

用于生成文本的方法和装置

相关申请的交叉引用

本专利申请要求于 2021 年 01 月 22 日提交的、申请号为
5 202110085890.0、发明名称为“用于生成文本的方法和装置”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。

技术领域

本公开的实施例涉及计算机技术领域，具体涉及深度学习领域，
10 尤其涉及一种用于生成文本的方法和装置。

背景技术

文本摘要是指将原始文本转换为只包含关键信息的精炼文本，主
要包括抽取式摘要和生成式摘要两种类型。抽取式摘要主要通过聚类
15 或序列标注等算法，从源文档中抽取关键的词语或短句。生成式摘要
主要采用深度学习的方法，利用序列到序列(Sequence2Sequence，
seq2seq)的框架，首先将源文档利用编码器编码为固定维度隐向量，然
后输入到解码器端进行逐步文字生成。

相关技术中，指针生成网络(Pointer Generator Network)在 seq2seq
20 模型的基础上，引入了复制机制和覆盖机制。复制机制在解码的每一
步都会计算源文档所有词语的注意力权重，可以选择性的将原文部分
拷贝到摘要中，进而可以生成出词汇表中未出现过的词语；覆盖机制
在解码的每一步都会累计计算当前步及之前所有步的注意力权重，避
免连续拷贝原文中已经获得高权重的部分。

25

发明内容

本公开的实施例提出了用于生成文本的方法和装置。

第一方面，本公开的实施例提供了一种用于生成文本的方法，该
方法包括：接收原始文本和要素标签；基于原始文本，采用预训练的
30 编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征

向量；采用预训练的解码器，执行如下解码步骤：基于解码器在当前时刻的隐状态向量和编码文本特征向量，确定原始文本中各个词在当前时刻的词注意力权重；基于解码器在当前时刻的隐状态向量、编码句特征向量和要素标签，确定原始文本中各个句子在当前时刻的句注意力权重；基于句注意力权重和词注意力权重，确定当前时刻原始文本中各个词归一化的词注意力权重；基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词；该方法还包括：基于解码器在各个时刻输出的目标词，生成目标文本。

10 在一些实施例中，基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量，包括：基于原始文本，生成文本特征矩阵；将文本特征矩阵输入预训练的第一编码器，得到编码文本特征向量；将文本特征矩阵输入预训练的第二编码器，得到原始文本中各个词的编码词特征向量；基于原始文本中的每个句子中的各个词的编码词特征向量，确定原始文本中各个句子的句特征向量；将原始文本中每个句子的句特征向量输入预训练的第三编码器，得到原始文本中各个句子的编码句特征向量。

20 在一些实施例中，基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词，包括：基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；基于上下文特征向量和编码文本特征向量，估计出当前时刻的目标词。

25 在一些实施例中，要素标签经由如下步骤生成：对原始文本聚类，得到聚类标签集；从聚类标签集中确定出目标聚类标签；将目标聚类标签确定为要素标签。

30 在一些实施例中，对原始文本聚类，得到聚类标签集，包括：将预设的标点符号作为切分边界，对原始文本进行切分，得到短句集；将短句集中不符合预设条件的短句删除，得到更新后的短句集；基于更新后的短句集，采用预训练的特征提取模型，提取每个短句中各个词的词特征，并基于词特征，确定出每个短句的句特征；基于更新后

的短句集中每个短句的句特征，采用预训练的聚类模型，确定出更新后的短句集中每个短句的聚类标签，得到聚类标签集。

第二方面，本公开的实施例提供了一种用于生成文本的装置，装置包括：接收单元，被配置成接收原始文本和要素标签；编码单元，
5 被配置成基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量；解码单元，被配置成采用预训练的解码器，执行如下解码步骤：基于解码器在当前时刻的隐状态向量和编码文本特征向量，确定原始文本中各个词在当前时刻的词注意力权重；基于解码器在当前时刻的隐状态向量、编码句特征向量
10 量和要素标签，确定原始文本中各个句子在当前时刻的句注意力权重；基于句注意力权重和词注意力权重，确定当前时刻原始文本中各个词归一化的词注意力权重；基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词；生成单元，被配置成基于解码器在各个时刻输出的目标词，生成目标文本。

15 在一些实施例中，编码单元进一步包括：特征矩阵生成模块，被配置成基于原始文本，生成文本特征矩阵；第一编码模块，被配置成将文本特征矩阵输入预训练的第一编码器，得到编码文本特征向量；第二编码模块，被配置成将文本特征矩阵输入预训练的第二编码器，得到原始文本中各个词的编码词特征向量；句特征向量生成模块，被
20 配置成基于原始文本中的每个句子中的各个词的编码词特征向量，确定原始文本中各个句子的句特征向量；第三编码模块，被配置成将原始文本中每个句子的句特征向量输入预训练的第三编码器，得到原始文本中各个句子的编码句特征向量。

25 在一些实施例中，解码单元进一步包括：上下文特征向量生成模块，被配置成基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；目标词输出模块，被配置成基于上下文特征向量和编码文本特征向量，估计出当前时刻的目标词。

30 在一些实施例中，该装置还包括要素标签生成单元，被配置成：对原始文本聚类，得到聚类标签集；从聚类标签集中确定出目标聚类

标签；将目标聚类标签确定为要素标签。

在一些实施例中，要素标签生成单元进一步包括：切分模块，被配置成将预设的标点符号作为切分边界，对原始文本进行切分，得到短句集；筛选模块，被配置成将短句集中不符合预设条件的短句删除，
5 得到更新后的短句集；特征提取模块，被配置成基于更新后的短句集，采用预训练的特征提取模型，提取每个短句中各个词的词特征，并基于词特征，确定出每个短句的句特征；标签生成模块，被配置成基于更新后的短句集中每个短句的句特征，采用预训练的聚类模型，确定出更新后的短句集中每个短句的聚类标签，得到聚类标签集。

10 第三方面，本公开的实施例提供了一种电子设备，该电子设备包括：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，本公开的实施例提供了一种计算机可读介质，其上存
15 储有计算机程序，该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。

第五方面，本公开的实施例提供了一种计算机程序产品，包括计算机程序，计算机程序在被处理器执行时实现如第一方面任一实现方式描述的方法。

20 本公开的实施例提供的用于生成文本的方法和装置，通过编码器对原始文本进行压缩，通过解码器确定出各个词的词注意力权重，以及基于原始文本中各个句子和要素标签的相关性，确定出各个句子的句注意力权重，进而基于词注意力权重和句注意力权重确定出归一化后的词注意力权重，并基于归一化后的词注意力权重估计出目标词，
25 最终基于目标词生成目标文本，使得目标文本中的句子和词与要素标签的相关性较高，提高了文本生成的准确度和可控程度。

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，
30 本公开的其它特征、目的和优点将会变得更明显：

图 1 是本公开的一些实施例可以应用于其中的示例性系统架构图；
图 2 是根据本公开的用于生成文本的方法的一个实施例的流程图；
图 3 是图 2 所示的用于生成文本的方法的一个实施例的应用场景示意图；

5 图 4 是根据本公开的用于生成文本的方法的又一个实施例的流程图；

图 5 是根据本公开的用于生成文本的装置的一个实施例的结构示意图；

图 6 是适于用来实现本公开的实施例的电子设备的结构示意图。

10

具体实施方式

下面结合附图和实施例对本公开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
15 有关发明相关的部分。

需要说明的是，在不冲突的情况下，本公开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公开。

图 1 示出了可以应用本公开的实施例的用于生成文本的方法或用于生成文本的装置的示例性系统架构 100。
20

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，例如有线、无线通信链路或者光纤电缆等等。

25 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，例如可以将原始文本发送至服务器，还可以从服务器接收目标文本。

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，可以是具备通信功能的电子设备，包括
30 但不限于智能手机、平板电脑、电子书阅读器、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。其可以实现成例如用来提供分布式服务的多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

- 5 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 上传的原始文本进行处理（例如基于原始文本生成目标文本）的后台数据服务器。后台数据服务器可以对接收到的原始文本进行分编码、解码等处理，并将处理结果（例如目标文本）反馈给终端设备。

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
10 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，可以实现成例如用来提供分布式服务的多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本公开的实施例所提供的用于生成文本的方法可
15 以由终端设备 101、102、103 执行，也可以由服务器 105 执行。相应地，用于生成文本的装置可以设置于终端设备 101、102、103 中，也可以设置于服务器 105 中。在此不做具体限定。

继续参考图 2，示出了根据本公开的用于生成文本的方法的一个实施例的流程 200。该用于生成文本的方法，包括以下步骤：

- 20 步骤 201，接收原始文本和要素标签。

在本实施例中，要素标签可以表征用户对目标文本内容的期望，可以由用户根据实际需求预先设置。

作为示例，原始文本可以是一个商品的详细信息，目标文本可以是该商品的信息摘要。若用户期望目标文本中可以包含与“外观”相关的
25 内容，可以将要素标签设置为“外观”，然后将该商品的详细信息和要素标签发送至本实施例的执行主体（例如可以是图 1 中所示的服务器 105）。再例如，若用户期望目标文本中可以包含与“销量”相关的内容，可以将要素标签设置为“销量”。

在本实施例的一些可选的实现方式中，要素标签经由如下步骤生成：
30 对原始文本聚类，得到聚类标签集；从聚类标签集中确定出目标

聚类标签；将目标聚类标签确定为要素标签。

在本实现方式中，聚类标签可以表征原始文本的内容特征。执行主体可以采用聚类算法从原始文本中提取出聚类标签，然后从中选取要素标签，如此，可以确保要素标签与原始文本内容的相关性。

- 5 作为示例，执行主体可以将原始文本输入预先训练的文本聚类模型，例如可以是 TF-IDF，word2vec 等，得到原始文本的聚类标签集。之后，执行主体可以将聚类标签集呈现给用户，由用户从中选取目标聚类标签，从而将用户选取的聚类标签确定为要素标签。

- 10 步骤 202，基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量。

- 在本实施例，编码器用于将原始文本转化成预设维度的编码文本特征向量，以及将原始文本中的每个句子转化成编码句特征向量。其中，编码文本特征向量包括原始文本中所有词的词向量，用于表征原始文本的内容特征。编码句特征向量包括该句子中所有词的词向量，
15 用于表征该句子的内容特征。词向量可以表征词语在预设的词汇表中的位置。

- 编码器可以采用多种类型的深度学习模型，例如可以是 CNN（Convolutional Neural Networks，卷积神经网络）、RNN（Recurrent Neural Network，循环神经网络）或 GRU（Gated Recurrent Unit，门控
20 循环单元）等。

作为示例，执行主体可以采用 LSTM（Long Short-Term Memory，长短期记忆）网络作为编码器。执行主体可以将原始文本输入编码器中，得到预设维度的编码文本特征向量，然后将原始文本中各个句子分别输入编码器中，得到每个句子的编码句特征向量。

- 25 在本实施例的一些可选的实现方式中，可以采用如下步骤得到编码文本特征向量以及原始文本中每个句子的编码句特征向量：基于原始文本，生成文本特征矩阵；将文本特征矩阵输入预训练的第一编码器，得到编码文本特征向量；将文本特征矩阵输入预训练的第二编码器，得到原始文本中各个词的编码词特征向量；基于原始文本中的每个
30 句子中的各个词的编码词特征向量，确定原始文本中各个句子的句

特征向量；将原始文本中每个句子的句特征向量输入预训练的第三编码器，得到原始文本中各个句子的编码句特征向量。

在本实现方式中，执行主体可以采用嵌入矩阵将原始文本转化为文本特征矩阵，嵌入矩阵表征词与词特征值的对应关系，词特征值表征词在词汇表中的位置。之后，执行主体通过第一编码器得到编码文本特征向量，通过第二编码器和第三编码器得到原始文本中每个句子的编码句特征向量。

作为示例，第一编码器可以是 LSTM 编码器，第二编码器和第三编码器可以是 GRU 编码器。执行主体将文本特征矩阵分别输入第一编码器，将文本特征矩阵编码成第一预设维度的编码文本特征向量。将文本特征矩阵输入第二编码器，将文本特征矩阵中的各个词向量编码成第二预设维度的编码词向量，之后执行主体将每个句子中的各个词的编码词向量的均值作为该句子的句特征向量，并将句特征向量输入第三编码器，得到原始文本中各个句子的编码句特征向量。再例如，执行主体还可以将每个句子中的各个词的编码词向量中的最大值作为该句子的句特征向量，还可以将每个句子中的排在句尾的词的编码词向量作为句子的句特征向量。

步骤 203，采用预训练的解码器，执行解码步骤：步骤 2031 至步骤 2034。

在本实施例，解码器用于从编码器输出的编码文本特征向量中估计出目标词。解码器可以采用与编码器同样的深度学习模型。

步骤 2031，基于解码器在当前时刻的隐状态向量和编码文本特征向量，确定原始文本中各个词在当前时刻的词注意力权重。

通常，解码器每进行一步解码，都会输出当前时刻的隐状态向量，隐状态向量用于表征上一步解码中解码器的各个隐藏层的状态参数。

在本实施例，由当前时刻的隐状态向量与编码文本特征向量确定出的词注意力权重，可以表征解码器在当前时刻选中该词语的概率。

作为示例，执行主体将步骤 202 中得到的编码文本向量 B 输入预先训练的 LSTM 解码器中，由 LSTM 解码器逐步对编码文本特征向量进行解码。假设 LSTM 完成第 n 步解码，当前时刻的隐状态向量为对

应于第 n 步解码的隐状态向量 A_n ，执行主体基于隐状态向量 A_n 与编码文本特征向量 B 计算得到当前时刻各个词的词注意力权重 $\{C_i^n\}$ ，其中， C 表示词注意力权重， i 表示原始文本中的第 i 个词。之后，LSTM 完成第 $n+1$ 步解码，则当前时刻的隐状态向量为对应于第 $n+1$ 步解码的隐状态向量 A_{n+1} ，执行主体基于 A_{n+1} 与编码文本特征向量 B 计算得到当前时刻各个词的词注意力权重 $\{C_i^{n+1}\}$ 。

步骤 2032，基于解码器在当前时刻的隐状态向量、编码句特征向量和要素标签，确定原始文本中各个句子在当前时刻的句注意力权重。

在本实施例中，句注意力权重表征句子的内容特征（例如语法结构、语义等特征）与要素标签的相关程度，相关程度越高则句注意力权重的数值就越高。

在一个具体的示例中，执行主体可以基于预设的要素标签与数值的对应关系列表，确定出要素标签的数值，并转化成向量形式，得到要素标签向量。之后，执行主体可以将隐状态向量与要素标签向量组合成一个组合向量。然后，执行主体计算每个句子的编码句特征向量分别与组合向量的相关度，得到每个句子的句注意力权重。

步骤 2033，基于句注意力权重和词注意力权重，确定当前时刻原始文本中各个词归一化的词注意力权重。

在本实施例中，句注意力权重表征句子的内容特征与要素标签的相关程度，如此一来，归一化后的词注意力权重也可以表征词与要素标签的相关程度。

作为示例，执行主体可以将句注意力权重和词注意力权重的乘积作为归一化后的词注意力权重，如此，归一化后的词注意力权重正相关于词与要素标签的相关程度。

步骤 2034，基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词。

在本实施例中，执行主体上装载的解码器基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，确定出原始文本中各个词的置信度，然后将置信度最高的词确定为当前时刻的目标词。之后，解码器进行下一步解码步骤，输出下一个时刻的目标词，

直至解码器将编码文本特征向量完全解码，可以得到目标词序列。

在本实施例的一些可选的实现方式中，基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；基于上下文特征向量和编码文本特征向量，估计出
5 当前时刻的目标词。

在本实现方式中，上下文特征向量可以表征原始文本中上下文语境对词的置信度的影响，如此，编码器确定当前时刻的目标词的过程中可以引入上下文语境对词的影响，可以进一步提高目标词与原始文本的内容的契合程度。

10 步骤 204，基于解码器在各个时刻输出的目标词，生成目标文本。

在本实施例中，执行主体可以按照时序将解码器输出的各个目标词串联，得到目标文本。由于每个目标词均来自原始文本，且与要素标签相关，因而目标文本可以表征执行主体按照要素标签从原始文本中提取出来的文本。

15 继续参见图 3，图 3 是如 2 所示方法的一个实施例的应用场景示意图。在图 3 中所示的场景 300 中，原始文本为商品信息 303，例如可以包括商品名称以及商品的详细信息，目标文本为商品信息摘要 305。终端设备 301 可以通过网络将商品信息 303 和要素标签 304 发送至服务器 302，服务器 302 上装载有预先训练的文本生成模型，文本生成模型可以是包括了注意力模块的 LSTM 编码器和 LSTM 解码器。
20 服务器 302 将接收到的商品信息和要素标签输入该文本生成模型，执行如下文本生成步骤：由 LSTM 编码器将商品信息编码成第三预设维度的编码文本特征向量，以及，将商品信息中的各个句子编码成第四预设维度的编码句特征向量。之后，由 LSTM 编码器逐步对将编码文本特征向量，输出当前时刻的隐状态向量，然后由注意力模块根据编
25 码文本特征向量和隐状态向量确定商品信息中各个词在当前时刻的词注意力权重，并根据编码句特征向量、要素标签和隐状态向量确定出商品信息中各个句子的句注意力权重，以及将词的词注意力权重与所有包括该词的句子的句注意力权重的乘积作为归一化后的词注意力权重，
30 得到原始文本中每个词在当前时刻的归一化后的词注意力权重。

再之后，由编码器根据编码文本特征向量和原始文本中每个词在当前时刻的归一化后的词注意力权重，估计出当前时刻的目标词，并按照时序将各个时刻的目标词串联成目标文本，即可得到商品信息的摘要，且摘要中的内容与要素标签的相关程度较高。

- 5 本公开的实施例提供的用于生成文本的方法和装置，通过编码器对原始文本进行压缩，通过解码器确定出各个词的词注意力权重，以及基于原始文本中各个句子和要素标签的相关性，确定出各个句子的句注意力权重，进而基于词注意力权重和句注意力权重确定出归一化后的词注意力权重，并基于归一化后的词注意力权重估计出目标词，
- 10 最终基于目标词生成目标文本，使得目标文本中的句子和词与要素标签的相关性较高，提高了文本生成的准确度和可控程度。

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

步骤 401，接收原始文本。

- 15 步骤 402，将预设的标点符号作为切分边界，对原始文本进行切分，得到短句集。

在本实施例中，执行主体可以将句号、感叹号或问号作为切分边界，将原始文本切分成若干短句，得到短句集。如此，可以保证每一个短句在语义上的完整性。

- 20 步骤 403，将短句集中不符合预设条件的短句删除，得到更新后的短句集。

- 作为示例，可以采用如下预设条件：短句中至少包括一个逗号且短句的长度处于预设长度区间内。执行主体基于该预设条件将步骤 402 中得到的短句进行筛选，将不符合条件的短句删除，如此，可以降低噪声数据的干扰。
- 25

步骤 404，基于更新后的短句集，采用预训练的特征提取模型，提取每个短句中各个词的词特征，并基于词特征，确定出每个短句的句特征。

- 在本实施例中，短句的句特征表征短句的内容特征，例如可以包括短句的语法结构特征和语义特征。
- 30

特征提取模型可以采用 TF-IDF 或 word2vec 等现有的或未来可能出现的模型，本申请对此不作限定。

作为示例，执行主体可以采用预训练的 BERT (Bidirectional Encoder Representation from Transformers, 来自变压器的双向编码器表示) 模型，从每个短句中提取词特征，将每个词特征进行加权平均后作为该短句的句特征

步骤 405，基于更新后的短句集中每个短句的句特征，采用预训练的聚类模型，确定出更新后的短句集中每个短句的聚类标签，得到聚类标签集。

10 作为示例，执行主体可以采用 K-means (K 均值聚类算法) 模型对短句的句特征进行聚类，得到每个短句的聚类标签，进而得到整个更新后的短句集对应的聚类标签集。

聚类模型还可以是变分自编码器或 DBSCAN (Density-Based Spatial Clustering of Applications with Noise, 基于密度的聚类算法) 等
15 现有的或未来可能出现的聚类模型。

步骤 406，从聚类标签集中确定出目标聚类标签。

步骤 407，将目标聚类标签确定为要素标签。

在本实施例中，聚类标签可以表征原始文本的内容特征，要素标签表征用户对目标文本的内容期望，从聚类标签集中确定出的要素标签可以兼顾原始文本的内容特征与用户的内容期望。
20

步骤 408，基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量。此步骤与前述步骤 202 相对应，此处不赘述。

步骤 409，采用预训练的解码器，执行如下解码步骤。此步骤与前述步骤 203 相对应，此处不再赘述。
25

步骤 410，基于解码器在各个时刻输出的目标词，生成目标文本。此步骤与前述步骤 204 相对应，此处不在赘述。

从图 4 中可以看出，本实施例中的用于生成文本的方法的流程 400 体现了基于聚类算法提取原始文本中的短句的聚类标签，并从聚类标签中确定出要素标签的步骤，可以提高要素标签与原始文本的相关性，
30

进而使得生成的目标文本中的句子和词语与要素标签的相关性更高，如此，可以进一步提高文本生成的准确度和可控程度。

进一步参考图 5，作为对上述各图所示方法的实现，本公开提供了一种用于生成文本的装置 500 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文本的装置 500 包括：接收单元 501，被配置成接收原始文本和要素标签；编码单元 502，被配置成基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量；解码单元 503，被配置成采用预训练的解码器，执行如下解码步骤：基于解码器在当前时刻的隐状态向量和编码文本特征向量，确定原始文本中各个词在当前时刻的词注意力权重；基于解码器在当前时刻的隐状态向量、编码句特征向量和要素标签，确定原始文本中各个句子在当前时刻的句注意力权重；基于句注意力权重和词注意力权重，确定当前时刻原始文本中各个词归一化的词注意力权重；基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词；生成单元 504，被配置成基于解码器在各个时刻输出的目标词，生成目标文本。

在本实施例中，编码单元 502 进一步包括：特征矩阵生成模块，被配置成基于原始文本，生成文本特征矩阵；第一编码模块，被配置成将文本特征矩阵输入预训练的第一编码器，得到编码文本特征向量；第二编码模块，被配置成将文本特征矩阵输入预训练的第二编码器，得到原始文本中各个词的编码词特征向量；句特征向量生成模块，被配置成基于原始文本中的每个句子中的各个词的编码词特征向量，确定原始文本中各个句子的句特征向量；第三编码模块，被配置成将原始文本中每个句子的句特征向量输入预训练的第三编码器，得到原始文本中各个句子的编码句特征向量。

在本实施例中，解码单元 503 进一步包括：上下文特征向量生成模块，被配置成基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；目标词输出模块，被配置成基于上下文特征向量和编码文本特征向量，估计出

当前时刻的目标词。

在本实施例中,该装置 500 还包括要素标签生成单元,被配置成:对原始文本聚类,得到聚类标签集;从聚类标签集中确定出目标聚类标签;将目标聚类标签确定为要素标签。

- 5 在本实施例中,要素标签生成单元进一步包括:切分模块,被配置成将预设的标点符号作为切分边界,对原始文本进行切分,得到短句集;筛选模块,被配置成将短句集中不符合预设条件的短句删除,得到更新后的短句集;特征提取模块,被配置成基于更新后的短句集,采用预训练的特征提取模型,提取每个短句中各个词的词特征,并基于词特征,确定出每个短句的句特征;标签生成模块,被配置成基于更新后的短句集中每个短句的句特征,采用预训练的聚类模型,确定出更新后的短句集中每个短句的聚类标签,得到聚类标签集。
- 10

- 下面参考图 6,其示出了适于用来实现本公开的实施例的电子设
备(例如图 1 中的服务器或终端设备)600 的结构示意图。本公开的
15 实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、
数字广播接收器、PDA(个人数字助理)、PAD(平板电脑)等等的移
动终端以及诸如数字 TV、台式计算机等等的固定终端。图 6 示出的终
端设备仅仅是一个示例,不应对本公开的实施例的功能和使用范围带
来任何限制。

- 20 如图 6 所示,电子设备 600 可以包括处理装置(例如中央处理器、
图形处理器等)601,其可以根据存储在只读存储器(ROM)602 中的
程序或者从存储装置 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,还存储有电子设备
600 操作所需的各种程序和数据。处理装置 601、ROM 602 以及 RAM
25 603 通过总线 604 彼此相连。输入/输出(I/O)接口 605 也连接至总线
604。

- 通常,以下装置可以连接至 I/O 接口 605:包括例如触摸屏、触摸
板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置
606;包括例如液晶显示器(LCD)、扬声器、振动器等的输出装置 607;
30 包括例如磁带、硬盘等的存储装置 608;以及通信装置 609。通信装置

609 可以允许电子设备 600 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 600, 但是应理解的是, 并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图 6 中示出的每个方框可以代表一个装置, 也可以根据
5 需要代表多个装置。

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,
10 该计算机程序可以通过通信装置 609 从网络上被下载和安装, 或者从存储装置 608 被安装, 或者从 ROM 602 被安装。在该计算机程序被处理装置 601 执行时, 执行本公开的实施例的方法中限定的上述功能。需要说明的是, 本公开的实施例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存储介质或者是上述两者的任意组合。
15 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)、
20 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。在本公开的实施例中, 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。而在本公开的实施例中, 计算机可读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, 该计算机可读信号介质可以发送、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
30

输，包括但不限于：电线、光缆、RF（射频）等等，或者上述的任意合适的组合。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。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电子设备执行时，使得该电子设备：接收原始文本和要素标签；基于原始文本，采用预训练的编码器生成编码文本特征向量以及原始文本中每个句子的编码句特征向量；采用预训练的解码器，执行如下解码步骤：基于解码器在当前时刻的隐状态向量和编码文本特征向量，确定原始文本中各个词在当前时刻的词注意力权重；基于解码器在当前时刻的隐状态向量、编码句特征向量和要素标签，确定原始文本中各个句子在当前时刻的句注意力权重；基于句注意力权重和词注意力权重，确定当前时刻原始文本中各个词归一化的词注意力权重；基于编码文本特征向量和当前时刻原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词；该方法还包括：基于解码器在各个时刻输出的目标词，生成目标文本。

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公开的实施例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”语言或类似的程序设计语言。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）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、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。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，

- 5 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，或者可以用专用硬件与计算机指令的组合来实现。

- 描述于本公开的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
10 器中，例如，可以描述为：一种处理器包括接收单元、编码单元、解码单元和生成单元。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，例如，接收单元还可以被描述为“接收原始文本和要素标签的单元”。

- 以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。
15 本领域技术人员应当理解，本公开的实施例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。例如上述特征与本公开的实施例中公开的（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
20 方案。

权 利 要 求

1、一种用于生成文本的方法，其中，包括：

接收原始文本和要素标签；

5 基于所述原始文本，采用预训练的编码器生成编码文本特征向量以及所述原始文本中每个句子的编码句特征向量；

采用预训练的解码器，执行如下解码步骤：

 基于所述解码器在当前时刻的隐状态向量和所述编码文本特征向量，确定所述原始文本中各个词在当前时刻的词注意力权重；基于所述解码器在当前时刻的隐状态向量、所述编码句特征向量和所述要素
10 标签，确定所述原始文本中各个句子在当前时刻的句注意力权重；基于所述句注意力权重和所述词注意力权重，确定当前时刻所述原始文本中各个词归一化的词注意力权重；基于所述编码文本特征向量和当前时刻所述原始文本中各个词归一化的词注意力权重，估计出当前时
15 刻的目标词；

 所述方法还包括：基于所述解码器在各个时刻输出的目标词，生成目标文本。

2、根据权利要求1所述的方法，其中，基于所述原始文本，采用
20 预训练的编码器生成编码文本特征向量以及所述原始文本中每个句子的编码句特征向量，包括：

 基于所述原始文本，生成文本特征矩阵；

 将所述文本特征矩阵输入预训练的第一编码器，得到所述编码文本特征向量；

25 将所述文本特征矩阵输入预训练的第二编码器，得到所述原始文本中各个词的编码词特征向量；

 基于所述原始文本中的每个句子中的各个词的编码词特征向量，确定所述原始文本中各个句子的句特征向量；

 将所述原始文本中每个句子的句特征向量输入预训练的第三编
30 码器，得到所述原始文本中各个句子的编码句特征向量。

3、根据权利要求1所述的方法，其中，基于所述编码文本特征向量和当前时刻所述原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词，包括：

- 5 基于所述编码文本特征向量和当前时刻所述原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；

 基于所述上下文特征向量和所述编码文本特征向量，估计出当前时刻的目标词。

- 10 4、根据权利要求1至3之一所述的方法，其中，所述要素标签经由如下步骤生成：

 对所述原始文本聚类，得到聚类标签集；

 从所述聚类标签集中确定出目标聚类标签；

 将所述目标聚类标签确定为所述要素标签。

15

- 5、根据权利要求4所述的方法，其中，对所述原始文本聚类，得到聚类标签集，包括：

 将预设的标点符号作为切分边界，对所述原始文本进行切分，得到短句集；

- 20 将所述短句集中不符合预设条件的短句删除，得到更新后的短句集；

 基于所述更新后的短句集，采用预训练的特征提取模型，提取每个短句中各个词的词特征，并基于所述词特征，确定出每个短句的句特征；

- 25 基于所述更新后的短句集中每个短句的句特征，采用预训练的聚类模型，确定出所述更新后的短句集中每个短句的聚类标签，得到所述聚类标签集。

6、一种用于生成文本的装置，其中，包括：

- 30 接收单元，被配置成接收原始文本和要素标签；

编码单元，被配置成基于所述原始文本，采用预训练的编码器生成编码文本特征向量以及所述原始文本中每个句子的编码句特征向量；

解码单元，被配置成采用预训练的解码器，执行如下解码步骤：
基于所述解码器在当前时刻的隐状态向量和所述编码文本特征向量，
5 确定所述原始文本中各个词在当前时刻的词注意力权重；基于所述解码器在当前时刻的隐状态向量、所述编码句特征向量和所述要素标签，
确定所述原始文本中各个句子在当前时刻的句注意力权重；基于所述句注意力权重和所述词注意力权重，确定当前时刻所述原始文本中各个词归一化的词注意力权重；基于所述编码文本特征向量和当前时刻
10 所述原始文本中各个词归一化的词注意力权重，估计出当前时刻的目标词；

生成单元，被配置成基于所述解码器在各个时刻输出的目标词，生成目标文本。

15 7、根据权利要求 6 所述的装置，其中，所述编码单元进一步包括：
特征矩阵生成模块，被配置成基于所述原始文本，生成文本特征矩阵；

第一编码模块，被配置成将所述文本特征矩阵输入预训练的第一编码器，得到所述编码文本特征向量；

20 第二编码模块，被配置成将所述文本特征矩阵输入预训练的第二编码器，得到所述原始文本中各个词的编码词特征向量；

句特征向量生成模块，被配置成基于所述原始文本中的每个句子中的各个词的编码词特征向量，确定所述原始文本中各个句子的句特征向量；

25 第三编码模块，被配置成将所述原始文本中每个句子的句特征向量输入预训练的第三编码器，得到所述原始文本中各个句子的编码句特征向量。

8、根据权利要求 6 所述的装置，其中，所述解码单元进一步包括：
30 上下文特征向量生成模块，被配置成基于所述编码文本特征向量

和当前时刻所述原始文本中各个词归一化的词注意力权重，生成当前时刻的上下文特征向量；

目标词输出模块，被配置成基于所述上下文特征向量和所述编码文本特征向量，估计出当前时刻的目标词。

5

9、根据权利要求 6 至 8 之一所述的装置，其中，所述装置还包括要素标签生成单元，被配置成：

对所述原始文本聚类，得到聚类标签集；

从所述聚类标签集中确定出目标聚类标签；

10 将所述目标聚类标签确定为所述要素标签。

10、根据权利要求 9 所述的装置，其中，要素标签生成单元进一步包括：

15 切分模块，被配置成将预设的标点符号作为切分边界，对所述原始文本进行切分，得到短句集；

筛选模块，被配置成将所述短句集中不符合预设条件的短句删除，得到更新后的短句集；

20 特征提取模块，被配置成基于所述更新后的短句集，采用预训练的特征提取模型，提取每个短句中各个词的词特征，并基于所述词特征，确定出每个短句的句特征；

标签生成模块，被配置成基于所述更新后的短句集中每个短句的句特征，采用预训练的聚类模型，确定出所述更新后的短句集中每个短句的聚类标签，得到所述聚类标签集。

25 11、一种电子设备，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 1-5 中任一所述的方法。

30

12、一种计算机可读介质，其上存储有计算机程序，其中，所述程序被处理器执行时实现如权利要求 1-5 中任一所述的方法。

13. 一种计算机程序产品，包括计算机程序，所述计算机程序在被
5 处理器执行时实现权利要求 1-5 中任一项所述的方法。