



(12)发明专利申请

(10)申请公布号 CN 109117485 A

(43)申请公布日 2019.01.01

(21)申请号 201811037203.2

(22)申请日 2018.09.06

(71)申请人 北京京东尚科信息技术有限公司

地址 100195 北京市海淀区杏石口路65号

西杉创意园四区11号楼东段1-4层西
段1-4层

申请人 北京京东世纪贸易有限公司

(72)发明人 宋阳 陈蒙

(74)专利代理机构 中国国际贸易促进委员会专

利商标事务所 11038

代理人 王云飞 刘剑波

(51)Int.Cl.

G06F 17/28(2006.01)

G06F 17/27(2006.01)

G06Q 30/02(2012.01)

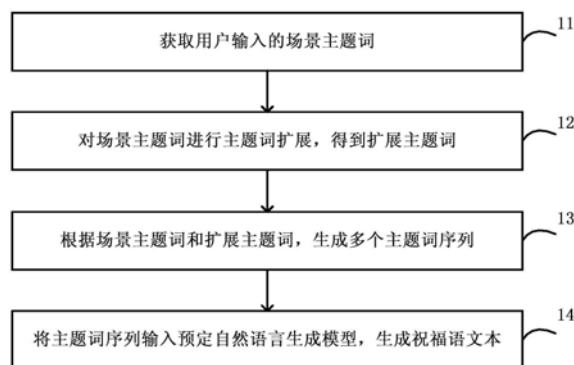
权利要求书2页 说明书11页 附图8页

(54)发明名称

祝福语文本生成方法和装置、计算机可读存储介质

(57)摘要

本发明公开一种祝福语文本生成方法和装置、计算机可读存储介质。该祝福语文本生成方法包括：获取用户输入的场景主题词；对场景主题词进行主题词扩展，得到扩展主题词；根据场景主题词和扩展主题词，生成多个主题词序列；将主题词序列输入预定自然语言生成模型，生成祝福语文本。本发明可以根据用户输入的主题词自动扩展主题词，并根据这些扩展出的主题词找出最优候选主题词序列。



1. 一种祝福语文本生成方法,其特征在于,包括:
获取用户输入的场景主题词;
对场景主题词进行主题词扩展,得到扩展主题词;
根据场景主题词和扩展主题词,生成多个主题词序列;
将主题词序列输入预定自然语言生成模型,生成祝福语文本。
2. 根据权利要求1所述的祝福语文本生成方法,其特征在于,还包括:
对预定自然语言生成模型生成的祝福语文本进行扩展,获取概率最高的N个祝福语文本,其中N为扩展宽度,N为大于1的自然数;
将N个祝福语文本输入预定语言模型,获取每个祝福语文本的困惑度;
将困惑度最低的祝福语文本作为最优祝福语文本反馈给用户。
3. 根据权利要求2所述的祝福语文本生成方法,其特征在于,还包括:
获取用户反馈的、用户修改后的祝福语文本;
将用户修改后的祝福语文本作为训练数据,对预定自然语言生成模型和预定语言模型进行优化。
4. 根据权利要求1-3中任一项所述的祝福语文本生成方法,其特征在于,所述对场景主题词进行主题词扩展,得到扩展主题词包括:
获取祝福语主题词和词向量模型;
对场景主题词利用词向量模型在祝福语主题词的范围内进行相似度扩展;
将与用户主题词的相似度大于预定值的祝福语主题词,作为扩展主题词。
5. 根据权利要求4所述的祝福语文本生成方法,其特征在于,所述获取祝福语主题词包括:
获取祝福语文本数据和通用数据;
对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果;
根据祝福语文本数据的切词结果和通用数据的切词结果获取一个词语出现祝福语文本数据的词频和出现在通用数据的词频;
若一个词语出现祝福语文本数据的词频大于出现在通用数据的词频、且所述词语出现在祝福语文本数据的次数大于预定值,则判定所述词语为祝福语主题词。
6. 根据权利要求5所述的祝福语文本生成方法,其特征在于,所述获取词向量模型包括:
利用祝福语文本数据的切词结果和通用数据的切词结果进行训练,获取词向量模型。
7. 根据权利要求1-3中任一项所述的祝福语文本生成方法,其特征在于,所述根据场景主题词和扩展主题词,生成多个主题词序列包括:
生成通用数据主题词序列语言模型和祝福语主题词序列语言模型;
将通用数据主题词序列语言模型和祝福语主题词序列语言模型进行插值处理,确定综合主题词序列语言模型;
将场景主题词和扩展主题词输入综合主题词序列语言模型,获取多个主题词序列。
8. 根据权利要求7所述的祝福语文本生成方法,其特征在于,所述生成通用数据主题词序列语言模型和祝福语主题词序列语言模型包括:

对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果;

对祝福语文本数据的切词结果和通用数据的切词结果中的每句话进行主题词提取;

根据每句话的主题词,确定每个祝福语文本的主题词序列;

利用祝福语文本的主题词序列进行训练,以生成通用数据主题词序列语言模型和祝福语主题词序列语言模型。

9. 根据权利要求1-3中任一项所述的祝福语文本生成方法,其特征在于,所述将主题词序列输入预定自然语言生成模型,生成祝福语文本包括:

预先采用训练数据进行训练,获取预定自然语言生成模型;

将主题词序列输入预定自然语言生成模型,生成祝福语文本。

10. 根据权利要求9所述的祝福语文本生成方法,其特征在于,所述预先采用训练数据进行训练,获取预定自然语言生成模型包括:

获取训练数据,其中,所述训练数据包括祝福语文本数据和通用数据;

从每行训练数据提出一个主题词;

将当前主题词和之前的训练结果输入预定自然语言生成模型,进行本轮训练。

11. 根据权利要求9所述的祝福语文本生成方法,其特征在于,所述预先采用训练数据进行训练,获取预定自然语言生成模型包括:

采用通用数据作为训练数据进行训练,得到通用模型;

采用祝福语文本数据作为训练数据继续训练,得到预定自然语言生成模型。

12. 一种祝福语文本生成装置,其特征在于,包括:

场景主题词获取模块,用于获取用户输入的场景主题词;

主题词扩展模块,用于对场景主题词进行主题词扩展,得到扩展主题词;

主题词序列生成模块,用于根据场景主题词和扩展主题词,生成多个主题词序列;

祝福语文本生成模块,用于将主题词序列输入预定自然语言生成模型,生成祝福语文本。

13. 根据权利要求12所述的祝福语文本生成装置,其特征在于,所述祝福语文本生成装置用于执行实现如权利要求1-12中任一项所述的祝福语文本生成方法的操作。

14. 一种祝福语文本生成装置,其特征在于,包括:

存储器,用于存储指令;

处理器,用于执行所述指令,使得所述祝福语文本生成装置执行实现如权利要求1-11中任一项所述的祝福语文本生成方法的操作。

15. 一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机指令,所述指令被处理器执行时实现如权利要求1-11中任一项所述的祝福语文本生成方法。

祝福语文本生成方法和装置、计算机可读存储介质

技术领域

[0001] 本发明涉及通信技术领域,特别涉及一种祝福语文本生成方法和装置、计算机可读存储介质。

背景技术

[0002] 随着电商行业的不断发展和人民生活水平的不断提高,越来越多的用户开始选择从电商平台网上购物作为礼品送给亲朋好友。

[0003] 在网购礼品赠送亲朋好友的同时,用户一般都会有为礼品配上几句祝福话语的需求,大部分用户会选择自己写祝福语文本,但还是会存在部分用户需要根据他们的需求生成一些个性化祝福语文本。

[0004] 相关技术商业化内容生成的应用场景很多,主要是通过以下三种方案进行内容生成:

[0005] 第一、基于人工生成的:人工预先为每个用户场景(送给爸爸,送给妈妈,送给老婆)写上几首通用的祝福语文本,然后由用户选择场景,挑选祝福语文本。

[0006] 第二、基于检索的:给定一组主题词,为每个主题词在预先构建好的祝福语文本倒排索引知识库当中检索出一句话,将检索出来的句子组合成祝福语文本。

[0007] 第三、基于NLG(Natural Language Generation,自然语言生成)模型的。

发明内容

[0008] 申请人认为:相关技术存在如下一些缺点。

[0009] 第一、基于人工生成的相关技术存在以下缺点:

[0010] 1、人工写的祝福语文本成本太高,需要耗费大量人力物力。

[0011] 2、用户只能在指定场景下选择祝福语文本,无法满足用户个性化需求。

[0012] 3、同一场景下可选结果也是有限的,缺乏多样性。

[0013] 第二、基于检索的相关技术存在以下缺点:

[0014] 1、相同主题词检索出来的内容是一样的缺乏个性化。

[0015] 2、检索出来的内容缺乏语义连贯性。

[0016] 第三、基于NLG模型的相关技术存在以下缺点:

[0017] 1、相关技术NLG模型对用户选择的主题词的质量和顺序都非常敏感,但目前还没有根据用户需求自动生成主题词序列的解决方案。

[0018] 2、训练自然语言生成模型通常都需要大量的训练数据,但是大多数情况下很难收集到大量符合要求的训练数据。

[0019] 3、同一组主题词输入到自然语言生成模型里以后,只能输出一个概率最大的通用结果,缺乏多样性。

[0020] 4、相关技术NLG模型是静态的,没有自动学习和更新的能力。

[0021] 鉴于以上技术问题中的至少一项,本发明提供了一种祝福语文本生成方法和装

置、计算机可读存储介质,可以根据用户输入的主题词自动扩展主题词,并根据这些扩展出的主题词找出最优候选主题词序列。

[0022] 根据本发明的一个方面,提供一种祝福语文本生成方法,包括:

[0023] 获取用户输入的场景主题词;

[0024] 对场景主题词进行主题词扩展,得到扩展主题词;

[0025] 根据场景主题词和扩展主题词,生成多个主题词序列;

[0026] 将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0027] 在本发明的一些实施例中,所述祝福语文本生成方法还包括:

[0028] 对预定自然语言生成模型生成的祝福语文本进行扩展,获取概率最高的N个祝福语文本,其中N为扩展宽度,N为大于1的自然数;

[0029] 将N个祝福语文本输入预定语言模型,获取每个祝福语文本的困惑度;

[0030] 将困惑度最低的祝福语文本作为最优祝福语文本反馈给用户。

[0031] 在本发明的一些实施例中,所述祝福语文本生成方法还包括:

[0032] 获取用户反馈的、用户修改后的祝福语文本;

[0033] 将用户修改后的祝福语文本作为训练数据,对预定自然语言生成模型和预定语言模型进行优化。

[0034] 在本发明的一些实施例中,所述对场景主题词进行主题词扩展,得到扩展主题词包括:

[0035] 获取祝福语主题词和词向量模型;

[0036] 对场景主题词利用词向量模型在祝福语主题词的范围内进行相似度扩展;

[0037] 将与用户主题词的相似度大于预定值的祝福语主题词,作为扩展主题词。

[0038] 在本发明的一些实施例中,所述获取祝福语主题词包括:

[0039] 获取祝福语文本数据和通用数据;

[0040] 对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果;

[0041] 根据祝福语文本数据的切词结果和通用数据的切词结果获取一个词语出现祝福语文本数据的词频和出现在通用数据的词频;

[0042] 若一个词语出现祝福语文本数据的词频大于出现在通用数据的词频、且所述词语出现在祝福语文本数据的次数大于预定值,则判定所述词语为祝福语主题词。

[0043] 在本发明的一些实施例中,所述获取词向量模型包括:

[0044] 利用祝福语文本数据的切词结果和通用数据的切词结果进行训练,获取词向量模型。

[0045] 在本发明的一些实施例中,所述根据场景主题词和扩展主题词,生成多个主题词序列包括:

[0046] 生成通用数据主题词序列语言模型和祝福语主题词序列语言模型;

[0047] 将通用数据主题词序列语言模型和祝福语主题词序列语言模型进行插值处理,确定综合主题词序列语言模型;

[0048] 将场景主题词和扩展主题词输入综合主题词序列语言模型,获取多个主题词序列。

[0049] 在本发明的一些实施例中,所述生成通用数据主题词序列语言模型和祝福语主题词序列语言模型包括:

[0050] 对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果;

[0051] 对祝福语文本数据的切词结果和通用数据的切词结果中的每句话进行主题词提取;

[0052] 根据每句话的主题词,确定每个祝福语文本的主题词序列;

[0053] 利用祝福语文本的主题词序列进行训练,以生成通用数据主题词序列语言模型和祝福语主题词序列语言模型。

[0054] 在本发明的一些实施例中,所述将主题词序列输入预定自然语言生成模型,生成祝福语文本包括:

[0055] 预先对训练数据进行训练,获取预定自然语言生成模型;

[0056] 将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0057] 在本发明的一些实施例中,所述预先采用训练数据进行训练,获取预定自然语言生成模型包括:

[0058] 获取训练数据,其中,所述训练数据包括祝福语文本数据和通用数据;

[0059] 从每行训练数据提出一个主题词;

[0060] 将当前主题词和之前的训练结果输入预定自然语言生成模型,进行本轮训练。

[0061] 在本发明的一些实施例中,所述预先采用训练数据进行训练,获取预定自然语言生成模型包括:

[0062] 采用通用数据作为训练数据进行训练,得到通用模型;

[0063] 采用祝福语文本数据作为训练数据继续训练,得到预定自然语言生成模型。

[0064] 根据本发明的另一方面,提供一种祝福语文本生成装置,包括:

[0065] 场景主题词获取模块,用于获取用户输入的场景主题词;

[0066] 主题词扩展模块,用于对场景主题词进行主题词扩展,得到扩展主题词;

[0067] 主题词序列生成模块,用于根据场景主题词和扩展主题词,生成多个主题词序列;

[0068] 祝福语文本生成模块,用于将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0069] 在本发明的一些实施例中,所述祝福语文本生成装置用于执行实现如上述任一实施例所述的祝福语文本生成方法的操作。

[0070] 根据本发明的另一方面,提供一种祝福语文本生成装置,包括:

[0071] 存储器,用于存储指令;

[0072] 处理器,用于执行所述指令,使得所述祝福语文本生成装置执行实现如上述任一实施例所述的祝福语文本生成方法的操作。

[0073] 根据本发明的另一方面,提供一种计算机可读存储介质,所述计算机可读存储介质存储有计算机指令,所述指令被处理器执行时实现如上述任一实施例所述的祝福语文本生成方法。

[0074] 本发明可以根据用户输入的主题词自动扩展主题词,并根据这些扩展出的主题词找出最优候选主题词序列。

附图说明

[0075] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0076] 图1为本发明祝福语文本生成方法一些实施例的示意图。

[0077] 图2为本发明一些实施例中主题词扩展的示意图。

[0078] 图3为本发明一些实施例中确定主题词序列的示意图。

[0079] 图4为本发明一些实施例中生成祝福语文本的示意图。

[0080] 图5为本发明一个实施例中seq2seq模型训练过程的示意图。

[0081] 图6为本发明一个实施例中seq2seq模型预测过程的示意图。

[0082] 图7为本发明一个实施例中自适应模型训练过程的示意图。

[0083] 图8为本发明祝福语文本生成方法另一个实施例的示意图。

[0084] 图9为一个实施例中生成结果扩展与评估流程的示意图。

[0085] 图10为本发明祝福语文本生成方法又一个实施例的示意图。

[0086] 图11为本发明祝福语文本生成装置一些实施例的示意图。

[0087] 图12为本发明祝福语文本生成装置另一些实施例的示意图。

具体实施方式

[0088] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。以下对至少一个示例性实施例的描述实际上仅仅是说明性的,决不作为对本发明及其应用或使用的任何限制。基于本发明中的实施例,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,都属于本发明保护的范围。

[0089] 除非另外具体说明,否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

[0090] 同时,应当明白,为了便于描述,附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

[0091] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论,但在适当情况下,所述技术、方法和设备应当被视为授权说明书的一部分。

[0092] 在这里示出和讨论的所有示例中,任何具体值应被解释为仅仅是示例性的,而不是作为限制。因此,示例性实施例的其它示例可以具有不同的值。

[0093] 应注意到:相似的标号和字母在下面的附图中表示类似项,因此,一旦某一项在一个附图中被定义,则在随后的附图中不需要对其进行进一步讨论。

[0094] 图1为本发明祝福语文本生成方法一些实施例的示意图。优选的,本实施例可由本发明祝福语文本生成装置执行。该方法包括以下步骤:

[0095] 步骤11,获取用户输入的场景主题词。

[0096] 步骤12,对场景主题词(用户主题词)进行主题词扩展,得到扩展主题词。

[0097] 步骤13,根据场景主题词和扩展主题词,生成多个主题词序列。

[0098] 步骤14,将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0099] 在本发明的一些实施例中,所述预定自然语言生成模型可以为语言模型、RNN (Recurrent Neural Network,循环神经网络)模型和seq2seq(sequence to sequence,序列到序列)模型等模型中的至少一项。其中,语言模型的主要思路为基于当前词序列的语言模型统计信息预测下一个词,进而生成一句话;RNN模型的主要思想也是基于当前词序列的神经网络状态信息预测下一个词;seq2seq模型的主要思想是通过编码器对输入文本进行编码得到主题向量,然后根据编码器输出的主题向量利用解码器解码生成一个预测文本序列。

[0100] 基于本发明上述实施例提供的祝福语文本生成方法,针对相关技术NLG模型对主题词质量顺序过于敏感的技术问题,可以根据用户输入的主题词自动扩展主题词,并根据这些扩展出的主题词找出最优候选主题词序列。

[0101] 图2为本发明一些实施例中主题词扩展的示意图。如图2所示,本发明主题词扩展流程(例如图1实施例的步骤12)可以包括:

[0102] 步骤121,获取祝福语文本数据和通用数据。

[0103] 在本发明的一些实施例中,祝福语文本数据可以从网上收集的一些节日祝福语,生日祝福语等文本,数据量小;通用数据可以从网上收集的公开的现代诗歌、散文、歌词、小说、台词等数据,数据量大。

[0104] 步骤122,对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果。

[0105] 步骤123,根据祝福语文本数据的切词结果和通用数据的切词结果获取一个词语出现祝福语文本数据的词频和出现在通用数据的词频。

[0106] 在本发明的一些实施例中,步骤123可以包括:根据公式(1)利用切词结果分别统计词频,训练贝叶斯模型找出祝福语文本数据主题词。

$$[0107] \quad \mathbf{TF}(\omega_i) = \frac{\text{count}(\omega_i)}{\sum_{j=1}^N \text{count}(\omega_j)} \quad (1)$$

[0108] 如公式(1)所示,定义单词在某个语料数据中的词频为 $\mathbf{TF}(\omega_i)$,他是由该单词在语料出现次数 $\text{count}(\omega_i)$ 除以该语料所有的单词数 $\sum_{j=1}^N \text{count}(\omega_j)$ 得到;定义 $\mathbf{TF}_{\text{common}}(\omega_i)$ 表示单词 ω_i 在通用数据的词频, $\mathbf{TF}_{\text{target}}(\omega_i)$ 表示单词 ω_i 在目标数据祝福语数据的词频。

[0109] 步骤124,将每一词语出现祝福语文本数据的词频和出现在通用数据的词频输入祝福语主题词提取模型,确定该词语是否为祝福语主题词。

[0110] 如果一个主题词 ω_i 出现在祝福语文本当中的词频 $\mathbf{TF}_{\text{target}}(\omega_i)$ 大于出现在通用数据中的词频 $\mathbf{TF}_{\text{common}}(\omega_i)$,那么这个词很有可能就是一个祝福语文本主题词。

[0111] 在本发明的一些实施例中,步骤124可以包括:若一个词语出现祝福语文本数据的词频大于出现在通用数据的词频、且所述词语出现在祝福语文本数据的次数大于预定值,则判定所述词语为祝福语主题词。

[0112] 在本发明的另一些实施例中,为了增加置信度,要求祝福语文本主题词必须同时满足两个条件:

[0113] 1) 出现在祝福语文本中的次数大于一定概率,通常取 $\text{count}(\omega_i) > 100$ 。

[0114] 2) 出现在祝福语文本当中的词频大于出现在通用数据中的词频,例如可以取:

$$[0115] \quad \frac{\text{TF}_{\text{target}}(\omega_i)}{\text{TF}_{\text{common}}(\omega_i)} > 2$$

[0116] 步骤125,利用祝福语文本数据的切词结果和通用数据的切词结果进行训练,获取词向量模型。

[0117] 步骤126,对用户输入的场景主题词利用词向量模型在祝福语主题词的范围内进行相似度扩展。

[0118] 在本发明的一些实施例中,相似度算法一般采用余弦相似度。

[0119] 在本发明的一些具体实施例中,如公式(2)所示,向量a与向量b的余弦相似度为:

$$[0120] \quad \text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \times \|\mathbf{b}\|} \quad (2)$$

[0121] 步骤127,将与用户主题词的相似度大于预定值的祝福语主题词,作为扩展主题词。

[0122] 在本发明的一些实施例中,步骤127可以包括:如果用户主题词和某个祝福语主题词的余弦相似度大于0.5,即可以判定为扩展主题词。

[0123] 本发明上述实施例通过以上步骤可以对用户输入的场景相关的主题词进行扩展,并且保证扩展出来的主题词都和祝福语相关。

[0124] 图3为本发明一些实施例中确定主题词序列的示意图。如图3所示,本发明主题词序列确定流程(例如图1实施例的步骤13)可以包括:

[0125] 步骤131,对祝福语文本数据和通用数据分别切词,得到祝福语文本数据的切词结果和通用数据的切词结果。

[0126] 步骤132,对祝福语文本数据的切词结果和通用数据的切词结果中的每句话进行主题词提取。

[0127] 在本发明的一些实施例中,步骤132可以包括:使用tf-idf或者textrank算法,得到每句话的主题词。

[0128] 步骤133,由于每个祝福语文本会包含多句话,因此根据每句话的主题词,确定每个祝福语文本的主题词序列。

[0129] 在本发明的一些实施例中,如表1实施例所述的文本所对应的主题词序列为:流星手心伤心世界梦爱心。

[0130] 表1

[0131]	文本	主题词
	当泪似流星	流星
	滑过手心握不紧	手心
	别伤心不要紧	伤心
	这个世界还年轻	世界
	我们的梦还年轻	梦
	爱没有翅膀	爱
[0132]	也能飞行到你心	心

[0133] 步骤134,利用祝福语文本的主题词序列进行训练,以生成通用数据主题词序列语言模型和祝福语主题词序列语言模型。

[0134] 步骤135,将通用数据主题词序列语言模型和祝福语主题词序列语言模型进行插值处理,确定综合主题词序列语言模型。

[0135] 由于祝福语主题词序列语言模型过于稀疏,本发明上述实施例可以采用插值的方法,将通用文本主题词序列语言模型和祝福语文本主题词序列语言模型按1:1的比例插值。

[0136] 下面举一个例子简单的介绍下插值的原理:

[0137] 假如语料1中: $P_1(\text{篮球}|\text{我喜欢})=0.7$,语料2中: $P_2(\text{篮球}|\text{我喜欢})=0.3$,那么插值后的概率为 $P(\text{篮球}|\text{我喜欢})=0.7*0.5+0.3*0.5=0.5$,由此本发明上述实施例可以解决主题词序列数据过于稀疏的问题。

[0138] 步骤136,将场景主题词和扩展主题词输入综合主题词序列语言模型,获取多个主题词序列。

[0139] 在本发明的一些实施例中,步骤136可以包括:序列生成时,首先在扩展主题词和用户主题词的合集(后面统称:主题词合集)当中根据语言模型概率选取一个或者多个适合作为主题词序列开始的主题词;然后依次根据语言模型概率在主题词合集选取下一个词,直到达到需要的长度或者根据语言模型概率达到结尾,这样就得到了一个或者多个主题词序列。

[0140] 图4为本发明一些实施例中生成祝福语文本的示意图。如图4所示,本发明祝福语文本生成流程(例如图1实施例的步骤14)可以包括:

[0141] 步骤141,预先对训练数据进行训练,获取预定自然语言生成模型。

[0142] 步骤142,将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0143] 在本发明的一些实施例中,所述预定自然语言生成模型可以为seq2seq模型。seq2seq模型是一种序列到序列的自然语言生成模型。

[0144] 图5为本发明一个实施例中seq2seq模型训练过程的示意图。如图5所示,seq2seq模型包含一个编码器和一个解码器,编码器负责对输入的文本按时序进行编码,通常都采用RNN模型,其中RNN模型包括但不限于包括:RNN模型、LSTM(Long Short-Term Memory,长短期记忆网络)模型、GRU(Gate Recurrent Unit,门控循环单元)。

[0145] 首先要准备好训练数据,训练数据为申请人从网上收集的祝福语数据和通用文本数据;利用tf-idf或者textrank算法为每行数据提取一个主题词;训练时将主题词送给seq2seq模型的编码器进行编码,解码器的输出结果就是对应的文本。为了增加上下文之间的语义连贯性,每一次训练时,都会把当前数据的前N行文本一起送给seq2seq模型的编码器进行编码。本发明上述实施例还采用了attention的机制,让输出的每一个词都会有侧重于前面输入文本的主题。

[0146] 图6为本发明一个实施例中seq2seq模型预测过程的示意图。如图6所示,seq2seq模型的预测流程和训练过程基本一致,区别主要在于预测过程中只需要用户输入主题词,模型的解码器会自动生成一个预测文本,下一轮训练时将当前主题词和之前N(这个值要和训练保持一致)轮的预测结果,一起送给编码器完成下一轮预测。

[0147] 图7为本发明一个实施例中自适应模型训练过程的示意图。如图7所示,图4实施例的步骤141可以包括:本发明自适应模型训练过程。

[0148] 具体而言,所述自适应模型训练过程可以包括:

[0149] 步骤71,采用通用数据作为训练数据进行训练,得到通用模型。

[0150] 在本发明的一些实施例中,步骤71可以包括:从互联网上收集一些通用文本数据,包括但不限于诗歌、散文、歌词、小说、台词等等;利用通用文本训练一个通用的seq2seq模型。

[0151] 步骤72,在通用模型的基础上,采用祝福语文本数据作为训练数据继续训练,得到预定自然语言生成模型。

[0152] 在本发明的一些实施例中,步骤72可以包括:在这个通用模型的基础上,训练数据换成祝福语文本继续训练,得到一个更符合祝福语语境的seq2seq模型作为本发明上述实施例的最终模型。

[0153] 通常本发明上述实施例收集到的祝福语数据都很少,训练seq2seq模型时数据太少会影响到模型的效果。为了解决祝福语文本过少的问题,本发明上述实施例提出了一种自适应模型的解决方案。

[0154] 针对相关技术很难收集到大量符合要求的训练数据的技术问题,本发明上述实施例提出了一种自适应学习的方案,利用大量的通用的数据训练一个基础模型,然后在这个基础模型的基础上,利用少量符合任务要求的数据继续训练,得到最终符合任务要求的模型。

[0155] 本发明上述实施例自适应模型训练后获取的模型既能保证预测结果的通顺性、多样性,又能最大限度的符合任务场景。

[0156] 图8为本发明祝福语文本生成方法另一个实施例的示意图。优选的,本实施例可由本发明祝福语文本生成装置执行。该方法包括以下步骤:

[0157] 步骤11,获取用户输入的场景主题词。

[0158] 步骤12,对场景主题词(用户主题词)进行主题词扩展,得到扩展主题词。

[0159] 步骤13,根据场景主题词和扩展主题词,生成多个主题词序列。

[0160] 步骤14,将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0161] 步骤15,对预定自然语言生成模型生成的祝福语文本进行扩展,获取概率最高的N个祝福语文本,其中N为扩展宽度,N为大于1的自然数。

[0162] 步骤16,将N个祝福语文本输入预定语言模型,获取每个祝福语文本的困惑度。

[0163] 步骤17,将困惑度最低的祝福语文本作为最优祝福语文本反馈给用户。

[0164] 本发明上述实施例在图1实施例步骤11-步骤14的基础上,增加了生成结果扩展与评估的步骤15-17。

[0165] 图9为一个实施例中生成结果扩展与评估流程的示意图。如图9所示,所述生成结果扩展与评估流程可以包括:

[0166] 步骤15,单纯的使用seq2seq模型,每输入一个主题词返回的都是概率最大的结果,而这个结果往往都是最通俗结果,缺乏新意。因此本发明上述实施例可以使用beamsearch算法对每一步的生成结果进行了扩展。

[0167] 例如:在解码时每一步不在只保留概率最大的top 1个结果,而是根据预定的beamsearch宽度N设置保留概率最大的top N个结果,这样假设beamsearch宽度为10,每个主题词序列都能得到10个祝候选预测本文。

[0168] 步骤16,利用祝福语数据文本和通用数据文本合并到一起训练了一个语言模型,对beamsearch所有的输出预测本文计算困惑度。

[0169] 步骤17,将困惑度最低的祝福语文本作为最优祝福语文本反馈给用户,由此用户得到可以根据得到的多个祝福语主题词序列对应的预测文本中选择一个自己满意的结果。

[0170] 本发明上述实施例针对同一主题词序列生成结果过于单一的问题,本发明提出了利用beamsearch算法针对同一主题词序列生成多个候选结果,然后利用语言模型每个候选结果进行评估,找出最优结果。

[0171] 图10为本发明祝福语文本生成方法又一个实施例的示意图。优选的,本实施例可由本发明祝福语文本生成装置执行。在图1、图7-图9任一实施例的基础上,该方法还可以包括以下步骤:

[0172] 步骤18,获取用户反馈的、用户修改后的祝福语文本。

[0173] 步骤19,将用户修改后的祝福语文本作为训练数据,对预定自然语言生成模型和预定语言模型进行优化。

[0174] 为了能够动态的优化本发明上述实施例的NLG模型,本发明上述实施例设计了一套根据用户反馈自动更新模型的机制。用户得到本发明上述实施例的预测文本结果以后,可以对本发明上述实施例的结果进行修改,这些修改结果都可以反过来指导本发明上述实施例的模型优化的方向。本发明上述实施例可以把用户修改后的结果作为训练数据继续优化本发明上述实施例的模型,由此本发明上述实施例的模型上线以后就获得了自动更新的能力,可以根据用户的反馈自动的自我优化和完善。

[0175] 针对模型不能自动学习和更新的技术问题,本发明上述实施例提出了基于用户反馈动态来优化NLG模型,使模型具有了自动学习和更新的能力。

[0176] 本发明上述实施例提出了一套完整的结合用户需求自动生成个性化礼品祝福语文本的方案。

[0177] 图11为本发明祝福语文本生成装置一些实施例的示意图。如图11所示,所述祝福语文本生成装置可以包括场景主题词获取模块111、主题词扩展模块112、主题词序列生成模块113和祝福语文本生成模块114,其中:

[0178] 场景主题词获取模块111,用于获取用户输入的场景主题词。

[0179] 主题词扩展模块112,用于对场景主题词进行主题词扩展,得到扩展主题词。

[0180] 主题词序列生成模块113,用于根据场景主题词和扩展主题词,生成多个主题词序列。

[0181] 祝福语文本生成模块114,用于将主题词序列输入预定自然语言生成模型,生成祝福语文本。

[0182] 基于本发明上述实施例提供的祝福语文本生成装置,针对相关技术NLG模型对主题词质量顺序过于敏感的技术问题,可以根据用户输入的主题词自动扩展主题词,并根据这些扩展出的主题词找出最优候选主题词序列。

[0183] 在本发明的一些实施例中,如图11所示,所述祝福语文本生成装置还可以包括生成结果扩展与评估模块115,其中:

[0184] 生成结果扩展与评估模块115,用于对预定自然语言生成模型生成的祝福语文本进行扩展,获取概率最高的N个祝福语文本,其中N为扩展宽度,N为大于1的自然数;将N个祝福语文本输入预定语言模型,获取每个祝福语文本的困惑度;将困惑度最低的祝福语文本作为最优祝福语文本反馈给用户。

[0185] 本发明上述实施例针对同一主题词序列生成结果过于单一的问题,本发明提出了利用beamsearch算法针对同一主题词序列生成多个候选结果,然后利用语言模型每个候选结果进行评估,找出最优结果。

[0186] 在本发明的一些实施例中,如图11所示,所述祝福语文本生成装置还可以包括用户反馈模块116,其中:

[0187] 用户反馈模块116,用于获取用户反馈的、用户修改后的祝福语文本;将用户修改后的祝福语文本作为训练数据,对预定自然语言生成模型和预定语言模型进行优化

[0188] 针对模型不能自动学习和更新的技术问题,本发明上述实施例提出了基于用户反馈动态来优化NLG模型,使模型具有了自动学习和更新的能力。由此本发明上述实施例的模型上线以后就获得了自动更新的能力,可以根据用户的反馈自动的自我优化和完善。

[0189] 在本发明的一些实施例中,所述祝福语文本生成装置可以用于采用通用数据作为训练数据进行训练,得到通用模型;在通用模型的基础上,采用祝福语文本数据作为训练数据继续训练,得到预定自然语言生成模型。

[0190] 针对相关技术很难收集到大量符合要求的训练数据的技术问题,本发明上述实施例提出了一种自适应学习的方案,利用大量的通用的数据训练一个基础模型,然后在这个基础模型的基础上,利用少量符合任务要求的数据继续训练,得到最终符合任务要求的模型。

[0191] 本发明上述实施例自适应模型训练后获取的模型既能保证预测结果的通顺性、多样性,又能最大限度的符合任务场景。

[0192] 在本发明的一些实施例中,所述祝福语文本生成装置用于执行实现如上述任一实施例(例如图1-图10任一实施例)所述的祝福语文本生成方法的操作。

[0193] 本发明上述实施例提出了一套完整的结合用户需求自动生成个性化礼品祝福语文本的方案。

[0194] 本发明上述实施例采用与主题词扩展模块112进行主题词扩展,然后利用扩展出的主题词通过主题词序列生成模块113生成多个候选主题词序列,将生成的主题词序列灌

倒祝福语文本生成模块114 (NLG模型) 中生成祝福语文本, 然后利用生成结果扩展与评估模块115和用户反馈模块116对整个系统的各个模型进行优化, 从而实现了模型的自动更新, 提高了用户体验。

[0195] 图12为本发明祝福语文本生成装置另一些实施例的示意图。如图12所示, 所述祝福语文本生成装置可以包括存储器121和处理器122, 其中:

[0196] 存储器121, 用于存储指令。

[0197] 处理器122, 用于执行所述指令, 使得所述祝福语文本生成装置执行实现如上述任一实施例 (例如图1-图10任一实施例) 所述的祝福语文本生成方法的操作。

[0198] 基于本发明上述实施例提供的祝福语文本生成装置, 针对相关技术NLG模型对主题词质量顺序过于敏感的技术问题, 可以根据用户输入的主题词自动扩展主题词, 并根据这些扩展出的主题词找出最优候选主题词序列。

[0199] 根据本发明的另一方面, 提供一种计算机可读存储介质, 所述计算机可读存储介质存储有计算机指令, 所述指令被处理器执行时实现如上述任一实施例所述的祝福语文本生成方法。

[0200] 基于本发明上述实施例提供的计算机可读存储介质, 针对相关技术NLG模型对主题词质量顺序过于敏感的技术问题, 可以根据用户输入的主题词自动扩展主题词, 并根据这些扩展出的主题词找出最优候选主题词序列。

[0201] 在上面所描述的祝福语文本生成装置可以实现为用于执行本申请所描述功能的通用处理器、可编程逻辑控制器 (PLC)、数字信号处理器 (DSP)、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件或者其任意适当组合。

[0202] 至此, 已经详细描述了本发明。为了避免遮蔽本发明的构思, 没有描述本领域所公知的一些细节。本领域技术人员根据上面的描述, 完全可以明白如何实施这里公开的技术方案。

[0203] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成, 也可以通过程序来指令相关的硬件完成, 所述的程序可以存储于一种计算机可读存储介质中, 上述提到的存储介质可以是只读存储器, 磁盘或光盘等。

[0204] 本发明的描述是为了示例和描述起见而给出的, 而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本发明的原理和实际应用, 并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

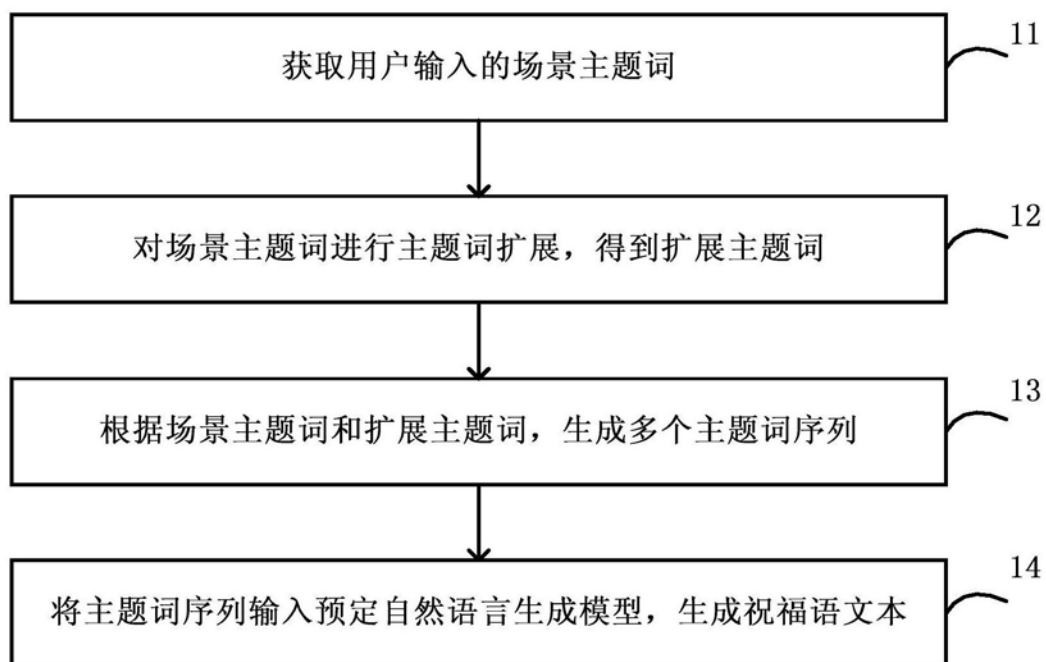


图1

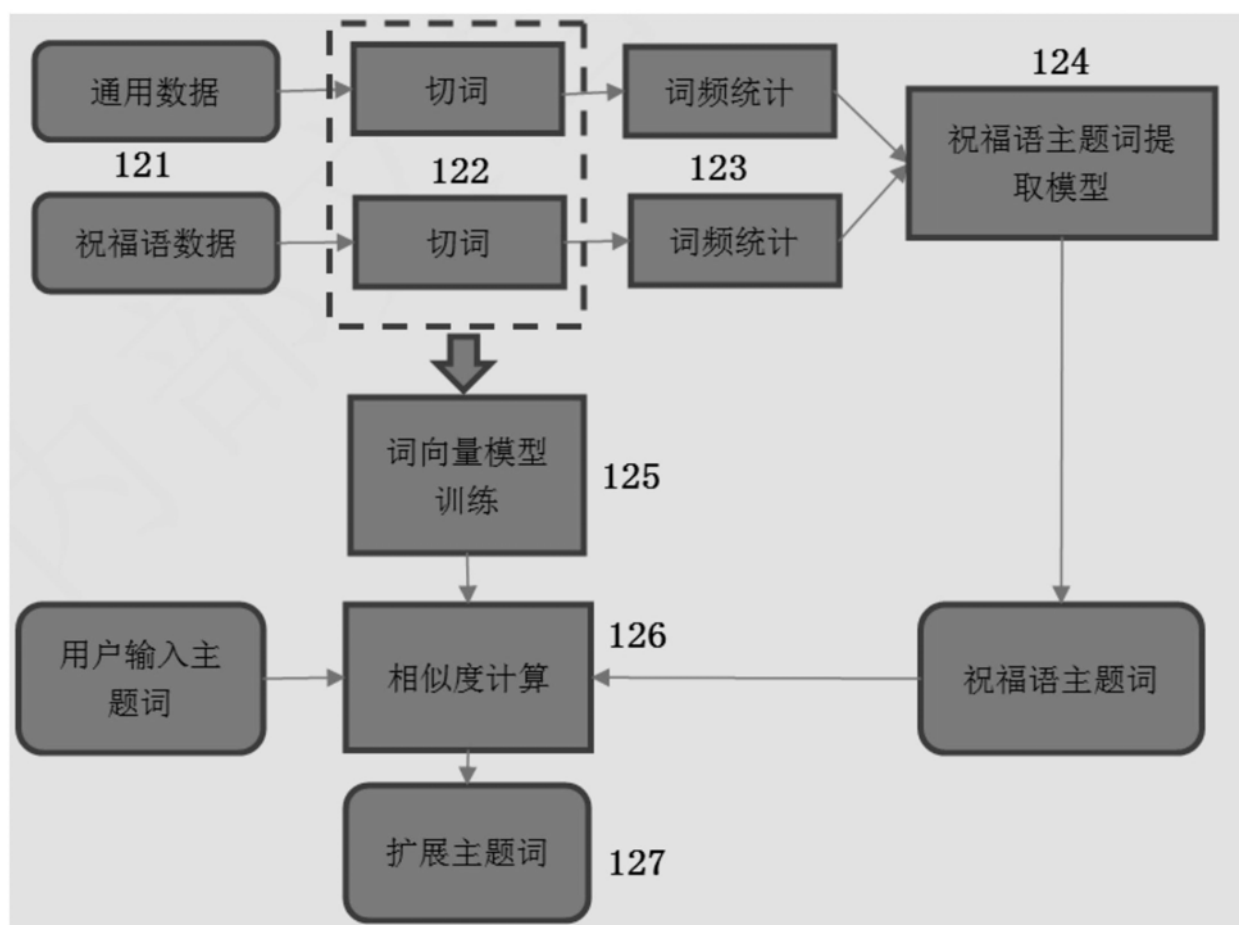


图2

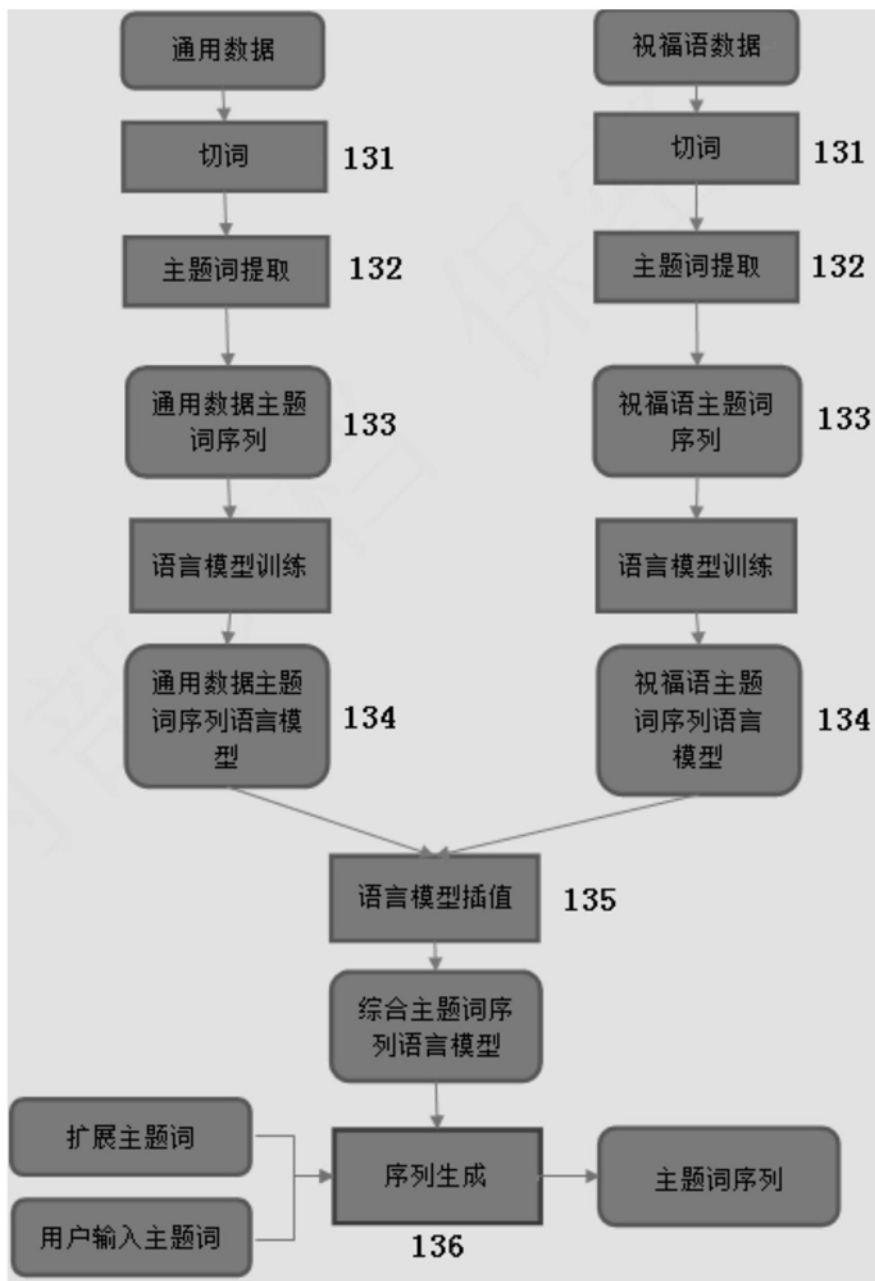


图3

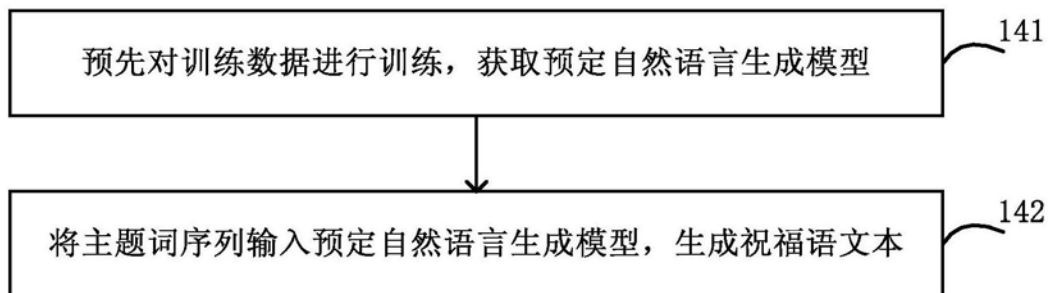


图4

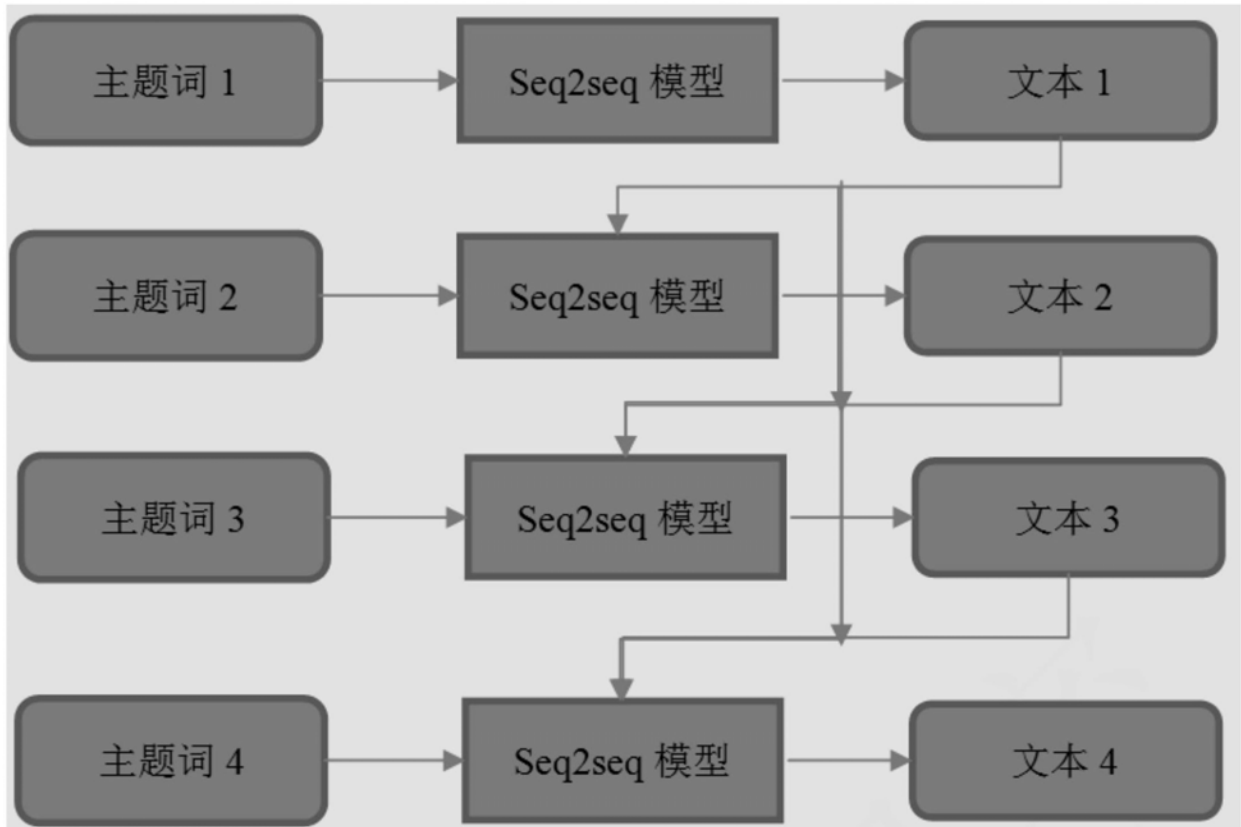


图5

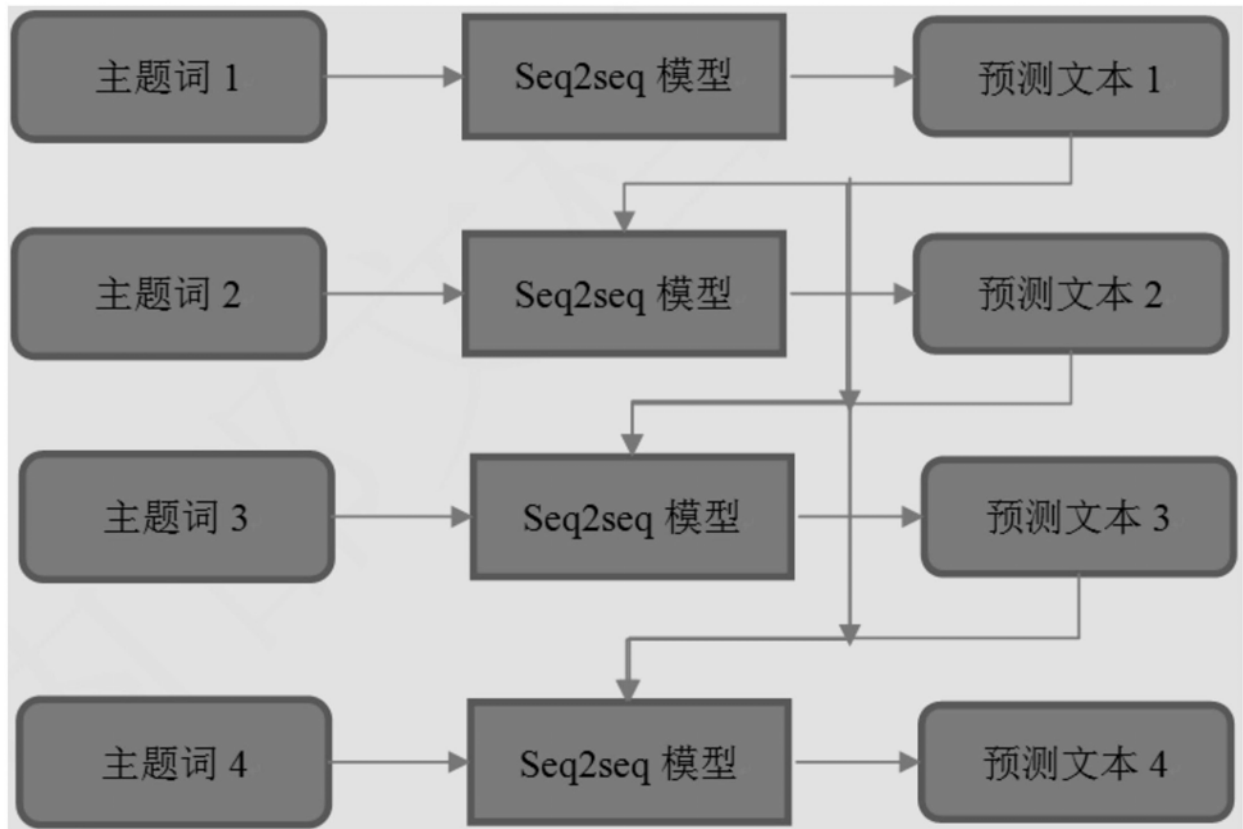


图6



图7

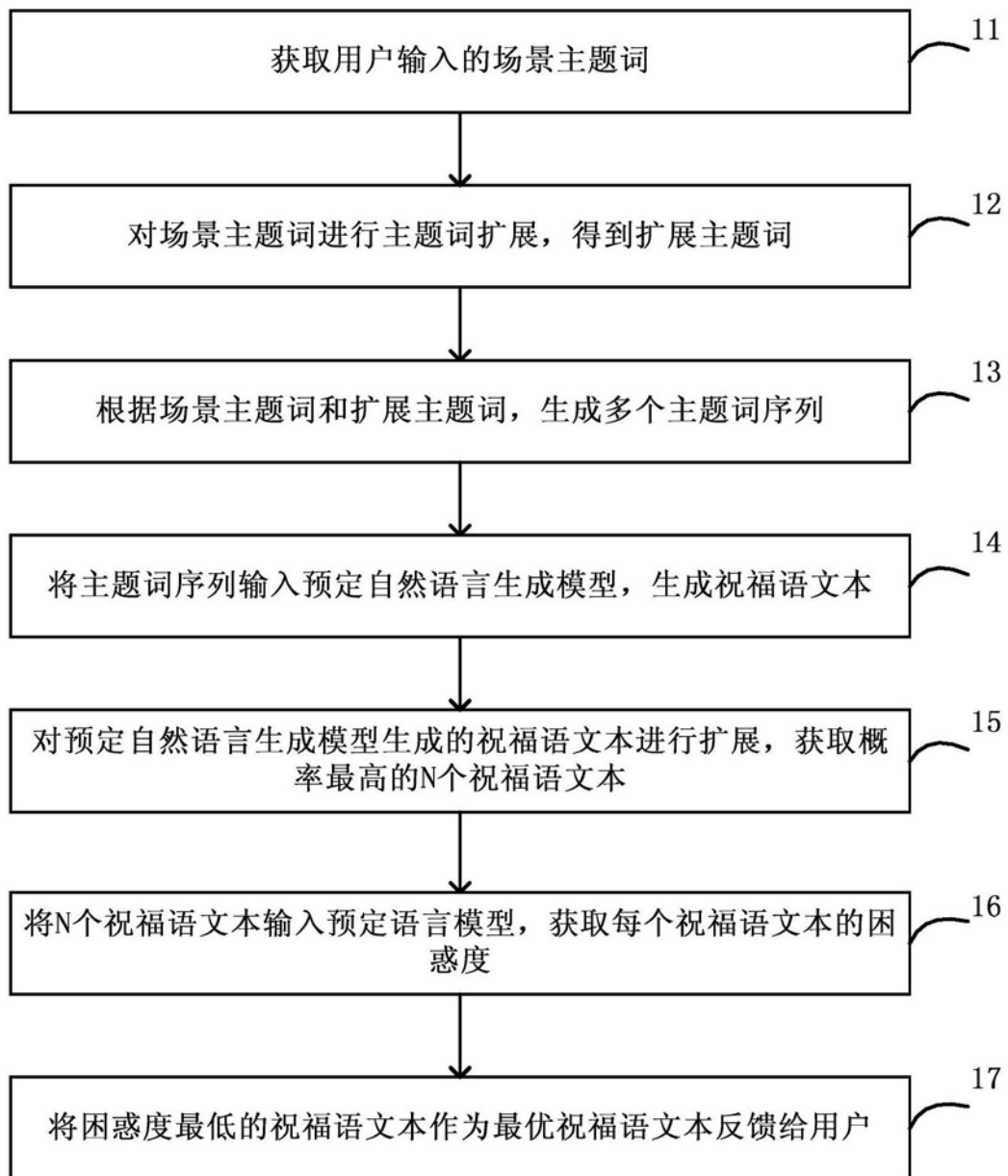


图8

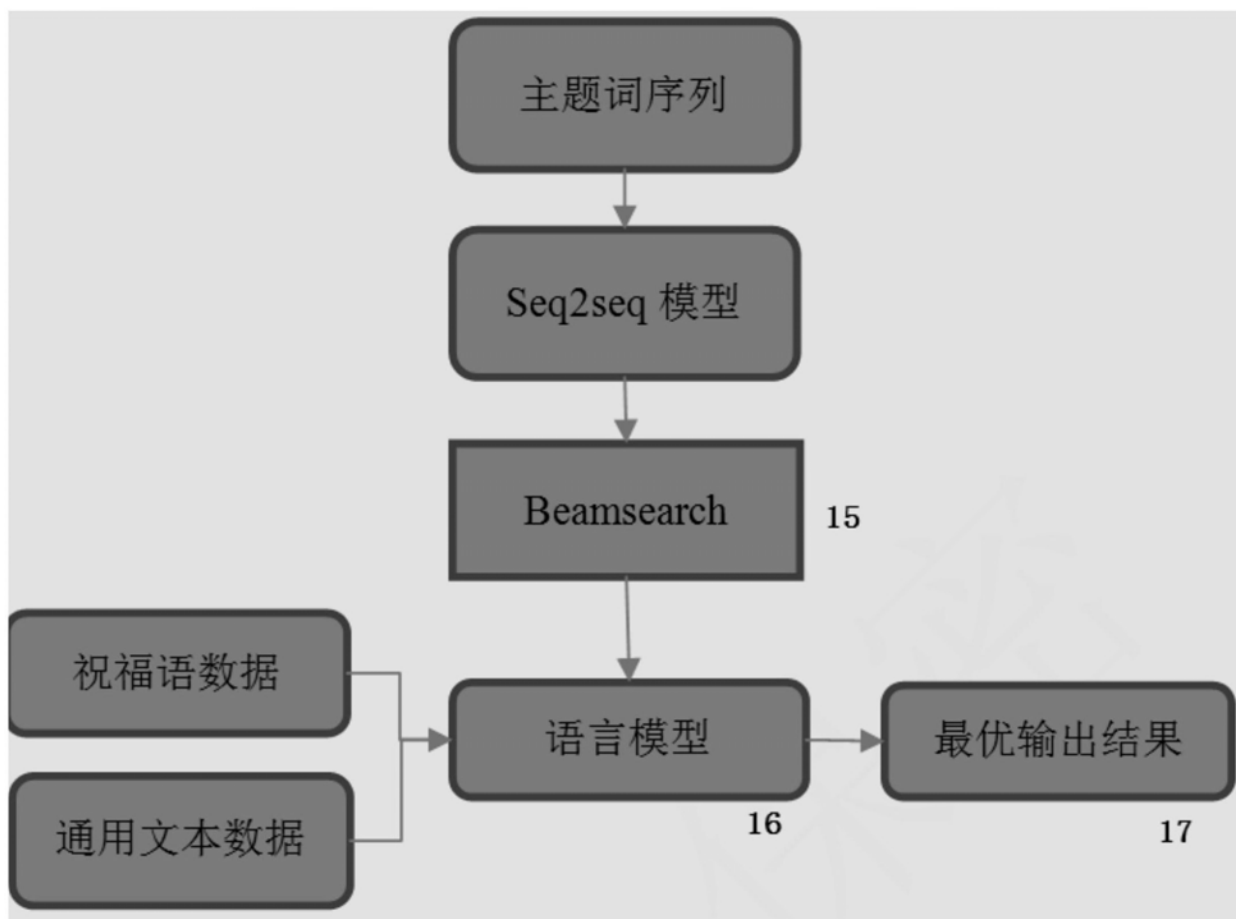


图9



图10

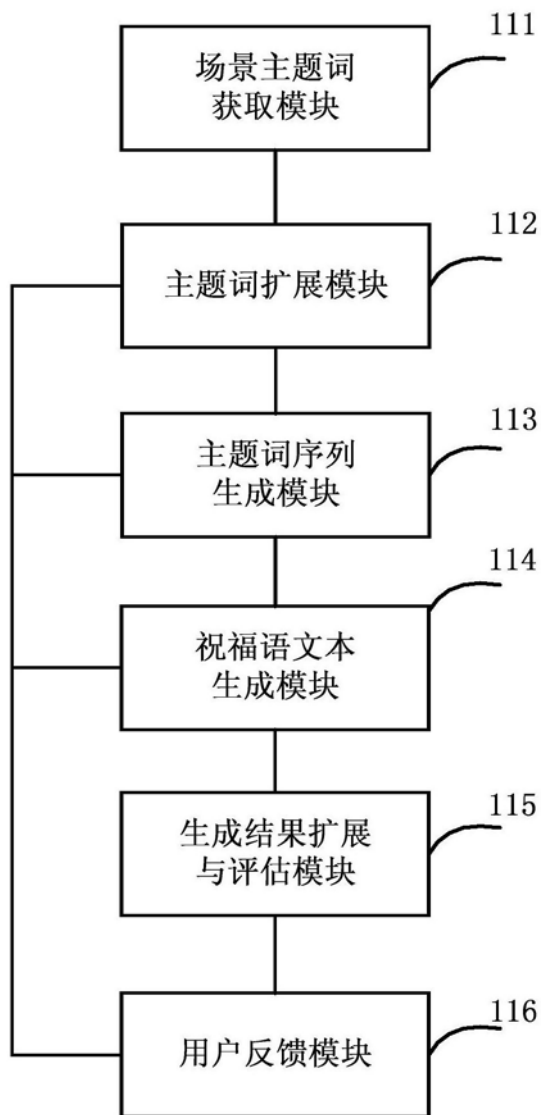


图11

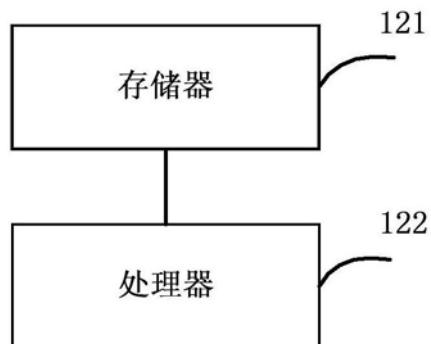


图12