

一种文本生成方法、装置、设备和存储介质

本申请要求在 2023 年 5 月 12 日提交中国专利局、申请号为 202310538686.9 的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及计算机技术，例如涉及一种文本生成方法、装置、设备和存储介质。

背景技术

随着计算机技术的快速发展，往往需要自动生成一些文本，从而满足不同的业务需求，比如生成的文本可以用于机器学习或者文本翻译等。

目前，通常是通过集束搜索方式（beam search）生成文本。集束搜索方式是用于搜索生成的文本序列的条件概率值最大的序列。

然而，在实现本申请过程中，发明人发现相关技术中至少存在如下问题：

相关技术中的集束搜索方式生成的文本相似度非常高，比如生成的两个文本长句中只有一个词不同或者文本语义完全相同，从而导致生成的文本多样性较低。

发明内容

本申请实施例提供了一种文本生成方法、装置、设备和存储介质，以提高生成文本的多样性。

第一方面，本申请实施例提供了一种文本生成方法，包括：

获取上一时刻生成的预设数量的第一文本；

确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

对所述预设数量的所述第一文本对应的多个第一候选文本进行遍历组合，

确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；

基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的所述预设数量的第二文本。

第二方面，本申请实施例还提供了一种文本生成装置，包括：

第一文本获取模块，设置为获取上一时刻生成的预设数量的第一文本；

第一候选文本生成模块，设置为确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

候选文本集合确定模块，设置为对所述预设数量的所述第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

目标分数确定模块，设置为基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

集合分数确定模块，设置为基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；

第二文本确定模块，设置为基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的所述预设数量的第二文本。

第三方面，本申请实施例还提供了一种电子设备，所述电子设备包括：

至少一个处理器；

存储器，设置为存储至少一个程序；

当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现如本申请任意实施例所提供的文本生成方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本申请任意实施例所提供的文本生成方法。

附图说明

图 1 是本申请一个实施例所提供的一种文本生成方法的流程图；

图 2 是本申请一个实施例所涉及的一种文本生成的示例；

图 3 是本申请一个实施例提供的另一种文本生成方法的流程图；

图 4 是本申请一个实施例所涉及的一种第二文本对应的多个第二候选词的示例；

图 5 是本申请一个实施例所涉及的一种第二文本对应的目标第二候选词的示例；

图 6 是本申请一个实施例提供的一种文本生成装置的结构示意图；

图 7 是本申请一个实施例提供的一种电子设备的结构示意图。

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本申请，而非对本申请的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本申请相关的部分而非全部结构。

图 1 为本申请一个实施例所提供的一种文本生成方法的流程图，本实施例可适用于生成具有多样性的文本情况。该方法可以由文本生成装置来执行，该

装置可以由软件和/或硬件的方式来实现，集成于电子设备中。如图 1 所示，该方法包括以下步骤：

S110、获取上一时刻生成的预设数量的第一文本。

其中，预设数量可以是预先设置的，每个时刻生成的文本数量。例如，预设数量也可以用于表征集束宽度。上一时刻可以是指上次生成文本的时刻。当前时刻可以是指当次生成文本的时刻。上一时刻可以是指与当前时刻相邻的前一时刻。例如，当前时刻为 t 时刻，上一时刻为 $t-1$ 时刻。第一文本可以是指上一时刻所生成的文本。本实施例中的文本可以利用至少一个词组成的词序列进行表征。完整的文本需要重复地在上一时刻生成的第一文本的基础上继续生成获得的。例如，在首次生成时，获得的上一时刻对应的第一文本可以为空文本，从而可以从头生成新文本；或者也可以为预设的已有文本，从而可以在已有文本的基础上继续生成文本。随着时间的推移，生成的文本长度越来越长。

S120、确定每个第一文本对应的多个第一候选词，并将每个第一候选词作为第一文本的后缀，生成多个第一候选文本。

其中，第一候选词是指可以作为第一文本后缀的词汇。第一候选文本可以作为当前时刻生成的文本。

示例性地，可以基于词典确定出每个第一文本对应的所有第一候选词，并将每个第一候选词作为相应的第一文本的后缀进行拼接组合，获得每个第一文本对应的所有第一候选文本。例如，图 2 给出了一种文本生成的示例。如图 2 所示，存在两个第一文本分别为：“这款”和“无线”，“这款”对应的第一候选词为：“耳机”、“耳麦”等。“无线”对应的第一候选词为：“耳机”、“蓝牙”等。将每个第一候选词作为相应的第一文本的后缀，获得的第一候选文本为：“这款耳机”、“这款耳麦”、“无线耳机”、“无线蓝牙”等。

S130、对预设数量的第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括预设数量的第一候选文本。

示例性地，可以对所有第一文本对应的所有第一候选文本进行遍历组合的

穷举，获得包含有预设数量的第一候选文本的所有候选文本集合。例如，若预设数量为 2，则对所有第一文本对应的所有第一候选文本进行两两组合，获得每种组合情况对应的候选文本集合，以便更加全面地生成文本。也可以从每个第一文本对应的所有第一候选文本中遍历选取一个第一候选文本进行组合，获得每种选取情况对应的候选文本集合，从而使得候选文本集合中的每个第一候选文本具有不同的文本前缀，保证文本生成的多样性。需要说明的是，两个候选文本集合中可以具有至少一个不同的第一候选文本。例如，图 2 仅给出了两个候选文本集合，分别为：{“这款耳机”，“无线耳机”}和{“这款耳机”，“无线蓝牙”}。

S140、确定每个候选文本集合中重复出现的重复词，并对重复词对应的词分数进行惩罚处理，确定重复词对应的目标分数，其中，重复词对应的词分数用于表征重复词作为文本后缀词的条件概率值。

其中，重复词可以是指候选文本集合中重复出现的词汇。重复词对应的词分数可以用于表征重复词作为文本后缀词的条件概率值。例如，可以直接将条件概率值作为词分数，也可以对条件概率值进行取对数等运算获得的得分作为词分数。词分数可以用于反映文本的可读性和流畅性。例如，词分数越高，利用该词生成的文本的可读性和流畅性越高。重复词对应的目标分数可以是指对词分数进行惩罚处理后所获得的分数。重复词对应的目标分数小于相应的词分数。

示例性地，针对每个候选文本集合而言，可以将该候选文本集合中的每个第一候选文本中的所有词进行比较，确定在该候选文本集合中重复出现的所有重复词，并对每个重复词对应的词分数进行惩罚处理，确定每个重复词对应的目标分数。例如，可以将小于重复词的词分数的任一分数作为该重复词对应的目标分数。例如，可以将每个重复词对应的词分数减去预设分数获得的分数作为该重复词对应的目标分数。

需要说明的是，每个词对应的词分数可以利用对条件概率值进行取对数等

运算获得,这种方式获得的词分数为负数。负数越大,则表明条件概率值越高,可读性和流畅性也越高。例如,图2中的“这款耳机”中的“耳机”对应的词分数为:-1.8,其表示的是“耳机”这个词作为“这款”后缀词的得分。“无线耳机”中的“耳机”对应的词分数为:-1.9,其表示的是“耳机”这个词作为“无线”后缀词的得分。

示例性地,S140中的“对重复词对应的词分数进行惩罚处理,确定重复词对应的目标分数”,可以包括:获取当前重复词在每个出现位置对应的词分数;将至少一个预设出现位置对应的词分数与预设惩罚系数进行相乘,获得当前重复词在预设出现位置对应的目标分数;将除预设出现位置之外的剩余出现位置对应的词分数作为当前重复词在剩余出现位置对应的目标分数。

其中,可以将每个重复词作为当前重复词进行词分数的惩罚处理。出现位置可以是指重复词在候选文本集合中的出现位置。出现位置不同,对应的词分数可能不同。例如,对图2中的{“这款耳机”,“无线耳机”}这个候选文本集合而言,重复词为“耳机”,该重复词出现两次,第一次出现位置为“这款耳机”中,相应的词分数为:-1.8;第二次出现位置为:“无线耳机”中,相应的词分数为:-1.9。预设出现位置可以是预先设置的,需要进行词分数的惩罚处理的出现位置。预设出现位置的数量可以为至少一个,其基于业务需求进行设置。例如,预设出现位置为第一次出现位置或者最后一次出现位置等。剩余出现位置是指所有出现位置中除了预设出现位置之外的其他出现位置。预设惩罚系数可以是预先设置的大于1的系数。例如,预设惩罚系数为10。

示例性地,如图2所示,在预设出现位置为最后一次出现位置时,确定重复词“耳机”在第一次出现位置对应的目标分数为:原有的词分数,即-1.8,在第二次出现位置,即最后一次出现位置对应的目标分数为:原有的词分数-1.9与预设惩罚系数10的乘积,即-19。通过仅对预设出现位置处的词分数进行惩罚处理,可以在保证多样性效果的同时提高惩罚处理效率。

S150、基于每个候选文本集合中的重复词对应的目标分数和非重复词对应

的词分数，确定每个候选文本集合对应的集合分数。

示例性地，可以对每个候选文本集合中的所有重复词对应的目标分数和所有非重复词对应的词分数进行整合，获得每个候选文本集合对应的集合分数。在本实施例中，可以仅对候选文本集合中的重复词进行词分数的惩罚处理，无需对非重复词进行词分数的惩罚处理，从而可以倾向于生成包含不重复词的文本，提高了生成文本的多样性。

示例性地，S150 可以包括：将当前候选文本集合中的所有重复词对应的目标分数和所有非重复词对应的词分数进行相加，获得的相加结果确定为当前候选文本集合对应的集合分数。

示例性地，可以将每个候选文本集合作为当前候选文本集合进行集合分数的计算。例如，图 2 中的{“这款耳机”，“无线耳机”}这个候选文本集合对应的集合分数为： $-3.1-3.2-1.8-19=-27.1$ 。

若当前候选文本集合中的词均为非重复词，不存在重复词，则可以直接将当前候选文本集合中的所有非重复词对应的词分数进行相加，获得的相加结果确定为当前候选文本集合对应的集合分数。图 2 中的{“这款耳机”，“无线蓝牙”}这个候选文本集合对应的集合分数为： $-3.1-3.2-1.8-2.2=-10.3$ 。

S160、基于集合分数，从多个候选文本集合中确定目标文本集合，并基于目标文本集合，确定当前时刻生成的预设数量的第二文本。

其中，目标文本集合可以是指当前时刻生成的最终文本集合。第二文本可以是指当前时刻生成的最终文本。

示例性地，基于每个候选文本集合对应的集合分数，选取最优的目标文本集合，并基于目标文本集合获得当前时刻生成的预设数量的第二文本。通过对重复词进行词分数惩罚处理，并基于集合分数进行选取，从而可以倾向于选取包含不重复词的第二文本，大大提高了生成的第二文本的多样性。

示例性地，S160 可以包括：将集合分数最高的候选文本集合确定为目标文本集合，并将目标文本集合中的目标文本确定为当前时刻生成的预设数量的第

二文本。可以基于集合分数对所有候选文本集合进行排序，将集合分数最高的候选文本集合作为最终的目标文本集合，并将目标文本集合中的每个目标文本作为当前时刻生成的第二文本，从而生成多样化的第二文本。例如，确定图 2 中的{“这款耳机”，“无线蓝牙”}这个候选文本集合作为目标文本集合，该集合中的两个目标文本即为当前时刻生成的两个第二文本。

需要说明的是，可以通过循环执行步骤 S110-S160 的方式，基于当前时刻生成的第二文本继续生成下一时刻的文本，直到生成次数等于预设次数或者生成的文本长度等于预设的完整文本长度为止。

本实施例的技术方案，通过在当前时间步长内，基于上一时刻生成的预设数量的第一文本生成多个第一候选文本，并对预设数量的第一文本对应的多个第一候选文本进行遍历组合，获得多个候选文本集合，对每个候选文本集合中重复出现的重复词进行词分数的惩罚处理，并基于惩罚后的目标分数确定每个候选文本集合对应的集合分数，基于集合分数，从多个候选文本集合中确定目标文本集合，并基于目标文本集合确定当前时刻生成的预设数量的第二文本，从而通过对每个候选文本集合中重复出现的重复词进行词分数的惩罚处理，可以倾向于生成包含不重复词的文本，提高了生成文本的多样性。

在上述技术方案的基础上，在 S160 之后，还可以包括如下步骤：

S170、若第二文本的长度大于或等于预设长度，则确定每个第二文本对应的多个第二候选词。

其中，预设长度可以是预先设置的，停止利用上述步骤 S110-S160 方式循环生成文本的最小长度。第二候选词是指可以作为第二文本后缀的词汇。

需要注意的是，在生成的文本达到一定长度时，上述提高多样性的生成方式会一定程度上影响生成文本的可读性，从而在超过一定长度后需要利用保证可读性的生成方式继续生成文本，以便在提高多样性的前提下保证文本可读性。

示例性地，在当前时刻生成第二文本后，可以检测当前时刻生成的第二文本的长度是否大于或等于预设长度。在当前时刻生成的第二文本的长度大于或

等于预设长度的情况下，利用步骤 S170-S190 的方式生成下一时刻的文本。在当前时刻生成的第二文本的长度小于预设长度的情况下，继续利用上述步骤 S110-S160 方式生成下一时刻的文本。在第二文本的长度大于或等于预设长度时，可以基于词典确定出每个第二文本对应的多个第二候选词。

S180、将每个第二候选词作为第二文本的后缀，生成多个第二候选文本，并基于每个第二候选文本中的每个词对应的词分数，确定每个第二候选文本对应的文本分数。

示例性地，可以将每个第二候选词作为相应的第二文本的后缀进行拼接组合，获得每个第二文本对应的所有第二候选文本。针对每个第二候选文本而言，可以将该第二候选文本中的每个词对应的词分数进行相加，获得的相加结果作为该第二候选文本对应的文本分数。文本分数可以用于表征第二候选文本的可读性，比如文本分数越高，第二候选文本的可读性越高。

S190、基于文本分数，从所述预设数量的第二文本对应的多个第二候选文本中确定下一时刻生成的预设数量的第三文本。

示例性地，可以基于文本分数对所有第二候选文本进行降序排列，获得文本分数依次降低的第二候选文本序列，并将第二候选文本序列中的前预设数量的第二候选文本确定为下一时刻生成的第三文本。通过选取文本分数越高的第二候选文本作为第三文本，从而可以在提高多样性的前提下保证文本可读性。

需要说明的是，在下一时刻之后，可以通过循环执行步骤 S170-S190 的方式，基于下一时刻生成的第三文本继续生成后续时刻的文本，直到生成次数等于预设次数或者生成的文本长度等于预设的完整文本长度为止。

图 3 为本申请一个实施例提供的另一种文本生成方法的流程图，本实施例在上述各实施例的基础上，对步骤“将每个第二候选词作为第二文本的后缀，生成多个第二候选文本，并基于每个第二候选文本中的每个词对应的词分数，确定每个第二候选文本对应的文本分数”进行了细化，并在此基础上，还对步

骤“基于文本分数，从多个第二候选文本中确定下一时刻生成的预设数量的第三文本”也进行了细化。其中与上述各实施例相同或相应的术语的解释在此不再赘述。

参见图 3，本实施例提供的另一种文本生成方法包括以下步骤：

S310、获取上一时刻生成的预设数量的第一文本。

S320、确定每个第一文本对应的多个第一候选词，并将每个第一候选词作为第一文本的后缀，生成多个第一候选文本。

S330、对预设数量的第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括预设数量的第一候选文本。

S340、确定每个候选文本集合中重复出现的重复词，并对重复词对应的词分数进行惩罚处理，确定重复词对应的目标分数。

S350、基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数。

S360、基于集合分数，从多个候选文本集合中确定目标文本集合，并基于目标文本集合，确定当前时刻生成的预设数量的第二文本。

S370、若第二文本的长度大于或等于预设长度，则确定每个第二文本对应的多个第二候选词。

示例性地，图 4 给出了一种第二文本对应的多个第二候选词的示例。图 4 中的预设数量为 2，即存在两个第二文本，即图 4 中第二文本 1 包含两个第二文本，第二文本 2 也包含两个第二文本，每个第二文本都是独立地确定出相应的第二候选词，并独立地执行后续步骤 S380-S391。

S380、将当前第二文本对应的每个当前第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本。

需要注意的是，可以将每个第二文本作为当前第二文本，并行地通过执行步骤 S380-S391 的方式确定出每个第二文本对应的第三候选文本。

示例性地，可以将当前第二文本对应的每个当前第二候选词作为当前第二

文本的后缀进行拼接组合，获得当前第二文本对应的所有的当前第二候选文本。

S390、基于每个当前第二候选文本中的每个词对应的词分数，确定每个当前第二候选文本对应的文本分数。

示例性地，针对每个当前第二候选文本而言，可以将该当前第二候选文本中的每个词对应的词分数进行相加，获得的相加结果作为该当前第二候选文本对应的文本分数。文本分数可以用于表征当前第二候选文本的可读性，比如文本分数越高，当前第二候选文本的可读性越高。

S391、基于文本分数，从多个当前第二候选文本中确定出当前第二文本对应的预设数量的当前第三候选文本。

示例性地，可以基于文本分数对所有当前第二候选文本进行降序排列，获得文本分数依次降低的当前第二候选文本序列，并将当前第二候选文本序列中的前预设数量的当前第二候选文本确定为当前第二文本对应的当前第三候选文本，从而可以彼此独立地针对每个第二文本生成相应的预设数量的第三候选文本。第二文本之间互不影响，从而提高不同第二文本之间的第三候选文本的多样性，进而在文本长度超过预设长度后，也可以在保证文本可读性的前提下进一步保证文本的多样性，避免出现生成的文本多样性下降的情况。

S392、从预设数量的第二文本对应的第三候选文本中确定下一时刻生成的预设数量的第三文本。

示例性地，可以基于每个第二文本对应的每个第三候选文本对应的文本分数，从所有的第三候选文本中选取最优的预设数量的第三文本。

示例性地，S392 可以包括：基于每个第二文本对应的每个第三候选文本对应的文本分数，对预设数量的第二文本对应的所有第三候选文本进行降序排列，获得第三候选文本序列；将第三候选文本序列中的前预设数量的第三候选文本确定为下一时刻生成的第三文本。通过将文本分数最高的前预设数量的第三候选文本作为下一时刻生成的第三文本，可以在保证文本可读性的前提下进一步保证文本的多样性。

本实施例的技术方案，通过彼此独立地针对每个第二文本生成相应的预设数量的第三候选文本，并从所有的第三候选文本中确定下一时刻生成的预设数量的第三文本，从而在文本长度超过预设长度后，也可以在保证文本可读性的前提下进一步保证文本的多样性。

在上述技术方案的基础上，S380 可以包括：获取预设数量的第二文本中的已用词；将当前第二文本对应的每个当前第二候选词与已用词进行比较，确定不同于已用词的目标第二候选词；将每个目标第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本。

其中，已用词可以是指第二文本中已使用的词汇。目标第二候选词可以是指在所有当前第二候选词中除去已用词后剩余的当前第二候选词。

图 5 给出了一种第二文本对应的目标第二候选词的示例。如图 5 所示，第二文本 1 和第二文本 2 中包含的所有词均为已用词。对于第二文本 1 而言，第二文本 1 对应的每个第二候选词均不是已用词，从而可以直接将第二文本 1 对应的每个第二候选词作为相应的目标第二候选词。对于第二文本 2 而言，第二候选词“轻量化”为已用词，此时可以将该“轻量化”进行删除，将剩余的其它第二候选词作为第二文本 2 对应的目标第二候选词。在确定出目标第二候选词之后，可以将当前第二文本对应的每个目标第二候选词作为当前第二文本的后缀进行拼接组合，获得当前第二文本对应的所有的当前第二候选文本。通过将第二文本对应的第二候选词中删除已经使用的已用词，从而实现去重处理，避免了同一词的重复出现，提高了不同第二文本生成的当前第二候选文本的多样性，进而也提高了第三文本生成的多样性，使得在文本长度超过预设长度后，也可以在保证文本可读性的前提下进一步提高文本的多样性。

以下是本申请实施例提供的文本生成装置的实施例，该装置与上述各实施例的文本生成方法属于同一个发明构思，在文本生成装置的实施例中未详细描述的细节内容，可以参考上述文本生成方法的实施例。

图 6 为本申请实施例提供的一种文本生成装置的结构示意图，本实施例可适用于生成具有多样性的文本情况。如图 6 所示，该装置包括：第一文本获取模块 610、第一候选文本生成模块 620、候选文本集合确定模块 630、目标分数确定模块 640、集合分数确定模块 650 和第二文本确定模块 660。

其中，第一文本获取模块 610，设置为获取上一时刻生成的预设数量的第一文本；第一候选文本生成模块 620，设置为确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；候选文本集合确定模块 630，设置为对所述预设数量的第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；目标分数确定模块 640，设置为基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；集合分数确定模块 650，设置为基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；第二文本确定模块 660，设置为基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的预设数量的第二文本。

本实施例的技术方案，通过在当前时间步长内，基于上一时刻生成的预设数量的第一文本生成多个第一候选文本，并对所述预设数量的第一文本对应的多个第一候选文本进行遍历组合，获得多个候选文本集合，对每个候选文本集合中重复出现的重复词进行词分数的惩罚处理，并基于惩罚后的目标分数确定每个候选文本集合对应的集合分数，基于集合分数，从多个候选文本集合中确定目标文本集合，并基于目标文本集合确定当前时刻生成的预设数量的第二文本，从而通过对每个候选文本集合中重复出现的重复词进行词分数的惩罚处理，可以倾向于生成包含不重复词的文本，提高了生成文本的多样性。

可选地，目标分数确定模块 640，设置为：

获取当前重复词在每个出现位置对应的词分数；将至少一个预设出现位置对应的词分数与预设惩罚系数进行相乘，获得当前重复词在所述预设出现位置对应的目标分数；将除预设出现位置之外的剩余出现位置对应的词分数作为当前重复词在所述剩余出现位置对应的目标分数。

可选地，集合分数确定模块 650，设置为：

将当前候选文本集合中的各个重复词对应的目标分数和各个非重复词对应的词分数进行相加，获得的相加结果确定为当前候选文本集合对应的集合分数。

可选地，第二文本确定模块 660，设置为：

将所述集合分数最高的候选文本集合确定为目标文本集合，并将所述目标文本集合中的各个目标文本确定为当前时刻生成的预设数量的第二文本。

可选地，该装置还包括：

第二候选词确定模块，设置为在确定当前时刻生成的预设数量的第二文本之后，若所述第二文本的长度大于或等于预设长度，则确定每个所述第二文本对应的多个第二候选词；

文本分数确定模块，设置为将每个所述第二候选词作为第二文本的后缀，生成多个第二候选文本，并基于每个第二候选文本中的每个词对应的词分数，确定每个第二候选文本对应的文本分数；

第三文本确定模块，设置为基于所述文本分数，从多个第二候选文本中确定下一时刻生成的预设数量的第三文本。

可选地，文本分数确定模块，包括：

当前第二候选文本确定单元，设置为将当前第二文本对应的每个当前第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本；

文本分数确定单元，设置为基于每个当前第二候选文本中的每个词对应的词分数，确定每个当前第二候选文本对应的文本分数；

第三文本确定模块，包括：

当前第三候选文本确定单元，设置为基于所述文本分数，从多个当前第二候选文本中确定出当前第二文本对应的预设数量的当前第三候选文本；

第三文本确定单元，设置为每个第二文本对应的第三候选文本中确定下一时刻生成的预设数量的第三文本。

可选地，当前第二候选文本确定单元，设置为：

获取每个第二文本中的已用词；将当前第二文本对应的每个当前第二候选词与所述已用词进行比较，确定不同于所述已用词的目标第二候选词；将每个所述目标第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本。

可选地，第三文本确定单元，设置为：

基于每个第二文本对应的每个第三候选文本对应的文本分数，对所述第三候选文本进行降序排列，获得第三候选文本序列；将所述第三候选文本序列中的前预设数量的第三候选文本确定为下一时刻生成的第三文本。

本申请实施例所提供的文本生成装置可执行本申请任意实施例所提供的文本生成方法，具备执行文本生成方法相应的功能模块。

值得注意的是，上述文本生成装置的实施例中，所包括的各个单元和模块只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能单元的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。

图 7 为本申请实施例提供的一种电子设备的结构示意图。图 7 示出了适于用来实现本申请实施方式的示例性电子设备 12 的框图。图 7 显示的电子设备 12 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 7 所示，电子设备 12 以通用计算设备的形式表现。电子设备 12 的组件可以包括但不限于：至少一个处理器或者处理单元 16，系统存储器 28，连接不同系统组件（包括系统存储器 28 和处理单元 16）的总线 18。

总线 18 表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构（Industry Standard Architecture, ISA）总线，微通道体系结构（Micro Channel Architecture, MCA）总线，增强型 ISA 总线、视频电子标准协会（Video Electronics Standards Association, VESA）局域总线以及外围组件互连（Peripheral Component Interconnect, PCI）总线。

电子设备 12 典型地包括多种计算机系统可读介质。这些介质可以是任何能够被电子设备 12 访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

系统存储器 28 可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器（Random Access Memory, RAM）30 和/或高速缓存存储器 32。电子设备 12 可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统 34 可以用于读写不可移动的、非易失性磁介质（图 7 未显示，通常称为“硬盘驱动器”）。尽管图 7 中未示出，可以提供用于对可移动非易失性磁盘（例如“软盘”）读写的磁盘驱动器，以及对可移动非易失性光盘（例如紧凑磁盘只读存储器（Compact Disc Read-Only Memory, CD-ROM），数字视盘（Digital Video Disc Read-Only Memory, DVD-ROM）或其它光介质）读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线 18 相连。系统存储器 28 可以包括至少一个程序产品，该程序产品具有一组（例如至少一个）程序模块，这些程序模块被配置以执行本申请各实施例的功能。

具有一组（至少一个）程序模块 42 的程序/实用工具 40，可以存储在例如系统存储器 28 中，这样的程序模块 42 包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块 42 通常执行本申请所描述的实施例中的功能

和/或方法。

电子设备 12 也可以与一个或多个外部设备 14（例如键盘、指向设备、显示器 24 等）通信，还可与一个或者多个使得用户能与该电子设备 12 交互的设备通信，和/或与使得该电子设备 12 能与一个或多个其它计算设备进行通信的任何设备（例如网卡，调制解调器等等）通信。这种通信可以通过输入/输出（Input/Output, I/O）接口 22 进行。并且，电子设备 12 还可以通过网络适配器 20 与一个或者多个网络（例如局域网（Local Area Network, LAN），广域网（Wide Area Network, WAN）和/或公共网络，例如因特网）通信。如图所示，网络适配器 20 通过总线 18 与电子设备 12 的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备 12 使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、磁盘阵列（Redundant Arrays of Independent Disks, RAID）系统、磁带驱动器以及数据备份存储系统等。

处理单元 16 通过运行存储在系统存储器 28 中的程序，从而执行各种功能应用以及数据处理，例如实现本发实施例所提供的一种文本生成方法步骤，该方法包括：

获取上一时刻生成的预设数量的第一文本；

确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

对所述预设数量的第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分

数，确定每个候选文本集合对应的集合分数；

基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的预设数量的第二文本。

当然，本领域技术人员可以理解，处理器还可以实现本申请任意实施例所提供的文本生成方法的技术方案。

本实施例提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本申请任意实施例所提供的文本生成方法步骤，该方法包括：

获取上一时刻生成的预设数量的第一文本；

确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

对所述预设数量的第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；

基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的预设数量的第二文本。

本申请实施例的计算机存储介质，可以采用一个或多个计算机可读的介质的任意组合。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存储介质。计算机可读存储介质例如可以是但不限于：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。计算机可读存储介质的更具体的例子（非穷举的列表）包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM（Erasable Programmable Read-Only Memory）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。在本文件中，计算机可读存储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。

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存储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、射频（Radio Frequency, RF）等等，或者上述的任意合适的组合。

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）。

本领域普通技术人员应该明白，上述的本申请的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个计算装置上，或者分布在多个计算装置所组成的网络上，可选地，他们可以用计算机装置可执行的程序代码来实现，从而可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本申请不限制于任何特定的硬件和软件的结合。

注意，上述仅为本申请的可选实施例及所运用技术原理。本领域技术人员会理解，本申请不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本申请的保护范围。因此，虽然通过以上实施例对本申请进行了较为详细的说明，但是本申请不仅仅限于以上实施例，在不脱离本申请构思的情况下，还可以包括更多其他等效实施例，而本申请的范围由所附的权利要求范围决定。

权 利 要 求 书

1、一种文本生成方法，包括：

获取上一时刻生成的预设数量的第一文本；

确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

对所述预设数量的所述第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；

基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的所述预设数量的第二文本。

2、根据权利要求1所述的方法，其中，所述对所述重复词对应的词分数进行惩罚处理，确定所述重复词对应的目标分数，包括：

获取当前重复词在每个出现位置对应的词分数；

将至少一个预设出现位置对应的词分数与预设惩罚系数进行相乘，获得当前重复词在所述至少一个预设出现位置对应的目标分数；

将除所述至少一个预设出现位置之外的剩余出现位置对应的词分数作为当前重复词在所述剩余出现位置对应的目标分数。

3、根据权利要求1所述的方法，其中，所述基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数，包括：

将当前候选文本集合中的所有重复词对应的目标分数和所有非重复词对应的词分数进行相加，获得的相加结果确定为当前候选文本集合对应的集合分数。

4、根据权利要求1所述的方法，其中，所述基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的所述预设数量的第二文本，包括：

将所述集合分数最高的候选文本集合确定为目标文本集合，并将所述目标文本集合中的目标文本确定为当前时刻生成的所述预设数量的所述第二文本。

5、根据权利要求1-4任一项所述的方法，在确定当前时刻生成的所述预设数量的第二文本之后，还包括：

若所述第二文本的长度大于或等于预设长度，则确定每个所述第二文本对应的多个第二候选词；

将每个所述第二候选词作为第二文本的后缀，生成多个第二候选文本，并基于每个第二候选文本中的每个词对应的词分数，确定每个第二候选文本对应的文本分数；

基于所述文本分数，从所述预设数量的所述第二文本对应的多个第二候选文本中确定下一时刻生成的所述预设数量的第三文本。

6、根据权利要求5所述的方法，其中，所述将每个所述第二候选词作为第二文本的后缀，生成多个第二候选文本，并基于每个第二候选文本中的每个词对应的词分数，确定每个第二候选文本对应的文本分数，包括：

将当前第二文本对应的每个当前第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本；

基于每个当前第二候选文本中的每个词对应的词分数，确定每个当前第二候选文本对应的文本分数；

所述基于所述文本分数，从所述预设数量的所述第二文本对应的多个第二候选文本中确定下一时刻生成的所述预设数量的第三文本，包括：

基于所述文本分数，从多个当前第二候选文本中确定出当前第二文本对应的所述预设数量的当前第三候选文本；

从所述预设数量的所述第二文本对应的第三候选文本中确定下一时刻生成

的所述预设数量的所述第三文本。

7、根据权利要求6所述的方法，其中，所述将当前第二文本对应的每个当前第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本，包括：

获取所述预设数量的所述第二文本中的已用词；

将当前第二文本对应的每个当前第二候选词与所述已用词进行比较，确定不同于所述已用词的目标第二候选词；

将每个所述目标第二候选词作为当前第二文本的后缀，生成当前第二文本对应的当前第二候选文本。

8、根据权利要求6所述的方法，其中，所述从所述预设数量的所述第二文本对应的第三候选文本中确定下一时刻生成的所述预设数量的所述第三文本，包括：

基于每个第二文本对应的每个第三候选文本对应的文本分数，对所述预设数量的所述第二文本对应的第三候选文本进行降序排列，获得第三候选文本序列；

将所述第三候选文本序列中的前预设数量的第三候选文本确定为下一时刻生成的第三文本。

9、一种文本生成装置，包括：

第一文本获取模块，设置为获取上一时刻生成的预设数量的第一文本；

第一候选文本生成模块，设置为确定每个所述第一文本对应的多个第一候选词，并将每个所述第一候选词作为第一文本的后缀，生成多个第一候选文本；

候选文本集合确定模块，设置为对所述预设数量的所述第一文本对应的多个第一候选文本进行遍历组合，确定多个候选文本集合，其中，每个候选文本集合包括所述预设数量的第一候选文本；

目标分数确定模块，设置为基于每个第一候选文本中的每个词，确定每个候选文本集合中重复出现的重复词，并对所述重复词对应的词分数进行惩罚处

理，确定所述重复词对应的目标分数，其中，所述重复词对应的词分数用于表征所述重复词作为文本后缀词的条件概率值；

集合分数确定模块，设置为基于每个候选文本集合中的重复词对应的目标分数和非重复词对应的词分数，确定每个候选文本集合对应的集合分数；

第二文本确定模块，设置为基于所述集合分数，从多个候选文本集合中确定目标文本集合，并基于所述目标文本集合，确定当前时刻生成的所述预设数量的第二文本。

10、一种电子设备，包括：

至少一个处理器；

存储器，设置为存储至少一个程序；

当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现如权利要求 1-8 中任一所述的文本生成方法。

11、一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1-8 中任一所述的文本生成方法。