(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111832309 A
(43)申请公布日 2020.10.27

(21)申请号 201910231583.1

(22)申请日 2019.03.26

(71)申请人 北京京东尚科信息技术有限公司
地址 100086 北京市海淀区知春路76号8层
申请人 北京京东世纪贸易有限公司

(72)发明人 刘瑞雪 陈蒙 刘晓华

(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
代理人 李昊 刘剑波

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/216(2020.01)
G06F 40/289(2020.01)
G06F 16/35(2019.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

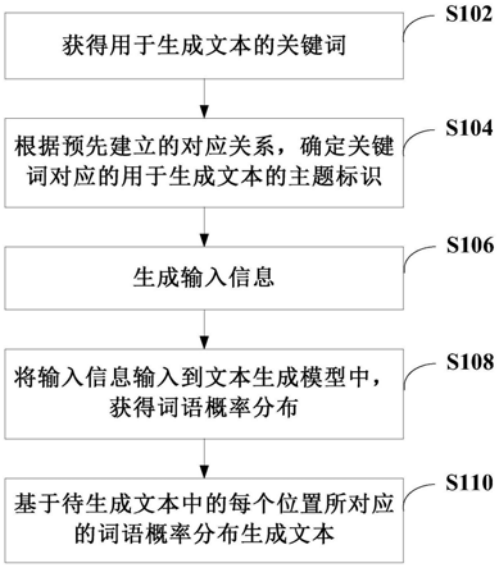
权利要求书2页 说明书10页 附图6页

(54)发明名称

文本生成方法、装置和计算机可读存储介质

(57)摘要

本发明公开了一种文本生成方法、装置和计算机可读存储介质，涉及自然语言处理领域。文本生成方法包括：获得用于生成文本的关键词；根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识；生成输入信息，其中，输入信息包括关键词和主题标识；将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布；基于待生成文本中的每个位置所对应的词语概率分布生成文本。本发明的实施例可以基于用于生成文本的关键词确定主题标识，使得文本生成的过程中能够参考主题标识对应的语义信息，从而可以使得生成的文本可读性更好。



CN 111832309 A

1.一种文本生成方法,包括:

获得用于生成文本的关键词;

根据预先建立的对应关系,确定所述关键词对应的用于生成文本的主题标识;

生成输入信息,其中,所述输入信息包括所述关键词和所述主题标识;

将输入信息输入到文本生成模型中,获得待生成文本中的每个位置所对应的词语概率分布;

基于所述待生成文本中的每个位置所对应的词语概率分布生成文本。

2.根据权利要求1所述的文本生成方法,其中,所述输入信息还包括用于生成文本的情感标识。

3.根据权利要求1或2所述的文本生成方法,其中,所述文本生成模型为基于循环神经网络模型的文本生成模型。

4.根据权利要求3所述的文本生成方法,其中,所述输入信息还包括前文文本。

5.根据权利要求4所述的文本生成方法,其中,在输入信息中,关键词、主题标识和前文文本为文本生成模型的编码器的输入,情感标识为文本生成模型的解码器的输入。

6.根据权利要求5所述的文本生成方法,其中,所述将输入信息输入到文本生成模型中,获得待生成文本中的每个位置所对应的词语概率分布包括:

将输入信息中的关键词、主题标识和前文文本输入到基于循环神经网络的文本生成模型的编码器中,获得多个第一隐藏状态;

将基于多个第一隐藏状态生成的、待生成文本中的每个位置所对应的上下文向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中,获得待生成文本中的每个位置所对应的第二隐藏状态;

根据待生成文本中的每个位置所对应的第二隐藏状态以及预设的权重向量矩阵,确定待生成文本中的每个位置所对应的词语概率分布。

7.根据权利要求6所述的文本生成方法,其中,所述获得待生成文本中的每个位置所对应的第二隐藏状态包括:

根据多个第一隐藏状态以及待生成文本中的第n-1个位置所对应的第二隐藏状态,确定待生成文本中的第n个位置所对应的、基于注意力机制的上下文向量,其中,n为大于1的正整数;

将待生成文本中的第n个位置所对应的上下文向量、已确定的第n-1个位置的词的词向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中,获得待生成文本中的第n个位置所对应的第二隐藏状态。

8.根据权利要求1所述的文本生成方法,还包括:

获取训练文本的语义特征;

通过对语义特征进行聚类,获得训练文本的语义分类结果,其中,属于同一语义类别的训练文本具有相同的主题,每个语义类别具有唯一的主题标识;

根据对每个语义类别的训练文本中关键词的统计结果,生成语义类别的主题标识与关键词之间的对应关系。

9.根据权利要求2所述的文本生成方法,还包括:

将训练文本输入到情感分类器中,获得训练文本的情感分类结果,其中,每个情感类别

具有唯一的情感标识；

将训练文本的关键词和相应的主题标识、训练文本的情感标识输入到循环神经网络模型中；

根据循环神经网络的输出结果与所述训练文本的差距，调整循环神经网络模型的参数，以便获得基于循环神经网络的文本生成模型。

10.一种文本生成装置，包括：

关键词获得模块，被配置为获得用于生成文本的关键词；

主题标识确定模块，被配置为根据预先建立的对应关系，确定所述关键词对应的用于生成文本的主题标识；

输入信息生成模块，被配置为生成输入信息，其中，所述输入信息包括所述关键词和所述主题标识；

概率分布获得模块，被配置为将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布；

文本生成模块，被配置为基于所述待生成文本中的每个位置所对应的词语概率分布生成文本。

11.一种文本生成装置，包括：

存储器；以及

耦接至所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行如权利要求1～9中任一项所述的文本生成方法。

12.一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权利要求1～9中任一项所述的文本生成方法。

文本生成方法、装置和计算机可读存储介质

技术领域

[0001] 本发明涉及自然语言处理领域，特别涉及一种文本生成方法、装置和计算机可读存储介质。

背景技术

[0002] 自然语言文本生成(Natural Language Generation,NLG)技术是自然语言处理领域的重要研究方向之一。NLG技术的可以自动生成文本资源，从而可以节省人工成本。

[0003] 在相关技术中，主要采用基于深度学习的方法生成文本。这种方法将丰富的大数据信息作为训练数据，并使用神经网络的算法设计不同复杂程度和结构的线性或非线性学习模型。神经网络的使用可以保证模型在训练数据中学习到相应的语义信息，并且通过分析已知的源语言与目标语言关系，来生成对应的文本。相关技术中的文本生成方法主要依赖于已知的关键词的信息。

发明内容

[0004] 发明人认识到，由于目前生成文本仅依赖于关键词，因此相关技术生成的文本存在文本结构不合理、内容单一的问题，从而导致生成的文本可读性很差。

[0005] 本发明实施例所要解决的一个技术问题是：如何提高自动生成的文本的可读性。

[0006] 根据本发明一些实施例的第一个方面，提供一种文本生成方法，包括：获得用于生成文本的关键词；根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识；生成输入信息，其中，输入信息包括关键词和主题标识；将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布；基于待生成文本中的每个位置所对应的词语概率分布生成文本。

[0007] 在一些实施例中，输入信息还包括用于生成文本的情感标识。

[0008] 在一些实施例中，文本生成模型为基于循环神经网络模型的文本生成模型。

[0009] 在一些实施例中，输入信息还包括前文文本。

[0010] 在一些实施例中，在输入信息中，关键词、主题标识和前文文本为文本生成模型的编码器的输入，情感标识为文本生成模型的解码器的输入。

[0011] 在一些实施例中，将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布包括：将输入信息中的关键词、主题标识和前文文本输入到基于循环神经网络的文本生成模型的编码器中，获得多个第一隐藏状态；将基于多个第一隐藏状态生成的、待生成文本中的每个位置所对应的上下文向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的每个位置所对应的第二隐藏状态；根据待生成文本中的每个位置所对应的第二隐藏状态以及预设的权重向量矩阵，确定待生成文本中的每个位置所对应的词语概率分布。

[0012] 在一些实施例中，获得待生成文本中的每个位置所对应的第二隐藏状态包括：根据多个第一隐藏状态以及待生成文本中的第n-1个位置所对应的第二隐藏状态，确定待生

成文本中的第n个位置所对应的、基于注意力机制的上下文向量，其中，n为大于1的正整数；将待生成文本中的第n个位置所对应的上下文向量、已确定的第n个位置的词的词向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的第n个位置所对应的第二隐藏状态。

[0013] 在一些实施例中，文本生成方法还包括：获取训练文本的语义特征；通过对语义特征进行聚类，获得训练文本的语义分类结果，其中，属于同一语义类别的训练文本具有相同的主题，每个语义类别具有唯一的主题标识；根据对每个语义类别的训练文本中关键词的统计结果，生成语义类别的主题标识与关键词之间的对应关系。

[0014] 在一些实施例中，文本生成方法还包括：将训练文本输入到情感分类器中，获得训练文本的情感分类结果，其中，每个情感类别具有唯一的情感标识；将训练文本的关键词和相应的主题标识、训练文本的情感标识输入到循环神经网络模型中；根据循环神经网络的输出结果与训练文本的差距，调整循环神经网络模型的参数，以便获得基于循环神经网络的文本生成模型。

[0015] 根据本发明一些实施例的第二个方面，提供一种文本生成装置，包括：关键词获得模块，被配置为获得用于生成文本的关键词；主题标识确定模块，被配置为根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识；输入信息生成模块，被配置为生成输入信息，其中，输入信息包括关键词和主题标识；概率分布获得模块，被配置为将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布；文本生成模块，被配置为基于待生成文本中的每个位置所对应的词语概率分布生成文本。

[0016] 根据本发明一些实施例的第三个方面，提供一种文本生成装置，包括：存储器；以及耦接至存储器的处理器，处理器被配置为基于存储在存储器中的指令，执行前述任意一种文本生成方法。

[0017] 根据本发明一些实施例的第四个方面，提供一种计算机可读存储介质，其上存储有计算机程序，其中，该程序被处理器执行时实现前述任意一种文本生成方法。

[0018] 上述发明中的一些实施例具有如下优点或有益效果：本发明的实施例可以基于用于生成文本的关键词确定主题标识，使得文本生成的过程中能够参考主题标识对应的语义信息，从而可以使得生成的文本可读性更好。

[0019] 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

附图说明

[0020] 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性的前提下，还可以根据这些附图获得其他的附图。

[0021] 图1为根据本发明一些实施例的文本生成方法的流程示意图。

[0022] 图2为根据本发明另一些实施例的文本生成方法的流程示意图。

[0023] 图3为根据本发明一些实施例的基于情感标识生成文本的方法的流程示意图。

[0024] 图4为根据本发明隐藏状态确定方法的流程示意图。

[0025] 图5为根据本发明一些实施例的主题标识与关键词之间的对应关系确定方法的流程示意图。

[0026] 图6为根据本发明一些实施例的文本生成模型训练方法的流程示意图。

[0027] 图7为根据本发明一些实施例的文本生成装置的结构示意图。

[0028] 图8为根据本发明另一些实施例的文本生成装置的结构示意图。

[0029] 图9为根据本发明又一些实施例的文本生成装置的结构示意图。

具体实施方式

[0030] 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。

[0031] 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。

[0032] 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。

[0033] 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。

[0034] 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。

[0035] 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。

[0036] 发明人经过分析后发现，自动生成的文本的质量主要依赖于生成过程中所使用的关键词的丰富程度。如果给定的关键词很少出现在训练数据中，或者给定的关键词不适合特定类型的文本，那么将无法生成合理的文本。并且，单纯地将单个关键词作为源语言来生成目标文本的过程只能参考单个词语的信息，因此无法生成语义丰富的文本，因此无法保证文本的完整性、流畅性和可读性。

[0037] 发明人想到，除了提供关键词以外，还可以根据关键词索引相关的主题。在生成文本的过程中，除了可以参考特定关键词所提供的信息，还可以参考该关键词所属的主题下的其他信息。从而，可以使得生成的文本语义更丰富，可读性更好。下面参考图1描述本发明一些实施例的文本生成方法。

[0038] 图1为根据本发明一些实施例的文本生成方法的流程示意图。如图1所示，该实施例的文本生成方法包括步骤S102～S110。

[0039] 在步骤S102中，获得用于生成文本的关键词。

[0040] 用于生成文本的关键词是根据需要而进行预设的。例如，想要生成的文本为描述炊烟的诗句，那么预设的关键词则为“炊烟”。根据需要，关键词可以有一个或多个。

[0041] 在步骤S104中，根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识。

[0042] 关键词与主题标识之间的对应关系可以通过对大量的文本进行分析获得，例如，可以通过聚类、语义统计等方法获得。每个主题均对应多个关键词，这些关键词来源于同一主题下的多个文本。从而，通过该对应关系，可以将输入的信息从关键词本身扩展到语义信息更丰富的主题。

[0043] 在步骤S106中，生成输入信息，其中，输入信息包括关键词和主题标识。

[0044] 例如，当输入信息中包括关键词“春天”和主题标识“LT2”时，则输入信息可以为“春天LT2”。在一些实施例中，可以将待输入信息通过词向量矩阵转换为输入向量，然后将输入向量输入到文本生成模型中。例如，可以分别确定“春天”、“LT2”所对应的词向量，然后采用这些词向量构建输入向量作为输入信息。

[0045] 在一些实施例中，输入信息中还可以包括前文文本。前文文本是指在待生成的文本之前出现的文本。例如，欲生成一个文本段落的第三个语句，那么段落中的第一个语句或者第二个语句可以作为用于生成文本的前文文本。在一些实施例中，前文文本为待生成的文本之前的、相邻的文本，从而可以使得前文文本与待生成的文本具有更紧密的语义关系。例如，待生成的文本是一个段落的第三个语句，那么一些实施例中可以仅将段落中的第二个语句作为前文文本。从而，生成文本时还可以参考前文文本的语义信息。

[0046] 在步骤S108中，将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布。

[0047] 在一些实施例中，文本生成模型可以是基于循环神经网络的模型。循环神经网络模型具有记忆性，并且是一种在序列的演进方向进行递归的递归神经网络。而自然语言的文本也可以被视为一种词语序列，并且上下文之间具有语义关联性。因此，可以借助循环神经网络实现文本的自动生成。循环神经网络在计算的过程中会逐一输出生成的文本中的每个位置所对应的概率分布。

[0048] 在一些实施例中，当文本生成模型为循环神经网络模型时，文本生成模型包括编码器和解码器。编码器用于将输入的向量转换为隐藏状态，解码器用于基于编码器的输出来确定生成的文本中各个位置对应的词语概率分布。编码器和解码器均是基于循环神经网络的计算模块。例如，可以将输入信息中的关键词、主题标识、前文文本等信息输入到基于循环神经网络的文本生成模型的编码器中，然后获得解码器输出的词语概率分布。

[0049] 待生成文本中的每个位置所对应的词语概率分布表示当前位置上出现每个词语的概率。从而，可以将出现概率最高的词语作为当前位置上的词语。当文本生成模型预测当前位置为结束符的概率最高时，表示不再继续当前文本中的词语预测，从而可以结束当前文本的生成过程。

[0050] 在步骤S110中，基于待生成文本中的每个位置所对应的词语概率分布生成文本。

[0051] 通过上述实施例的方法，可以基于用于生成文本的关键词确定主题标识，使得文本生成的过程中能够参考主题标识对应的语义信息，从而可以使得生成的文本可读性更好。

[0052] 发明人经过进一步分析后认识到，除了提供关键词、主题标识以外，在文本生成过程中，可以融合文本情感信息，从而使得生成的文本具有更多文学色彩的表达，进一步提升自动生成的文本的可读性。下面参考图2描述本发明文本生成方法的实施例。

[0053] 图2为根据本发明另一些实施例的文本生成方法的流程示意图。如图2所示，该实

施例的文本生成方法包括步骤S202～S210。

[0054] 在步骤S202中，获得用于生成文本的关键词和情感标识。

[0055] 在步骤S204中，根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识。

[0056] 在步骤S206中，生成输入信息，其中，输入信息包括关键词、主题标识和情感标识。根据需要，输入信息中还可以包括前文文本。

[0057] 例如，一种示例性的输入信息包括“河流+LT23+这被暴风雨所打击的土地+Sad”，其中，“河流”为关键词，“LT23”为主题标识，“这被暴风雨所打击的土地”为前文文本，“Sad”为表示悲伤的情感标识。根据需要，可以逐一将“河流”、“LT23”、“这”、“被”、“暴风雨”、“所”、“打击”、“的”、“土地”、“Sad”以及结束符转换为相应的向量，并根据这些向量生成输入信息。

[0058] 在步骤S208中，将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布。

[0059] 在步骤S210中，基于待生成文本中的每个位置所对应的词语概率分布生成文本。

[0060] 例如，待生成文本中的首个位置对应的词语概率分布中，“春天”的概率为3％、“阳光”的概率为0.85％、“我”的概率为0.5％……，其中“春天”的概率最高。因此，可以将“春天”作为待生成文本的首个位置对应的词。类似地，对第二个位置对应的词语概率分布进行解码而得到的词为“在”，第三、四、五、六个位置的解码结果依次为“诗人”、“的”、“眼中”、结束符。则生成的文本为“春天在诗人的眼中”。

[0061] 通过上述实施例的方法，可以在生成文本的过程中参考特定的情感信息，从而使得生成的文本更为丰富，更具备可读性。

[0062] 在一些实施例中，可以将输入信息中的关键词、主题标识和前文文本输入到文本生成模型的编码器，将输入信息中的情感标识输入到文本生成模型的解码器。从而，关键词、主题标识和前文文本可以在编码阶段提供更多的语义信息，情感标识可以在解码阶段令输出结果具有相应的情感倾向。下面参考图3描述本发明基于情感标识生成文本的方法的实施例。

[0063] 图3为根据本发明一些实施例的基于情感标识生成文本的方法的流程示意图。如图3所示，该实施例的文本生成方法包括步骤S302～S310。

[0064] 在步骤S302中，将输入信息中的关键词、主题标识和前文文本输入到基于循环神经网络的文本生成模型的编码器中，获得多个第一隐藏状态。

[0065] 在步骤S304中，将基于多个第一隐藏状态生成的、待生成文本中的每个位置所对应的上下文向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的每个位置所对应的第二隐藏状态。上下文向量是用于将输入和输出进行关联的中间变量。

[0066] 在步骤S306中，根据待生成文本中的每个位置所对应的第二隐藏状态以及预设的权重向量矩阵，确定待生成文本中的每个位置所对应的词语概率分布。

[0067] 在步骤S308中，根据待生成文本中的每个位置所对应的词语概率分布，确定待生成文本中的每个位置所对应的词语。

[0068] 在步骤S310中，根据待生成文本中的每个位置所对应的词语生成文本。

[0069] 通过上述实施例的方法，可以在编码阶段输入关键词、主题标识和前文文本以丰富文本生成过程中的语义信息，并且在解码阶段输入情感标识、以使得解码后的结果的情感更为丰富。从而，进一步提升了生成的文本的可读性。

[0070] 下面结合图4进一步描述第一隐藏状态和第二隐藏状态的确定方法。

[0071] 图4为根据本发明隐藏状态确定方法的流程示意图。如图4所示，该实施例的隐藏状态确定方法包括步骤S402～S406。

[0072] 在步骤S402中，将输入信息中的关键词、主题标识和前文文本输入到基于循环神经网络的文本生成模型的编码器中，获得多个第一隐藏状态。编码器的编码过程例如可以如公式(1)所示。

[0073] $h_m = LSTM(h_{m-1}, [x_m; T; P])$ (1)

[0074] 其中，m表示输入信息中的各个单位元素所在的位置，单位元素是指一个词或者一个标志，例如对于输入信息“河流+LT23+这被暴风雨所打击的土地+Sad”，“河流”、“LT23”、“暴风雨”、“土地”、“Sad”等均可以视为一个单位元素；h_m表示位置m所对应的第一隐藏状态；LSTM(•)表示编码器采用长短记忆网络进行计算(Long Short-Term Memory Neural Network，简称：LSTM)；x_m表示将输入信息中的第m个单位元素所对应的向量，其中，输入信息$X = [x_1, x_2, \cdots, x_m, \cdots] = [K; T; P]$，并且K表示关键词、T表示主题标识、P表示前文文本、[•]表示将中括号中的多个项进行合并后的结果。

[0075] 当m＝1时，可以令循环神经网络的输入为0，以确定输入信息的首个位置所对应的第一隐藏状态h_1。

[0076] 根据需要，还可以采用其他类型的循环神经网络，例如递归神经网络(Recurrent Neural Network，简称：RNN)、门控循环单元(Gated Recurrent Unit，简称：GRU)等等。

[0077] 在步骤S404中，根据多个第一隐藏状态以及待生成文本中的第n-1个位置所对应的第二隐藏状态，确定待生成文本中的第n个位置所对应的、基于注意力机制的上下文向量，其中，n为大于1的正整数。

[0078] 例如，可以采用公式(2)～(4)计算上下文向量。

[0079] $\alpha_{mn} = f(s_{n-1}, h_m)$ (2)

[0080] $\alpha'_{mn} = sof\ tmax(\alpha_{mn})$ (3)

[0081] $$c_n = \sum_{t'=0}^{M} \alpha'_{mt'} * h_{t'} \quad (4)$$

[0082] 其中，m表示输入信息中的各个单位元素所在的位置；n表示待生成的文本中的各个词所在的位置；α_{mn}和α'_{mn}为中间变量，用于反映输入的内容中各个位置所对应的得分信息；f(•)为预设的冲激函数，用于表示函数中的两个项的匹配程度，本领域技术人员可以根据需要来进行设置；s_{n-1}表示待生成文本中的第n-1个位置所对应的第二隐藏状态；h_m表示输入的内容中位置m所对应的第一隐藏状态；c_n表示待生成文本中的第n个位置所对应的上下文向量，其作用是将输出与输入进行关联；M表示表示输入的内容中的位置总数，即词的总数。一个主题标识也可以被视为一个词。

[0083] 在步骤S406中，将待生成文本中的第n个位置所对应的上下文向量、已确定的第n-1个位置的词的词向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的第n个位置所对应的第二隐藏状态。例如，可以采用公

式(5)计算第二隐藏状态。

[0084] $s_n = LSTM(s_{n-1}, [c_n; 0; e(y_{n-1})])$ (5)

[0085] 其中，n表示待生成的文本中的各个词所在的位置；s_n表示待生成文本中的第n个位置所对应的第二隐藏状态；LSTM(•)表示解码器采用长短记忆网络进行计算；c_n表示待生成文本中的第n个位置所对应的上下文向量；0表示情感标识；y_{n-1}表示待生成文本中的第n-1个位置上所对应的词语概率分布；$e(y_{n-1})$表示基于y_{n-1}确定的相应词语的词向量；[•]表示将中括号中的多个项进行合并后的结果。

[0086] 当n＝1时，可以令循环神经网络的输入为0，以确定输入信息的首个位置所对应的第二隐藏状态s_1。

[0087] 在确定了第二隐藏状态后，可以计算位置n所对应的词语概率分布y_t。例如，可以参考公式(6)。

[0088] $y_n = P(y_n | y_1, y_2, y_3, \cdots, y_{n-1}, c_n) = softmax(Ws_n)$ (6)

[0089] 其中，$P(y_n | y_1, y_2, y_3, \cdots, y_{n-1}, c_n)$表示在待生成文本中的前n-1个位置分别对应词语概率分布y_1、y_2、y_3、…、y_{n-1}并且上下文向量为c_n的情况下，第n个位置的词语概率分布为y_n的概率；W表示预设的权重向量；s_n表示待生成文本中的第n个位置所对应的第二隐藏状态。

[0090] 通过上述实施例的方法，可以基于关键词、主题标识和前文文本以及在之前生成的第一隐藏状态确定当前的第一隐藏状态，并且基于第一隐藏状态、情感标识和之前的解码结果确定当前的第二隐藏状态。从而，可以生成语义和情感信息更丰富的文本，提高了生成的文本的可读性。

[0091] 下面描述本发明一些实施例的训练过程。

[0092] 主题标识与关键词之间的对应关系可以是预先通过分析大量文本数据获得的。例如，每个主题标识所对应的关键词为同一类文本中的关键词。下面参考图5描述本发明主题标识与关键词之间的对应关系确定方法的实施例。

[0093] 图5为根据本发明一些实施例的主题标识与关键词之间的对应关系确定方法的流程示意图。如图5所示，该实施例的对应关系确定方法包括步骤S502～S506。

[0094] 在步骤S502中，获取训练文本的语义特征。语义特征可以是通过向量表示的。例如，可以使用句向量工具Sent2Vec方法获得每个训练文本的语义特征向量；又例如，可以首先获取训练文本中每个词的词向量，然后通过对词向量进行加权平均等方式获得训练文本的语义特征向量。

[0095] 在步骤S504中，通过对语义特征进行聚类，获得训练文本的语义分类结果，其中，属于同一语义类别的训练文本具有相同的主题，每个语义类别具有唯一的主题标识。

[0096] 在一些实施例中，可以使用亲和传播聚类算法(又称AP聚类算法，全称为：Affinity Propagation Clustering Algorithm)。根据需要，本领域技术人员也可以采用其他聚类算法，这里不再赘述。

[0097] 在对训练文本的语义特征进行聚类后，会生成若干针对不同隐含主题的句子簇，每个句子簇对应一个语义分类结果。每个分类可以具有唯一的主题标识，例如LT1、LT2……LT1000分别用于表示1000个类中每个类的主题标识。

[0098] 在步骤S506中，根据对每个语义类别的训练文本中关键词的统计结果，生成语义类别的主题标识与关键词之间的对应关系。

[0099] 在一些实施例中，可以将属于同一类中的训练文本的所有关键词与该类对应的主题标识进行对应，也可以将将属于同一类中的训练文本中出现频率高于预设值、或出现频率的排名高于预设值的关键词与该类对应的主题标识进行对应。

[0100] 上述实施例的方法基于文本的语义特征进行聚类，使得具有相同或相近语义信息的文本被分配到了同一类，以确定不同的文本所属的主题。从而，可以更准确地确定关键词与主题标识之间的对应关系。进一步地，在文本生成阶段，可以基于该对应关系提供更丰富、更准确的语义信息，使得生成的文本可读性更高。

[0101] 在一些实施例中，可以采用训练数据预先对循环神经网络进行训练，以获得文本生成模型。下面参考图6描述本发明一些实施例的文本生成模型训练方法的实施例。

[0102] 图6为根据本发明一些实施例的文本生成模型训练方法的流程示意图。如图6所示，该实施例的文本生成模型训练方法包括步骤S602～S606。

[0103] 在步骤S602中，将训练文本输入到情感分类器中，获得训练文本的情感分类结果，其中，每个情感类别具有唯一的情感标识。

[0104] 在一些实施例中，情感分类器可以是预先训练的。例如，可以获取具有情感标识的标注的数据集；然后，采用数据集对多种分类器进行训练；最后，对每种分类器的预测结果进行测试，以将准确率最高的分类器确定为情感分类器。

[0105] 在步骤S604中，将训练文本的关键词和相应的主题标识、情感标识输入到循环神经网络模型中。训练数据的一个示例如表1所示。

[0106] 表1

[0107]

训练文本	关键词	情感标签	隐含主题
假如我是一只鸟	鸟	中性	LT19
我也应该用嘶哑的喉咙歌唱	歌唱	悲伤	LT19
这被暴风雨所打击的土地	暴风雨	中性	LT40
这永远汹涌着我们的悲愤的河流	河流	悲伤	LT23
……	……	……	……
为什么我的眼里常含泪水	泪水	悲伤	LT21
因为我对这土地爱得深沉	土地	中性	LT34

[0108] 在步骤S606中，根据循环神经网络的输出结果与训练文本的差距，调整循环神经网络模型的参数，以便获得基于循环神经网络的文本生成模型。

[0109] 例如，将“歌唱+悲伤+LT19+假如我是一只鸟”输入到循环神经网络训练模型中，生成的文本为“我用美妙的喉咙歌唱”。根据生成的文本与训练文本的原内容“我也应该用嘶哑的喉咙歌唱”之间的差距，可以对神经网络模型进行调整。本领域技术人员应当清楚，上述调整过程仅为示例性的说明。在实际的训练过程中，通常基于批量的输入信息的输出结果对模型的参数进行更新。

[0110] 通过上述实施例的方法，可以基于情感标识、主题标识对循环神经网络模型进行训练，从而可以使得训练的文本生成模型能够输出语义更丰富、可读性更好的文本。

[0111] 本发明的实施例可以应用于多种场景。例如，自动生成网络中的产品的描述文本、自动生成对来自网络用户的问题的应答文本等等。由于本发明的实施例生成的文本具备更

好的可读性，用户可以快速地理解自动生成的文本所表达的内容，降低了用户由于不清楚文本内容而与文本的提供方进行反复沟通、确认的次数，节约了网络资源。

[0112] 当然，本发明的实施例还可以应用于生成网络文本以外的文本。本领域技术人员可以根据需要进行选择。

[0113] 下面参考图7描述本发明文本生成装置的实施例。

[0114] 图7为根据本发明一些实施例的文本生成装置的结构示意图。如图7所示，该实施例的文本生成装置70包括：关键词获得模710，被配置为获得用于生成文本的关键词；主题标识确定模块720，被配置为根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识；输入信息生成模块730，被配置为生成输入信息，其中，输入信息包括关键词和主题标识；概率分布获得模块740，被配置为将输入信息输入到文本生成模型中，获得待生成文本中的每个位置所对应的词语概率分布；文本生成模块750，被配置为基于待生成文本中的每个位置所对应的词语概率分布生成文本。

[0115] 在一些实施例中，输入信息还包括用于生成文本的情感标识。

[0116] 在一些实施例中，文本生成模型为基于循环神经网络模型的文本生成模型。

[0117] 在一些实施例中，输入信息还包括前文文本。

[0118] 在一些实施例中，在输入信息中，关键词、主题标识和前文文本为文本生成模型的编码器的输入，情感标识为文本生成模型的解码器的输入。

[0119] 在一些实施例中，概率分布获得模块740进一步被配置为将输入信息中的关键词、主题标识和前文文本输入到基于循环神经网络的文本生成模型的编码器中，获得多个第一隐藏状态；将基于多个第一隐藏状态生成的、待生成文本中的每个位置所对应的上下文向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的每个位置所对应的第二隐藏状态；根据待生成文本中的每个位置所对应的第二隐藏状态以及预设的权重向量矩阵，确定待生成文本中的每个位置所对应的词语概率分布。

[0120] 在一些实施例中，概率分布获得模块740进一步被配置为根据多个第一隐藏状态以及待生成文本中的第n-1个位置所对应的第二隐藏状态，确定待生成文本中的第n个位置所对应的、基于注意力机制的上下文向量，其中，n为大于1的正整数；将待生成文本中的第n个位置所对应的上下文向量、已确定的第n个位置的词的词向量以及输入信息中的情感标识输入到基于循环神经网络的文本生成模型的解码器中，获得待生成文本中的第n个位置所对应的第二隐藏状态。

[0121] 在一些实施例中，文本生成装置70还包括：对应关系生成模块760，被配置为获取训练文本的语义特征；通过对语义特征进行聚类，获得训练文本的语义分类结果，其中，属于同一语义类别的训练文本具有相同的主题，每个语义类别具有唯一的主题标识；根据对每个语义类别的训练文本中关键词的统计结果，生成语义类别的主题标识与关键词之间的对应关系。

[0122] 在一些实施例中，文本生成装置70还包括：模型训练模块770，被配置为将训练文本输入到情感分类器中，获得训练文本的情感分类结果，其中，每个情感类别具有唯一的情感标识；将训练文本的关键词和相应的主题标识、训练文本的情感标识输入到循环神经网络模型中；根据循环神经网络的输出结果与训练文本的差距，调整循环神经网络模型的参

数,以便获得基于循环神经网络的文本生成模型。

[0123] 图8为根据本发明另一些实施例的文本生成装置的结构示意图。如图8所示,该实施例的文本生成装置80包括:存储器810以及耦接至该存储器810的处理器820,处理器820被配置为基于存储在存储器810中的指令,执行前述任意一个实施例中的文本生成方法。

[0124] 其中,存储器810例如可以包括系统存储器、固定非易失性存储介质等。系统存储器例如存储有操作系统、应用程序、引导装载程序(Boot Loader)以及其他程序等。

[0125] 图9为根据本发明又一些实施例的文本生成装置的结构示意图。如图9所示,该实施例的文本生成装置90包括:存储器910以及处理器920,还可以包括输入输出接口930、网络接口940、存储接口950等。这些接口930,940,950以及存储器910和处理器920之间例如可以通过总线960连接。其中,输入输出接口930为显示器、鼠标、键盘、触摸屏等输入输出设备提供连接接口。网络接口940为各种联网设备提供连接接口。存储接口950为SD卡、U盘等外置存储设备提供连接接口。

[0126] 本发明的实施例还提供一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现前述任意一种文本生成方法。

[0127] 本领域内的技术人员应当明白,本发明的实施例可提供为方法、系统、或计算机程序产品。因此,本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用非瞬时性存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0128] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解为可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0129] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0130] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0131] 以上所述仅为本发明的较佳实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

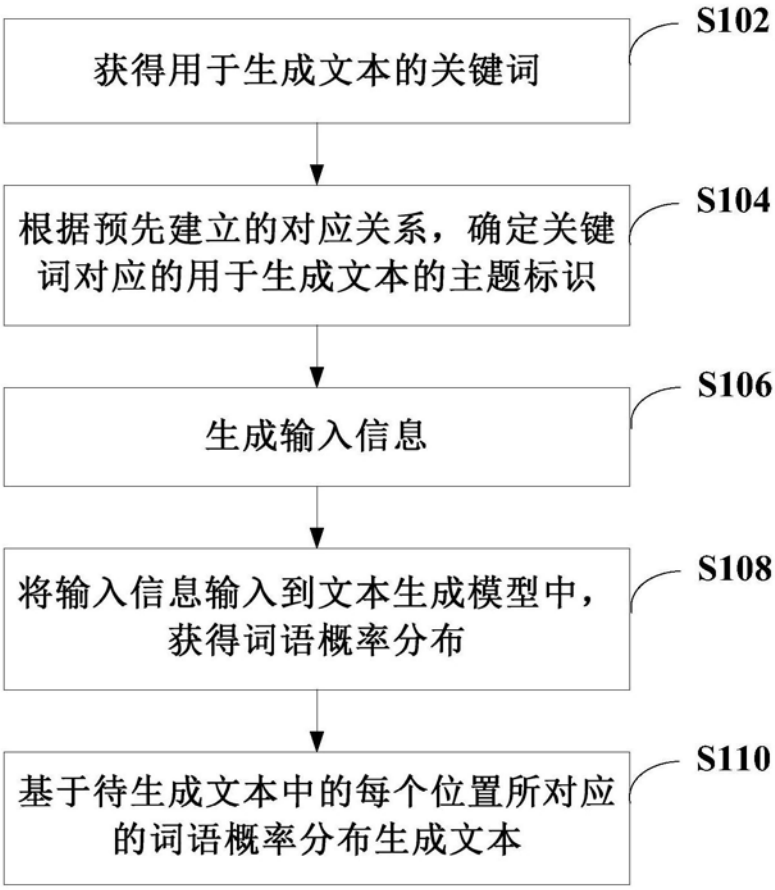

S102
获得用于生成文本的关键词
S104
根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识
S106
生成输入信息
S108
将输入信息输入到文本生成模型中，获得词语概率分布
S110
基于待生成文本中的每个位置所对应的词语概率分布生成文本

图1

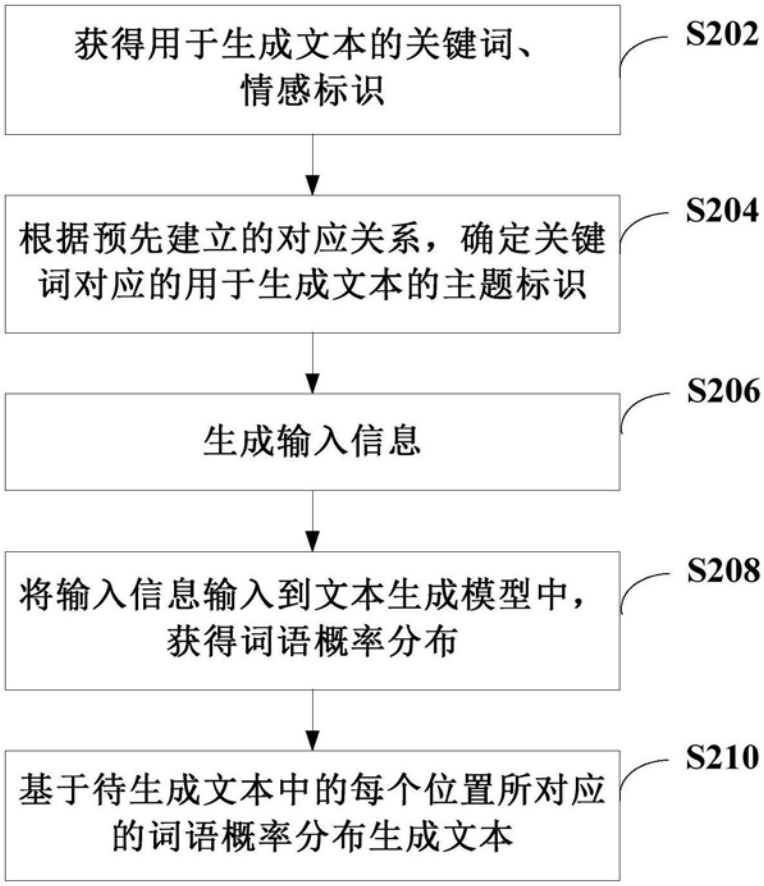

获得用于生成文本的关键词、情感标识
S202
根据预先建立的对应关系，确定关键词对应的用于生成文本的主题标识
S204
生成输入信息
S206
将输入信息输入到文本生成模型中，获得词语概率分布
S208
基于待生成文本中的每个位置所对应的词语概率分布生成文本
S210

图2

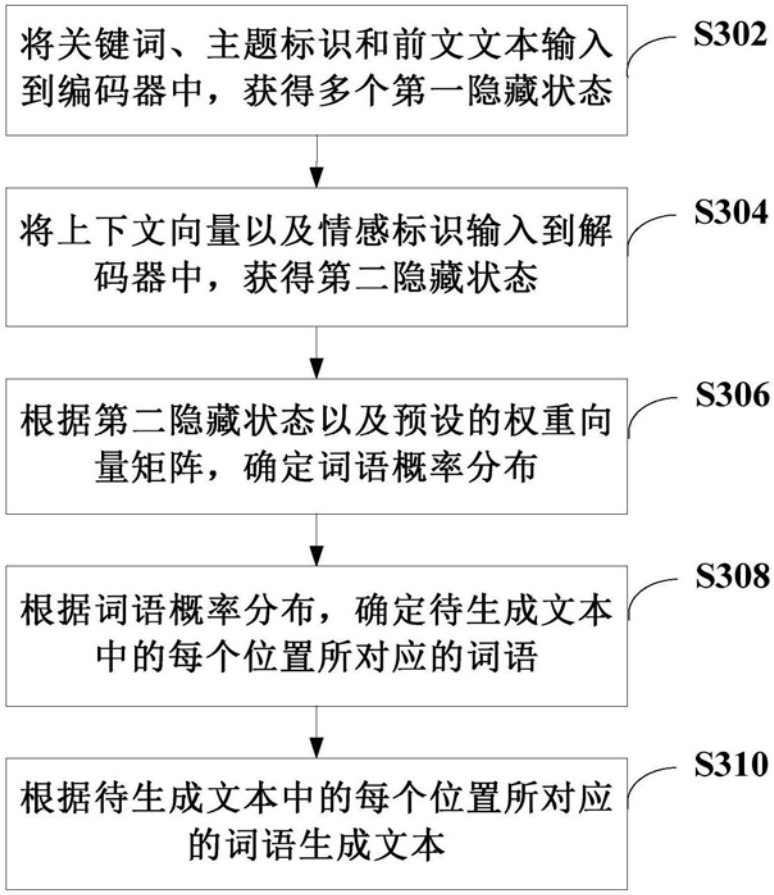

将关键词、主题标识和前文文本输入到编码器中，获得多个第一隐藏状态
S302
将上下文向量以及情感标识输入到解码器中，获得第二隐藏状态
S304
根据第二隐藏状态以及预设的权重向量矩阵，确定词语概率分布
S306
根据词语概率分布，确定待生成文本中的每个位置所对应的词语
S308
根据待生成文本中的每个位置所对应的词语生成文本
S310

图3

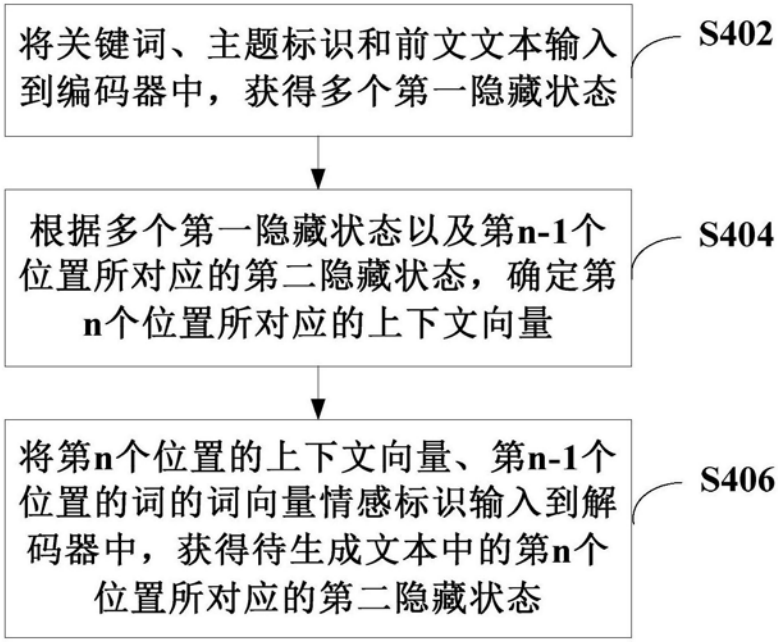

将关键词、主题标识和前文文本输入到编码器中，获得多个第一隐藏状态
S402
根据多个第一隐藏状态以及第n-1个位置所对应的第二隐藏状态，确定第n个位置所对应的上下文向量
S404
将第n个位置的上下文向量、第n-1个位置的词的词向量情感标识输入到解码器中，获得待生成文本中的第n个位置所对应的第二隐藏状态
S406

图4

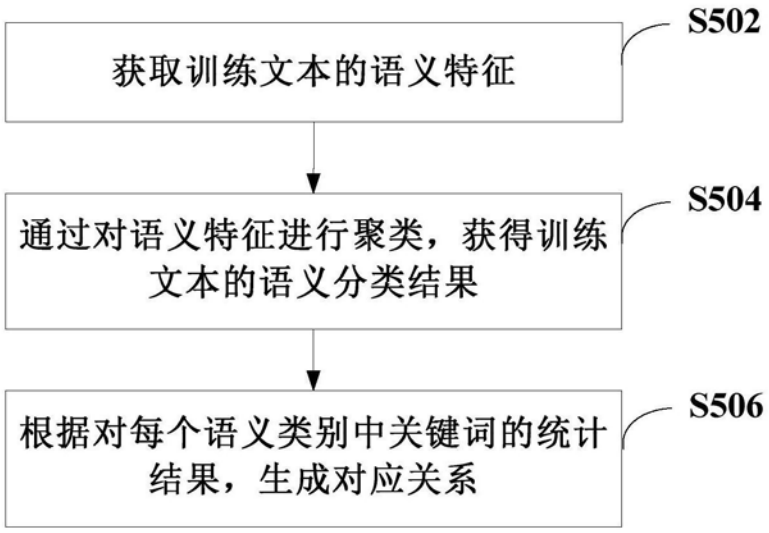

S502
获取训练文本的语义特征
S504
通过对语义特征进行聚类，获得训练文本的语义分类结果
S506
根据对每个语义类别中关键词的统计结果，生成对应关系

图5

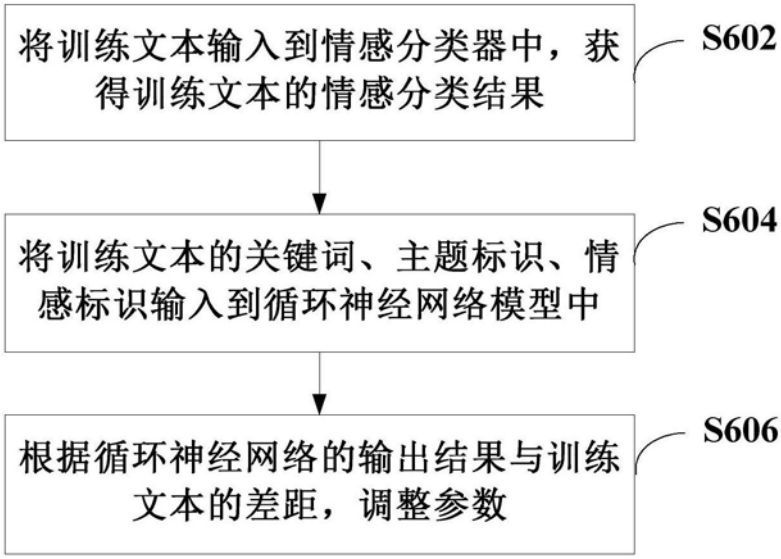

将训练文本输入到情感分类器中，获得训练文本的情感分类结果
S602
将训练文本的关键词、主题标识、情感标识输入到循环神经网络模型中
S604
根据循环神经网络的输出结果与训练文本的差距，调整参数
S606

图6

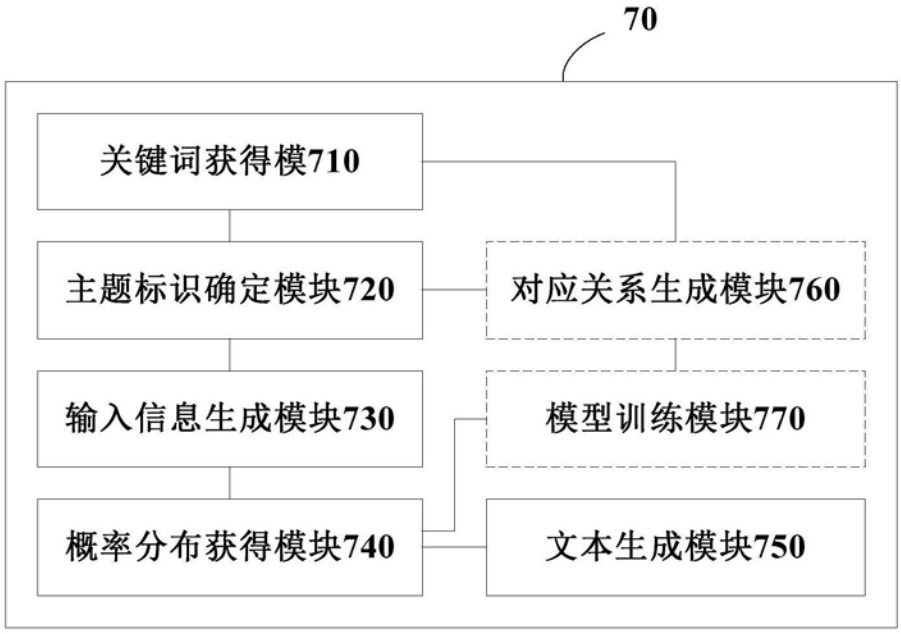

70
关键词获得模710
主题标识确定模块720
对应关系生成模块760
输入信息生成模块730
模型训练模块770
概率分布获得模块740
文本生成模块750

图7

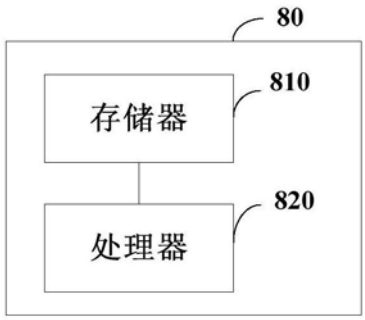

80
存储器
810
处理器
820

图8

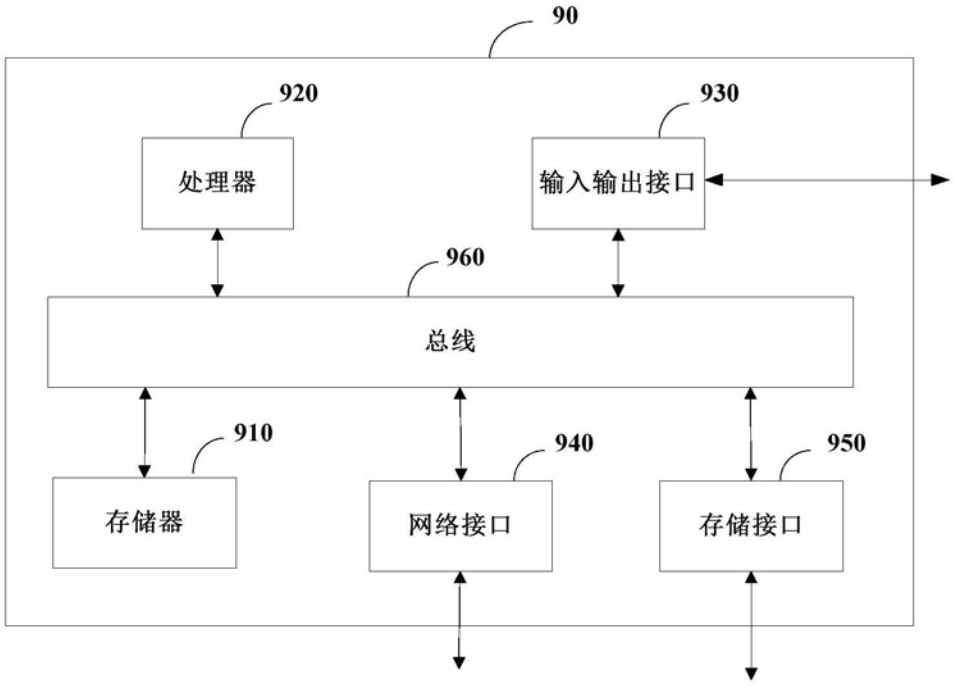


图9