中所述一元词典中保存多个词条，每个词条对应的拼音，以及每个词条对应的权重，所述二元词典中保存词组，以及每个词组的权重；

将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；其中，确定该候选词对应的分数包括：判断所述候选词是否为所述语句的第一个候选词，当所述候选词为第一个候选词时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数，否则，判断所述候选词与所述候选词之前的候选词组成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

2. 如权利要求 1 所述的方法，其特征在于，确定每个语句的总分数之前所述方法进一步包括：

根据每个语句中已确定分数的候选词，及该已确定分数的候选词对应的分数，确定每个语句的子分数；

根据所述每个语句的子分数，按照子分数由大到小的顺序选择设定数量的语句作为准备确定总分数的语句。

3. 如权利要求 1 所述的方法，其特征在于，确定所述每个语句的总分数包括：

根据所述每个语句中每个候选词的分数，将所述每个候选词的分数进行乘积或累加运算，将每个候选词的分数进行乘积或累加运算得到的分数，作为该语句的总分数。

4. 一种语句生成装置，其特征在于，包括：

匹配模块，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

存储模块，用于保存一元词典及二元词典，其中所述一元词典中保存词条，每个词条对应的拼音，以及每个词条对应的权重，所述二元词典中保存词组，以及每个词组的权重；

分数确定模块，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；

其中，所述分数确定模块包括判断单元、第一分数确定单元、第二分数确定单元；所述判断单元，用于判断所述候选词是否为所述语句的第一个候选词；所述第一分数确定单元，用于确定所述候选词为所述语句的第一个候选词时，在一元词典中查找与所述候选词匹配

的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；所述第二分数确定单元，用于确定所述候选词非所述语句中第一个候选词时，判断所述候选词与所述候选词之前的候选词组成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

语句生成模块，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

5. 如权利要求 4 所述的装置，其特征在于，所述语句生成模块还用于，

根据每个语句中已确定分数的候选词，及该已确定分数的候选词对应的分数，确定每个语句对应的子分数，按照子分数由大到小的顺序选择选择设定数量的语句作为准备确定总分数的语句。

6. 如权利要求 4 所述的装置，其特征在于，所述语句生成模块在确定每个语句的总分数时具体用于，

根据所述每个语句中每个候选词的分数，将所述每个候选词的分数进行乘积或累加运算，将每个候选词的分数进行乘积或累加运算得到的分数，作为该语句的总分数。

7. 一种语句生成方法，其特征在于，所述方法包括：

将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；其中，确定该候选词对应的分数包括：判断所述候选词是否为所述语句的最后一个候选词，当所述候选词为最后一个候选词时，在所述词典的一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数，否则，判断所述候选词与所述候选词之后的候选词组成的词组是否在所述词典的二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

8. 一种语句生成装置，其特征在于，所述装置包括：

匹配模块，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

分数确定模块，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；

其中，所述分数确定模块包括判断单元、第一分数确定单元、第二分数确定单元；所述判断单元，用于判断所述候选词是否为所述语句的最后一个候选词；所述第一分数确定单元，用于确定所述候选词为最后一个候选词时，在所述词典的一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；所述第二分数确定单元，用于确定所述候选词非最后一个候选词时，判断所述候选词与所述候选词之后的候选词组成的词组是否在所述词典的二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

语句生成模块，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

一种语句生成方法及装置

技术领域

[0001] 本申请涉及汉字输入技术领域，尤其涉及一种语句生成方法及装置。

背景技术

[0002] 输入法(Input Method Editor，IME)是利用键盘，根据一定的编码规则，实现汉字输入的一种方法，而拼音输入法则是利用键盘输入拼音，从而实现汉字输入的方法。在通过拼音输入法进行汉字输入的过程中，针对用户输入的拼音流，需要将该拼音流进行切分，切分为多个合法的拼音音节序列，并将切分后的每个拼音音节转换为对应的汉字，从而实现语句的输出。

[0003] 当把用户输入的汉字切分为拼音音节序列时，由于每个拼音音节对应的候选词方案很多，因此根据用户输入的拼音流可能得到很多的语句。在现有技术中一般采用最大概率法从众多的语句中选择一个输出，即在多个候选词的组合中确定概率最大的一个组合方法，作为最后的语句输出结果。

[0004] 如图1所示根据拼音流确定的多个候选词组合方案，当输入拼音流“dongtianhaoleng”并将拼音流切分为多个拼音音节序列时，每个音频音节对应不同的候选词，如图1所示，对于拼音音节“dong”其对应的候选词包括：动、懂……东等，对于拼音音节“tian”其对应的候选词包括：添、填……天等，对于拼音音节“hao”其对应的候选词包括：豪、号……好等，对于拼音音节“leng”其对应的候选词包括：棱、楞……冷等，并且对于两个拼音音节“冬天”其本身也对应很多候选词例如冬天、洞天……动天等。因此在根据最大概率法确定输出的语句时，一般选择概率较大的候选词组合，如图1虚线所示即为选择的概率最大的候选词组合“冬天好冷”。

[0005] 由于在采用最大概率法进行语句输出时，选择概率最大的候选词组合，但 是即使每个候选词的权重都很大，多个权重很大的候选词组合成的语句也可能并不是用户所需的语句，从而导致语句的生成结果准确性较低。

发明内容

[0006] 有鉴于此，本申请实施例提供一种语句生成方法及装置，用以解决现有技术中拼音输入法生成的语句准确性低的问题。

[0007] 本申请实施例提供的一种语句生成方法，包括：

[0008] 将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；其中，所述词典包括一元词典和二元词典，其中所述一元词典中保存多个词条，每个词条对应的拼音，以及每个词条对应的权重，所述二元词典中保存词组，以及每个词组的权重；

[0009] 将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及所述词典中每个词组对应的权

重，确定该候选词对应的分数；其中，确定该候选词对应的分数包括：判断所述候选词是否为所述语句的第一个候选词，当所述候选词为第一个候选词时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数，否则，判断所述候选词与所述候选词之前的候选词组成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

[0010] 根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0011] 本申请实施例提供的一种语句生成装置，包括：

[0012] 匹配模块，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

[0013] 存储模块，用于保存一元词典及二元词典，其中所述一元词典中保存词条，每个词条对应的拼音，以及每个词条对应的权重，所述二元词典中保存词组，以及每个词组的权重；

[0014] 分数确定模块，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；

[0015] 其中，所述分数确定模块包括判断单元、第一分数确定单元、第二分数确定单元；所述判断单元，用于判断所述候选词是否为所述语句的第一个候选词；所述第一分数确定单元，用于确定所述候选词为所述语句的第一个候选词时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；所述第二分数确定单元，用于确定所述候选词非所述语句中第一个候选词时，判断所述候选词与所述候选词之前的候选词组成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

[0016] 语句生成模块，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0017] 本申请实施例提供的一种语句生成方法，包括：

[0018] 将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

[0019] 将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；其中，确定该候选词对应的分数包括：判断所述候选词是否为所述语句的最后一个候选词，当所述候选词为最后一个候选词时，在所述词典的一元词

典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数，否则，判断所述候选词与所述候选词之后的候选词组成的词组是否在所述词典的二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

[0020] 根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0021] 本申请实施例提供的一种语句生成装置，包括：

[0022] 匹配模块，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

[0023] 分数确定模块，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；

[0024] 其中，所述分数确定模块包括判断单元、第一分数确定单元、第二分数确定单元；所述判断单元，用于判断所述候选词是否为所述语句的最后一个候选词；所述第一分数确定单元，用于确定所述候选词为最后一个候选词时，在所述词典的一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；所述第二分数确定单元，用于确定所述候选词非最后一个候选词时，判断所述候选词与所述候选词之后 的候选词组成的词组是否在所述词典的二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

[0025] 语句生成模块，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0026] 本申请实施例提供了一种语句生成方法及装置，该方法包括：将拼音流切分后的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词，将每个候选词组成对应的语句，针对每个语句的每个候选词与该候选词之前或之后的候选词组成的词组，及词典中每个词组对应的权重，确定该候选词对应的分数，根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。由于只有经常出现的词组对应的权重才会比较高，即经常出现的词组一定是用户经常使用，或满足语言规则的词组，因此采用该方法可以使生成的语句更加的准确。

附图说明

[0027] 图 1 为现有技术中根据拼音流确定的多个候选词组合方案；

[0028] 图 2 为本申请实施例提供的语句生成的过程；

[0029] 图 3 为本申请实施例提供的语句生成的详细过程；

[0030] 图 4 为本申请实施例提供的语句生成的另一详细过程；

[0031] 图 5 为本申请实施例提供的语句生成的装置结构示意图；

具体实施方式

[0032] 图 6 为本申请实施例提供的另一语句生成的装置结构示意图。

[0033] 本申请实施例为了有效的提高语句输出的准确性，提供了一种语句生成的方法，在该方法中充分考虑了构成语句的每两个候选词组成的词组出现的权重，确定相应的分数，并进而确定语句的总分数，根据确定的语句的总分数，选择总分数最大的语句作为生成的语句输出。因为只有经常出现的词组对应的权重才会比较高，即经常出现的词组一定是用户经常使用，或满足语言规则的词组，因此采用该方法可以使生成的语句更加的准确。本申请实施例中的语句生成方法可以适用于生成一个完整的句子，也可以适用于生成一个完整句子的组成部分，且该语句可以是长句也可以是短句，本申请对此并不做限定。

[0034] 下面结合说明书附图，对本申请实施例进行详细说明。

[0035] 图 2 为本申请实施例提供的语句生成的过程，该过程包括以下步骤：

[0036] S201：对用户输入的拼音流进行切分，切分为包括至少两个拼音音节子序列的拼音音节序列，其中每个拼音子序列中包括至少一个拼音音节。

[0037] 对用户输入的拼音流进行切分，将其切分为合法的拼音音节序列，其中在该切分后获得的合法的拼音音节序列中包括至少两个拼音音节子序列。

[0038] S202：将拼音音节序列中的各拼音音节子序列与数据库中字典保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词。

[0039] 在本申请实施例中为了便于查询每个候选词，在数据库中保存有一元词典，该一元词典中保存有多个词条，每个词条对应的拼音，以及每个词条对应的权重。

[0040] S203：将每个拼音音节子序列对应的每个候选词组成对应的语句。

[0041] 根据拼音音节序列中每个拼音音节子序列对应的每个候选词，组成对应的每个语句，在本申请实施例中由于每个拼音音节子序列对应多个候选词，因此也会组成多个语句。

[0042] 例如拼音音节序列中包括拼音音节子序列 1、2、3，其中拼音音节子序列 1 对应的候选词分别为 A，拼音音节子序列 2 对应的候选词为 D、E，拼音音节 子序列 3 对应的候选词为 F、G，则该拼音音节序列 123 组成的对应语句包括 ADF，ADG，AEF，AEG。

[0043] S204：针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及数据库保存的词典中每个词组对应的权重，确定该候选词对应的分数。

[0044] 在本申请实施例中为了充分考虑不同词条之间的共同出现的关系，在数据库中保存了二元词典，在二元词典中保存有多个词组，并且保存有每个词组对应的权重，其中每个词组包括两个词条。同时由于每个语句由对应的候选词构成，针对每个语句中的每个候选词，由于每个候选词对应的分数的确定过程相同，因此针对语句中的每个候选词，在确定该候选词的分数时，根据该候选词与该候选词之前的候选词组成的词组，确定该候选词的对应的分数。

[0045] S205：根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0046] 由于在本申请实施例中考虑了词组出现的权重，在确定分数时，根据每个词组出

现的权重，以及设置的权重系数，确定每个候选词的分数。

[0047] 本申请实施例的词典中包括一元词典和二元词典，其中一元词典中保存有多个词条，每个词条对应的拼音，并且保存有每个词条对应的权重，表 1 为本申请实施例中一元词典的存储结构示意。

[0048]

词条	拼音	权重
冬天	D ong’t ian	100
洞天	D ong’t ian	54
朝阳	Zh ao’y ang	280
朝阳	Ch ao’y ang	89
朝野	Ch ao’y e	752
…	…	…
阿里巴巴	A’l I’b a’b a	189

[0049] 表 1

[0050] 二元词典中保存有多个词组，并且保存有每个词组对应的权重，其中每个词组包括两个词条，表 2 为二元词典的存储结构示意。其中在本申请实施例中 一元词典和二元词典中保存的信息，根据对大量的数据信息学习获取，即通过对大量数据信息的扫描、分词，并统计分词后的每个词条的权重，以及每个词组的权重，将统计的信息分别保存即可获取一元词典和二元词典。

[0051]

第一词条	第二词条	权重
打	酱油	300
天气	真好	56
举行	会议	765
词典	大小	32
…	…	…
淘宝	卖家	650

[0052] 表 2

[0053] 在一元词典中保存了词条信息，并在二元词典中保存了词组的信息后，当对用户输入的拼音流进行转换生成语句时，由于各拼音音节子序列与一元词典中拼音匹配成功时，匹配成功的拼音对应的词条很多，在本申请实施例中将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词。由于每个拼音音节子序列对应的候选词很多，按照拼音音节序列中的各拼音音节子序列组合而成的语句也会很多，针对每个语句中每两个候选词组成的词组是否在二元词典中出现，可以确定语句中每个候选词对应的分数，从而可以确定语句的总分数。

[0054] 在本申请实施例中为了提高语句生成的效率，减小语句生成的工作量，在每个语句中，当确定了部分候选词的分数后，可以根据该已确定分数的候选词，及该已确定分数的候选词对应的分数，确定每个语句的子分数，根据确定的每个语句的子分数，按照子分数由大到小的顺序选择设定数量的语句作为准备确定总分数的语句。在该准备确定总分数的语句中，每确定一个候选词的分数，即可计算该语句的子分数，根据确定的子分数及设定数量，进行准备确定总分数的语句的选择。由于选择了设定数量的语句作为后续确定总分数的语句，进行计算的语句数量减小，从而减小了存储空间，进而提高了语句生成的效率。

[0055] 下面通过具体的实施例详细说明，确定每个语句的总分数的过程。当确定 了拼音

音节序列对应的每个语句后，针对每个语句中的每个候选词，判断该候选词是否为该语句的第一个候选词，当该候选词为该语句的第一个候选词时，由于第一个候选词之前不存在其他的候选词，因此在确定第一个候选词的分数时，在一元词典中查找与该候选词匹配的词条对应的权重，根据该权重及保存的第二权重系数，确定该候选词的分数。其中，第二权重系数为不能与其他候选词组成词组的候选词对应的权重系数，可以为 0 和 1 之间的数。

[0056] 当该候选词非第一个候选词时，该候选词之前的候选词存在，因此在确定该候选词对应的分数时，将该候选词与该候选词之前的候选词组成词组，判断在二元词典中是否存在该词组，当二元词典中存在该词组时，查找该词组对应的权重，根据查找的权重及保存的第一权重系数，确定该候选词对应的分数。其中第一权重系数为能够组成词组的候选词对应的权重系数，可以为 0 和 1 之间的数，并且每次在生成语句的过程中，第一权重系数大于第二权重系数。

[0057] 当在二元词典中不存在该词组时，在一元词典中查找与该该候选词匹配的词条对应的权重，根据查找的该权重及保存的该第二权重系数，确定该该候选词对应的分数。

[0058] 当依据上述方法确定了每个语句中每个候选词的分数后，可以将每个候选词对应的分数进行乘积运算，或进行累加运算，根据该乘积或累加运算得到的分数，作为该语句的总分数。例如语句包括 A、B、C 三个候选词，其中候选词 A 对应的分数为 W1，根据候选词 A 和 B 组成的词组确定候选词 B 对应的分数为 W2，根据候选词 B 和 C 组成的词组确定候选词 C 对应的分数为 W3，则该语句的总分数为 W1+W2+W3，或者该语句的总分数为 W1×W2×W3。

[0059] 为了提高语句生成的效率，减小语句生成的计算工作量，在本申请实施例中当根据切分后的拼音音节序列中的第一个拼音音节子序列，与一元词典保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为第一个拼音音节子序列的每个候选词，当根据一元词典中保存的每个词条的权重，以及保存的第二权重系数确定了每个候选词的分数后，可以根据分数计算的结果，按 照分数由大到小的顺序选择设定数量的候选词作为待生成的语句中的第一个拼音音节子序列对应的候选词。

[0060] 之后，将第二个拼音音节子序列对应的每个候选词及选择的第一个拼音音节对应的每个候选词，分别组成词组，针对每个词组，确定第二个拼音音节子序列对应的候选词的分数，将该第二个拼音音节子序列对应的候选词的分数，及该词组中第一个拼音音节子序列对应的候选词的分数进行乘积或累加运算，确定由该词组组成的语句的子分数，根据该子分数，按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句。

[0061] 在进行后续计算过程中，依次确定每个语句的子分数，按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句，因此组成的语句的数量会相应的减小，从而减小在计算过程中由于存储每个语句而占用的存储空间，并且同样也可以减小后续确定每个语句的总分数的工作量，从而有效的提高语句生成的效率。

[0062] 本申请实施例中在根据拼音音节序列中的各拼音音节子序列，与一元词典中保存的各词条的拼音进行匹配，获取每个拼音音节子序列对应的每个候选词时，由于拼音音节序列中与一元词典中各词条的拼音匹配的拼音音节的数量不同，即拼音音节子序列包含的拼音音节的数量不同，因此获取的候选词包含的字节的数量也不同。

[0063] 例如对于拼音音节序列“dong’tian’hao’leng”，当拼音音节序列中的拼音音节子

序列“dong”与一元词典中各词条的拼音匹配时，匹配成功的为拼音“dong”的词条，该词条可能是“东”，“动”“懂”等。当然在匹配的过程中，也可能是拼音音节序列中的拼音音节子序列“dong’tian”与一元词典中各词条的拼音匹配时，匹配成功的为拼音为“dong’tian”的词条，该词条可能是“冬天”，“洞天”“动天”等。

[0064] 因此由于获取的每个候选词的长度不同，在根据权重及保存的权重系数确定每个语句的子分数时，可以针对候选词构成的语句的长度进行选择。例如当 拼音音节子序列对应的候选词的长度为 2 时，例如为“dong’tian”，则可以确定该候选词组成的语句对应的子分数，即确定“dong’tian”对应的候选词组成的语句对应的子分数，当然也可以根据拼音音节子序列“dong”和“tian”分别对应的候选词确定组成的语句的子分数，根据该“dong”和“tian”组成的语句的子分数以及“dong’tian”组成的语句的子分数，按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句。

[0065] 图 3 为本申请实施例提供的语句生成的详细过程，该过程包括以下步骤：

[0066] S301：对用户输入的拼音流“dongtianleng”进行切分，切分为包括 3 个拼音音节的拼音音节序列“dong’tian’leng”。

[0067] S302：将拼音音节序列中的每个拼音音节子序列，与一元词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词。

[0068] 例如对于拼音音节子序列“dong”，在一元词典中匹配与拼音音节子序列“dong”匹配的拼音，将匹配成功的拼音对应的每个词条“动”、“东”、“冬”等，作为该拼音音节子序列“dong”对应的每个候选词。当拼音音节子序列为“dong’tian”时，根据该拼音音节子序列“dong’tian”，在一元词典中匹配与拼音音节子序列“dong’tian”匹配的拼音，将匹配成功的拼音对应的每个词条“冬天”、“洞天”、“动天”，作为该拼音音节子序列“dong’tian”对应的每个候选词。

[0069] S303：在确定的每个拼音音节子序列对应的每个候选词中，根据拼音音节序列中各拼音音节子序列的顺序，将每个拼音音节子序列对应的每个候选词组成对应的语句。

[0070] 例如，获取拼音音节序列“dong’tian’leng”中与每个拼音音节子序列对应的每个候选词包括，与拼音音节子序列“dong”对应的候选词包括“东”，“动”，与拼音音节子序列“tian”对应的候选词包括“田”，与拼音音节子序列“leng”对应的候选词包括“冷”，“棱”，以及与拼音音节子序列“dong’tian” 对应的候选词包括“冬天”，“洞天”，则可以组成的语句包括“东田冷”、“东田棱”、“动田棱”、“动田冷”、“冬天冷”、“洞天棱”等。

[0071] S304：针对每个语句中的每个候选词，判断当前进行分数确定的候选词是否为该语句的第一个拼音音节子序列对应的候选词，即判断该候选词是否为该语句的第一个候选词，当判断结果为是时，进行步骤 S305，否则，进行步骤 S306。

[0072] 例如针对语句“动田冷”，当前进行判断的候选词为“动”时，则可以确定该候选词为第一个拼音音节子序列“dong”对应的候选词，即该候选词为该语句的第一个候选词。当针对语句“冬天冷”，当前进行判断的候选词为“冬天”时，则可以确定该候选词为第一个拼音音节子序列“dong’tian”对应的候选词，即该候选词为该语句的第一个候选词。

[0073] S305：在一元词典中查找与该候选词匹配的词条对应的权重，根据查找的该权重以及保存的第二权重系数 R2，确定该该候选词对应的分数。

[0074] 其中确定该候选词对应的分数的过程包括：计算该候选词对应的权重，及第二权

重系数 R2 的乘积，将乘积结果确定为该候选词对应的分数。

[0075] S306：确定该语句中该候选词之前的候选词，将该候选词与该候选词之前的候选词组合，根据组合后获得的词组，判断该词组是否在二元词典中存在，当二元词典中不存在该词组时，进行步骤 S307，否则，进行步骤 S308。

[0076] S307：在一元词典中查找与该候选词匹配的词条对应的权重，根据查找的所述权重，以及保存的第二权重系数 R2，确定该候选词对应的分数。

[0077] S308：查找该候选词和该候选词之前的候选词组成的词组在二元词典中对应的权重，根据查找的权重，以及保存的第一权重系数 R1，确定该候选词对应的分数。

[0078] 例如该候选词为“冷”，该候选词之前的候选词为“洞天”，则该候选词和该候选词之前的候选词组成词组“洞天冷”，在二元词典中查找是否存在“洞天冷”的词组。当二元词典中不存在“洞天冷”时，在一元词典中查找与该候 选词“冷”对应的词条“冷”对应的权重，根据该权重以及保存的第二权重系数 R2，确定该候选词“冷”对应的分数。当二元词典中存在“洞天冷”时，则在二元词典中查找“洞天冷”对应的权重，根据该权重以及保存的第一权重系数 R1，确定该候选词“冷”对应的分数。

[0079] S309：针对每个语句，根据每个语句中每个候选词对应的分数，确定每个语句的总分数，根据每个语句的总分数，将总分数最大的语句作为生成的语句。

[0080] 在本申请实施例中当至少两个语句的总分数都最大时，在该至少两个语句中任意选择一个作为生成的语句。

[0081] 在本申请实施例中还可以根据图 4 所示的语句的生成方法，进行语句的生成，该生成方法包括：

[0082] S401：对用户输入的拼音流“dongtianleng”进行切分，切分为包括 3 个拼音音节的拼音音节序列“dong’tian’leng”。

[0083] S402：将拼音音节序列中的第一个拼音音节子序列，与一元词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为对应第一个拼音音节子序列的每个第一个候选词。

[0084] 其中，该第一个拼音音节子序列可以为第一个拼音音节，也可以为前几个拼音音节组成的第一拼音音节子序列。例如，该第一拼音音节子序列可以为第一个拼音音节“dong”，也可以为第一个拼音音节和第二个拼音音节组成的第一拼音音节子序列“dong’tian”，根据每个第一个拼音子序列可以在一元词典中确定每个第一个拼音子序列对应的每个第一个候选词。

[0085] S403：在一元词典中查找与每个第一个候选词对应的词条的权重，根据该权重，以及保存的第二权重系数，确定由该每个第一个候选词组成的语句的子分数。

[0086] S404：根据该每个语句的子分数，按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句。例如按照子分数由大到小的顺序选择子分数较大的 20 个或 30 个第一个候选词组成的语句作为准备确定 总分数的语句。

[0087] 在本申请实施例中由于第一个选候选词长度不同，例如可以为“东”，“动”“懂”等或“冬天”，“洞天”，“动天”等，因此在选择第一个候选词组成的语句时，也可以根据第一个候选词长度的不同选择对应数量的第一个候选词组成的语句进行后续计算，例如当选择 20 个第一个候选词时，可以选择第一个候选词长度为 1，构成的语句的子分数较大的 10 个语

句作为准备确定总分数的语句，选择第一个候选词长度为 2，构成的语句的子分数较大的 10 个语句作为准备确定总分数的语句，具体选择可以根据需要灵活设定。

[0088] S405：将第二个拼音音节子序列与一元词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为第二个拼音音节子序列的每个第二个候选词。

[0089] 当第一个拼音音节子序列为“dong”时，第二个拼音音节子序列为“tian”，当第一个拼音音节子序列为“dong' tian”时，第二个拼音音节子序列为“leng”。

[0090] S406：将选择出的每个第一个候选词，和根据匹配确定的每个第二个候选词组成语句，并根据每个语句中第一个候选词和第二个候选词组成词组，确定二元词典中是否存在该词组，当确定存在时，进行步骤 S407，否则，进行步骤 S408。

[0091] S407：在二元词典中查找该词组对应的权重，并根据保存的第一权重系数 R1，确定该第二个候选词对应的分数。

[0092] S408：在一元词典中查找与该第二个候选词匹配的词条对应的权重，根据该权重以及保存的第二权重系数 R2，确定该第二个候选词对应的分数。

[0093] S409：根据每个语句中第一个候选词对应的分数，以及第二个候选词对应的分数，确定第一个候选词与第二个候选词组成的该语句的子分数，根据所述子分数，按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句。

[0094] S410：判断该第二个拼音音节子序列是或否为拼音音节序列中最后一个拼 音音节子序列，当判断结果为是时，进行步骤 S411，否则，将第三个拼音音节子序列作为第二个拼音音节，将选择的每个语句中第二个拼音音节子序列作为第一个拼音音节子序列，进行步骤 S405，在后续确定每个语句的子分数时，根据该语句中每个候选词对应的分数，确定由对应候选词构成的语句的子分数，并按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句。

[0095] S411：根据第一个拼音音节子序列与第二拼音音节子序列组成的每个语句中，每个候选词的分数，确定每个语句的总分数，根据该总分数，选择总分数最大语句作为生成的语句。

[0096] 本申请中对用户输入的拼音流进行切分，切分为包括至少两个拼音音节子序列的拼音音节序列，其中每个拼音音节子序列中包括至少一个拼音音节。将拼音音节序列中的各拼音音节子序列与数据库中字典保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词。将每个拼音音节子序列对应的每个候选词组成对应的语句。根据每个语句确定每个语句中的每个候选词对应的分数时，还可以包括：针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及数据库保存的词典中每个词组对应的权重，确定该候选词对应的分数。当确定了每个语句中每个候选词对应的分数后，根据每个语句中每个候选词对应的分数，确定每个语句的总分数，将总分数最大的语句作为生成的语句。

[0097] 上述实施过程中，将每个候选词与该候选词之后的候选词进行组合，构成词组，从而确定候选词对应的分数，其具体过程包括：判断所述候选词是否为该语句的最后一个候选词，当该候选词为该语句的最后一个候选词时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重以及保存的第二权重系数，确定所述候选词对应的分数，当确定该候选词非该语句的最后一个候选词时，判断所述候选词与该候选词之后的候选词组

成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权 重，以及保存的第一权重系数确定所述候选词对应的权重，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重以及保存的第二权重系数，确定所述候选词对应的分数。

[0098] 同样，确定每个语句的总分数之前该方法进一步包括：根据每个语句中已确定分数的候选词，及该已确定分数的候选词对应的分数，确定每个语句的子分数；根据所述每个语句的子分数，按照子分数由大到小的顺序选择设定数量的语句作为准备确定总分数的语句。

[0099] 上述在确定每个语句的每个候选词对应的分数的过程中，根据每个候选词与该候选词之后的候选词组成的词组，以及数据库的词典中每个词组对应的权重的过程，与根据每个候选词与该候选词之前的候选词组成的词组，以及数据库的词典中每个词组对应的权重过程类似，相信本领域技术人员根据本申请实施例的描述，可以确定具体的分数确定过程，在这里就不一一赘述。

[0100] 图 5 为本申请实施例提供的一种语句生成装置，该装置包括以下结构：

[0101] 匹配模块 51，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词；

[0102] 分数确定模块 52，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之前的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数；

[0103] 语句生成模块 53，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0104] 所述装置还包括：

[0105] 存储模块 54，用于保存一元词典及二元词典，其中所述一元词典中保存词条，每个词条对应的拼音，以及每个词条对应的权重，所述二元词典中保存词组，以及每个词组的权重。

[0106] 所述分数确定模块 52 包括：

[0107] 判断单元 521，用于判断所述候选词是否为所述语句的第一个候选词；

[0108] 第一分数确定单元 522，用于确定所述候选词为所述语句的第一个候选词时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数；

[0109] 第二分数确定单元 523，用于确定所述候选词非所述语句中第一个候选词时，判断所述候选词与所述候选词之前的候选词组成的词组是否在二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数。

[0110] 所述语句生成模块 53 还用于，

[0111] 根据每个语句中已确定分数的候选词，及该已确定分数的候选词对应的分数，确定每个语句对应的子分数，按照子分数由大到小的顺序选择选择设定数量的语句作为准备

确定总分数的语句。

[0112] 所述语句生成模块 53 在确定每个语句的总分数时具体用于，

[0113] 根据所述每个语句中每个候选词的分数，将所述每个候选词的分数进行乘积或累加运算，将每个候选词的分数进行乘积或累加运算得到的分数，作为该语句的总分数。

[0114] 图 6 为本申请实施例提供的一种语句生成的装置结构示意图，该装置包括 ：

[0115] 匹配模块 61，用于将用户输入的拼音流切分后获取的拼音音节序列中的各拼音音节子序列，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词 ；

[0116] 分数确定模块 62，用于将每个拼音音节子序列的每个候选词组成对应的语句，针对每个语句的每个候选词，根据该候选词与该候选词之后的候选词组成的词组，及所述词典中每个词组对应的权重，确定该候选词对应的分数 ；

[0117] 语句生成模块 63，用于根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。

[0118] 所述分数确定模块 62 包括 ：

[0119] 判断单元 621，用于判断所述候选词是否为所述语句的最后一个候选词 ；

[0120] 第一分数确定单元 622，用于确定所述候选词为最后一个候选词时，在所述词典的一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数 ；

[0121] 第二分数确定单元 623，用于确定所述候选词非最后一个候选词时，判断所述候选词与所述候选词之后的候选词组成的词组是否在所述词典的二元词典中存在，当判断存在时，根据二元词典中与所述词组匹配的词组对应的权重，及保存的第一权重系数确定所述候选词对应的分数，当判断不存在时，在一元词典中查找与所述候选词匹配的词条对应的权重，根据所述权重及保存的第二权重系数，确定所述候选词对应的分数。

[0122] 所述装置中还包括存储模块，与图 5 所示的装置中的存储模块的功能相同，在这里就不一一赘述。

[0123] 本申请实施例提供了一种语句生成方法及装置，该方法包括 ：将拼音流切分后的拼音音节序列中的各拼音音节，与词典中保存的各词条的拼音进行匹配，将匹配成功的拼音对应的每个词条作为对应拼音音节的每个候选词，将每个候选词组成对应的语句，针对每个语句的每个候选词与该候选词之前的候选词组成的词组，及词典中每个词组对应的权重，确定该候选词对应的分数，根据所述每个语句中每个候选词的分数，确定所述每个语句的总分数，并根据确定的总分数，将总分数最大的语句作为生成的语句。由于只有经常出现的词组对应的权重才会比较高，即经常出现的词组一定是用户经常使用，或满足语言规则的词组，因此采用该方法可以使生成的语句更加的准确。

[0124] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申 请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

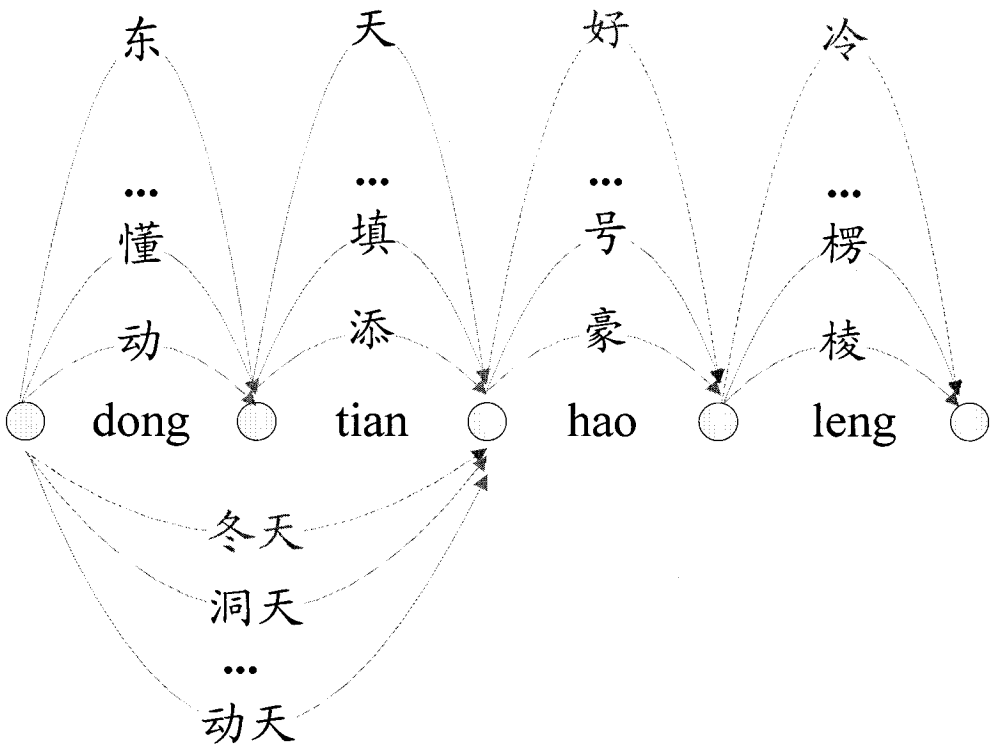


图 1

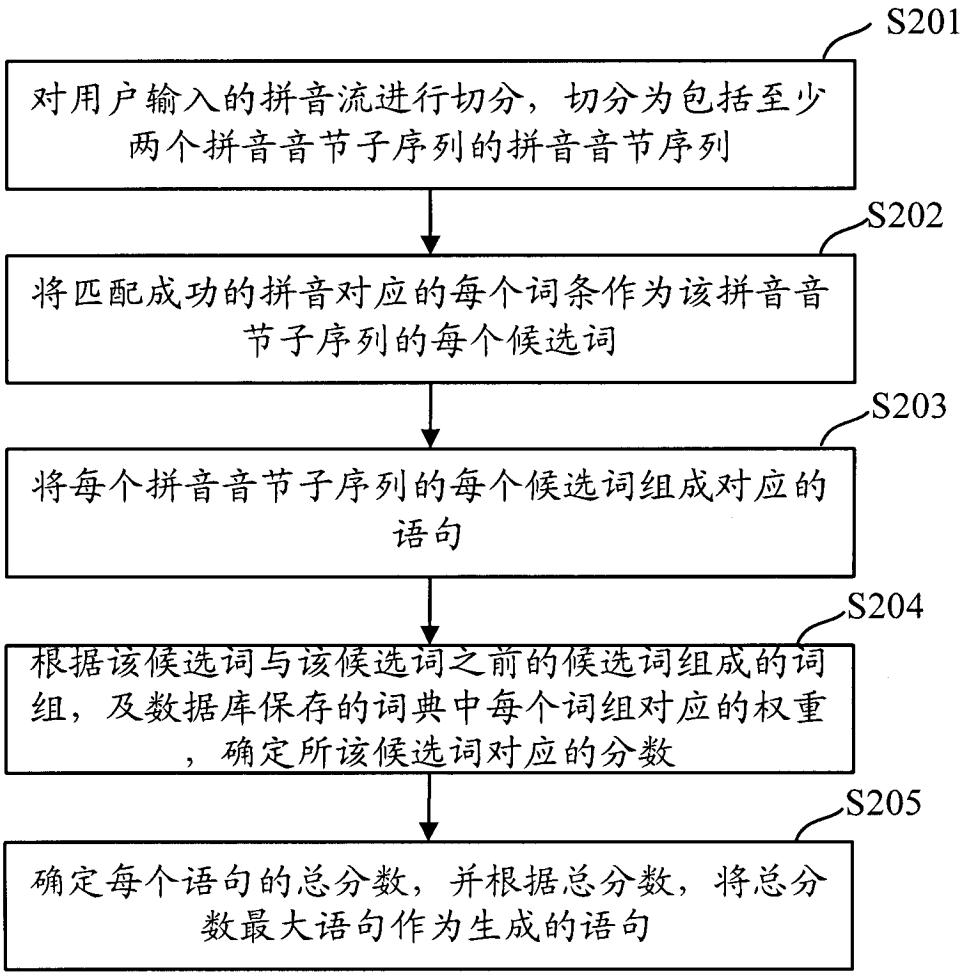

S201
对用户输入的拼音流进行切分，切分为包括至少两个拼音音节子序列的拼音音节序列
S202
将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词
S203
将每个拼音音节子序列的每个候选词组成对应的语句
S204
根据该候选词与该候选词之前的候选词组成的词组，及数据库保存的词典中每个词组对应的权重，确定所该候选词对应的分数
S205
确定每个语句的总分数，并根据总分数，将总分数最大语句作为生成的语句

图 2

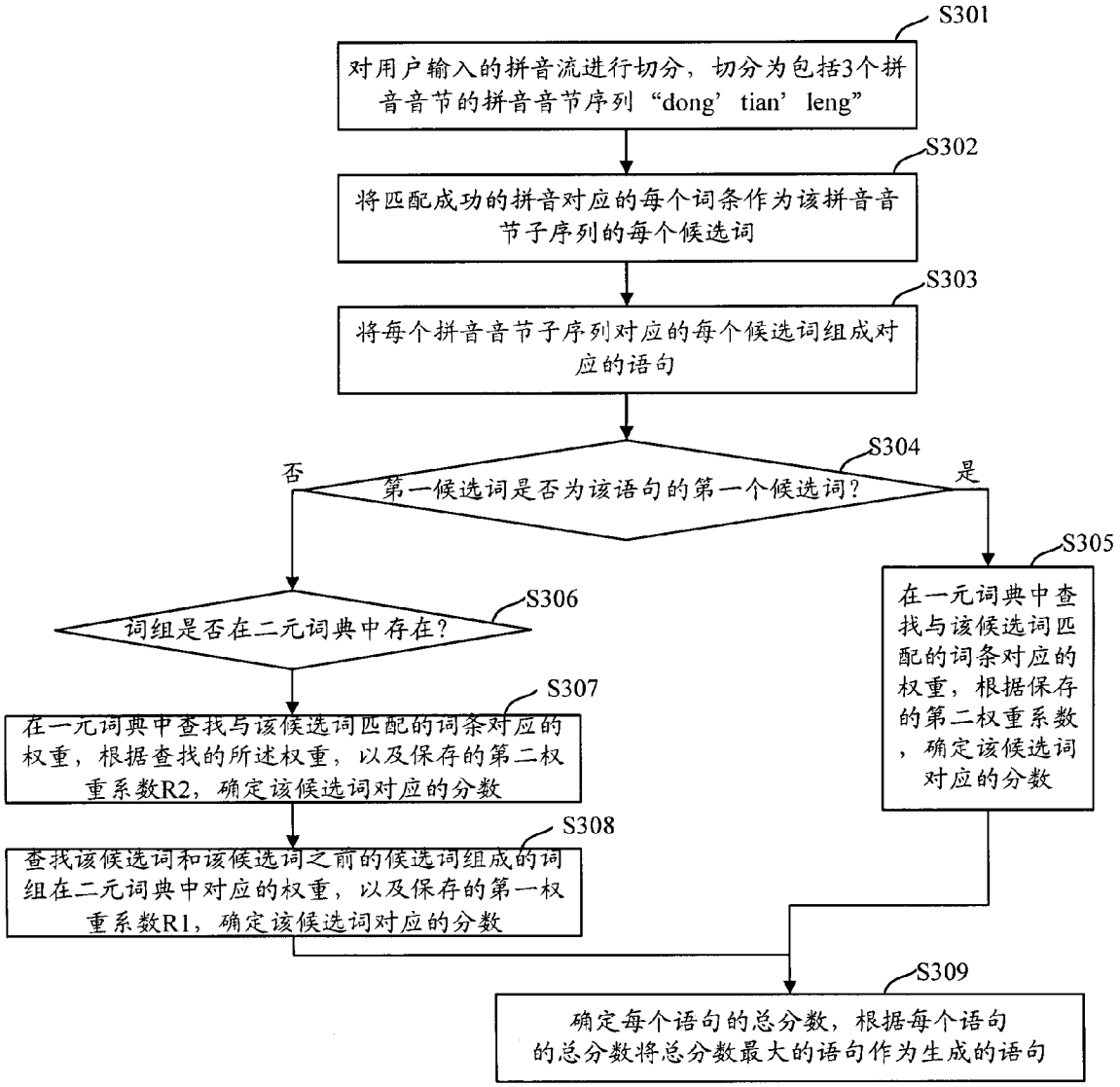

S301
对用户输入的拼音流进行切分，切分为包括3个拼音音节的拼音音节序列“dong’ tian’ leng”
S302
将匹配成功的拼音对应的每个词条作为该拼音音节子序列的每个候选词
S303
将每个拼音音节子序列对应的每个候选词组成对应的语句
S304
第一候选词是否为该语句的第一个候选词？
否
是
S305
在一元词典中查找与该候选词匹配的词条对应的权重，根据保存的第二权重系数，确定该候选词对应的分数
S306
词组是否在二元词典中存在？
S307
在一元词典中查找与该候选词匹配的词条对应的权重，根据查找的所述权重，以及保存的第二权重系数R2，确定该候选词对应的分数
S308
查找该候选词和该候选词之前的候选词组成的词组在二元词典中对应的权重，以及保存的第一权重系数R1，确定该候选词对应的分数
S309
确定每个语句的总分数，根据每个语句的总分数将总分数最大的语句作为生成的语句

图 3

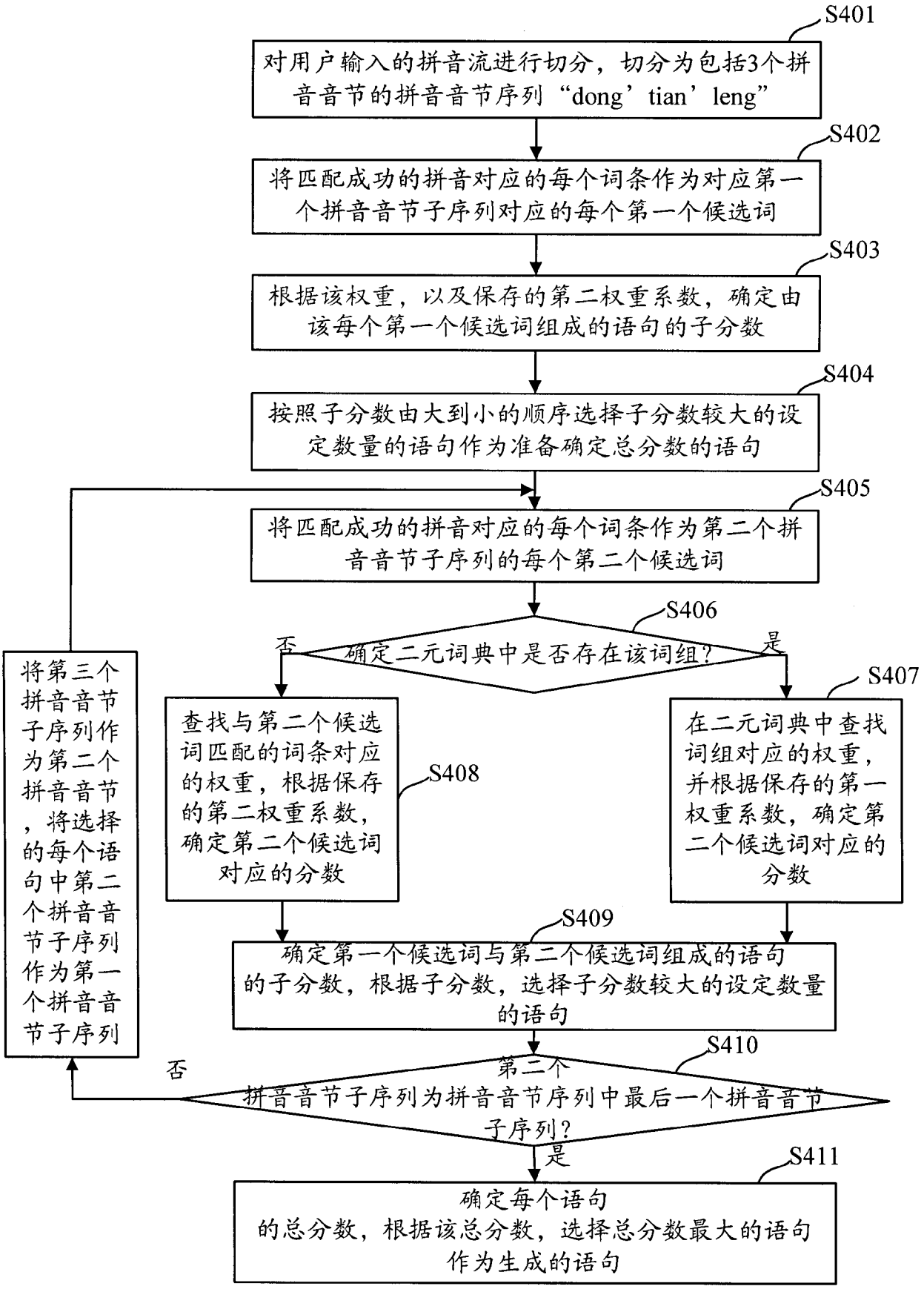

S401
对用户输入的拼音流进行切分，切分为包括3个拼音音节的拼音音节序列“dong’tian’leng”
S402
将匹配成功的拼音对应的每个词条作为对应第一个拼音音节子序列对应的每个第一个候选词
S403
根据该权重，以及保存的第二权重系数，确定由该每个第一个候选词组成的语句的子分数
S404
按照子分数由大到小的顺序选择子分数较大的设定数量的语句作为准备确定总分数的语句
S405
将匹配成功的拼音对应的每个词条作为第二个拼音音节子序列的每个第二个候选词
S406
确定二元词典中是否存在该词组？
否
是
S407
在二元词典中查找词组对应的权重，并根据保存的第一权重系数，确定第二个候选词对应的分数
S408
查找与第二个候选词匹配的词条对应的权重，根据保存的第二权重系数，确定第二个候选词对应的分数
S409
确定第一个候选词与第二个候选词组成的语句的子分数，根据子分数，选择子分数较大的设定数量的语句
S410
第二个拼音音节子序列为拼音音节序列中最后一个拼音音节子序列？
否
是
将第三个拼音音节子序列作为第二个拼音音节，将选择的每个语句中第二个拼音音节子序列作为第一个拼音音节子序列
S411
确定每个语句的总分数，根据该总分数，选择总分数最大的语句作为生成的语句

图 4

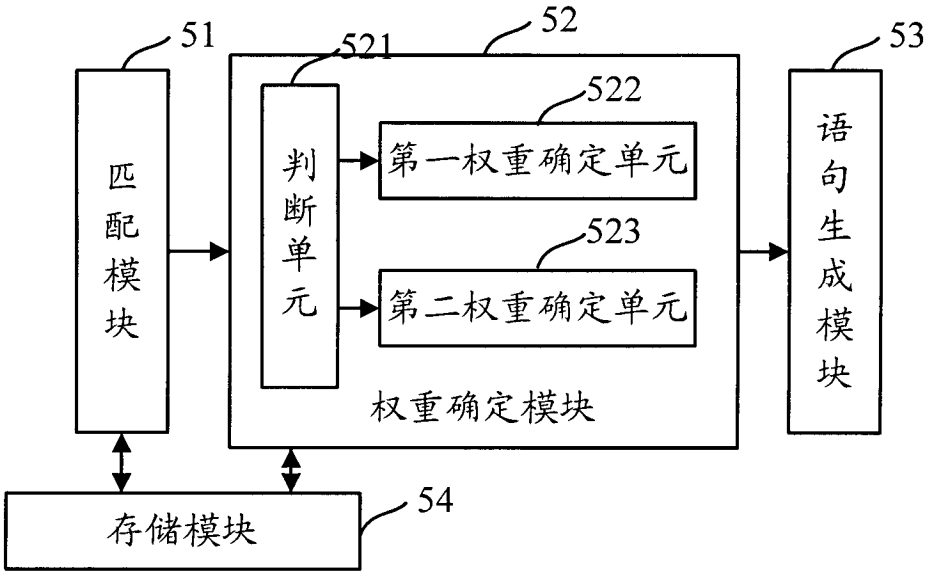

51
52
521
522
523
53
匹配模块
判断单元
第一权重确定单元
第二权重确定单元
权重确定模块
语句生成模块
存储模块
54

图 5

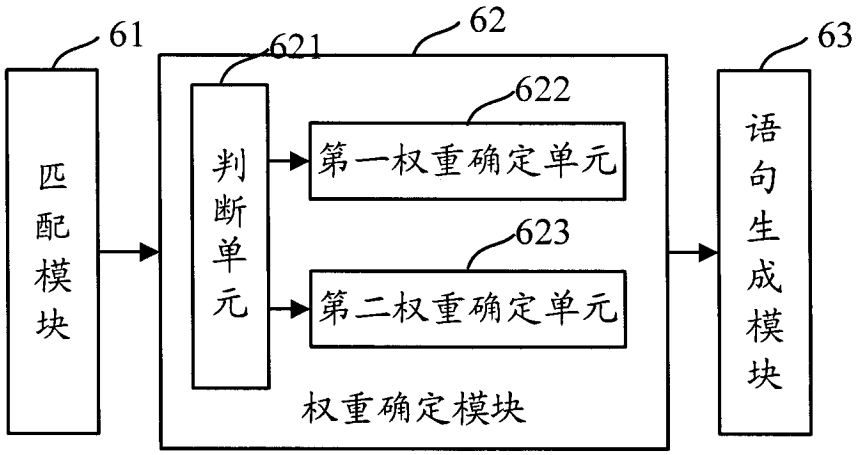

61
62
621
622
623
63
匹配模块
判断单元
第一权重确定单元
第二权重确定单元
权重确定模块
语句生成模块

图 6