

发明名称：文摘自动提取方法、装置、计算机设备及存储介质

本申请要求于2018年3月8日提交中国专利局、申请号为201810191506.3、申请名称为“文摘自动提取方法、装置、计算机设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及文摘提取技术领域，尤其涉及一种文摘自动提取方法、装置、计算机设备及存储介质。

背景技术

目前，对文章概括文摘时，采用的是基于抽取式的方法。抽取式文摘是提取文章中最有代表性的关键句作为该文章的文摘。具体如下：

- 1) 首先，对文章进行分词，去停用词，获得的组成文章的基本词组。
- 2) 然后，根据计算词频获取高频词，并把高频词所在的句子作为关键句。
- 3) 最后，指定若干数量的关键句即可组合成文摘。

上述抽取式方法比较适用于新闻、议论文等在文中往往出现总结性长句子的文体。例如财经文章，高频词往往是“现金”、“股票”、“央行”、“利息”等，抽取结果就往往是“央行加息导致股价下跌，现金为上已成股民众识”之类的长句子。抽取式方法有很大的局限性，如果处理的文本中缺失代表性的“关键句”，那抽取结果很可能毫无意义，尤其是对话类的文本。

发明内容

本申请提供了一种文摘自动提取方法、装置、计算机设备及存储介质，旨在解决现有技术中采用抽取式方法提取文章中的文摘仅适用于新闻、议论文等在文中出现总结性长句子的文体，对无关键句的文本提取摘要提取结果不准确的问题。

第一方面，本申请提供了一种文摘自动提取方法，其包括：依序获取目标文本所包括的字符，将字符按顺序输入至LSTM模型中的第一层LSTM结构进

行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短记忆神经网络；将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列；将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列；根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的概率分布，将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

第二方面，本申请提供了一种文摘自动提取装置，其包括：

第一输入单元，用于依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短记忆神经网络；

第二输入单元，用于将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列；

第三输入单元，用于将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列；

上下文向量获取单元，用于根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；

摘要获取单元，用于根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的概率分布，将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

第三方面，本申请又提供了一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现本申请提供的任一项所述的文摘自动提取方法。

第四方面，本申请还提供了一种存储介质，其中所述存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行本申请提供的任一项所述的文摘自动提取方法。

本申请提供一种文摘自动提取方法、装置、计算机设备及存储介质。该方法采用 LSTM 模型对目标文本进行编码和解码后，并结合上下文变量，得到目标文本的摘要，采取了概括的方式来总结获取目标文本的摘要，提高了文摘获

取的准确性。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，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。

图1 为本申请实施例提供的一种文摘自动提取方法的示意图；

图2 为本申请实施例提供的一种文摘自动提取方法的另一示意图；

图3 为本申请实施例提供的一种文摘自动提取方法的子流程示意图；

图4 为本申请实施例提供的一种文摘自动提取装置的示意性框图；

图5 为本申请实施例提供的一种文摘自动提取装置的另一示意性框图；

图6 为本申请实施例提供的一种文摘自动提取装置的子单元示意性框图；

图7 为本申请实施例提供的一种计算机设备的示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

还应当理解，在此本申请说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本申请。如在本申请说明书和所附权利要求书中所使用的那样，除非上下文清楚地指明其它情况，否则单数形式的“一”、“一个”及“该”意在包括复数形式。

还应当进一步理解，在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

请参阅图 1, 图 1 是本申请实施例提供的一种文摘自动提取方法的示意图。该方法应用于台式电脑、手提电脑、平板电脑等终端中。如图 1 所示, 该方法包括步骤 S101~S105。

S101、依序获取目标文本所包括的字符, 将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得到隐含状态组成的序列; 其中 LSTM 模型为长短记忆神经网络。

在本实施例中, 先是通过分词来获取目标文本所包括的字符, 所获取的字符为中文字符或英文字符, 经过上述处理后将目标文本拆分成了多个字符。例如, 对一篇中文文章进行分词时, 采用如下步骤:

1)、对一个待分词的子串 S, 按照从左到右的顺序取出全部候选词 $w_1, w_2, \dots, w_i, \dots, w_n$;

2)、到词典中查出每个候选词的概率值 $P(w_i)$, 并记录每个候选词的全部左邻词;

3)、计算每个候选词的累计概率, 同时比较得到每个候选词的最佳左邻词;

4)、如果当前词 w_n 是字串 S 的尾词, 且累计概率 $P(w_n)$ 最大, 则 w_n 就是 S 的终点词;

5)、从 w_n 开始, 按照从右到左顺序, 依次将每个词的最佳左邻词输出, 即 S 的分词结果。

依序获取了目标文本所包括的字符后, 将其按顺序输入至已根据历史数据训练得到的 LSTM 模型, 就能从多个分词中提炼出能构成摘要的词语组成最终的文摘。具体处理时, 可以是以自然段为单位进行上述分词处理, 提取当前自然段的关键句, 最后将每段的关键句组合形成摘要(本申请中优选这一分词处理方式)。也可以是直接以一整篇文章为单位进行上述分词处理, 提取多个关键词后组合成摘要。

在获取了目标文本所包括的字符后, 输入 LSTM 模型进行处理。LSTM 模型即长短记忆神经网络, 其中 LSTM 的全称是 Long Short-Term Memory, 是一种时间递归神经网络, LSTM 适合于处理和预测时间序列中间隔和延迟非常长的重要事件。通过 LSTM 模型能目标文本所包括的字符进行编码, 进行文本的摘要提取的前序处理。

为了更清楚的理解 LSTM 模型, 下面对 LSTM 模型进行介绍。

LSTM 的关键是元胞状态(Cell State), 其可以视为横穿整个元胞顶部的水平线。元胞状态类似于传送带, 它直接穿过整个链, 同时只有一些较小的线性交互。元胞状态上承载的信息可以很容易地流过而不改变, LSTM 有能力对元胞状态添加或者删除信息, 上述能力通过门的结构来控制, 即门可以选择性让信息通过, 其中门结构是由一个 Sigmoid 神经网络层和一个元素级相乘操作组成。Sigmoid 层输出 0~1 之间的值, 每个值表示对应的部分信息是否应该通过。0 值表示不允许信息通过, 1 值表示让所有信息通过。一个 LSTM 有 3 个门, 来保护和控制元胞状态。

LSTM 中至少包括三个门, 分别如下:

- 1) 遗忘门, 其决定了上一时刻的单元状态有多少保留到当前时刻;
- 2) 输入门, 其决定了当前时刻网络的输入有多少保存到单元状态;
- 3) 输入门, 其决定了单元状态有多少输出到 LSTM 的当前输出值。

在一实施例中, 所述 LSTM 模型为门限循环单元, 所述门限循环单元的模型如下:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned}$$

其中, W_z 、 W_r 、 W 是训练得到的权值参数值, x_t 是输入, h_{t-1} 是隐含状态, z_t 是更新状态, r_t 是重置信号, $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆, h_t 是输出, $\sigma()$ 是 sigmoid 函数, $\tanh()$ 是双曲正切函数。

目标文本所包括的字符通过了第一层 LSTM 结构进行编码, 就转化成隐含状态组成的序列, 对其继续进行解码就能获取初次处理后的序列, 实现了对待选分词的精准提取。

在一实施例中, 如图 2 所示, 所述步骤 S101 之前还包括:

S101a、将语料库中的多篇历史文本置入第一层 LSTM 结构, 并将历史文本对应的文摘置入第二层 LSTM 结构, 进行训练得到 LSTM 模型。

LSTM 模型的整体框架是固定的, 只需要设置其输入层、隐藏层、输出层等各层的参数, 就可以得到模型, 其中设置输入层、隐藏层、输出层等各层的参

数可以通过实验多次来得到最优的参数值。譬如，隐藏层节点有 10 个节点，那每个节点的数值可以从 1 取到 10，那么就会尝试 100 种组合来得到 100 个训练模型，然后用大量数据去训练这 100 个模型，根据准确率等来得到一个最优的训练模型，这个最优的训练模型对应的节点值等参数就是最优参数（可以理解为上述 GRU 模型中的 W_z 、 W_r 、 W 就为此处的最优参数）。用最优的训练模型来应用到本方案中作为 LSTM 模型，这样能确保所提取的文摘更为准确。

S102、将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列。

如图 3 所示，该步骤 S102 包括以下子步骤：

S1021、获取隐含状态组成的序列中概率最大的词，将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语；

S1022、将初始位词语中的每个字输入至第二层 LSTM 结构，与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列；

S1023、重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构，与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤，直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，并将隐含状态组成的序列作为摘要的字词序列。

在本实施例中，上述过程也即 Beam Search 算法（Beam Search 算法即集束搜索算法），是用于解码隐含状态组成的序列的方法之一，其具体过程如下：

1) 获取隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语；2) 将初始位词语中的每个字与词表中的字进行组合得到第一次组合后序列，获取第一次组合后序列中概率最大的词作第一次更新后序列；重复上述过程直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，最终输出摘要的字词序列。

Beam Search 算法只在实际使用过程中（即 test 过程中）的时候需要，在训练过程中并不需要。训练的时候由于知道正确答案，并不需要再进行这个搜索。而在实际使用的时候，假设词表大小为 3，内容为 a, b, c。beam search 算法最终输出序列个数（可用 size 表示最终输出序列个数）是 2，decoder（第二层 LSTM

结构可以视为解码器 decoder) 解码的时候:

生成第 1 个词的时候, 选择概率最大的 2 个词, 假设为 a,c,那么当前序列就是 a c; 生成第 2 个词的时候, 我们将当前序列 a 和 c, 分别与词表中的所有词进行组合, 得到新的 6 个序列 aa、ab、ac、ca、cb、cc, 然后从其中选择 2 个得分最高的作为当前序列, 假如为 aa cb; 后面会不断重复这个过程, 直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止, 最终输出 2 个得分最高的序列。将目标文本经过编码和解码后输出摘要的字词序列, 此时还未组成一段完整的摘要文字。为了将摘要的字词序列组成一段完整的摘要, 需要进行进一步的处理。

在一实施例中, 将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列中, 所述摘要的字词序列为与词表大小相同的多项式分布层, 并输出向量 $y^t \in R^K$; 其中 y^t 中的第 k 维代表生成第 k 个词语的概率, t 的取值为正整数, K 为历史文本所对应词表的大小。

其中, 将目标文本 x^t 设置结束标志 (如文本末尾的句号), 每次将目标文本中的一个词输入到第一层 LSTM 结构, 当到达目标文本 x^t 的末尾时, 则表示目标文本 x^t 编码得到的隐含状态组成的序列 (即 hidden state vector) 将作为第二层 LSTM 结构的输入进行解码, 第二层 LSTM 结构输出与词表大小相同的 softmax 层 (softmax 层即多项式分布层), softmax 层中的分量代表每个词语的概率; 当 LSTM 的输出层为 softmax 时, 每个时刻输出会产生向量 $y^t \in R^K$, K 即为词表的大小, y^t 向量中的第 k 维代表生成第 k 个词语的概率。通过向量来表示摘要的字词序列中每一词语的概率, 更利于其作为下一次数据处理的输入的参考。

S103、将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得到更新后隐含状态组成的序列。

在本实施例中, 将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 是为了二次进行处理, 以从摘要的字词序列选取最有可能的字词作为摘要的组成词。

S104、根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值, 获取与编码器隐藏状态的贡献值相对应的上下文向量。

在本实施例中, 编码器隐藏状态的贡献值代表了他的所有隐藏状态的加权

和，其中最高的权重对应了解码器在决定下一个词是考虑的增强隐藏状态的最大贡献以及最重要的隐藏状态。通过这一方式，能更准确的获取能代表文摘的上下文向量。

例如，将更新后隐含状态组成的序列转化为特征向量 a ，其中 $a=\{a_1, a_2, \dots, a_L\}$ ，则上下文向量 Z_t 用下式表示：

$$Z_t = \sum_{i=1}^L \alpha_{t,i} \alpha_i$$

其中， $\alpha_{t,i}$ 就是衡量生成第 t 个词语时，第 i 个位置的特征向量所占的权重， L 为更新后隐含状态组成的序列中字符的个数。

S105、根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的的概率分布，将字词的的概率分布中概率最大的字词输出作为目标文本的摘要。

在本实施例中，对目标文本的每一段文字进行处理，每一段都通过上述步骤来概括摘要，最后组合成一个完成的摘要。

可见，该方法采用 LSTM 对目标文本进行编码解码后，结合上下文变量得到目标文本的摘要，采取概括方式获取摘要，提高获取准确性。

本申请实施例还提供一种文摘自动提取装置，该文摘自动提取装置用于执行前述任一项文摘自动提取方法。具体地，请参阅图 4，图 4 是本申请实施例提供的一种文摘自动提取装置的示意性框图。文摘自动提取装置 100 可以安装于台式电脑、平板电脑、手提电脑、等终端中。

如图 4 所示，文摘自动提取装置 100 包括第一输入单元 101、第二输入单元 102、第三输入单元 103、上下文向量获取单元 104、摘要获取单元 105。

第一输入单元 101，用于依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短记忆神经网络。

在本实施例中，先是通过分词来获取目标文本所包括的字符，所获取的字符为中文字符或英文字符，经过上述处理后将目标文本拆分成了多个字符。例如，对一篇中文文章进行分词时，采用如下步骤：

1)、对一个待分词的子串 S ，按照从左到右的顺序取出全部候选词 $w_1, w_2, \dots, w_i, \dots, w_n$ ；2)、到词典中查出每个候选词的概率值 $P(w_i)$ ，并记录每个

候选词的全部左邻词；3）、计算每个候选词的累计概率，同时比较得到每个候选词的最佳左邻词；4）、如果当前词 w_n 是字串 S 的尾词，且累计概率 $P(w_n)$ 最大，则 w_n 就是 S 的终点词；5）、从 w_n 开始，按照从右到左顺序，依次将每个词的最佳左邻词输出，即 S 的分词结果。

依序获取了目标文本所包括的字符后，将其按顺序输入至已根据历史数据训练得到的 LSTM 模型，就能从多个分词中提炼出能构成摘要的词语组成最终的文摘。具体处理时，可以是以自然段为单位进行上述分词处理，提取当前自然段的关键句，最后将每段的关键句组合形成摘要（本申请中优选这一分词处理方式）。也可以是直接以一整篇文章为单位进行上述分词处理，提取多个关键词后组合成摘要。

在获取了目标文本所包括的字符后，输入 LSTM 模型进行处理。LSTM 模型即长短记忆神经网络，其中 LSTM 的全称是 Long Short-Term Memory，是一种时间递归神经网络，LSTM 适合于处理和预测时间序列中间隔和延迟非常长的重要事件。通过 LSTM 模型能目标文本所包括的字符进行编码，进行文本的摘要提取的前序处理。

为了更清楚的理解 LSTM 模型，下面对 LSTM 模型进行介绍。

LSTM 的关键是元胞状态(Cell State)，其可以视为横穿整个元胞顶部的水平线。元胞状态类似于传送带，它直接穿过整个链，同时只有一些较小的线性交互。元胞状态上承载的信息可以很容易地流过而不改变，LSTM 有能力对元胞状态添加或者删除信息，上述能力通过门的结构来控制，即门可以选择性让信息通过，其中门结构是由一个 Sigmoid 神经网络层和一个元素级相乘操作组成。Sigmoid 层输出 0~1 之间的值，每个值表示对应的部分信息是否应该通过。0 值表示不允许信息通过，1 值表示让所有信息通过。一个 LSTM 有 3 个门，来保护和控制元胞状态。

LSTM 中至少包括三个门，分别如下：

1) 遗忘门，其决定了上一时刻的单元状态有多少保留到当前时刻；2) 输入门，其决定了当前时刻网络的输入有多少保存到单元状态；3) 输出门，其决定了单元状态有多少输出到 LSTM 的当前输出值。

在一实施例中，所述 LSTM 模型为门限循环单元，所述门限循环单元的模型如下：

$$\begin{aligned}
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
\tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}$$

其中, W_z 、 W_r 、 W 是训练得到的权值参数值, x_t 是输入, h_{t-1} 是隐含状态, z_t 是更新状态, r_t 是重置信号, $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆, h_t 是输出, $\sigma()$ 是 sigmoid 函数, $\tanh()$ 是双曲正切函数。

目标文本所包括的字符通过了第一层 LSTM 结构进行编码, 就转化成隐含状态组成的序列, 对其继续进行解码就能获取初次处理后的序列, 实现了对待选分词的精准提取。

在一实施例中, 如图 5 所示, 所述文摘自动提取装置 100 还包括:

历史数据训练单元 101a、将语料库中的多篇历史文本置入第一层 LSTM 结构, 并将历史文本对应的文摘置入第二层 LSTM 结构, 进行训练得到 LSTM 模型。

LSTM 模型的整体框架是固定的, 只需要设置其输入层、隐藏层、输出层等各层的参数, 就可以得到模型, 其中设置输入层、隐藏层、输出层等各层的参数可以通过实验多次来得到最优的参数值。譬如, 隐藏层节点有 10 个节点, 那每个节点的数值可以从 1 取到 10, 那么就会尝试 100 种组合来得到 100 个训练模型, 然后用大量数据去训练这 100 个模型, 根据准确率等来得到一个最优的训练模型, 这个最优的训练模型对应的节点值等参数就是最优参数 (可以理解在上述 GRU 模型中的 W_z 、 W_r 、 W 就为此处的最优参数)。用最优的训练模型来应用到本方案中作为 LSTM 模型, 这样能确保所提取的文摘更为准确。

第二输入单元 102, 用于将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列。

如图 6 所示, 所述第二输入单元 102 包括以下子单元:

初始化单元 1021, 用于获取隐含状态组成的序列中概率最大的词, 将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语;

更新单元 1022, 用于将初始位词语中的每个字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序

列中概率最大的词作为隐含状态组成的序列；

重复执行单元 1023，用于重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构，与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤，直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，并将隐含状态组成的序列作为摘要的字词序列。

在本实施例中，上述过程也即 Beam Search 算法（Beam Search 算法即集束搜索算法），是用于解码隐含状态组成的序列的方法之一，其具体过程如下：

1) 获取隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语；2) 将初始位词语中的每个字与词表中的字进行组合得到第一次组合后序列，获取第一次组合后序列中概率最大的词作第一次更新后序列；重复上述过程直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，最终输出摘要的字词序列。

Beam Search 算法只在实际使用过程中（即 test 过程中）的时候需要，在训练过程中并不需要。训练的时候由于知道正确答案，并不需要再进行这个搜索。而在实际使用的时候，假设词表大小为 3，内容为 a, b, c。beam search 算法最终输出序列个数（可用 size 表示最终输出序列个数）是 2，decoder（第二层 LSTM 结构可以视为解码器 decoder）解码的时候：

生成第 1 个词的时候，选择概率最大的 2 个词，假设为 a,c,那么当前序列就是 a c；生成第 2 个词的时候，我们将当前序列 a 和 c，分别与词表中的所有词进行组合，得到新的 6 个序列 aa、ab、ac、ca、cb、cc，然后从其中选择 2 个得分最高的作为当前序列，假如为 aa cb；后面会不断重复这个过程，直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，最终输出 2 个得分最高的序列。

将目标文本经过编码和解码后输出摘要的字词序列，此时还未组成一段完整的摘要文字。为了将摘要的字词序列组成一段完整的摘要，需要进行进一步的处理。

在一实施例中，将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列中，所述摘要的字词序列为与词表大小相同的多项式分布层，并输出向量 $y^t \in R^K$ ；其中 y^t 中的第 k 维代表生成第 k 个词

语的概率， t 的取值为正整数， K 为历史文本所对应词表的大小。

其中，将目标文本 x^l 设置结束标志（如文本末尾的句号），每次将目标文本中的一个词输入到第一层 LSTM 结构，当到达目标文本 x^l 的末尾时，则表示目标文本 x^l 编码得到的隐含状态组成的序列（即 hidden state vector）将作为第二层 LSTM 结构的输入进行解码，第二层 LSTM 结构输出与词表大小相同的 softmax 层（softmax 层即多项式分布层），softmax 层中的分量代表每个词语的概率；当 LSTM 的输出层为 softmax 时，每个时刻输出会产生向量 $y^t \in R^K$ ， K 即为词表的大小， y^t 向量中的第 k 维代表生成第 k 个词语的概率。通过向量来表示摘要的字词序列中每一词语的概率，更利于其作为下一次数据处理的输入的参考。

第三输入单元 103，用于将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列。

在本实施例中，将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，是为了二次进行处理，以从摘要的字词序列选取最有可能的字词作为摘要的组成词。

上下文向量获取单元 104，用于根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量。

在本实施例中，编码器隐藏状态的贡献值代表了他的所有隐藏状态的加权和，其中最高的权重对应了解码器在决定下一个词是考虑的增强隐藏状态的最大贡献以及最重要的隐藏状态。通过这一方式，能更准确的获取能代表文摘的上下文向量。

例如，将更新后隐含状态组成的序列转化为特征向量 a ，其中 $a = \{a_1, a_2, \dots, a_L\}$ ，则上下文向量 Z_t 用下式表示：

$$Z_t = \sum_{i=1}^L \alpha_{t,i} a_i$$

其中， $a_{t,i}$ 就是衡量生成第 t 个词语时，第 i 个位置的特征向量所占的权重， L 为更新后隐含状态组成的序列中字符的个数。

摘要获取单元 105，用于根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的的概率分布，将字词的的概率分布中概率最大的字词输出作为目标文本的摘要。

在本实施例中，对目标文本的每一段文字进行处理，每一段都通过上述步骤来概括摘要，最后组合成一个完成的摘要。

可见，该装置采用 LSTM 对目标文本进行编码解码后，结合上下文变量得到目标文本的摘要，采取概括方式获取摘要，提高获取准确性。

上述文摘自动提取装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 7 所示的计算机设备上运行。

请参阅图 7，图 7 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 500 设备可以是终端。该终端可以是平板电脑、笔记本电脑、台式电脑、个人数字助理等电子设备。

参阅图 7，该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存存储器 504。

该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 包括程序指令，该程序指令被执行时，可使得处理器 502 执行一种文摘自动提取方法。该处理器 502 用于提供计算和控制能力，支撑整个计算机设备 500 的运行。该内存存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境，该计算机程序 5032 被处理器 502 执行时，可使得处理器 502 执行一种文摘自动提取方法。该网络接口 505 用于进行网络通信，如发送分配的任务等。本领域技术人员可以理解，图 7 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备 500 的限定，具体的计算机设备 500 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 502 用于运行存储在存储器中的计算机程序 5032，以实现如下功能：依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短记忆神经网络；将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列；将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列；根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；根据更新后隐含状态组成的序列及上下文向

量, 获取更新后隐含状态组成的序列中字词的概率分布, 将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

在一实施例中, 处理器 502 还执行如下操作: 将语料库中的多篇历史文本置入第一层 LSTM 结构, 并将历史文本对应的文摘置入第二层 LSTM 结构, 进行训练得到 LSTM 模型。

在一实施例中, 所述 LSTM 模型为门限循环单元, 所述门限循环单元的模型如下:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned}$$

其中, W_z 、 W_r 、 W 是训练得到的权值参数值, x_t 是输入, h_{t-1} 是隐含状态, z_t 是更新状态, r_t 是重置信号, $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆, h_t 是输出, $\sigma()$ 是 sigmoid 函数, $\tanh()$ 是双曲正切函数。

在一实施例中, 所述摘要的字词序列为与词表大小相同的多项式分布层, 并输出向量 $y^t \in R^K$; 其中 y^t 中的第 k 维代表生成第 k 个词语的概率, t 的取值为正整数, K 为历史文本所对应词表的大小。

在一实施例中, 处理器 502 还执行如下操作: 获取隐含状态组成的序列中概率最大的词, 将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语; 将初始位词语中的每个字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列; 重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤, 直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止, 并将隐含状态组成的序列作为摘要的字词序列。

本领域技术人员可以理解, 图 7 中示出的计算机设备的实施例并不构成对计算机设备具体构成的限定, 在其他实施例中, 计算机设备可以包括比图示更多或更少的部件, 或者组合某些部件, 或者不同的部件布置。例如, 在一些实

施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 7 所示实施例一致，在此不再赘述。

应当理解，在本申请实施例中，处理器 502 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器 502 还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供一种存储介质。该存储介质可以为非易失性的计算机可读存储介质。该存储介质存储有计算机程序，其中计算机程序包括程序指令。该程序指令被处理器执行时实现本申请实施例的文摘自动提取方法。

所述存储介质可以是前述设备的内部存储单元，例如设备的硬盘或内存。所述存储介质也可以是所述设备的外部存储设备，例如所述设备上配备的插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡，闪存卡 (Flash Card) 等。进一步地，所述存储介质还可以既包括所述设备的内部存储单元也包括外部存储设备。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的设备、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1. 一种文摘自动提取方法，其特征在于，包括：

依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短期记忆神经网络；

将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列；

将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列；

根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；

根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的概率分布，将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

2. 根据权利要求 1 所述的文摘自动提取方法，其特征在于，所述依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列之前，还包括：

将语料库中的多篇历史文本置入第一层 LSTM 结构，并将历史文本对应的文摘置入第二层 LSTM 结构，进行训练得到 LSTM 模型。

3. 根据权利要求 1 所述的文摘自动提取方法，其特征在于，所述 LSTM 模型为门限循环单元，所述门限循环单元的模型如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

其中， W_z 、 W_r 、 W 是训练得到的权值参数值， x_t 是输入， h_{t-1} 是隐含状态， z_t 是更新状态， r_t 是重置信号， $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆， h_t 是输出， σ

() 是 sigmoid 函数, $\tanh()$ 是双曲正切函数。

4. 根据权利要求 3 所述的文摘自动提取方法, 其特征在于, 所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列中, 所述摘要的字词序列为与词表大小相同的多项式分布层, 并输出向量 $y^t \in R^K$; 其中 y^t 中的第 k 维代表生成第 k 个词语的概率, t 的取值为正整数, K 为历史文本所对应词表的大小。

5. 根据权利要求 2 所述的文摘自动提取方法, 其特征在于, 所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列, 包括:

获取隐含状态组成的序列中概率最大的词, 将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语;

将初始位词语中的每个字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列;

重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤, 直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止, 并将隐含状态组成的序列作为摘要的字词序列。

6. 一种文摘自动提取装置, 其特征在于, 包括:

第一输入单元, 用于依序获取目标文本所包括的字符, 将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得到隐含状态组成的序列; 其中 LSTM 模型为长短记忆神经网络;

第二输入单元, 用于将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列;

第三输入单元, 用于将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得到更新后隐含状态组成的序列;

上下文向量获取单元, 用于根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值, 获取与编码器隐藏状态的贡献值相对应的上下文向量;

摘要获取单元, 用于根据更新后隐含状态组成的序列及上下文向量, 获取

更新后隐含状态组成的序列中字词的概率分布，将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

7.根据权利要求6所述的文摘自动提取装置，其特征在于，还包括：

历史数据训练单元，用于将语料库中的多篇历史文本置入第一层LSTM结构，并将历史文本对应的文摘置入第二层LSTM结构，进行训练得到LSTM模型。

8.根据权利要求7所述的文摘自动提取装置，其特征在于，所述第二输入单元，包括：

初始化单元，用于获取隐含状态组成的序列中概率最大的词，将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语；

更新单元，用于将初始位词语中的每个字输入至第二层LSTM结构，与第二层LSTM结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列；

重复执行单元，用于重复执行隐含状态组成的序列中每一字输入至第二层LSTM结构，与第二层LSTM结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤，直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，并将隐含状态组成的序列作为摘要的字词序列。

9.根据权利要求6所述的文摘自动提取装置，其特征在于，所述LSTM模型为门限循环单元，所述门限循环单元的模型如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

其中， W_z 、 W_r 、 W 是训练得到的权值参数值， x_t 是输入， h_{t-1} 是隐含状态， z_t 是更新状态， r_t 是重置信号， $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆， h_t 是输出， $\sigma()$ 是sigmoid函数， $\tanh()$ 是双曲正切函数。

10.根据权利要求9所述的文摘自动提取装置，其特征在于，所述将隐含状态组成的序列输入至LSTM模型中的第二层LSTM结构进行解码，得到摘要的

字词序列中，所述摘要的字词序列为与词表大小相同的多项式分布层，并输出向量 $y^t \in \mathbb{R}^K$ ；其中 y^t 中的第 k 维代表生成第 k 个词语的概率， t 的取值为正整数， K 为历史文本所对应词表的大小。

11. 一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现以下步骤：

依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列；其中 LSTM 模型为长短期记忆神经网络；

将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列；

将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到更新后隐含状态组成的序列；

根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；

根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中词语的概率分布，将词语的概率分布中概率最大的词语输出作为目标文本的摘要。

12. 根据权利要求 11 所述的计算机设备，其特征在于，所述依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列之前，还包括：

将语料库中的多篇历史文本置入第一层 LSTM 结构，并将历史文本对应的文摘置入第二层 LSTM 结构，进行训练得到 LSTM 模型。

13. 根据权利要求 11 所述的计算机设备，其特征在于，所述 LSTM 模型为门限循环单元，所述门限循环单元的模型如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

其中, W_z 、 W_r 、 W 是训练得到的权值参数值, x_t 是输入, h_{t-1} 是隐含状态, z_t 是更新状态, r_t 是重置信号, $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆, h_t 是输出, σ () 是 sigmoid 函数, $\tanh$ () 是双曲正切函数。

14. 根据权利要求 13 所述的计算机设备, 其特征在于, 所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列中, 所述摘要的字词序列为与词表大小相同的多项式分布层, 并输出向量 $y^t \in \mathbb{R}^K$; 其中 y^t 中的第 k 维代表生成第 k 个词语的概率, t 的取值为正整数, K 为历史文本所对应词表的大小。

15. 根据权利要求 12 所述的计算机设备, 其特征在于, 所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列, 包括:

获取隐含状态组成的序列中概率最大的词, 将隐含状态组成的序列中概率最大的词作为摘要的字词序列中的初始位词语;

将初始位词语中的每个字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列;

重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构, 与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列, 获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤, 直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止, 并将隐含状态组成的序列作为摘要的字词序列。

16. 一种存储介质, 其特征在于, 所述存储介质存储有计算机程序, 所述计算机程序包括程序指令, 所述程序指令当被处理器执行时使所述处理器执行以下操作:

依序获取目标文本所包括的字符, 将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得到隐含状态组成的序列; 其中 LSTM 模型为长短记忆神经网络;

将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码, 得到摘要的字词序列;

将摘要的字词序列输入至 LSTM 模型中的第一层 LSTM 结构进行编码, 得

到更新后隐含状态组成的序列；

根据更新后隐含状态组成的序列中编码器隐藏状态的贡献值，获取与编码器隐藏状态的贡献值相对应的上下文向量；

根据更新后隐含状态组成的序列及上下文向量，获取更新后隐含状态组成的序列中字词的概率分布，将字词的概率分布中概率最大的字词输出作为目标文本的摘要。

17. 根据权利要求 16 所述的存储介质，其特征在于，所述依序获取目标文本所包括的字符，将字符按顺序输入至 LSTM 模型中的第一层 LSTM 结构进行编码，得到隐含状态组成的序列之前，还包括：

将语料库中的多篇历史文本置入第一层 LSTM 结构，并将历史文本对应的文摘置入第二层 LSTM 结构，进行训练得到 LSTM 模型。

18. 根据权利要求 16 所述的存储介质，其特征在于，所述 LSTM 模型为门限循环单元，所述门限循环单元的模型如下：

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned}$$

其中， W_z 、 W_r 、 W 是训练得到的权值参数值， x_t 是输入， h_{t-1} 是隐含状态， z_t 是更新状态， r_t 是重置信号， $\tilde{h}_t$ 是与隐含状态 h_{t-1} 对应的新记忆， h_t 是输出， $\sigma()$ 是 sigmoid 函数， $\tanh()$ 是双曲正切函数。

19. 根据权利要求 18 所述的存储介质，其特征在于，所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列中，所述摘要的字词序列为与词表大小相同的多项式分布层，并输出向量 $y^t \in R^K$ ；其中 y^t 中的第 k 维代表生成第 k 个词语的概率， t 的取值为正整数， K 为历史文本所对应词表的大小。

20. 根据权利要求 17 所述的存储介质，其特征在于，所述将隐含状态组成的序列输入至 LSTM 模型中的第二层 LSTM 结构进行解码，得到摘要的字词序列，包括：

获取隐含状态组成的序列中概率最大的词，将隐含状态组成的序列中概率

最大的词作为摘要的字词序列中的初始位词语；

将初始位词语中的每个字输入至第二层 LSTM 结构，与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列；

重复执行隐含状态组成的序列中每一字输入至第二层 LSTM 结构，与第二层 LSTM 结构的词表中每一字进行组合得到组合后序列，获取组合后序列中概率最大的词作为隐含状态组成的序列的步骤，直至检测到隐含状态组成的序列中的每一字与词表中的终止符组合时停止，并将隐含状态组成的序列作为摘要的字词序列。