(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 111324728 A
(43)申请公布日 2020.06.23

(21)申请号 202010073600.6

(22)申请日 2020.01.22

(71)申请人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 陈增健 容毅峰 廖梦 徐进 王志平

(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
代理人 崔晓岚 张颖玲

(51)Int.Cl.
G06F 16/34(2019.01)
G06F 40/211(2020.01)
G06F 40/289(2020.01)
G06F 40/258(2020.01)

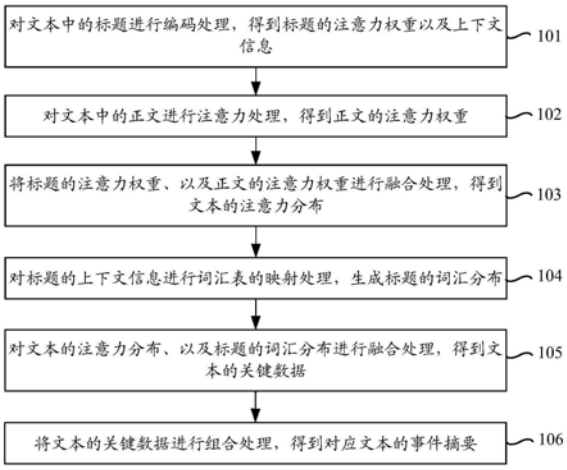
权利要求书3页 说明书27页 附图10页

(54)发明名称

文本事件摘要的生成方法、装置、电子设备及存储介质

(57)摘要

本发明提供了一种文本事件摘要的生成方法、装置、电子设备及存储介质;方法包括:对文本中的标题进行编码处理,得到所述标题的注意力权重以及上下文信息;对所述文本中的正文进行注意力处理,得到所述正文的注意力权重;将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理,得到所述文本的注意力分布;对所述标题的上下文信息进行词汇表的映射处理,生成所述标题的词汇分布;对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理,得到所述文本的关键数据;将所述文本的关键数据进行组合处理,得到对应所述文本的事件摘要。通过本发明,能够融合文本中的正文以及标题,精确地抽取文本的事件摘要。



CN 111324728 A

1.一种文本事件摘要的生成方法,其特征在于,所述方法包括:

对文本中的标题进行编码处理,得到所述标题的注意力权重以及上下文信息;

对所述文本中的正文进行注意力处理,得到所述正文的注意力权重;

将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理,得到所述文本的注意力分布;

对所述标题的上下文信息进行词汇表的映射处理,生成所述标题的词汇分布;

对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理,得到所述文本的关键数据;

将所述文本的关键数据进行组合处理,得到对应所述文本的事件摘要。

2.根据权利要求1所述的方法,其特征在于,所述对文本中的标题进行编码处理,得到所述标题的注意力权重以及上下文信息,包括:

对所述文本中的标题进行隐状态转换处理,得到所述标题的隐状态;

对所述标题的隐状态进行注意力处理,得到所述标题的注意力权重;

基于所述标题的注意力权重,对所述标题的隐状态进行加权求和,得到所述标题的上下文信息。

3.根据权利要求2所述的方法,其特征在于,所述对所述文本中的标题进行隐状态转换处理,得到所述标题的隐状态,包括:

将所述文本中的标题进行词向量转换处理,得到所述标题的词向量;

对所述标题的词向量进行前向编码处理,得到对应所述标题的前向隐向量;

对所述标题的词向量进行后向编码处理,得到对应所述标题的后向隐向量;

将所述前向隐向量以及所述后向隐向量进行拼接处理,得到所述标题的隐状态。

4.根据权利要求2所述的方法,其特征在于,所述对所述标题的隐状态进行注意力处理,得到所述标题的注意力权重,包括:

对所述标题的隐状态、解码隐状态以及可学习参数进行双曲正切处理,得到处理结果;

对所述处理结果进行非线性映射处理,得到所述标题的注意力权重。

5.根据权利要求1所述的方法,其特征在于,所述对所述文本中的正文进行注意力处理,得到所述正文的注意力权重之前,所述方法还包括:

对所述文本中的正文进行筛选处理,得到简化的正文序列;

所述对所述文本中的正文进行注意力处理,得到所述正文的注意力权重,包括:

对所述简化的正文序列进行隐状态转换处理,得到所述正文序列的隐状态;

对所述正文序列的隐状态进行注意力处理,得到所述正文的注意力权重。

6.根据权利要求5所述的方法,其特征在于,所述对所述文本中的正文进行筛选处理,得到简化的正文序列,包括:

对所述文本中的正文进行句子粒度提取处理,得到所述正文中的目标句子;

对所述正文进行词粒度提取处理,得到所述正文中的目标词语;

将所述目标词语对齐到所述目标句子中,得到所述目标句子中未被对齐的词语;

基于所述目标句子中未被对齐的词语的词性,对所述目标句子中的词语进行过滤处理,得到简化的正文序列。

7.根据权利要求6所述的方法,其特征在于,所述对所述文本中的正文进行句子粒度提

取处理，得到所述正文中的目标句子，包括：

对所述文本中的正文进行分句处理，得到多个候选句子；

对所述候选句子进行向量转换处理，得到所述候选句子的句子向量；

确定所述候选句子的句子向量、与所述标题的句子向量的第一相似度，确定所述候选句子的句子向量、与已提取句子的句子向量的第二相似度；

将所述第一相似度以及所述第二相似度进行加权求和，并对加权求和结果进行映射处理，得到所述正文中的目标句子。

8.根据权利要求6所述的方法，其特征在于，所述对所述正文进行词粒度提取处理，得到所述正文中的目标词语，包括：

对所述文本中的正文进行分词处理，得到对应所述正文的词语；

根据所述词语的词性，对所述对应所述正文的词语进行过滤处理，得到多个所述正文的候选词语；

将所述多个所述正文的候选词语组合成所述候选词语的序列，并基于所述候选词语的序列，构建候选目标词图；

基于所述候选目标词图中节点权重，确定所述正文中的目标词语。

9.根据权利要求1-8任一项所述的方法，其特征在于，所述将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理，得到所述文本的注意力分布，包括：

确定对应所述标题的第一融合权重以及对应所述正文的第二融合权重；

确定所述标题的注意力权重与所述第一融合权重的第一乘积、以及所述正文的注意力权重与所述第二融合权重的第二乘积，并

将所述第一乘积与所述第二乘积的求和结果确定为所述文本的注意力分布。

10.根据权利要求9所述的方法，其特征在于，所述确定对应所述标题的第一融合权重之前，所述方法还包括：

基于所述正文的注意力权重，对正文的隐状态进行加权求和，得到所述正文的上下文信息；

所述确定对应所述标题的第一融合权重，包括：

对所述正文的上下文信息、所述标题的上下文信息、解码隐状态、已生成的所述文本的关键数据以及可学习参数进行非线性映射处理，得到对应所述标题的第一融合权重。

11.根据权利要求1-8任一项所述的方法，其特征在于，所述对所述标题的上下文信息进行词汇表的映射处理，生成所述标题的词汇分布，包括：

将所述标题的上下文信息、与解码隐状态进行拼接处理，得到拼接数据；

对所述拼接数据进行第一线性映射处理，得到第一线性映射结果；

对所述第一线性映射结果进行第二线性映射处理，得到第二线性映射结果；

对所述第二线性映射结果进行词汇表的非线性映射处理，生成所述标题的词汇分布。

12.根据权利要求1-8任一项所述的方法，其特征在于，所述对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理，得到所述文本的关键数据，包括：

确定对应所述词汇分布的第一生成权重、以及对应所述注意力分布的第二生成权重；

确定所述标题的词汇分布与所述第一生成权重的第三乘积，并确定所述文本的注意力分布与所述第二生成权重的第四乘积；

将所述第三乘积与所述第四乘积的求和结果确定为所述文本的候选关键数据分布；

将所述候选关键数据分布中的最大概率对应的候选关键数据，确定为所述文本的关键数据。

13.一种文本事件摘要的生成装置，其特征在于，所述装置包括：

编码模块，用于对文本中的标题进行编码处理，得到所述标题的注意力权重以及上下文信息；

注意力模块，用于对所述文本中的正文进行注意力处理，得到所述正文的注意力权重；

第一融合模块，用于将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理，得到所述文本的注意力分布；

映射模块，用于对所述标题的上下文信息进行词汇表的映射处理，生成所述标题的词汇分布；

第二融合模块，用于对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理，得到所述文本的关键数据；

组合模块，用于将所述文本的关键数据进行组合处理，得到对应所述文本的事件摘要。

14.一种电子设备，其特征在于，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现权利要求1至12任一项所述的文本事件摘要的生成方法。

15.一种计算机可读存储介质，其特征在于，存储有可执行指令，用于引起处理器执行时，实现权利要求1至12任一项所述的文本事件摘要的生成方法。

文本事件摘要的生成方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及人工智能的自然语言处理技术，尤其涉及一种文本事件摘要的生成方法、装置、电子设备及存储介质。

背景技术

[0002] 人工智能(Artificial Intelligence,AI)是计算机科学的一个综合技术，通过研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科，涉及领域广泛，例如自然语言处理技术以及机器学习/深度学习等几大方向，随着技术的发展，人工智能技术将在更多的领域得到应用，并发挥越来越重要的价值。

[0003] 自然语言处理(Nature Language Processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向，能实现人与计算机之间用自然语言进行有效通信。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，该领域将涉及自然语言，即人们日常使用的语言，所以它与语言学有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0004] 事件摘要生成系统是自然语言处理领域的重要应用之一，事件摘要生成系统是指将文本所包含的核心事件以精炼的语言进行概括描述，生成对应文本的事件摘要。事件摘要生成系统在搜索系统、推荐系统等中都有广泛的应用，即事件摘要生成系统是这些复杂系统的基础组件。

[0005] 但是，传统的事件摘要生成系统生成的事件摘要的准确性低，即事件摘要不通顺、且不能精确表达文本所包含的核心事件。

发明内容

[0006] 本发明实施例提供一种文本事件摘要的生成方法、装置、电子设备及存储介质，能够融合文本中的正文以及标题，精确地抽取文本的事件摘要。

[0007] 本发明实施例的技术方案是这样实现的：

[0008] 本发明实施例提供一种文本事件摘要的生成方法，包括：

[0009] 对文本中的标题进行编码处理，得到所述标题的注意力权重以及上下文信息；

[0010] 对所述文本中的正文进行注意力处理，得到所述正文的注意力权重；

[0011] 将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理，得到所述文本的注意力分布；

[0012] 对所述标题的上下文信息进行词汇表的映射处理，生成所述标题的词汇分布；

[0013] 对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理，得到所述文本的关键数据；

[0014] 将所述文本的关键数据进行组合处理，得到对应所述文本的事件摘要。

[0015] 本发明实施例提供一种文本事件摘要的生成装置，包括：

[0016] 编码模块，用于对文本中的标题进行编码处理，得到所述标题的注意力权重以及上下文信息；

[0017] 注意力模块，用于对所述文本中的正文进行注意力处理，得到所述正文的注意力权重；

[0018] 第一融合模块，用于将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理，得到所述文本的注意力分布；

[0019] 映射模块，用于对所述标题的上下文信息进行词汇表的映射处理，生成所述标题的词汇分布；

[0020] 第二融合模块，用于对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理，得到所述文本的关键数据；

[0021] 组合模块，用于将所述文本的关键数据进行组合处理，得到对应所述文本的事件摘要。

[0022] 上述技术方案中，所述编码模块还用于对所述文本中的标题进行隐状态转换处理，得到所述标题的隐状态；

[0023] 对所述标题的隐状态进行注意力处理，得到所述标题的注意力权重；

[0024] 基于所述标题的注意力权重，对所述标题的隐状态进行加权求和，得到所述标题的上下文信息。

[0025] 上述技术方案中，所述编码模块还用于将所述文本中的标题进行词向量转换处理，得到所述标题的词向量；

[0026] 对所述标题的词向量进行前向编码处理，得到对应所述标题的前向隐向量；

[0027] 对所述标题的词向量进行后向编码处理，得到对应所述标题的后向隐向量；

[0028] 将所述前向隐向量以及所述后向隐向量进行拼接处理，得到所述标题的隐状态。

[0029] 上述技术方案中，所述编码模块还用于对所述标题的隐状态、解码隐状态以及可学习参数进行双曲正切处理，得到处理结果；

[0030] 对所述处理结果进行非线性映射处理，得到所述标题的注意力权重。

[0031] 上述技术方案中，所述装置还包括：

[0032] 筛选模块，用于对所述文本中的正文进行筛选处理，得到简化的正文序列；

[0033] 所述注意力模块还用于对所述简化的正文序列进行隐状态转换处理，得到所述正文序列的隐状态；

[0034] 对所述正文序列的隐状态进行注意力处理，得到所述正文的注意力权重。

[0035] 上述技术方案中，所述筛选模块还用于对所述文本中的正文进行句子粒度提取处理，得到所述正文中的目标句子；

[0036] 对所述正文进行词粒度提取处理，得到所述正文中的目标词语；

[0037] 将所述目标词语对齐到所述目标句子中，得到所述目标句子中未被对齐的词语；

[0038] 基于所述目标句子中未被对齐的词语的词性，对所述目标句子中的词语进行过滤处理，得到简化的正文序列。

[0039] 上述技术方案中，所述筛选模块还用于对所述文本中的正文进行分句处理，得到多个候选句子；

[0040] 对所述候选句子进行向量转换处理，得到所述候选句子的句子向量；

[0041] 确定所述候选句子的句子向量、与所述标题的句子向量的第一相似度，确定所述候选句子的句子向量、与已提取句子的句子向量的第二相似度；

[0042] 将所述第一相似度以及所述第二相似度进行加权求和，并对加权求和结果进行映射处理，得到所述正文中的目标句子。

[0043] 上述技术方案中，所述筛选模块还用于对所述候选句子进行词向量转换处理，得到所述候选句子的词向量；

[0044] 基于所述词向量的词频以及逆文本频率指数，确定所述词向量的权重；

[0045] 基于所述词向量的权重，对所述候选句子的词向量进行加权平均处理，得到所述候选句子的句子向量。

[0046] 上述技术方案中，所述筛选模块还用于对所述文本中的正文进行分词处理，得到对应所述正文的词语；

[0047] 根据所述词语的词性，对所述对应所述正文的词语进行过滤处理，得到多个所述正文的候选词语；

[0048] 将所述多个所述正文的候选词语组合成所述候选词语的序列，并基于所述候选词语的序列，构建候选目标词图；

[0049] 基于所述候选目标词图中节点权重，确定所述正文中的目标词语。

[0050] 上述技术方案中，所述筛选模块还用于将所述候选词语的序列中的候选词语确定为所述候选目标词图的节点；

[0051] 当任意两节点在所述候选词语的序列中的距离小于或者等于距离阈值时，连接所述任意两节点的边；

[0052] 将所述两节点在所述序列中出现的频率确定为所述边的节点权重；

[0053] 根据所述节点、所述边以及所述节点权重，构建所述候选目标词图。

[0054] 上述技术方案中，所述第一融合模块还用于确定对应所述标题的第一融合权重以及对应所述正文的第二融合权重；

[0055] 确定所述标题的注意力权重与所述第一融合权重的第一乘积、以及所述正文的注意力权重与所述第二融合权重的第二乘积，并

[0056] 将所述第一乘积与所述第二乘积的求和结果确定为所述文本的注意力分布。

[0057] 上述技术方案中，所述装置还包括：

[0058] 处理模块，用于基于所述正文的注意力权重，对正文的隐状态进行加权求和，得到所述正文的上下文信息；

[0059] 所述第一融合模块还用于对所述正文的上下文信息、所述标题的上下文信息、解码隐状态、已生成的所述文本的关键数据以及可学习参数进行非线性映射处理，得到对应所述标题的第一融合权重。

[0060] 上述技术方案中，所述映射处理还用于将所述标题的上下文信息、与解码隐状态进行拼接处理，得到拼接数据；

[0061] 对所述拼接数据进行第一线性映射处理，得到第一线性映射结果；

[0062] 对所述第一线性映射结果进行第二线性映射处理，得到第二线性映射结果；

[0063] 对所述第二线性映射结果进行词汇表的非线性映射处理，生成所述标题的词汇分布。

[0064] 上述技术方案中，所述第二融合模块还用于确定对应所述词汇分布的第一生成权重、以及对应所述注意力分布的第二生成权重；

[0065] 确定所述标题的词汇分布与所述第一生成权重的第三乘积，并确定所述文本的注意力分布与所述第二生成权重的第四乘积；

[0066] 将所述第三乘积与所述第四乘积的求和结果确定为所述文本的候选关键数据分布；

[0067] 将所述候选关键数据分布中的最大概率对应的候选关键数据，确定为所述文本的关键数据。

[0068] 上述技术方案中，所述第二融合模块还用于对所述标题的上下文信息、解码隐状态、已生成的所述文本的关键数据以及可学习参数进行非线性映射处理，得到对应所述词汇分布的第一生成权重。

[0069] 本发明实施例提供一种用于生成文本事件摘要的电子设备，所述电子设备包括：

[0070] 存储器，用于存储可执行指令；

[0071] 处理器，用于执行所述存储器中存储的可执行指令时，实现本发明实施例提供的文本事件摘要的生成方法。

[0072] 本发明实施例提供一种计算机可读存储介质，存储有可执行指令，用于引起处理器执行时，实现本发明实施例提供的文本事件摘要的生成方法。

[0073] 本发明实施例具有以下有益效果：

[0074] 通过将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布，实现生成的事件摘要融合了标题和文本，使得事件摘要更加完整；通过对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布，实现生成的事件摘要融合了词汇表，使得事件摘要更加准确。

附图说明

[0075] 图1是本发明实施例提供的文本事件摘要的生成系统10的应用场景示意图；

[0076] 图2是本发明实施例提供的文本事件摘要的电子设备500的结构示意图；

[0077] 图3-6是本发明实施例提供的文本事件摘要的生成方法的流程示意图；

[0078] 图7是本发明实施例提供的新闻事件query展示示意图；

[0079] 图8是本发明实施例提供的新闻事件query搜索提示示意图；

[0080] 图9是本发明实施例提供的文本事件摘要的生成方法的流程示意图；

[0081] 图10是本发明实施例提供的文本事件摘要生成模型的结构示意图；

[0082] 图11是本发明实施例提供的层次交互式内容信息提取器的流程示意图；

[0083] 图12是本发明实施例提供的对齐、剪枝的效果示意图；

[0084] 图13是本发明实施例提供的模型微调的示意图。

具体实施方式

[0085] 为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作进一步地详细描述，所描述的实施例不应视为对本发明的限制，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它实施例，都属于本发明保护的范围。

[0086] 在以下的描述中,所涉及的术语“第一\第二\第三\第四”仅仅是是区别类似的对象,不代表针对对象的特定排序,可以理解地,“第一\第二\第三\第四”在允许的情况下可以互换特定的顺序或先后次序,以使这里描述的本发明实施例能够以除了在这里图示或描述的以外的顺序实施。

[0087] 除非另有定义,本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本发明实施例的目的,不是旨在限制本发明。

[0088] 对本发明实施例进行进一步详细说明之前,对本发明实施例中涉及的名词和术语进行说明,本发明实施例中涉及的名词和术语适用于如下的解释。

[0089] 1)事件摘要:又称事件短描述,文本所包含的核心事件的概括性描述,即不加评论和补充解释,简明、确切地记述文本的实质内容,甚至用几个词来总结文本中的关键事件。例如,某文本为“欧洲足球锦标赛,简称“欧锦赛”,也称“欧洲杯”,是一项由欧洲足球协会联盟举办,欧洲足协成员国间参加的最高级别国家级足球赛事。1960年举行第一届,其后每四年举行一届,已举办15届……”,则对应的事件摘要为“欧洲杯简介”。

[0090] 2)分词:将连续的字序列按照一定的规范重新组合成词序列的过程。通过让计算机模拟人对句子的理解,达到识别词的效果。

[0091] 3)召回(Recall):从文档库中检索出待推荐的候选的文档。

[0092] 4)词到向量(word2vec)模型:用来产生词向量的模型,能够将所有的词向量化,使得词与词之间语义上的距离量化为对应向量之间的距离,可以定量地度量词与词之间的关系,从而挖掘词与词之间的联系。

[0093] 5)实体词:实体是指能够独立存在的、作为一切属性的基础和万物本原的东西,即实体词是指能够表示实体的词语。名词与代词为实体词,例如“小红”、“地点”为实体词。

[0094] 6)未登录词(OOV,Out Of Vocabulary):没有被收录在分词词表中,但必须从文本中切分出来的词,包括各类专有名词(人名、地名、企业名等)、缩写词、新增词汇等等。

[0095] 本发明实施例提供一种文本事件摘要的生成方法、装置、电子设备和存储介质,能够根据文本中的标题以及正文,自动并准确地抽取文本的事件摘要。下面说明本发明实施例提供的文本事件摘要的电子设备的示例性应用,本发明实施例提供的文本事件摘要的电子设备可以是服务器,例如部署在云端的服务器,根据其他设备或者用户提供的文本(包括标题和正文),对该文本进行一系列处理,抽取对应文本的事件摘要,并向用户展示该事件摘要,例如,服务器根据其他设备获得文本,通过文本事件摘要模型对该文本进行编码、融合、映射等处理,抽取准确的事件摘要;也可是笔记本电脑,平板电脑,台式计算机,移动设备(例如,移动电话,个人数字助理)等各种类型的用户终端,例如手持终端,根据用户在手持终端上输入的文本,获得准确的事件摘要,并显示在手持终端的显示界面上。

[0096] 在一个实施场景中,对于搜索应用(Application,APP),服务器或者终端可以根据输入的部分检索信息,获得完整的检索信息。在搜索之前,服务器或者终端预先通过文本事件摘要模型对文本中的标题进行编码处理,得到标题的注意力权重以及上下文信息;对文本中的正文进行注意力处理,得到正文的注意力权重;将标题的注意力权重、以及正文的注意力权重进行融合处理,得到文本的注意力分布;对标题的上下文信息进行词汇表的映射处理,生成标题的词汇分布;对文本的注意力分布、以及标题的词汇分布进行融合处理,得

到文本的关键数据;将文本的关键数据进行组合处理,得到对应文本的事件摘要,并将所有对应文本的事件摘要存储至数据库。在搜索时,用户在搜索应用中输入部分检索信息后,服务器或者终端可以根据该部分检索信息对数据库中的事件摘要进行匹配,并匹配成功的事件摘要呈现给用户,使得用户根据部分检索信息,得到完整的检索信息、即准确的事件摘要,并根据简短的事件摘要,对应获得与事件摘要对应的完整文本,包括标题和正文。例如文本的标题为"欧洲杯-国家级足球赛事"、文本的正文为"欧洲足球锦标赛,简称"欧锦赛",也称"欧洲杯",是一项由欧洲足球协会联盟举办,欧洲足协成员国间参加的最高级别国家级足球赛事。1960年举行第一届,其后每四年举行一届,已举办15届……",则对应文本的事件摘要为"欧洲杯简介",当用户在搜索应用中输入"欧洲杯"后,即可获得事件摘要"欧洲杯简介"。

[0097] 在一个实施场景中,对于推荐应用,服务器或者终端可以根据时新的文本(例如新闻),获得对应的事件摘要,并推荐给用户。服务器或者终端预先召回一些时新的文本,通过文本事件摘要模型对文本中的标题进行编码处理,得到标题的注意力权重以及上下文信息;对文本中的正文进行注意力处理,得到正文的注意力权重;将标题的注意力权重、以及正文的注意力权重进行融合处理,得到文本的注意力分布;对标题的上下文信息进行词汇表的映射处理,生成标题的词汇分布;对文本的注意力分布、以及标题的词汇分布进行融合处理,得到文本的关键数据;将文本的关键数据进行组合处理,得到对应文本的事件摘要,并将所有对应文本的事件摘要存储至数据库。在用户打开推荐应用后,服务器或者终端可以将时新的文本对应的事件摘要呈现给用户,使得用户根据呈现的事件摘要,以便用户了解最新的文本事件、即新闻,并对应获得与事件摘要对应的完整文本,包括标题和正文。例如文本的标题为"欧洲杯-国家级足球赛事"、文本的正文为"欧洲足球锦标赛,简称"欧锦赛",也称"欧洲杯",是一项由欧洲足球协会联盟举办,欧洲足协成员国间参加的最高级别国家级足球赛事。1960年举行第一届,其后每四年举行一届,已举办15届……",则对应文本的事件摘要为"欧洲杯简介",在用户打开推荐应用后,在推荐应用中呈现事件摘要"欧洲杯简介"。

[0098] 作为示例,参见图1,图1是本发明实施例提供的文本事件摘要的生成系统10的应用场景示意图,终端200通过网络300连接服务器100,网络300可以是广域网或者局域网,又或者是二者的组合。

[0099] 终端200可以被用来获取文本,例如,当用户通过输入界面输入文本,输入完成后,终端自动获取用户输入的文本。

[0100] 在一些实施例中,终端200本地执行本发明实施例提供的文本事件摘要的生成方法来完成根据用户输入的文本,得到准确的事件摘要,例如,在终端200上安装事件摘要助手,用户在事件摘要助手中,输入文本,终端200根据输入的文本,通过文本事件摘要模型对该文本进行编码、融合、映射等处理,得到准确的事件摘要,并将准确的事件摘要显示在终端200的显示界面210上。

[0101] 在一些实施例中,终端200也可以通过网络300向服务器100发送用户在终端200上输入的文本,并调用服务器100提供的文本事件摘要的生成功能,服务器100通过本发明实施例提供的文本事件摘要的生成方法获得对应事件摘要,例如,在终端200上安装事件摘要助手,用户在事件摘要助手中,输入文本,终端通过网络300向服务器100发送文本,服务器

100接收到该文本后，通过文本事件摘要模型对该文本进行编码、融合、映射等处理，得到准确的事件摘要，并将准确的事件摘要返回至事件摘要助手，将事件摘要显示在终端200的显示界面210上，或者，服务器100直接给出事件摘要。

[0102] 继续说明本发明实施例提供的文本事件摘要的电子设备的结构，文本事件摘要的电子设备可以是各种终端，例如手机、电脑等，也可以是如图1示出的服务器100。

[0103] 参见图2，图2是本发明实施例提供的用于生成文本事件摘要的电子设备500的结构示意图，图2所示的文本事件摘要的电子设备500包括：至少一个处理器510、存储器550、至少一个网络接口520和用户接口530。文本事件摘要的电子设备500中的各个组件通过总线系统540耦合在一起。可理解，总线系统540用于实现这些组件之间的连接通信。总线系统540除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，在图2中将各种总线都标为总线系统540。

[0104] 处理器510可以是一种集成电路芯片，具有信号的处理能力，例如通用处理器、数字信号处理器(DSP，Digital Signal Processor)，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任何常规的处理器等。

[0105] 用户接口530包括使得能够呈现媒体内容的一个或多个输出装置531，包括一个或多个扬声器和/或一个或多个视觉显示屏。用户接口530还包括一个或多个输入装置532，包括有助于用户输入的用户接口部件，比如键盘、鼠标、麦克风、触屏显示屏、摄像头、其他输入按钮和控件。

[0106] 存储器550包括易失性存储器或非易失性存储器，也可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器(ROM，Read Onl y Memory)，易失性存储器可以是随机存取存储器(RAM，Random Access M emory)。本发明实施例描述的存储器550旨在包括任意适合类型的存储器。存储器550可选地包括在物理位置上远离处理器510的一个或多个存储设备。

[0107] 在一些实施例中，存储器550能够存储数据以支持各种操作，这些数据的示例包括程序、模块和数据结构或者其子集或超集，下面示例性说明。

[0108] 操作系统551，包括用于处理各种基本系统服务和执行硬件相关任务的系统程序，例如框架层、核心库层、驱动层等，用于实现各种基础业务以及处理基于硬件的任务；

[0109] 网络通信模块552，用于经由一个或多个(有线或无线)网络接口520到达其他计算设备，示例性的网络接口520包括：蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB，Universal Serial Bus)等；

[0110] 显示模块553，用于经由一个或多个与用户接口530相关联的输出装置531(例如，显示屏、扬声器等)使得能够呈现信息(例如，用于操作外围设备和显示内容和信息的用户接口)；

[0111] 输入处理模块554，用于对一个或多个来自一个或多个输入装置532之一的一个或多个用户输入或互动进行检测以及翻译所检测的输入或互动。

[0112] 在一些实施例中，本发明实施例提供的文本事件摘要的生成装置可以采用软硬件结合的方式实现，作为示例，本发明实施例提供的文本事件摘要的生成装置可以是采用硬件译码处理器形式的处理器，其被编程以执行本发明实施例提供的文本事件摘要的生成方

法，例如，硬件译码处理器形式的处理器可以采用一个或多个应用专用集成电路(ASIC，Application Specific Integrated Circui t)、DSP、可编程逻辑器件(PLD，Programmable Logic Device)、复杂可编程逻辑器件(CPLD，Complex Programmable Logic Device)、现场可编程门阵列(FPGA，Field-Programmable Gate Array)或其他电子元件。

[0113] 在另一些实施例中，本发明实施例提供的文本事件摘要的生成装置可以采用软件方式实现，图2示出了存储在存储器550中的文本事件摘要的生成装置555，其可以是程序和插件等形式的软件，并包括一系列的模块，包括编码模块5551、注意力模块5552、第一融合模块5553、映射模块5554、第二融合模块5555、组合模块5556、筛选模块5557以及处理模块5558；其中，编码模块5551、注意力模块5552、第一融合模块5553、映射模块5554、第二融合模块5555、组合模块5556、筛选模块5557以及处理模块5558用于实现本发明实施例提供的文本事件摘要的生成方法。

[0114] 根据上文可以理解，本发明实施例提供的文本事件摘要的生成方法可以由各种类型的文本事件摘要的电子设备实施，例如智能终端和服务器等。

[0115] 下面结合本发明实施例提供的服务器的示例性应用和实施，说明本发明实施例提供的文本事件摘要的生成方法。参见图3，图3是本发明实施例提供的文本事件摘要的生成方法的流程示意图，结合图3示出的步骤进行说明。

[0116] 在步骤101中，对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息。

[0117] 用户可以在终端的输入界面上输入文本信息，当输入完成后，终端可以将文本信息转发至服务器，服务器接收到文本信息后，可以对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息，以便后续融合标题和正文。

[0118] 参见图4，图4是本发明实施例提供的一个可选的流程示意图，在一些实施例中，图4示出图3中步骤101可以通过图4示出的步骤1011-1013实现。在步骤1011中，对文本中的标题进行隐状态转换处理，得到标题的隐状态；在步骤1012中，对标题的隐状态进行注意力处理，得到标题的注意力权重；在步骤1013中，基于标题的注意力权重，对标题的隐状态进行加权求和，得到标题的上下文信息。

[0119] 通过指针生成网络(Pointer Generator Network)中的编码器对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息。为了得到标题的上下文信息，可以通过编码器对文本中的标题序列进行隐状态转换处理，得到标题的隐状态，并对标题的隐状态进行注意力处理，得到标题的注意力权重，并基于标题的注意力权重，对标题的隐状态进行加权求和，从而得到标题的上下文信息，以便根据标题的上下文信息，生成标题的词汇分布，以生成词汇表中的词语。

[0120] 在一些实施例中，对文本中的标题进行隐状态转换处理，得到标题的隐状态，包括：将文本中的标题进行词向量转换处理，得到标题的词向量；对标题的词向量进行前向编码处理，得到对应标题的前向隐向量；对标题的词向量进行后向编码处理，得到对应标题的后向隐向量；将前向隐向量以及后向隐向量进行拼接处理，得到标题的隐状态。

[0121] 作为示例，服务器将文本中的标题进行词向量转换处理，得到标题的词向量，将词向量输入至双向长短时记忆网络(BLSTM或BiLSTM，Bidirectional Long Short-term Memory)编码器的隐层，并通过BLSTM编码器的隐层对标题的词向量分别进行前向编码和后

向编码处理，从而得到标题的前向隐向量以及标题的后向隐向量，并对前向隐向量以及标题的后向隐向量进行拼接处理，从而得到标题的隐状态，其中，标题的前向隐向量包含标题的前向所有信息，标题的后向隐向量包含标题的后向所有信息。因此，拼接标题的前向隐向量以及标题的后向隐向量后的标题的隐状态包含标题所有的信息。

[0122] 其中，可以通过BLSTM编码器对标题的词向量中的第i向量进行前向编码处理，得到标题的第i前向隐向量；对标题的词向量中的第i向量进行后向编码处理，得到标题的第i后向隐向量；将第i前向隐向量、第i后向隐向量进行拼接处理，得到包含标题的第i隐状态。其中，$0<i\leqslant N$，且i、N为正整数，N为词向量中向量的总数目。当词向量中有N个向量，则对N个向量按照前向方向进行编码，依次得到在前向方向的N个隐向量，例如对词向量按照前向方向进行编码处理后，得到在前向方向的隐向量为$\{h_{1l},h_{2l},...h_{il}...,h_{Nl}\}$，其中，$h_{il}$表示第i向量在前向方向的第i隐向量。对N个向量按照后向方向进行编码，依次得到在后向方向的N个隐向量，例如对词向量按照后向方向进行编码处理后，得到在后向方向的隐向量为$\{h_{1r},h_{2r},...h_{ir}...,h_{Nr}\}$，其中，$h_{ir}$表示第i向量在后向方向的第i隐向量。将在前向方向的隐向量为$\{h_{1l},h_{2l},...h_{il}...,h_{Nl}\}$以及在后向方向的隐向量为$\{h_{1r},h_{2r},...h_{ir}...,h_{Nr}\}$进行拼接，得到标题的隐状态$\{[h_{1l},h_{1r}],[h_{2l},h_{2r}],...[h_{il},h_{ir}]...,[h_{Nl},h_{Nr}]\}$，例如，将第i向量在前向方向的第i隐向量$h_{il}$、第i向量在后向方向的第i隐向量$h_{ir}$进行拼接处理，得到包含上下文信息的第i编码信息$\{h_{il},h_{ir}\}$。为了节约计算过程，由于前向方向的最后一个隐向量包含前向方向的大部分信息、后向方向的最后一个隐向量包含后向方向的大部分信息，因此，可以直接对前向方向的最后一个隐向量以及后向方向的最后一个隐向量进行融合，得到包含标题的隐状态。

[0123] 在一些实施例中，对标题的隐状态进行注意力处理，得到标题的注意力权重，包括：对标题的隐状态、解码隐状态以及可学习参数进行双曲正切处理，得到处理结果；对处理结果进行非线性映射处理，得到标题的注意力权重。

[0124] 承接上述示例，在得到标题的隐状态后，可以对标题的隐状态、解码隐状态以及可学习参数进行双曲正切处理，得到处理结果，并对处理结果进行非线性映射处理，得到标题的注意力权重，其中，标题的注意力权重的计算公式为$a_{t,i}=\mathrm{softmax}(v^{T}\tanh(W_{h}h_{i}+W_{d}d_{t}+b))$，$a_{t,i}$表示标题的注意力权重，v、$W_{h}$、$W_{d}$、b表示可学习参数、即用于训练的参数，$h_{i}$表示标题的一个隐状态，$d_{t}$表示解码的隐状态，softmax函数表示逻辑回归函数、即非线性映射函数。

[0125] 在步骤102中，对文本中的正文进行注意力处理，得到正文的注意力权重。

[0126] 为了后续能够融合标题和正文，可以对文本中的正文进行注意力处理，得到正文的注意力权重，以便后续融合标题的注意力权重、以及正文的注意力权重。

[0127] 参见图5，图5是本发明实施例提供的一个可选的流程示意图，在一些实施例中，图5示出图3中在步骤102之前，还可以在步骤107中，对文本中的正文进行筛选处理，得到简化的正文序列；对文本中的正文进行注意力处理，得到正文的注意力权重，包括：对简化的正文序列进行隐状态转换处理，得到正文序列的隐状态；对正文序列的隐状态进行注意力处理，得到正文的注意力权重。

[0128] 在一些实施例中，由于文本的正文存在很多冗余的信息，为了减小对正文的处理量，可以先对文本中的正文进行筛选处理，从而得到简化的正文序列，后续可以对简化的正

文序列进行处理，可以对简化的正文序列进行隐状态转换处理，得到正文序列的隐状态；对正文序列的隐状态进行注意力处理，得到正文的注意力权重。通过指针生成网络中的编码器对简化的正文序列进行隐状态转换处理，得到正文序列的隐状态，并对正文序列的隐状态进行注意力处理，得到正文的注意力权重，以便后续融合标题的注意力权重、以及正文的注意力权重，得到文本的注意力分布。

[0129] 在一些实施例中，对文本中的正文进行筛选处理，得到简化的正文序列，包括：对文本中的正文进行句子粒度提取处理，得到正文中的目标句子；对正文进行词粒度提取处理，得到正文中的目标词语；将目标词语对齐到目标句子中，得到目标句子中未被对齐的词语；基于目标句子中未被对齐的词语的词性，对目标句子中的词语进行过滤处理，得到简化的正文序列。

[0130] 例如，在服务器获得文本的正文后，可以通过层次交互式内容提取器(HI CS，Hierarchical Interaction Content Selector)分别句子粒度以及词粒度对文本中的正文进行提取处理，得到正文中的目标句子以及目标词语，并采用对齐手段将目标词语对齐到目标句子中，得到目标句子中未被对齐的词语，采用剪枝手段基于目标句子中未被对齐的词语的词性，对目标句子中的词语进行过滤处理，从而得到简化的正文序列。通过句子粒度和词粒度对正文的关键信息进行提取，并采用对齐和剪枝的手段，弥补了正文句子的信息冗余性以及单纯词粒度上的语序信息缺失，能够有效地整合两种粒度信息提取的优势，对于后续的事件摘要生成上，能够有效地降低信息提取的难度，同时保留了语序的信息，能够更精准利用正文信息。

[0131] 在一些实施例中，对文本中的正文进行句子粒度提取处理，得到正文中的目标句子，包括：对文本中的正文进行分句处理，得到多个候选句子；对候选句子进行向量转换处理，得到候选句子的句子向量；确定候选句子的句子向量、与标题的句子向量的第一相似度，确定候选句子的句子向量、与已提取句子的句子向量的第二相似度；将第一相似度以及第二相似度进行加权求和，并对加权求和结果进行映射处理，得到正文中的目标句子。

[0132] 粒度(granularity)指的是信息单元的相对大小或粗糙程度。在服务器获得文本的正文后，可以对文本中的正文进行句子粒度提取处理，得到正文中的目标句子，即对文本中的正文以句子为单位进行提取，得到正文中的目标句子。为了得到目标句子，可以先对文本中的正文进行分句处理，得到多个候选句子，例如可以以“。”、“?”、“!”等作为截断句子的符号，以便后续根据候选句子得到目标句子，对候选句子进行向量转换处理，得到候选句子的句子向量，确定候选句子的句子向量、与标题的句子向量的第一相似度，确定候选句子的句子向量、与已提取句子的句子向量的第二相似度，并将第一相似度以及第二相似度进行加权求和，并对加权求和结果进行映射处理，得到正文中的目标句子，使得所提取目标句子考虑到与标题的关联程度以及与各个所提取句子之间的不同程度，这样所提取的目标句子，能够保证所提取的目标句子为关键句的同时，还能够保证所提取句子之间的多样性。

[0133] 在一些实施例中，对候选句子进行向量转换处理，得到候选句子的句子向量，包括：对候选句子进行词向量转换处理，得到候选句子的词向量；基于词向量的词频以及逆文本频率指数，确定词向量的权重；基于词向量的权重，对候选句子的词向量进行加权平均处理，得到候选句子的句子向量。

[0134] 为了能够得到准确的候选句子的句子向量，需要先对候选句子进行词向量转换处

理，得到候选句子的词向量，并通过TF-IDF(Term Frequency-Inverse Document Frequency)方法获得词向量的词频以及逆文本频率指数，其中，TF-IDF方法是一种用于信息检索与数据挖掘的常用加权技术，TF表示词频(Term Frequency)，IDF表示逆文本频率指数(Inverse Document Frequency)，将词向量在候选句子中出现的次数确定为词向量的词频，将总候选句子的数目除以包含该词向量的候选句子的数目，再将得到的商取以10为底的对数得到词向量的逆文本频率指数。将词向量的词频与逆文本频率指数相乘，得到词向量的权重。并基于词向量的权重，对候选句子的词向量进行加权平均处理，从而得到准确的候选句子的句子向量。

[0135] 在一些实施例中，对正文进行词粒度提取处理，得到正文中的目标词语，包括：对文本中的正文进行分词处理，得到对应正文的词语；根据词语的词性，对对应正文的词语进行过滤处理，得到多个正文的候选词语；将多个正文的候选词语组合成候选词语的序列，并基于候选词语的序列，构建候选目标词图；基于候选目标词图中节点权重，确定正文中的目标词语。

[0136] 在服务器获得文本的正文后，可以对正文进行词粒度提取处理，得到正文中的目标词语，即对文本中的正文以词语为单位进行提取，得到正文中的目标词语。为了得到目标词语，可以先对文本中的正文进行分词处理，得到对应正文的词语，以便后续得到目标词语，根据词语的词性，对对应正文的词语进行过滤处理，得到多个正文的候选词语，例如，可以将介词、副词等信息增益低的词语过滤掉，以粗略简化正文。将多个正文的候选词语组合成候选词语的序列，并基于候选词语的序列，构建候选目标词图，并对候选目标词图中节点权重进行降序排序，将前M个节点权重对应的节点的候选词语确定为正文中的目标词语，其中，M可以是预先设定的数值，M为自然数。

[0137] 在一些实施例中，基于候选词语的序列，构建候选目标词图，包括：将候选词语的序列中的候选词语确定为候选目标词图的节点；当任意两节点在候选词语的序列中的距离小于或者等于距离阈值时，连接任意两节点的边；将两节点在序列中出现的频率确定为边的节点权重；根据节点、边以及节点权重，构建候选目标词图。

[0138] 作为示例，在确定了候选词语的序列后，可以基于候选词语的序列，构建候选目标词图，将候选词语的序列中的候选词语确定为候选目标词图的节点，例如候选词语的序列为$S=[t_1,t_2,...,t_i,...,t_n]$，其中，$t_i$表示序列中的第i个候选词语，当任意两节点在候选词语的序列中的距离小于或者等于距离阈值时，连接任意两节点的边，例如，候选词语t_2与候选词语t_i节点的距离小于或者等于距离阈值5时，连接候选词语t_2与候选词语t_i节点的边，且将两节点在序列中出现的频率确定为边的节点权重，例如，候选词语t_2与候选词语t_i节点的距离小于或者等于距离阈值5时出现的频率记为t_2与t_i边的节点权重，最后根据节点、边以及节点权重，构建候选目标词图，以便后续根据候选目标词图确定正文中的目标词语。

[0139] 在步骤103中，将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布。

[0140] 作为示例，在服务器获得标题的注意力权重、以及正文的注意力权重后，可以对标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布，即通过拷贝的机制，可以从输入标题或者正文中复制最终的生成部分事件摘要。对应的，该注意力分

布是在文本上的概率分布，以便决定生成的部分事件摘要选择拷贝生成哪一个词。由于是以标题和正文的注意力权重为源进行融合处理的，因此得到注意力分布也可以称为多源注意力分布。

[0141] 参见图6，图6是本发明实施例提供的一个可选的流程示意图，在一些实施例中，图6示出图3中步骤103可以通过图6示出的步骤1031-1032实现。

[0142] 在步骤1031中，确定对应标题的第一融合权重以及对应正文的第二融合权重；在步骤1032中，确定标题的注意力权重与第一融合权重的第一乘积、以及正文的注意力权重与第二融合权重的第二乘积，并将第一乘积与第二乘积的求和结果确定为文本的注意力分布。

[0143] 为了得到文本的注意力分布，需要先确定标题以及正文的注意力权重，然后进行融合处理，例如，确定对应标题的第一融合权重以及对应正文的第二融合权重，并确定标题的注意力权重与第一融合权重的第一乘积、以及正文的注意力权重与第二融合权重的第二乘积，则第一乘积与第二乘积的求和结果为文本的注意力分布，即将不同权重分配给标题和文本，以获得注意力分布的。

[0144] 在一些实施例中，为了确定对应标题的第一融合权重，可以在确定对应标题的第一融合权重之前，基于正文的注意力权重，对正文的隐状态进行加权求和，得到正文的上下文信息；从而，在确定对应标题的第一融合权重时，可以对正文的上下文信息、标题的上下文信息、解码隐状态、已生成的文本的关键数据以及可学习参数进行非线性映射处理，得到对应标题的第一融合权重。

[0145] 作为示例，在确定对应标题的第一融合权重之前，需要先通过指针生成网络的编码器基于正文的注意力权重，对正文的隐状态进行加权求和，得到正文的上下文信息。在得到正文的上下文信息后，可以对正文的上下文信息、标题的上下文信息、解码隐状态、已生成的文本的关键数据以及可学习参数进行非线性映射处理，从而得到对应标题的第一融合权重，其中，对应标题的第一融合权重的计算公式为 $\eta=\sigma(w_d^T d_t+w_y^T y_{t-1}+w_c^T c_t+w_{c'}^T c_t')$，$w_d$、$w_y$、$w_c$、$w_{c'}$表示可学习参数、即用于训练的参数，$c'_t$表示正文的上下文信息，$c_t$表示标题的上下文信息，$d_t$表示解码隐状态，$y_{t-1}$表示已生成的文本的关键数据，σ表示S型生长曲线(sigmoid)函数、即非线性映射函数。

[0146] 在步骤104中，对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布。

[0147] 生成的事件摘要可能是新产生的词语，即可以不是正文中的已存在的词语。考虑到需要新生成的词语，因此对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布，以根据词汇分布确定词汇表中的新词。为了节约计算量，由于标题的词语比较少，可以对标题进行词汇表的映射处理，生成标题的词汇分布。

[0148] 在一些实施例中，对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布，包括：将标题的上下文信息、与解码隐状态进行拼接处理，得到拼接数据；对拼接数据进行第一线性映射处理，得到第一线性映射结果；对第一线性映射结果进行第二线性映射处理，得到第二线性映射结果；对第二线性映射结果进行词汇表的非线性映射处理，生成标题的词汇分布。

[0149] 为了得到标题的词汇分布，可以通过指针生成网络中的解码器(单向长短时记忆

网络)将标题的上下文信息、与解码隐状态进行拼接处理,得到拼接数据,并对拼接数据依次进行两层线性层,即先对拼接数据进行第一线性映射处理,得到第一线性映射结果,再对第一线性映射结果进行第二线性映射处理,得到第二线性映射结果,其中,第一线性映射与第二线性映射的参数数值不同,最后对第二线性映射结果进行词汇表的非线性映射处理,从而生成标题的词汇分布,其中非线性映射可以是softmax函数。

[0150] 在步骤105中,对文本的注意力分布、以及标题的词汇分布进行融合处理,得到文本的关键数据。

[0151] 在服务器得到文本的注意力分布、以及标题的词汇分布后,可以对文本的注意力分布、以及标题的词汇分布进行融合处理,得到文本的关键数据,使得关键数据考虑到复制的标题信息、复制的正文信息以及词汇表信息。

[0152] 在一些实施例中,对文本的注意力分布、以及标题的词汇分布进行融合处理,得到文本的关键数据,包括:确定对应词汇分布的第一生成权重、以及对应注意力分布的第二生成权重;确定标题的词汇分布与第一生成权重的第三乘积,并确定文本的注意力分布与第二生成权重的第四乘积;将第三乘积与第四乘积的求和结果确定为文本的候选关键数据分布;将候选关键数据分布中的最大概率对应的候选关键数据,确定为文本的关键数据。

[0153] 为了得到文本的关键数据,首先需要确定对应词汇分布的第一生成权重、以及对应注意力分布的第二生成权重,以便确定标题的词汇分布与第一生成权重的第三乘积,并确定文本的注意力分布与第二生成权重的第四乘积,从而将第三乘积与第四乘积的求和结果确定为文本的候选关键数据分布,并将候选关键数据分布中的最大概率对应的候选关键数据,确定为文本的关键数据。

[0154] 在一些实施例中,确定对应词汇分布的第一生成权重,包括:对标题的上下文信息、解码隐状态、已生成的文本的关键数据以及可学习参数进行非线性映射处理,得到对应词汇分布的第一生成权重。

[0155] 作为示例,对应词汇分布的第一生成权重是根据标题的上下文信息、解码隐状态、已生成的文本的关键数据确定的,即对标题的上下文信息、解码隐状态、已生成的文本的关键数据以及可学习参数进行非线性映射处理,得到对应词汇分布的第一生成权重,其中,第一生成权重的计算公式为 $p_{gen}=\sigma(w_c^* c_t + w_d^* d_t + w_y^* y_{t-1} + b_{gen})$,$w_c^*$、$w_d^*$、$w_y^*$、$b_{gen}$表示可学习参数、即用于训练的参数,$c_t$表示标题的上下文信息,$d_t$表示解码隐状态,$y_{t-1}$表示已生成的文本的关键数据,σ表示S型生长曲线(sigmoid)函数、即非线性映射函数。

[0156] 在步骤106中,将文本的关键数据进行组合处理,得到对应文本的事件摘要。

[0157] 在服务器得到准确的关键数据后,可以对文本的关键数据进行组合处理,得到对应文本的事件摘要。例如,在依次得到关键数据$k_1,k_2,...,k_i,...,k_n$后,可以按照生成的顺序组合这些关键数据,得到事件摘要$F=[k_1,k_2,...,k_i,...,k_n]$,其中,$k_i$表示第i个生成的关键数据,n表示关键数据的总数,i、n为自然数。

[0158] 至此已经结合本发明实施例提供的服务器的示例性应用和实施,说明本发明实施例提供的文本事件摘要的生成方法,下面继续说明本发明实施例提供的文本事件摘要的生成装置555中各个模块配合实现文本事件摘要的生成的方案。

[0159] 编码模块5551,用于对文本中的标题进行编码处理,得到所述标题的注意力权重以及上下文信息;注意力模块5552,用于对所述文本中的正文进行注意力处理,得到所述正

文的注意力权重;第一融合模块5553,用于将所述标题的注意力权重、以及所述正文的注意力权重进行融合处理,得到所述文本的注意力分布;映射模块5554,用于对所述标题的上下文信息进行词汇表的映射处理,生成所述标题的词汇分布;第二融合模块5555,用于对所述文本的注意力分布、以及所述标题的词汇分布进行融合处理,得到所述文本的关键数据;组合模块5556,用于将所述文本的关键数据进行组合处理,得到对应所述文本的事件摘要。

[0160] 在一些实施例中,所述编码模块5551还用于对所述文本中的标题进行隐状态转换处理,得到所述标题的隐状态;对所述标题的隐状态进行注意力处理,得到所述标题的注意力权重;基于所述标题的注意力权重,对所述标题的隐状态进行加权求和,得到所述标题的上下文信息。

[0161] 在一些实施例中,所述编码模块5551还用于将所述文本中的标题进行词向量转换处理,得到所述标题的词向量;对所述标题的词向量进行前向编码处理,得到对应所述标题的前向隐向量;对所述标题的词向量进行后向编码处理,得到对应所述标题的后向隐向量;将所述前向隐向量以及所述后向隐向量进行拼接处理,得到所述标题的隐状态。

[0162] 在一些实施例中,所述编码模块5551还用于对所述标题的隐状态、解码隐状态以及可学习参数进行双曲正切处理,得到处理结果;对所述处理结果进行非线性映射处理,得到所述标题的注意力权重。

[0163] 在一些实施例中,所述文本事件摘要的生成装置555还包括:筛选模块5557,用于对所述文本中的正文进行筛选处理,得到简化的正文序列;所述注意力模块5552还用于对所述简化的正文序列进行隐状态转换处理,得到所述正文序列的隐状态;对所述正文序列的隐状态进行注意力处理,得到所述正文的注意力权重。

[0164] 在一些实施例中,所述筛选模块5557还用于对所述文本中的正文进行句子粒度提取处理,得到所述正文中的目标句子;对所述正文进行词粒度提取处理,得到所述正文中的目标词语;将所述目标词语对齐到所述目标句子中,得到所述目标句子中未被对齐的词语;基于所述目标句子中未被对齐的词语的词性,对所述目标句子中的词语进行过滤处理,得到简化的正文序列。

[0165] 在一些实施例中,所述筛选模块5557还用于对所述文本中的正文进行分句处理,得到多个候选句子;对所述候选句子进行向量转换处理,得到所述候选句子的句子向量;确定所述候选句子的句子向量、与所述标题的句子向量的第一相似度,确定所述候选句子的句子向量、与已提取句子的句子向量的第二相似度;将所述第一相似度以及所述第二相似度进行加权求和,并对加权求和结果进行映射处理,得到所述正文中的目标句子。

[0166] 在一些实施例中,所述筛选模块5557还用于对所述候选句子进行词向量转换处理,得到所述候选句子的词向量;基于所述词向量的词频以及逆文本频率指数,确定所述词向量的权重;基于所述词向量的权重,对所述候选句子的词向量进行加权平均处理,得到所述候选句子的句子向量。

[0167] 在一些实施例中,所述筛选模块5557还用于对所述文本中的正文进行分词处理,得到对应所述正文的词语;根据所述词语的词性,对所述对应所述正文的词语进行过滤处理,得到多个所述正文的候选词语;将所述多个所述正文的候选词语组合成所述候选词语的序列,并基于所述候选词语的序列,构建候选目标词图;基于所述候选目标词图中节点权重,确定所述正文中的目标词语。

[0168] 在一些实施例中，所述筛选模块5557还用于将所述候选词语的序列中的候选词语确定为所述候选目标词图的节点；当任意两节点在所述候选词语的序列中的距离小于或者等于距离阈值时，连接所述任意两节点的边；将所述两节点在所述序列中出现的频率确定为所述边的节点权重；根据所述节点、所述边以及所述节点权重，构建所述候选目标词图。

[0169] 在一些实施例中，所述第一融合模块5553还用于确定对应所述标题的第一融合权重以及对应所述正文的第二融合权重；确定所述标题的注意力权重与所述第一融合权重的第一乘积、以及所述正文的注意力权重与所述第二融合权重的第二乘积，并将所述第一乘积与所述第二乘积的求和结果确定为所述文本的注意力分布。

[0170] 在一些实施例中，所述文本事件摘要的生成装置555还包括：处理模块5558，用于基于所述正文的注意力权重，对正文的隐状态进行加权求和，得到所述正文的上下文信息；所述第一融合模块5553还用于对所述正文的上下文信息、所述标题的上下文信息、解码隐状态、已生成的所述文本的关键数据以及可学习参数进行非线性映射处理，得到对应所述标题的第一融合权重。

[0171] 在一些实施例中，所述映射处理5554还用于将所述标题的上下文信息、与解码隐状态进行拼接处理，得到拼接数据；对所述拼接数据进行第一线性映射处理，得到第一线性映射结果；对所述第一线性映射结果进行第二线性映射处理，得到第二线性映射结果；对所述第二线性映射结果进行词汇表的非线性映射处理，生成所述标题的词汇分布。

[0172] 在一些实施例中，所述第二融合模块5555还用于确定对应所述词汇分布的第一生成权重、以及对应所述注意力分布的第二生成权重；确定所述标题的词汇分布与所述第一生成权重的第三乘积，并确定所述文本的注意力分布与所述第二生成权重的第四乘积；将所述第三乘积与所述第四乘积的求和结果确定为所述文本的候选关键数据分布；将所述候选关键数据分布中的最大概率对应的候选关键数据，确定为所述文本的关键数据。

[0173] 在一些实施例中，所述第二融合模块5555还用于对所述标题的上下文信息、解码隐状态、已生成的所述文本的关键数据以及可学习参数进行非线性映射处理，得到对应所述词汇分布的第一生成权重。

[0174] 这里需要指出的是：以上涉及装置的描述，与上述方法描述是类似的，同方法的有益效果描述，不做赘述，对于本发明实施例所述装置中未披露的技术细节，请参照本发明方法实施例的描述。

[0175] 本发明实施例还提供一种存储有可执行指令的存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本发明实施例提供的文本事件摘要的生成方法，例如，如图3-6示出的文本事件摘要的生成方法。

[0176] 在一些实施例中，存储介质可以是FRAM、ROM、PROM、EPROM、EE PROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

[0177] 在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言(包括编译或解释语言，或者声明性或过程性语言)来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

[0178] 作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言(H TML，Hyper Text

Markup Language)文档中的一个或多个脚本中,存储在专用于所讨论的程序的单个文件中,或者,存储在多个协同文件(例如,存储一个或多个模块、子程序或代码部分的文件)中。

[0179] 作为示例,可执行指令可被部署为在一个计算设备上执行,或者在位于一个地点的多个计算设备上执行,又或者,在分布在多个地点且通过通信网络互连的多个计算设备上执行。

[0180] 下面,将说明本发明实施例在一个实际的应用场景中的示例性应用。

[0181] 文本事件摘要生成模型旨在将文章所包含的核心事件以精炼的语言概括描述。由于序列到序列(S2S,Sequence to Sequence)模型的未登录词和不可控性问题,限制了神经摘要模型的工业界应用。同时,由于网络新闻文章的多样性、多源性、训练语料少,进一步加大了事件级别短摘要生成的难度。准确地提炼核心事件信息和合适的语言组织是一个重要的技术点。

[0182] 相关技术中,句子压缩模型可以通过截断、整数线性规划、指针生成网络对标题进行句子压缩,但是句子压缩模型仅通过简单的规则截断、以及将句子压缩转化为整数线性规划问题来提炼句子的关键信息,由于句子压缩模型的精度不够,生成的句子的通顺性、完整性也相对较差;神经句子压缩模型可以融合神经网络的句子压缩方法,能够有效的提升压缩后句子的完整性和通顺性,基于删减式的序列到序列(S2S-del)模型和基于生成式的序列到序列(S2S-ge n)模型都在单句信息压缩上表现优异,但是网络文章通常是多源的(包括正文、标题),神经句子压缩模型在网络文章压缩上表现不好,同时神经句子压缩模型的效果高度依赖训练数据的大小;文本摘要生成模型旨在对文本生成几句话的长摘要,在对于事件要素的提取和概括上表现不足,并且文本事件摘要生成任务需要更加高层次的事件概括,文本摘要生成模型无论是生成式还是抽取式的,都存在着摘要信息冗余的问题;商品标题压缩模型旨在对电商网站上的商品标题进行有效压缩,这类模型主要是针对特定领域上的压缩,而事件通常是开放性的,具有开放域的特点,同时网络文章的标题,特别是自媒体文章标题更加的多样化、不规则化,商品标题压缩模型并不能很好解决文章事件摘要生成任务的具体效果要求。

[0183] 本发明实施例为了解决上述问题,以指针生成网络为基础模型,同时设计了一个多源(包括正文、标题)网络将文章的标题信息和正文信息融合,另一方面引入常用词词库(词汇表)来保证事件描述的完整性和通顺性。最后,基于海量的用户搜索日志,文本事件摘要生成模型进行端到端的预训练,可以在开放域场景下的准确识别事件要素。

[0184] 本发明实施例可应用新闻事件查询词(query)搜索和展示场景中。如图7所示,图7是本发明实施例提供的新闻事件query展示示意图,图7所示的是热点事件文章自动化、高时效性地生成事件的短摘要,服务器或者终端可以将时新的文本(热点事件)对应的事件摘要呈现给用户,使得用户根据呈现的事件摘要,以便用户了解最新的文本事件、即新闻,并对应获得与事件摘要对应的完整文本,包括标题和正文。例如文本的标题为“李小明与蒋小夫合作……”、文本的正文为“李小明于2019年与……,蒋小夫也于2019年……”,则对应文本的事件摘要为“李小明与蒋小夫”,并在服务器或者终端的显示界面701上呈现事件摘要“李小明与蒋小夫”。如图8所示,图8是本发明实施例提供的新闻事件query搜索提示示意图,在用户进行搜索时,用户输入部分检索信息后,服务器或者终端可以根据该部分检索信息对数据库中的事件摘要进行匹配,并将匹配成功的事件摘要呈现给用户,为用户提供热

点事件query的推荐，使得用户根据部分检索信息，得到完整的检索信息、即准确的事件摘要，并根据简短的事件摘要，对应获得与事件摘要对应的完整文本，包括标题和正文。例如，当用户输入“欧洲杯”，则在服务器或者终端的显示界面801显示“欧洲杯抽签”、“欧洲杯2020”、“欧洲杯预选赛”、“欧洲杯赛程2019赛程表”、“欧洲杯海选”，以为用户提供热点事件query的推荐，用户点击任意的“欧洲杯抽签”、“欧洲杯2020”、“欧洲杯预选赛”、“欧洲杯赛程2019赛程表”、“欧洲杯海选”，都可以阅读对应的完整文本。基于文章自动化生成事件短摘要，一方面可以节省运营的人力成本，同时相对人工运营编辑具有高时效性和多样性的特点。

[0185] 对于热点事件的监控和外显展示一直是新闻场景下的最重要任务，基于文章自动化生成事件短摘要能够极大地节省人力成本，同时自动化的生成流程能够提升时效性和多样性。文本事件摘要生成模型旨在对于单篇文章，通过文章标题信息的提炼，以及正文事件要素信息的补充，以一种基于注意力机制的融合方法，生成通顺、完整性强的事件短摘要，同时通过大量的预训练数据提升文本事件摘要生成模型对于关键事件要素的识别能力，能够对海量的热点文章高效地生成事件短摘要、并且应用于搜索查询词推荐(QS)&相关搜索提示(HINT)的时新性打造、热门话题短描述、时新query匹配等多个场景中。

[0186] 另外，文本事件摘要生成模型不仅可以根据提供的文章的标题和正文信息，生成事件短描述，同时也支持单纯从文章标题中生成事件短描述。如表1所示，其中，细分来源是指标题所来源的地址：

[0187] 表1

[0188]

标题	发表时间	摘要	摘要质量等级	热度	质量分	细分来源
厦门警方披露XX案信息:长期隐姓埋名、使用虚假身份	201912011447	厦门警方披露XX	3	5.286	0.5	hot_topic
AA地毯式扫楼，只为拿绩效?	201912011133	AA地毯式扫楼	3	4.651	0.35	hot_topic
伦敦桥恐袭案:BB宣称对此负责—遇难者为剑桥大学毕业生	201912011037	伦敦桥恐袭案	3	3.341	0.58	hot_topic
从多伦多万锦起飞的空难,飞行员一家五口全部遇难	201911301933	从多伦多万锦起飞的空难	3	3.176	0.6	hot_topic
贵州织金煤矿事故搜救结束7人遇难	201912011434	贵州织金煤矿事故	3	2.469	0.63	hot_topic
CC评DD去世:一边是痛心，一边是担心	201912011348	DD去世	3	2.469	0.48	hot_topic
EE第2次输FF暴露隐患,奥运单打名额仍有变,许昕—优势巨大	201912011333	EE第2次输FF	3	2.367	0.57	hot_topic

[0189] 下面说明本发明实施例提供的文本事件摘要的生成方法，参见图9和图10，图9是本发明实施例提供的文本事件摘要的生成方法的流程示意图，图10是本发明实施例提供的文本事件摘要生成模型的结构示意图，示出了文本事件摘要的生成流程，下面结合图9示出的步骤进行说明。

[0190] 如图9所示，将文本的正文长序列输入至层次交互式内容信息提取器中，层次交互

式内容信息提取器可以对正文长序列进行提取处理，得到正文事件要素序列(简化的正文序列)，并将正文事件要素序列输入至指针生成网络中的双向LSTM编码器进行编码处理，并将文本的标题短序列输入至指针生成网络中的双向LSTM编码器进行编码处理，将处理结果进行注意力融合，得到文本的多源注意力分布，并通过单向LSTM解码器对多源注意力分布进行解码处理，对输出结果进行质量控制，得到事件短描述。

[0191] 如图10所示，文本事件摘要生成模型采用指针生成网络对标题以及正文进行编码等处理。其中，双向LSTM编码器可以对正文或者标题进行编码，在相应的位置产生顺序隐藏状态(前向隐向量$\overrightarrow{h_1},...,\overrightarrow{h_i},...,\overrightarrow{h_N}$和后向隐向量$\overleftarrow{h_1},...,\overleftarrow{h_i},...,\overleftarrow{h_N}$，其中i、N表示自然数)，将前向隐向量以及后向隐向量进行拼接处理，得到对应正文或者标题的隐状态(例如，$h_i=[\overrightarrow{h_i},\overleftarrow{h_i}]$)，其中，每个隐状态的计算过程如公式(1)所示：

[0192] $h_n=f(d_n,h_{n-1},\theta)$ (1)

[0193] 其中，h_n表示第n个输入词(标题或者正文)的隐状态，f()函数表示一个非线性函数，d_n表示解码器隐状态。h_n是双向LSTM编码器的输出，同时也是接下来单向LSTM解码器的新初始状态。

[0194] 下面，说明双向LSTM编码器对标题的处理过程：

[0195] 如图10所示，通过双向LSTM编码器对标题进行编码处理，得到标题的隐状态h＝{$h_1...h_i$}，并通过注意力机制对标题的隐状态进行处理，得到标题的注意力权重，其计算公式如公式(2)所示：

[0196] $a_{t,i}=\mathrm{softmax}(v^T\tanh(W_hh_i+W_dd_t+b))$ (2)

[0197] 其中，$a_{t,i}$表示标题的一个注意力权重，v、W_h、W_d、b表示可学习参数、即用于训练的参数，h_i表示标题的一个隐状态，d_t表示解码器的隐状态，softma x函数表示逻辑回归函数、即非线性映射函数。

[0198] 当从词汇表中生成词语时，由标题的注意力权重$a_{t,i}$和h_i加权和，得到标题的上下文向量，如公式(3)所示：

[0199] $$c_t=\sum_{i=1}^{N}a_{t,i}h_i \quad (3)$$

[0200] 其中，$a_{t,i}$表示标题的注意力权重，h_i表示标题的隐状态，c_t表示标题的上下文向量，N表示隐状态的数量。c_t可以看成是该时间步时通读了标题的固定尺寸的表征(t时刻通读了上下文之后标题的信息表征)。

[0201] 将c_t和d_t(t时刻的解码器隐状态)经过两层线性层得到标题的词汇分布P_{vocab}(与解码器状态d_t连接，并通过两个线性层进行馈送，以生成词汇分布P_{vocab}，P_{vocab}表示词汇表中所有单词的概率分布，或者最终经过softmax多分类产生的各个单词的概率分布)，[]表示拼接，标题的词汇分布P_{vocab}的计算公式如公式(4)所示：

[0202] $$\begin{aligned}P_{vocab}&=P(y_t\mid y<t,S_{title};\theta)\\&=\mathrm{softmax}\left(W_2\left(W_1\left[d_t,c_t\right]+b_1\right)+b_2\right)\end{aligned} \quad (4)$$

[0203] 其中，S_{title}表示标题序列，θ、W_1、b_1、W_2、b_2表示模型参数。

[0204] 下面，结合图11说明层次交互式内容信息提取器对正文的处理过程，图11是本发明实施例提供的层次交互式内容信息提取器的流程示意图：

[0205] 尽管指针生成网络根据标题可以生成事件摘要，特别是当标题是规则的，且包含完整的事件信息时，但是根据标题得到的事件摘要仍然会丢失一些只出现在正文中的关键事件。例如，标题为“她为什么在25岁自杀?韩国娱乐中的舆论压力是最后一根稻草”，关键的一步是要从正文中正确地提取出谁是“她”。因此，如图10所示，除了通过双向LSTM编码器对标题进行处理之外，还引入了层次交互式内容信息提取器和另一个双向LSTM编码器来分层提取正文中有价值的辅助信息。

[0206] 句子粒度提取：本发明实施例采用最大边界相关法或者最大边缘相关(MMR，Maximal Marginal Relevance)算法对正文信息进行句子粒度的提取。MMR算法是一种文章长摘要的提取方法，其中MMR算法综合考虑了所提取句子跟标题的关联程度以及各个所提取句子之间的不同程度，这样通过MMR算法所提取的关键句子，能够保证提取的句子为关键句的同时，还能保证句子之间的多样性。通过MMR提取关键句(目标句子)的算法如公式(5)所示：

[0207] $MMR=\arg\max_{D_i\in R_c\backslash R_s}[\eta Sim1(D_i,S_{title})+$ (5)

[0208] $(1-\eta)\max_{D_j\in R_s}Sim2(D_i,D_j)]$

[0209] 其中，R_c表示抽取的候选句子的集合，S_{title}表示标题序列，R_s表示已提取的句子的集合，$R_c\backslash R_s$表示未被提取的句子集合，Sim1表示候选句子D_i的句子向量与标题的句子向量的相似度，Sim1表示候选句子的句子向量D_i与已提取句子D_j的句子向量的相似度，该相似度可以是余弦相似度，η表示对应Sim1的权重，1-η表示对应Sim2的权重。

[0210] 为了得到句子间的相似度Sim1和Sim2，其计算的方法为：1)将正文进行词性过滤，信息增益低的词性(例如，介词、副词等)将会被过滤掉；2)将剩下的词通过word2vec得到每个词的词向量表达，不在词表内的词则被过滤掉；3)对每个词进行tf-idf计算，最终以tf-idf作为权重计算各个词向量的加权和，得到候选句子的句子向量，计算的公式如公式(6)所示：

[0211] $vec(D)=1/L\sum[tf\text{-}idf(w_i)*word2vec(w_i)]$ (6)

[0212] 其中，L表示句子的长度，w_i表示句子D中的词语，tf表示w_i的词频，idf表示w_i的逆文本频率指数，$word2vec(w_i)$表示w_i的词向量

[0213] 词粒度提取：通过TextRank对正文进行关键词提取，同时，对于也会通过命名实体识别，针对关键的事件要素，例如时间、人名、地名、机构名等进行加权，凸显出正文中关键的事件要素信息。TextRank对正文进行关键词提取的过程如下：

[0214] (1)对文本中的正文进行分词处理，得到对应正文的词语。

[0215] (2)根据词语的词性，对对应正文的词语进行过滤处理，过滤掉停用词、介词等，得到多个正文的候选词语，将多个正文的候选词语组合成候选词语的序列$T=[s_1,s_2,\ldots,s_m]$。

[0216] (3)基于候选词语的序列，构建候选目标词图$G=(V,E)$，其中，V表示节点(由候选词语组成)，采用共现关系构建边E，两节点之间存在边且仅当节点对应的词语在距离为K的窗口出现，K表示窗口大小、即最多共现K个词语。

[0217] (4)根据TextRank，迭代传播个节点的权重，即节点出现的频率。

[0218] (5)基于候选目标词图中节点权重，确定正文中的目标词语，对候选目标词图中节点权重进行降序排序，将前M个节点权重对应的节点的候选词语确定为正文中的目标词语，

其中,M可以是预先设定的数值,M为自然数。

[0219] 对齐、剪枝:由句子粒度和词粒度得到正文中的目标句子和目标词语,目标句子属于语法比较通顺的序列,但是会存在信息冗余的问题,对于目标词语,虽然信息比较精炼,但是失去了语法信息和语序信息。因此,在这一步,可以将目标词语对齐到所提取的目标句子中,然后根据词性对目标句子进行无用词过滤(介词、助词等)。原来目标句子中,能够被目标词语映射中以及没被词性过滤步骤过滤的词,将会组成正文事件要素信息序列、即简化的正文序列。

[0220] 示例性地,如图12所示,图12是本发明实施例提供的对齐、剪枝的效果示意图,当获得目标句子为“今年11号台风“白鹿”正逐渐逼近中国,并将再次登陆福建晋江沿海地区...”,目标词语为“台风”、“白鹿”、“福建”、“登陆”以及“暴雨”,则将目标词语对齐到目标句子上,并进行剪枝处理,得到组成正文事件要素信息序列“11号台风“白鹿”登陆晋江福建,暴雨预警…”。

[0221] 下面,说明双向LSTM编码器对简化的正文序列的处理过程:

[0222] 如图10所示,通过双向LSTM编码器对简化的正文序列进行编码处理,得到正文的隐状态$h'=\{h'_1...h'_j\}$,并通过注意力机制对正文的隐状态进行处理,得到正文的注意力权重a'_t,其a'_t的计算公式与公式(2)类似。

[0223] 由正文的注意力权重a'_t和h'加权和,得到正文的上下文信息c'_t,其c'_t的计算公式与公式(3)类似。

[0224] 下面,说明文本和标题的融合过程:

[0225] 为了融合上述提取的正文信息和标题信息,一种直观的方法是将正文信息和标题信息串联起来,生成关键事件摘要。然而,这样方法缺乏处理标题和正文之间动态关系的灵活性,因为有时标题在生成事件摘要过程中起着关键作用,但有时正文在生成事件摘要过程中起着关键作用。

[0226] 本发明实施例通过引入融合权重η,将标题的注意力权重和正文的注意力权重合并为文本的多源注意力分布,多源注意力分布的计算公式如公式(7)所示:

[0227]
$$\begin{aligned}\hat{P}_{multi}(w) &= P(y_t = w \mid S_{title}, S_{klg}, y < t) \\ &= \eta \sum_{i:w_i=w} a_{ti} + (1-\eta) \sum_{j:w_j=w} a'_{tj}\end{aligned} \qquad (7)$$

[0228] 其中,$\hat{P}_{multi}(w)$表示多源注意力分布,y_t表示解码器t时刻的输出,S_{title}表示标题,S_{klg}表示简化的正文序列,η表示对应标题的融合权重,1-η表示对应正文的融合权重,a_{ti}表示标题的注意力权重,a'_{tj}表示正文的注意力权重。

[0229] 因此,在从输入序列(标题或者正文)中拷贝词汇时,即所得到的的注意力权重将不是计算固定词汇表中的概率分布来生成新词汇,而是通过拷贝的机制,直接从输入序列中复制最终的生成词汇。对应的,这部分得到的多源注意力分布将是输入序列上的概率分布,以便决定选择拷贝生成标题或者正文中的哪一个词。

[0230] 其中,η的计算公式如公式(8)所示:

[0231]
$$\eta = \sigma(w_d^T d_t + w_y^T y_{t-1} + w_c^T c_t + w_{c'}^T c'_t) \qquad (8)$$

[0232] 其中,w_d、w_y、w_c、$w_{c'}$表示可学习参数、即用于训练的参数,c'_t表示S_{klg}的上下文信

息，c_t表示标题的上下文信息，d_t表示解码隐状态，y_{t-1}表示已生成的文本的关键数据，σ表示S型生长曲线(sigmoid)函数、即非线性映射函数。

[0233] 单向LTSM解码器可以通过从标题和正文复制单词来生成简短的事件摘要，也可以从标题的词汇分布来生成简短的事件摘要。其中，对文本的多源注意力分布、以及标题的词汇分布进行融合处理，最终的文本分布、即文本的关键数据，其中，$P_{final}(w)$的计算公式如公式(9)所示：

[0234] $$P_{final}(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen})\hat{P}_{multi}(w) \quad (9)$$

[0235] 其中，$P_{vocab}(w)$表示标题的词汇分布，$\hat{P}_{multi}(w)$表示文本的多源注意力分布，p_{gen}表示对应词汇分布的生成权重，$1-p_{gen}$表示对应源注意力分布的生成权重。

[0236] 其中，p_{gen}的计算公式如公式(10)所示：

[0237] $$p_{gen} = \sigma(w_c^* c_t + w_d^* d_t + w_y^* y_{t-1} + b_{gen}) \quad (10)$$

[0238] 其中，w_c^*、w_d^*、w_y^*、b_{gen}表示可学习参数、即用于训练的参数，c_t表示标题的上下文信息，d_t表示解码隐状态，y_{t-1}表示已生成的文本的关键数据，σ表示S型生长曲线(sigmoid)函数、即非线性映射函数。

[0239] 对于构建的文本事件摘要生成模型的损失函数，由于命名实体是事件摘要的核心元素，错误的实体将导致严重的错误。因此，本发明实施例引入一个实体偏置权重，使每个关键数据共享不同的权重，损失函数的计算公式如公式(11)所示：

[0240] $L_t = -w_{bis}(w^*_t)\log p(y_t = w^*_t | S, y < t)$ (11)

[0241] 其中，w^*_t表示关键数据，$w_{bis}(w^*_t)$表示关键数据w^*_t的实体偏置权重，S表示输入数据(正文、标题)，p表示最终的文本分布中词语的概率。

[0242] 其中，实体偏置权重$w_{bis}(w^*_t)$的计算公式如公式(12)所示：

[0243] $$w_{bis}(w^*_t) = \begin{cases} 1, w^*_t \notin S_E \\ 1 + \frac{1}{|S_E|}(\frac{|S_E|}{|S_T|}), w^*_t \in S_E \end{cases} \quad (12)$$

[0244] 其中，S_E表示实体词集合，$|S_E|$表示实体词的数量，S_T表示标题，$|S_T|$表示标题中词语的数量。非实体词的$w_{bis}(w^*_t)$被设置为1，而实体词增加额外的权重。实体偏置权重$w_{bis}(w^*_t)$在区间$[\frac{1}{|S_T|}, \frac{1}{|S_E|}]$浮动，并且随着p的减小而增加，从而使得文本事件摘要生成模型更加关注错误的实体。特别是，$\frac{1}{|S_E|}$旨在针对多实体情况，避免多实体情况下过高的损失。

[0245] 本发明实施例还可以采用用户搜索日志进行预训练，由于相对于其他的生成任务，文本事件短摘要生成的任务的训练数据相对更少，而且对于网络事件的发生，通常是一个开放域的问题。因此，需要通过辅助数据来提升文本事件摘要生成模型对于事件的提取能力，弥补训练数据不足的问题。由于用户的搜索日志数据，通常是表征用户对于文本的关心点，也就是事件的关键要素。通过在搜索平台拉取的160万文章-搜索query的数据，并对

文本事件摘要生成模型进行端到端的预训练，使得文本事件摘要生成模型在文章正文关键信息提取和解码器识别关键信息上具备了更多的知识背景。由于预训练数据更大、而且涵盖的数据种类更广，因此能够缓解开放域问题带来的oov问题。

[0246] 下面介绍文本事件摘要生成模型：

[0247] 文本事件摘要生成模型主要由三个阶段组成：1)数据预处理；2)事件摘要生成；3)质量控制。

[0248] 1)数据预处理阶段

[0249] 首先导入已经训练好的Word2Vec词表，以便于在提取正文事件要素序列时，将对应的词映射成词向量(128维)。对于分词上，采用以结巴分词为基础、qq分词实体识别相融合的方案进行分词，结巴分词是一种长度优先的分词方案，相对其他分词方案更粗，能够降低生成分词的难度。另一方面，引入qq分词的命名实体识别，可以提升分词的精度，同时保证命名实体能够有效地识别出来。对于文本来源，可以通过新闻性，对一些新闻性差的文章进行过滤，以避免类似鸡汤文章、养生类文章作为文本事件摘要生成模型的输入。最后，对于某些外链的文章，也会通过解析和清洗，从而保证文本的质量。

[0250] 其中，结巴分词包括两个步骤，如下所示：

[0251] A、基于前缀词典(前缀词典是指在统计词典中一个词语最后一个字之前的所有部分的循环)实现高效的词图扫描，生成句子中词语所有可能成词情况所构成的有向无环图(DAG)，例如“财经大学”，其在统计词典中的前缀分别是“财”、“财经”、“财经大”，词语“大学”的前缀是“大”。例如，句子“在财经大学读书”，利用前缀词典进行文本切分，“在”一字没有前缀，只有一种划分方式；“财”一字，则有“财”、“财经”、“财经大学”三种划分方式；“经”一字，也只有一种划分方式；“大”一字，则有“大”、“大学”两种划分方式，通过这样的划分方式，就可以得到每个字开始的前缀词的划分方式。

[0252] B、采用动态规划查找最大概率路径，找出基于词频的最大切分组合。其中，最大概率路径的计算为从某一位置到一定距离后的另外一个位置存在多条路径，即有多种分词的结果，这时，就需要计算出最大概率的路径，从而获得概率最大的切分词结果，由于有向无环图的每个节点都是带权的，权重为前缀词典中每个词的词频，采用动态规划依次到达一个节点，通过得到前面的节点到终点的最大路径概率，最终通过求出最大权重得到切词结果。

[0253] 2)事件摘要生成阶段

[0254] 文章摘要生成阶段具体可以分为训练和预测两个阶段：

[0255] A、训练阶段分为两个部分：预训练和模型微调(fine tune)，在预训练阶段，可以将用户搜索日志中的搜索query作为事件摘要，文本的标题信息和正文信息也同时输入文本事件摘要生成模型，进行端到端的预训练，MMR关键句子提取的η也会得到调整，进而提升文本事件摘要生成模型在正文内容提取时，对于事件要素的提取精度，同时也改进编码时对关键信息的把握。在模型fine tune阶段，采用标注数据进行训练，生成词表、word2vec词表都需要跟预训练保持一致。隐藏状态的维度可以设为256维，对标题编码的最大的编码步数可以设为50，对正文编码的最大编码步数可以设为100，生成词表预训练和fine tune阶段都可以设为5000，学习率可以设为0.15进行学习。在训练阶段，解码器的词输入为标注的事件摘要对应的词语，在预测阶段，解码器对应的输入为上一步所生成的词语。同时，解码

的最大步数可以设为12。

[0256] B、预测阶段:解码可以采用集束搜索(beam search)(一种启发式图搜索算法)进行解码,beam search的大小可以设为8,这是为了在最后质量控制上能够有多个候选,以满足不同场景对事件短摘要长度、实体词、通顺程度的需求。预测时,解码器的输入将不再是标注的事件摘要里面的词语,而是文本事件摘要生成模型上一步所预测的词语。

[0257] 其中,集束搜索是维特比算法的贪心形式,集束搜索使用beam size参数来限制在每一步保留下来的可能性词的数量。假设序列为[a,b,c],beam size选择2,则在生成第1个词的时候,选择概率最大的2个词,例如,当前序列就是a或b;生成第2个词的时候,将当前序列a或b,分别与序列中所有词进行组合,得到新的6个序列aa、ab、ac、ba、bb、bc,然后从其中选择2个概率最高的,作为当前序列,例如ab或bb;不断重复这个过程,直到遇到结束符为止,最终输出2个概率最高的序列。

[0258] 为了生成更流畅的摘要并解决开放领域的问题,可以采用额外的惩罚修改评分函数,其中评分函数的计算公式如公式(13)所示:

[0259] $s(S,Y)=[\log p(Y|S)]/lp(Y)+cp(S;Y)$ (13)

[0260] 其中,s(S,Y)表示评分,并选择概率最大的作为输出,p(Y|S)表示beam s earch解码后所有候选事件摘要的概率值,lp表示长度惩罚,cp表示重复惩罚,Y表示候选的事件摘要,S表示输入(标题或者正文)。

[0261] 其中,cp的计算公式如公式(14)所示,lp的计算公式如公式(15)所示:

[0262] $$cp(S;Y)=\beta(-N+\sum_{i=1}^{N}\max(1.0,\sum_{t=1}^{K}a_{ti})) \quad (14)$$

[0263] $$lp(Y)=\frac{(5+|Y|)^{\alpha}}{(5+1)^{\alpha}} \quad (15)$$

[0264] 其中,N表示S的长度(词数量),K表示解码的步数长度,α和β表示设定调节的参数。

[0265] 3)质量控制

[0266] 由于现实的业务场景中,某些场景,比如热门事件排行榜、热门话题短描述显示上,对于文本生成的事件短摘要精度要求非常高,可以通过卷积神经网络(CNN,Convolutional Neural Networks)判别所生成的摘要是否为好描述,通过卷积神经网络的质量打分,可以在指针生成网络所生成的多个事件摘要中挑选出最好的摘要;另一方面,在这个阶段,也会对输入序列和生成的事件摘要进行命名实体匹配的判断(例如,通顺程度、语法结构等),进而提升所生成事件摘要对于文章关键实体信息提取的准确性。

[0267] 以下说明文本事件摘要生成模型对文本的处理过程:

[0268] 1、读取word2vec词表、qq实体识别模型、摘要生成模型。

[0269] 2、读取、解析待生成的文本。

[0270] 3、基于新闻性、敏感词、语法分析等对部分质量差的文本进行过滤。

[0271] 4、若文本信息只有标题,则通过单源指针生成网络生成事件摘要;若文本信息完整,则通过多源指针网络生成事件摘要:

[0272] A)通过已经预训练并且fine tune的文本事件摘要生成模型,分别采用两个双向LSTM编码器对输入的标题序列和正文序列进行编码,得到隐含表示。

[0273] B)解码阶段,将单向LSTM解码器输出的隐含状态和上一步解码所得到的词作为输

入，进而进行beam search解码，得到最优的8个候选事件摘要。

[0274] 5、根据质量控制模型的打分，以及对输入序列和输出序列的命名实体匹配，进一步提升事件短摘要的生成精度。

[0275] 6、输出文本事件摘要生成模型的识别结果。

[0276] 7、根据实际业务需求，筛选业务场景所需要的事件短摘要。

[0277] 对于数据集，由于文本事件摘要(AES)任务不存在现成的基准数据集，因此本发明实施例创建了一个新闻文章事件摘要(NAES)数据集和多文档事件摘要(MDES)数据集。数据集的所有文章都是从广泛使用的移动社交应用程序中收集的，在该应用程序中，组织或个人用户都可以为发布新闻和文章建立账户。其中，数据集的统计数据如表2所示：

[0278] 表2

[0279]

	NAES	MDES
数据集大小	31326	22002
文本标题平均长度	24.48	25.08
文本正文平均长度	1332.89	1522.52
文本事件摘要平均长度	11.27	8.31
文章簇大小	1	5.81

[0280] 为了评估本发明实施例中事件摘要的性能，可以通过一些基线模型进行验证。其中，基线模型如下所示：

[0281] 1)Trunc模型：最简单的基线模型，词语按其原始顺序进行保存，直到达到长度限制；

[0282] 2)ILP模型：采用一种无监督的方法，依赖于输入序列的预处理(即NER，术语加权)结果；

[0283] 3)Seq2seq-att模型：包含一个两层BLSTM编码器和一个单层LSTM解码器，并关注抽象总结；

[0284] 4)Seq2seq-del模型：通过删除压缩感知，并预测二进制标签；

[0285] 5)Ptr-Net模型：直接使用注意机制作为指针，从输入中选择标记作为输出；

[0286] 6)Ptr-Gen模型：一种混合模型，将Seq2seq-att与指针网络结合在一起；

[0287] 7)E-Pg-T模型：本发明实施例的基线模型，一个以标题序列为输入的文本事件摘要生成模型。

[0288] 8)E-Pg-Concat模型：通过序列拼接合并标题序列和正文的文本事件摘要生成模型；

[0289] 9)MS-E-Pg模型：通过两个单独的编码器以多源方式融合标题序列和正文。

[0290] 本发明实施例可以使用ROUGE作为自动评估指标，该指标通过计算生成的摘要和参考摘要之间的重叠词汇元素来衡量摘要的质量。具体来说，用RUG-1(单克)、RUG-2(双克)和RUG-L(最长公共子序列)的F1得分来评估文本事件摘要生成模型。

[0291] 对于事件摘要，有时可以用不同的方式来描述，可以对生成的事件摘要进行人工评估，以提高质量评估的正确性。由于繁重的评估过程(阅读长新闻文档)，可以从测试集中随机抽取了1000篇文章，并要求五名参与者注释生成的短文的质量。在手动评估过程中考虑三个方面：1)关键事件信息保留，事件信息是否正确保存在事件摘要中；2)可读性，事件

摘要的流畅性，语法是否正确；3)信息量，简短摘要的信息量有多大。

[0292] 对各模型的评估结果如表3所示，其中，Accu表示事件摘要的平均准确度，Read表示事件摘要的平均可读性，Info表示事件摘要的平均信息量：

[0293] 表3

[0294]

方法	NAES					
	RG-1	RG-2	RG-L	Accu	Read	Info
Trunc	39.19	33.11	41.55	16.12%	2.26	2.67
ILP	51.23	37.17	49.68	55.78%	4.06	3.95
Seq2seq-att	56.73	40.15	53.66	59.34%	4.28	3.91
Seq2seq-del	55.31	39.56	50.55	58.32%	3.83	3.79
Ptr-Net	63.78	53.53	57.73	67.26%	4.40	4.21
Ptr-Gen	63.11	53.21	57.62	66.54%	4.51	4.38
E-Pg-T	64.35	55.01	59.41	70.16%	4.56	4.41
E-Pg-Concat	61.32	50.02	57.31	68.33%	4.22	4.43
MS-E-Pg	66.37	55.71	61.88	75.00%	4.51	4.52
MS-E-P（预处理）	67.14	56.07	63.18	77.31%	4.57	4.56
方法	MDES					
	RG-1	RG-2	RG-L	Accu	Read	Info
Trunc	33.22	23.54	35.01	13.31%	2.33	2.78
ILP	47.21	30.31	44.54	46.52%	4.03	3.86
Seq2seq-att	49.66	34.54	46.31	55.34%	4.35	3.99
Seq2seq-del	46.49	34.41	45.11	53.29%	3.86	3.81
Ptr-Net	53.55	42.06	48.21	63.71%	4.32	4.19
Ptr-Gen	53.12	41.89	47.95	63.54%	4.48	4.34
E-Pg-T	55.51	43.06	50.15	64.35%	4.57	4.38
E-Pg-Concat	52.71	4003	48.86	61.05%	4.41	4.40
MS-E-Pg	56.91	43.32	50.94	69.11%	4.45	4.49
MS-E-P（预处理）	58.61	44.81	52.93	71.68%	4.52	4.52

[0295] 根据表3可知，将方法分为三组：传统方法(Trunc、ILP)、Seq2Seq基线方法(Seq2seq-att、Seq2seq-del、Ptr-Net、Ptr-Gen)和本发明实施例的方法(E-Pg-T、E-Pg-Concat、MS-E-Pg、MS-E-P(预处理))。对于自动评估，1)由于Trunc模型无法从文本标题的尾部提取信息量大的词语，Trunc模型在所有指标上表现最差，而作为一种强有力的传统句子压缩方法，ILP模型的性能明显优于Trunc模型，两者的ROUGE得分差距明显；2)每种形式的Seq2Seq变异模型明显优于ILP模型，表明Seq2Seq模型比无监督方法更能模拟编辑的事件短摘要，Seq2seq-del模型表现明显优于传统方法，但仍比Seq2seq-att模型、Ptr-Net模型和Ptr-Gen模型等其他Seq2seq模型差很多，这主要是由于事件短摘要中的重新排序问题；3)E-Pg-T模型具有额外的实体偏置损失和推断损失，比原始的Ptr-Net模型和Ptr-Gen模型能够获得更好的性能，为了利用正文信息，E-Pg-Concat模型以正文序列作为输入，引入注意力融合机制的两个编码器的多源框架可以显著提高性能，与作为一种训练前的学习程序，MS-E-Pg(预训练)在所有ROUGE指标上获得最佳性能。对于手动评估，根据“可读性”度量的结果表明，所有建立在Seq2Seq体系结构(不包括Seq2seq-del)上的模型都可以生成更流畅的摘要，而E-Pg-Concat模型和MS-E-Pg模型之间的差异揭示了多源框架在整合正文时可以保证可读性；另一方面，在“准确性”和“信息量”度量上表明，借助于正文编码器和预处理

过程,MS-E-Pg(预处理)能够更好地保留关键信息。考虑到所有这三个指标,MS-E-Pg(预处理)产生了更易读和信息丰富的事件摘要,证明引入正文编码器和预处理过程的优势。

[0296] 进一步分析本发明实施例中模型的有效性,表4详细分析了MS-E-Pg和基线模型的生成来源,其中“标题”或“正文”显示了摘要中的词在文本标题或正文中的百分比。“内容\标题”是指摘要中包含正文但不包含标题的词的百分比,“词表”是指生成的百分比,不是来自源文本。一般来说,手动编写的事件摘要由标题、正文以及生成的新词(不是来自源文本)组成,其中86.59%的词可以从标题序列中提取,10.85%的词从正文序列中提取,2.56%的新词是手动生成的。对于ILP、Seq2seq-del和Ptr-Net等提取方法,生成的词100%在源文本标题中,而Seq2seq-att的提取方法倾向于生成更多的新词(30.36%),MS-E-Pg基于标题和正文生成目标序列。

[0297] 表4

[0298]

数据	标题	正文	标题\正文	新词
手动摘要	86.59	88.07	10.85	2.56
ILP	100.00	-	-	-
Seq2seq-del	100.00	-	-	-
Seq2seq-att	63.66	72.13	4.98	30.36
Ptr-Net	100.00	-	-	-
Ptr-Gen	95.38	91.22	1.13	3.49
E-Pg-Concat	92.54	83.31	3.56	3.90
MS-E-Pg	90.67	85.44	7.87	1.46

[0299] 如图13所示,图13是本发明实施例提供的模型微调的示意图,只有20K训练数据,该文本事件摘要生成模型的ROUGE-1得分超过66.61。ROUGE得分曲线可分为两个阶段:显著提升阶段和缓慢提升阶段,当训练集的规模从0K增加到20K时,ROUGE得分明显提高,当继续将训练数据从20K增加到30K时,提高趋势会减弱,尤其是在Rouge-L分数上,表示文本事件摘要生成模型可以在一定数量的数据下获得相当好的性能,即现实应用程序的可伸缩性。另外,更多的训练数据有利于生成更流畅的摘要(ROUGE-2),但是当微调数据增加到一定量时,提取关键信息的能力(ROUGE-1)没有多大提高。

[0300] 综上所述,通过本发明实施例具有以下有益效果:

[0301] 1)层次交互式内容提取器分别在句子粒度和词粒度对文章的关键信息进行提取,采用对齐和剪枝的手段,分别弥补了正文句子的信息冗余性以及单纯词粒度上的语序信息缺失,能够有效的整合两种粒度信息提取的优势,对于后面的事件短摘要生成上,能够有效地降低信息提取的难度,同时保留了语序的信息,使得对于正文信息的利用上能够更精准;

[0302] 2)对于标题信息和正文信息的融合上,本发明实施例中的文本事件摘要生成模型采用了基于注意力机制的融合方法,使得编码器解码时以一种数据导向的方式生成事件摘要,即自适应地分别从标题序列和正文序列提取事件信息(关键数据),同时通过在生成词表补充词汇保证了事件表述的完整性;

[0303] 3)为缓解生成类任务对于训练数据大需求,以及提升文本事件摘要生成模型对于开放域的事件提取能力,文本事件摘要生成模型采用用户的搜索日志数据进行预训练,该方法不仅能提升文本事件摘要生成模型的泛化能力,而且能够有效提升事件关键要素的识

别和提炼能力；

[0304] 4)本发明实施例中的解码阶段，文本事件摘要生成模型引入了限制重复词过量生成、保证事件摘要通顺程度、命名实体匹配对齐等技术，有效地保证了所生成的事件摘要的质量；

[0305] 5)文本事件摘要生成模型提出了层次交互式的内容信息提取器、多源编码器、注意力融合机制、搜索日志预训练、beam search解码质量控制等技术，有效地保证了事件短摘要生成的完整性和通顺性，在实际应用中取得较高的精度和召回率。

[0306] 以上所述，仅为本发明的实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和范围之内所作的任何修改、等同替换和改进等，均包含在本发明的保护范围之内。

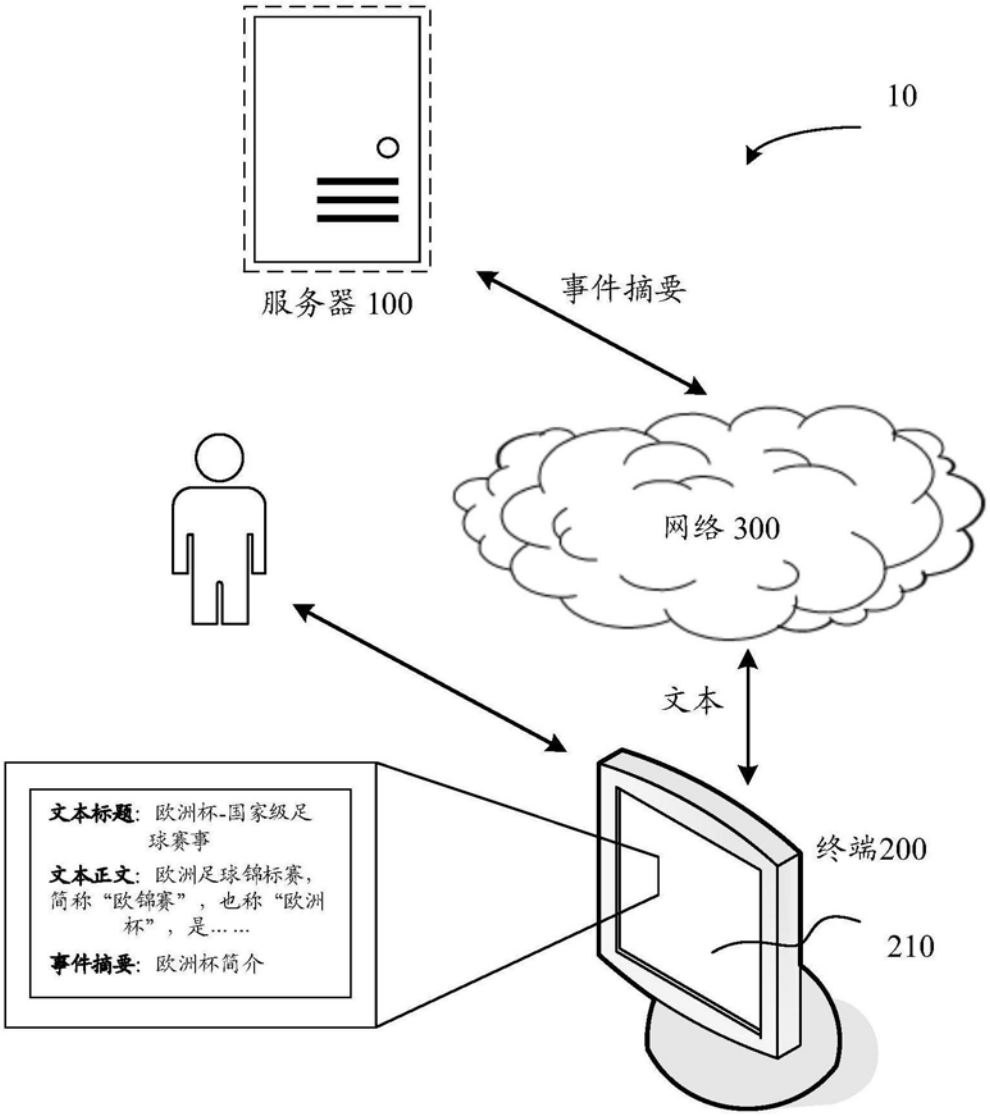

10
服务器 100
事件摘要
网络 300
文本
文本标题：欧洲杯-国家级足球赛事
文本正文：欧洲足球锦标赛，简称“欧锦赛”，也称“欧洲杯”，是……
事件摘要：欧洲杯简介
终端200
210

图1

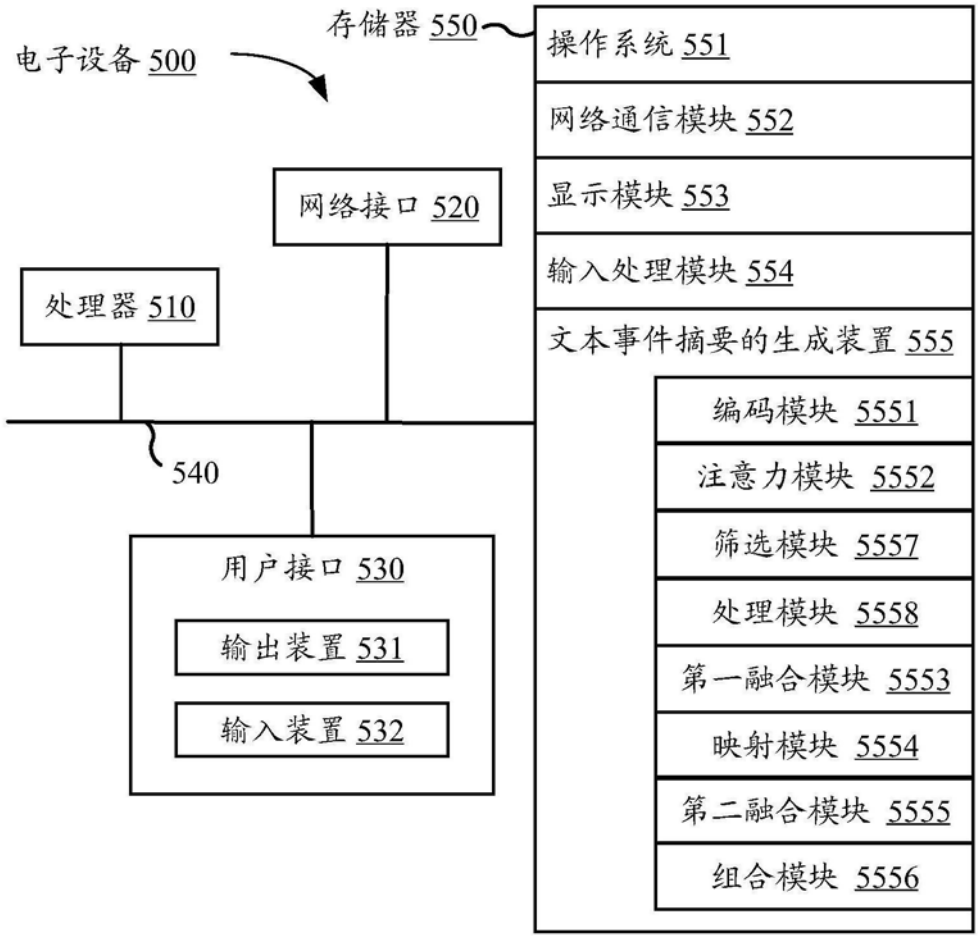

电子设备 500
存储器 550
操作系统 551
网络通信模块 552
显示模块 553
输入处理模块 554
文本事件摘要的生成装置 555
编码模块 5551
注意力模块 5552
筛选模块 5557
处理模块 5558
第一融合模块 5553
映射模块 5554
第二融合模块 5555
组合模块 5556
网络接口 520
处理器 510
540
用户接口 530
输出装置 531
输入装置 532

图2

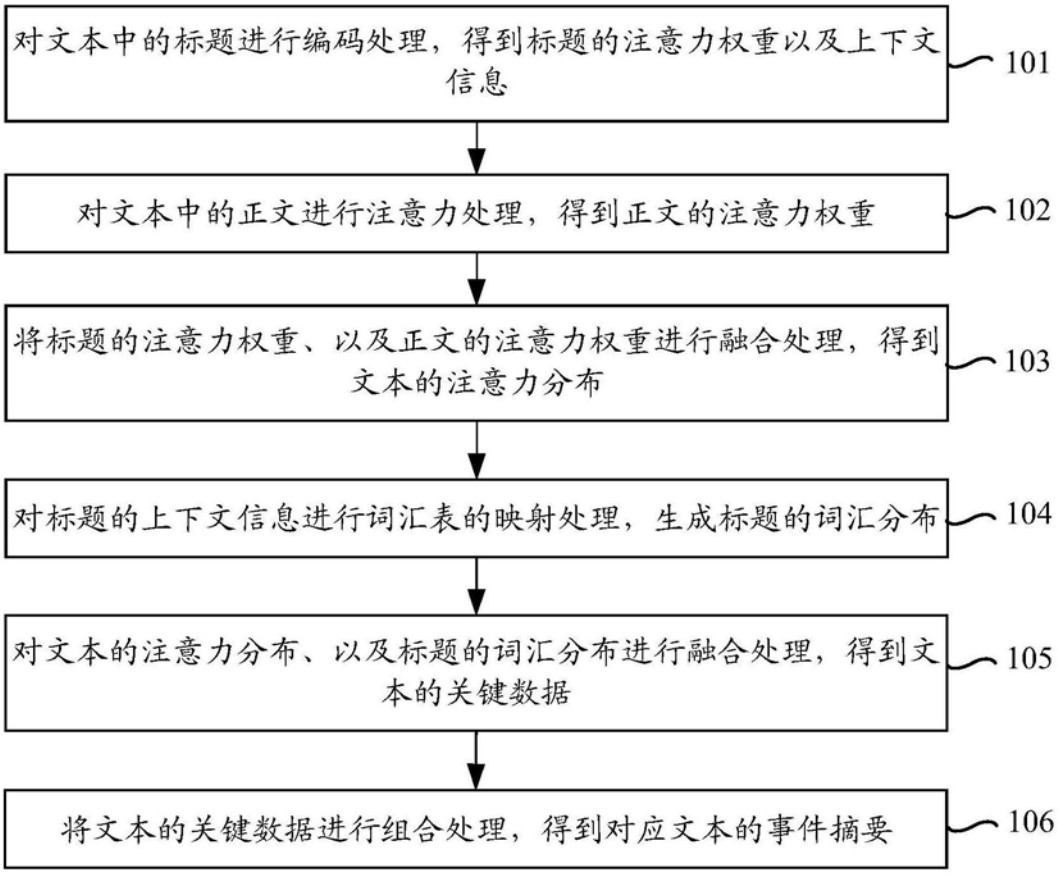

对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息
101
对文本中的正文进行注意力处理，得到正文的注意力权重
102
将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布
103
对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布
104
对文本的注意力分布、以及标题的词汇分布进行融合处理，得到文本的关键数据
105
将文本的关键数据进行组合处理，得到对应文本的事件摘要
106

图3

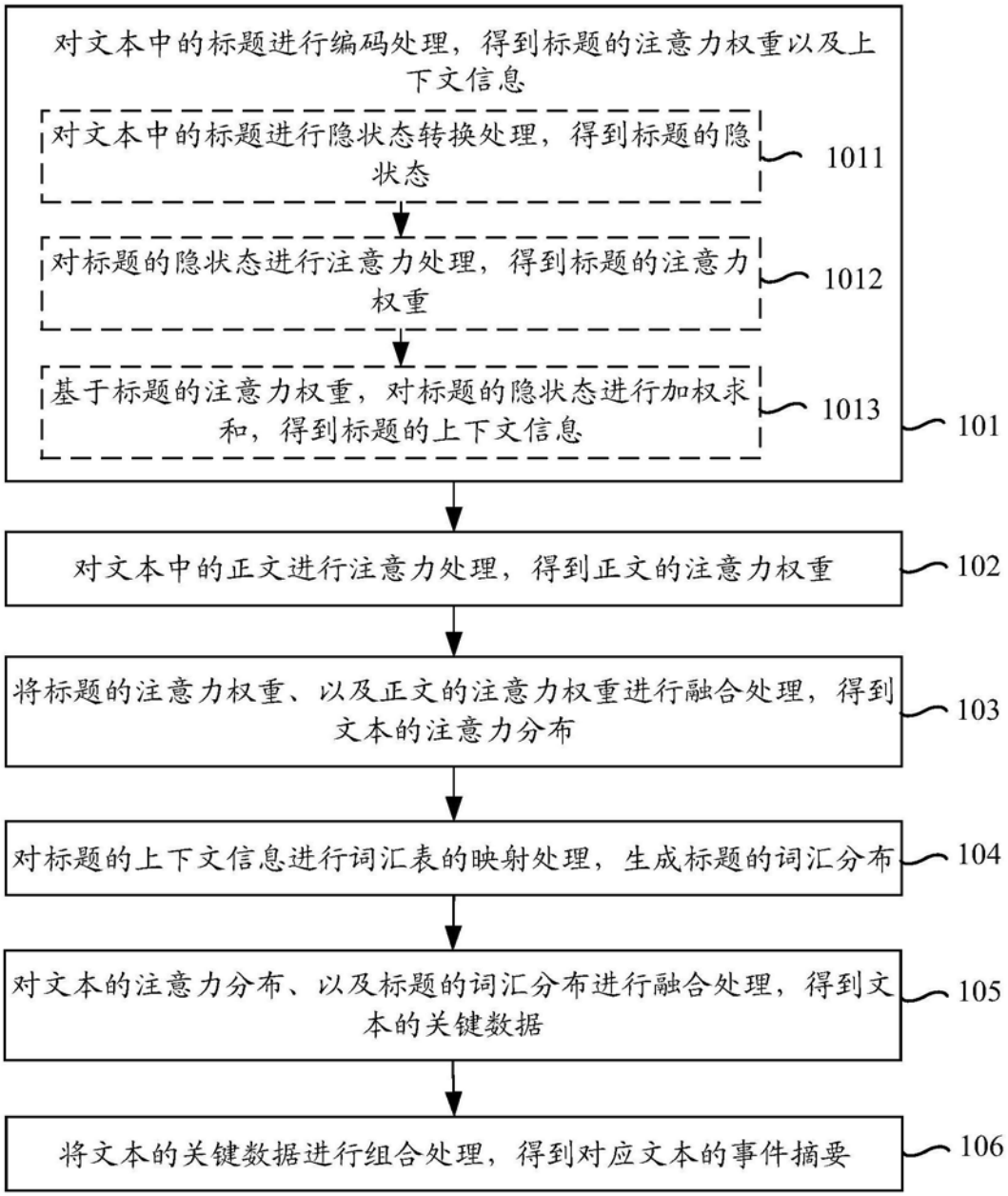

对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息
对文本中的标题进行隐状态转换处理，得到标题的隐状态
1011
对标题的隐状态进行注意力处理，得到标题的注意力权重
1012
基于标题的注意力权重，对标题的隐状态进行加权求和，得到标题的上下文信息
1013
101
对文本中的正文进行注意力处理，得到正文的注意力权重
102
将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布
103
对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布
104
对文本的注意力分布、以及标题的词汇分布进行融合处理，得到文本的关键数据
105
将文本的关键数据进行组合处理，得到对应文本的事件摘要
106

图4

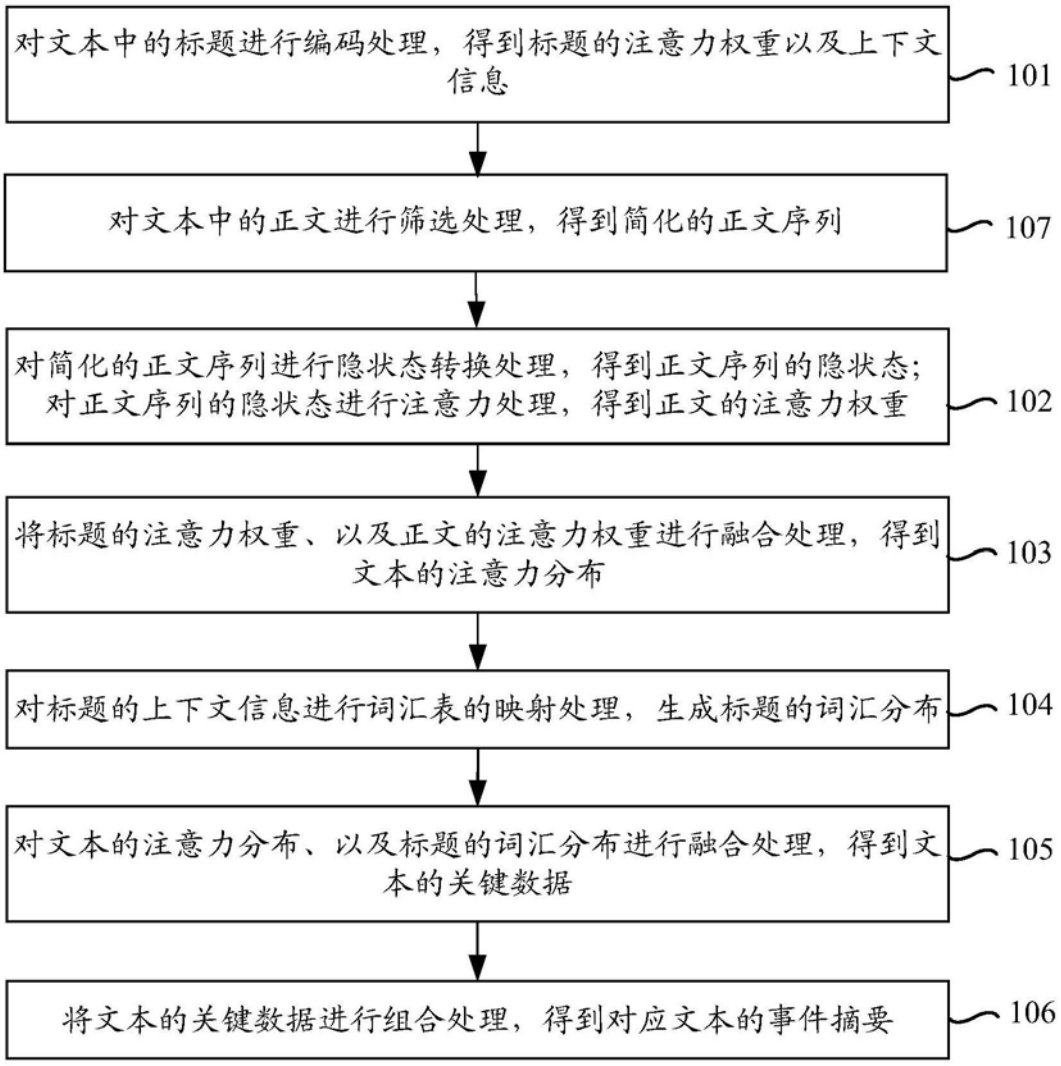

对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息
101
对文本中的正文进行筛选处理，得到简化的正文序列
107
对简化的正文序列进行隐状态转换处理，得到正文序列的隐状态；对正文序列的隐状态进行注意力处理，得到正文的注意力权重
102
将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布
103
对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布
104
对文本的注意力分布、以及标题的词汇分布进行融合处理，得到文本的关键数据
105
将文本的关键数据进行组合处理，得到对应文本的事件摘要
106

图5

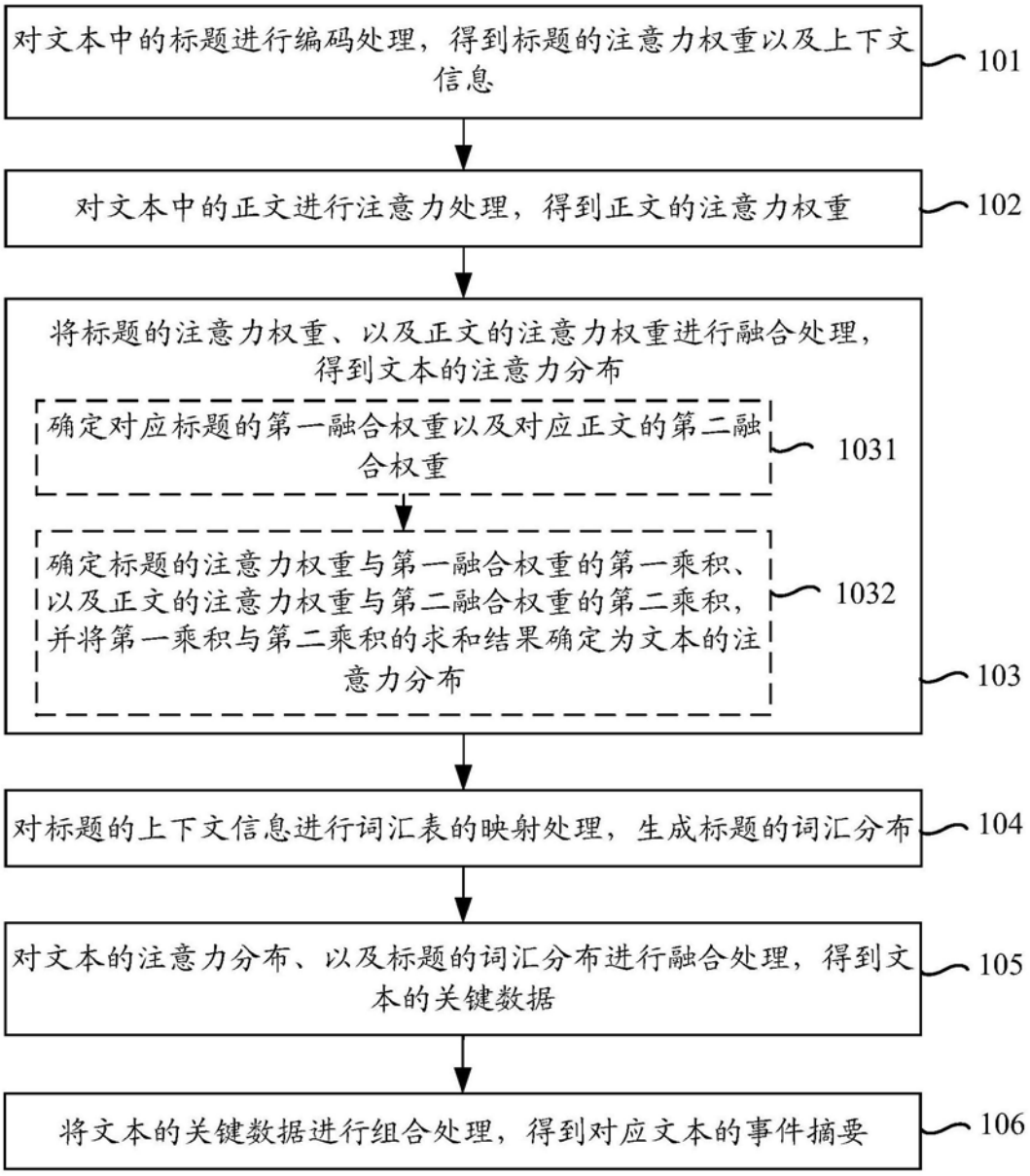

对文本中的标题进行编码处理，得到标题的注意力权重以及上下文信息
101
对文本中的正文进行注意力处理，得到正文的注意力权重
102
将标题的注意力权重、以及正文的注意力权重进行融合处理，得到文本的注意力分布
确定对应标题的第一融合权重以及对应正文的第二融合权重
1031
确定标题的注意力权重与第一融合权重的第一乘积、以及正文的注意力权重与第二融合权重的第二乘积，并将第一乘积与第二乘积的求和结果确定为文本的注意力分布
1032
103
对标题的上下文信息进行词汇表的映射处理，生成标题的词汇分布
104
对文本的注意力分布、以及标题的词汇分布进行融合处理，得到文本的关键数据
105
将文本的关键数据进行组合处理，得到对应文本的事件摘要
106

图6


中国 15:26 88%
搜一搜
搜索
朋友圈 文章 公众号
小程序 音乐 表情
701
微信热点
李小明蒋小夫 热议
天价施救费 热议
12月新规
欧洲杯抽签
海南高院张XX

图7

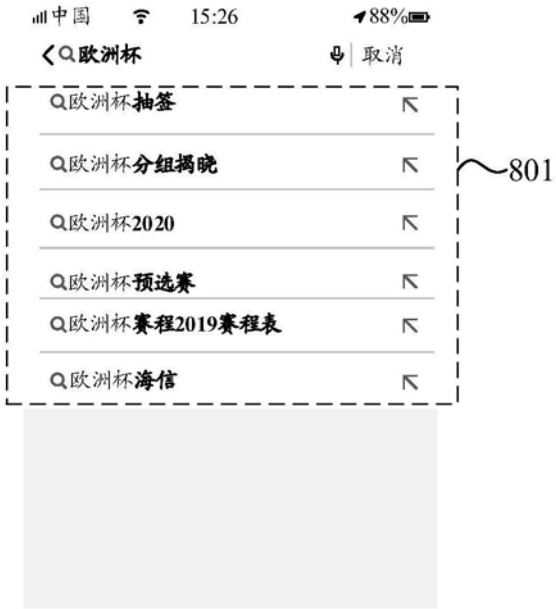

中国 15:26 88%
欧洲杯 取消
欧洲杯抽签
欧洲杯分组揭晓
欧洲杯2020
欧洲杯预选赛
欧洲杯赛程2019赛程表
欧洲杯海信
801

图8

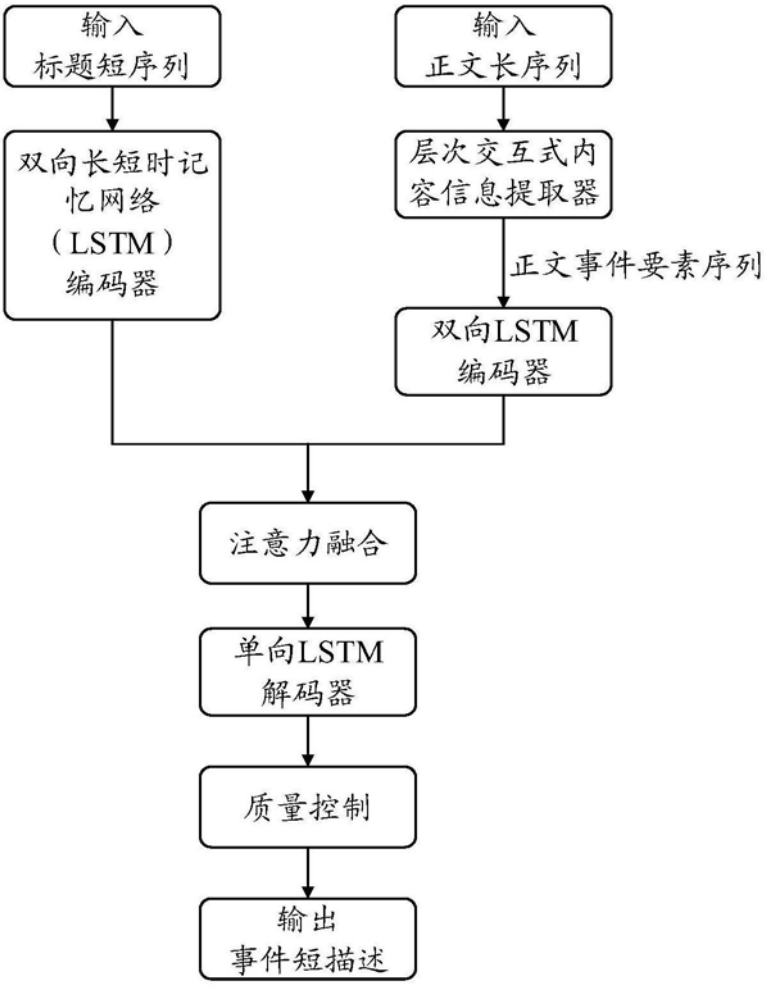

输入
标题短序列
输入
正文长序列
双向长短时记忆网络（LSTM）编码器
层次交互式内容信息提取器
正文事件要素序列
双向LSTM
编码器
注意力融合
单向LSTM
解码器
质量控制
输出
事件短描述

图9

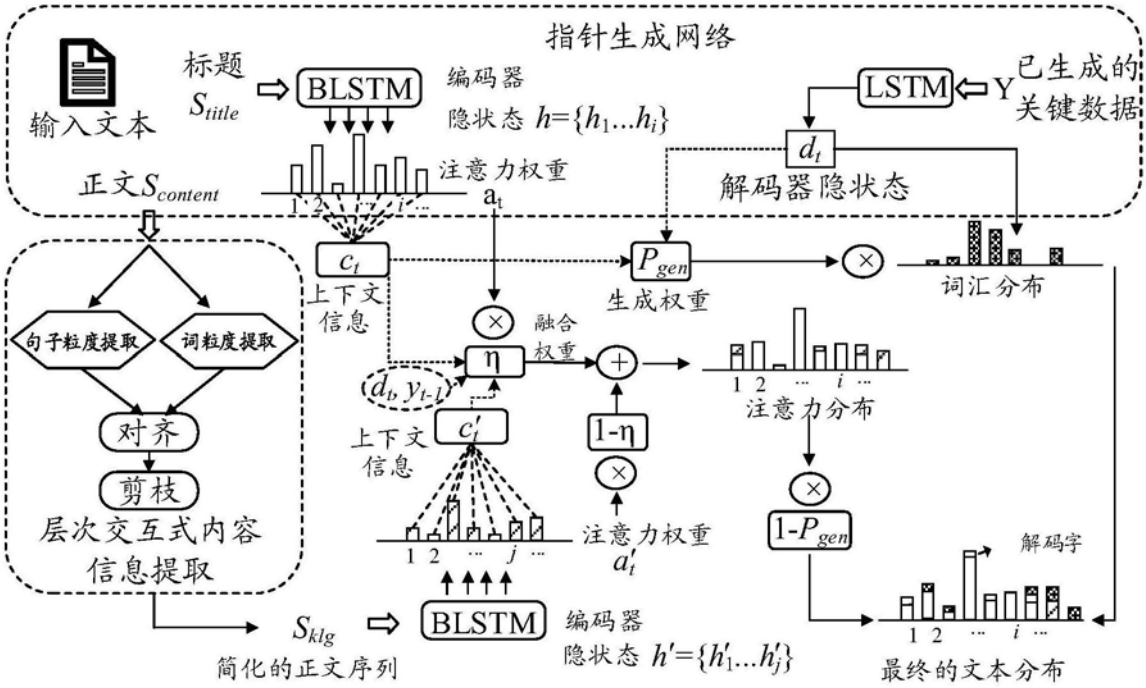

指针生成网络
输入文本
标题
S_{title}
BLSTM
编码器
隐状态 $h=\{h_1...h_i\}$
LSTM
Y
已生成的关键数据
正文$S_{content}$
注意力权重
a_t
d_t
解码器隐状态
c_t
上下文信息
P_{gen}
生成权重
词汇分布
句子粒度提取
词粒度提取
融合权重
η
(d_t, y_{t-1})
c_t'
上下文信息
1-η
注意力分布
对齐
剪枝
层次交互式内容信息提取
注意力权重
a_t'
1-P_{gen}
解码字
S_{klg}
BLSTM
编码器
隐状态 $h'=\{h_1'...h_j'\}$
简化的正文序列
最终的文本分布

图10

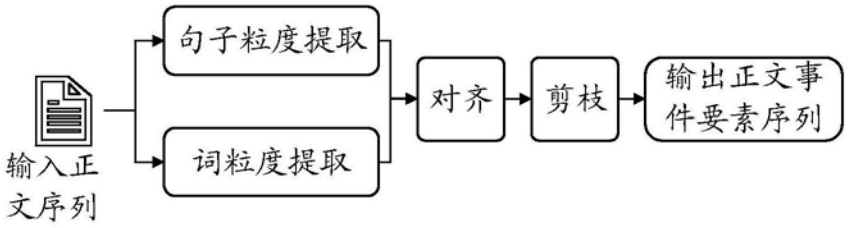

输入正文序列
句子粒度提取
词粒度提取
对齐
剪枝
输出正文事件要素序列

图11

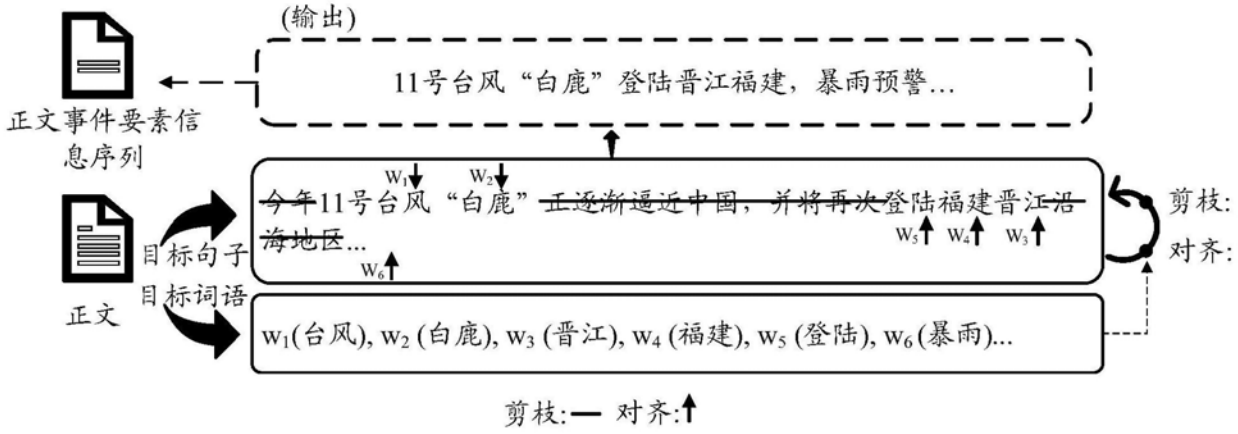

(输出)
11号台风“白鹿”登陆晋江福建，暴雨预警...
正文事件要素信息序列
W_1 W_2
今年11号台风“白鹿”正逐渐逼近中国，并将再次登陆福建晋江沿海地区...
W_5 W_4 W_3 W_6
剪枝:
对齐:
目标句子
目标词语
正文
w_1(台风), w_2 (白鹿), w_3 (晋江), w_4 (福建), w_5 (登陆), w_6 (暴雨)...
剪枝:— 对齐:↑

图12

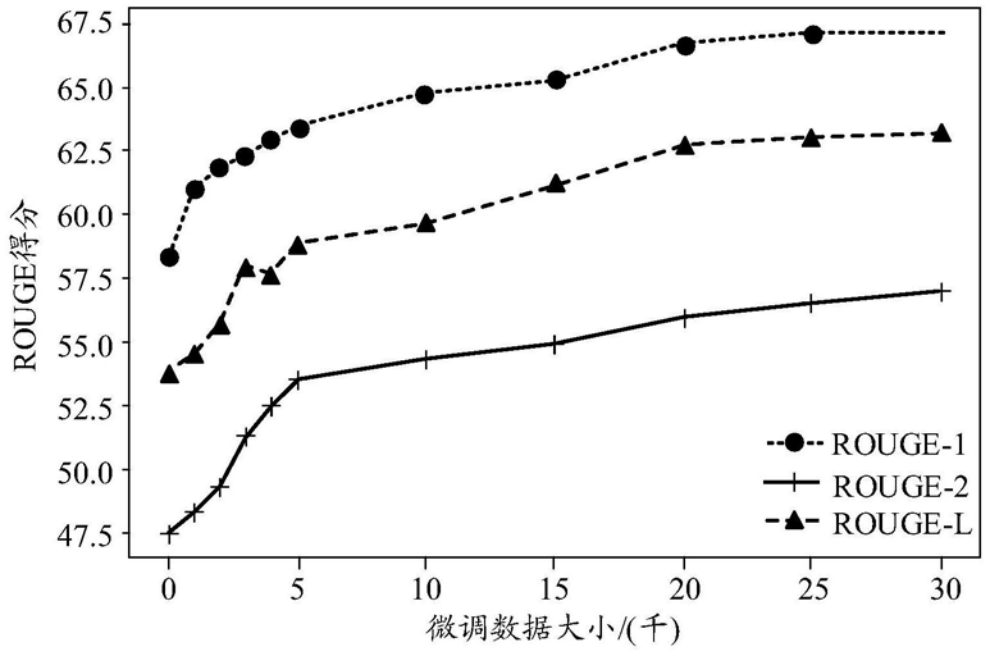

67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
47.5
ROUGE得分
0
5
10
15
20
25
30
微调数据大小/(千)
ROUGE-1
ROUGE-2
ROUGE-L

图13