(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113111663 B
(45)授权公告日 2024.09.06

(21)申请号 202110467022.9

(22)申请日 2021.04.28

(65)同一申请的已公布的文献号
申请公布号 CN 113111663 A

(43)申请公布日 2021.07.13

(73)专利权人 东南大学
地址 210096 江苏省南京市玄武区四牌楼2号

(72)发明人 杨鹏 周华健 任炳先 于晓潭

(74)专利代理机构 南京众联专利代理有限公司 32206
专利代理师 杜静静

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/216(2020.01)
G06F 16/953(2019.01)
G06N 3/0464(2023.01)
G06N 3/049(2023.01)
G06N 3/084(2023.01)

(56)对比文件
WO 2020082560 A1,2020.04.30
WO 2020107878 A1,2020.06.04

审查员 王慧林

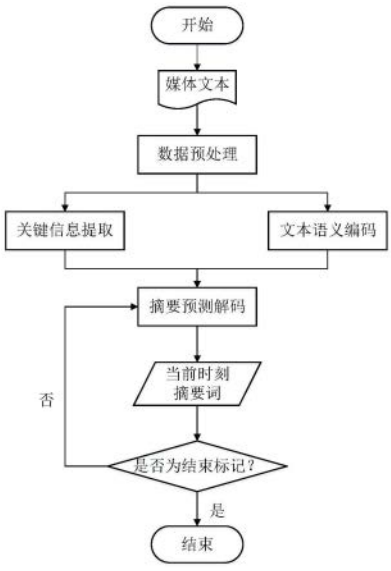
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种融合关键信息的摘要生成方法

(57)摘要

本发明公开了一种融合关键信息的摘要生成方法,能够为媒体文本自动地生成摘要。本发明首先通过互联网采集媒体文本,并在媒体文本基础上按人工规则生成标准摘要构造出样本数据集;然后对数据集进行预处理,生成训练模型的输入数据;接着构建基于关键信息融合的seq2seq摘要生成模型,并联合三元组损失与交叉熵损失进行模型训练;最后基于训练完毕后的模型进行输出模型的构建,并利用输出模型对待进行摘要的媒体文本进行自动摘要生成。与现有技术相比,本发明联合关键词与主题信息作为关键信息,能够多层次地对摘要生成过程进行引导,从而提高摘要结果的信息覆盖度与主题一致性。



CN 113111663 B

1.一种融合关键信息的摘要生成方法,其特征在于,所述方法包括以下步骤:

步骤1,积累样本数据集:通过互联网采集多篇媒体文本,积累样本数据集,具体如下:首先从互联网上搜集大量媒体文本,并按人工规则生成标准摘要,它们共同构成样本数据集D,标准摘要的人工生成规则为:40%的媒体文本将标题作为标准摘要、40%的媒体文本使用前三句话作为标准摘要、20%的媒体文本采用人工编写的摘要作为标准摘要;

步骤2,数据预处理:首先对数据集D的每篇媒体文本进行TextRank来提取关键词,选择TextRank打分最高的M个关键词作为媒体文本最终的关键词,并按原文中出现的位置顺序组织成关键词序列;对数据集D的每份样本构建三元组数据(A、P、N),其中基准实例A为所属样本的标准摘要词序列、正实例P为所属样本的媒体文本原文词序列,负实例N为另一份样本的媒体文本原文词序列;

步骤3,利用步骤2处理后的数据集D对关键信息融合的seq2seq摘要生成模型进行训练,首先利用TextRank方法提取样本中媒体文本的关键词,然后基于BiLSTM循环神经网络的关键信息抽取模块抽取出文本的全局主题信息与关键词的局部要素信息并联合为关键信息表示,seq2seq摘要生成模块通过融合关键信息的注意力机制引导摘要生成过程,最后利用三元组损失与交叉熵损失联合训练所述模型,具体分为以下子步骤:

步骤3-1,构建输入层,输入层接收关键词序列与三元组数据作为输入,利用预训练的word2vec模型将每个词序列转化为词向量序列,分别得到映射后的关键词向量序列E_K、基准实例词向量序列E_A、正实例词向量序列E_P与负实例词向量序列E_N;

步骤3-2,构建文本编码层,采用一个两层BiLSTM循环神经网络对正实例词向量序列E_P进行语义编码提取,得到正实例词向量序列E_P的隐层状态向量BiLSTM(E_P);

步骤3-3,构建关键信息提取层,关键信息提取层分为全局主题信息提取子层与局部要素信息提取子层,全局主题信息提取子层采用一个两层BiLSTM循环神经网络分别提取三元组词向量序列(E_A、E_P、E_N)的主题信息,将最后一层BiLSTM循环神经网络中前向LSTM与后向LSTM各自最后一个时刻的输出隐状态向量进行拼接后作为三元组词向量序列的全局主题信息表示(T_A、T_P、T_N);局部要素信息提取子层采用一个单层BiLSTM循环神经网络对关键词向量序列E_K进行消岐,得到要素词向量序列Elim(E_K);

步骤3-4,构建摘要解码层,采用一个两层LSTM循环神经网络与注意力机制进行摘要的解码,首先利用两层LSTM循环神经网络得到当前摘要词的隐状态向量H,并将其作为查询向量Query与要素词向量序列Elim(E_K)进行注意力计算,得到局部要素信息向量表示K,然后将局部要素信息向量表示K、全局主题信息表示T_P、解码层隐状态向量H进行维度拼接后与子步骤3-2得到的隐层状态向量BiLSTM(E_P)进行注意力计算得到上下文向量c,公式如下:

$$c=\text{Attention}(H\oplus K\oplus T_P,\text{BiLSTM}(E_P)) \quad (1)$$

其中代表维度拼接运算;

步骤3-5,构建摘要概率化层,使用一个线性映射函数fc与softmax激活函数,将上下文向量c与解码层隐状态向量H转化为摘要词的预测概率分布P,计算公式如下所示:

$$P=\text{softmax}(fc(H,c)) \quad (2)$$

$$fc(H,c)=W_H H+W_c c+b \quad (3)$$

其中,W_H、W_c和b是模型待训练的参数;

步骤3-6,构建损失函数层,本层联合主题信息表示的三元组损失函数L_T与摘要词的交叉熵损失函数L_S作为seq2seq摘要生成模型训练的总损失函数,具体如下:

$$L_T=\max\{d(T_A,T_P)-d(T_A,T_N)+Margin,0\} \quad (4)$$

$$d(T_A,T_P)=1-\cos(T_A,T_P) \quad (5)$$

$$d(T_A,T_N)=1-\cos(T_A,T_N) \quad (6)$$

$$L_{total}=\alpha L_S+\beta L_T \quad (7)$$

其中L_T为三元组损失,Margin为边界距离,取值为1,以保证正实例与负实例在主题语义上存在差异性;$d(T_A,T_P)$代表基准实例A与正实例P的主题向量语义距离,$d(T_A,T_N)$代表基准实例A与负实例N的主题向量语义距离;cos函数用于计算两个主题向量夹角的余弦值,用以衡量主题向量间的语义相似度;α与β为超参数,代表两个损失各自的权重系数;L_S为摘要词预测的交叉熵损失;L_{total}为本组样本的总体训练损失;

步骤3-7,训练所述seq2seq摘要生成模型,采用随机初始化的方式初始化所有待训练参数,在训练过程中采用Adam优化器进行梯度反向传播来更新模型参数,初始学习率设置为0.001,当训练损失不再下降或训练轮数超过50轮时,模型训练结束;

步骤4,利用训练完毕的模型构建输出模型生成摘要,具体如下,对于待进行摘要生成的媒体文本,首先用TextRank方法提取关键词,将媒体文本原文与文本关键词输入到步骤3中训练好的seq2seq摘要生成模型中,生成媒体文本摘要。

一种融合关键信息的摘要生成方法

技术领域

[0001] 本发明涉及一种融合关键信息的摘要生成方法,属于互联网技术领域。

背景技术

[0002] 随着互联网技术的高速发展,网络媒体成为人们快速获取和发布信息的重要平台,这使得各式各样的媒体新闻数量呈爆炸性增长。因此,对媒体文本进行全面分析,抽取、提炼出重要信息,并聚合成简短清晰的摘要呈现给读者,可以有效地帮助读者迅速、方便地了解媒体报道的主要内容,提高读者的信息获取效率。

[0003] 序列到序列(sequence-to-sequence,seq2seq)的生成式摘要模型是当前文本摘要生成领域的主流模型。该模型由一个编码器与一个解码器构成,通过编码器将输入文本序列编码为隐层状态向量,然后通过解码器将隐层状态向量解码为摘要进行输出。然而,传统的seq2seq模型通过注意力机制来对重要编码信息进行聚焦,但摘要生成任务中原始文本与目标摘要在长度上往往存在明显差距,注意力权重容易分散到大量冗余信息上,致使生成摘要存在重要信息缺失、上下文主题不一致的问题。为此,本发明在seq2seq模型的基础上,引入主题抽取任务训练基于三元组损失的文本主题表示,利用TextRank方法抽取文本的关键字作为文本要素信息,结合文本的主题表示与要素信息生成文本的关键信息并融入到解码过程中,对摘要的生成进行有效引导。

发明内容

[0004] 针对现有技术中存在的问题与不足,本发明提供一种融合关键信息的摘要生成方法,可以提取出媒体文本全局主题与局部要素两个层次的关键信息,并通过融合关键信息改善摘要生成过程缺乏有效控制的问题,提高摘要结果的主题一致性与信息覆盖度。

[0005] 为实现上述发明目的,本发明所述的一种融合关键信息的摘要生成方法,首先利用TextRank方法提取出文本的关键词;然后构建基于BiLSTM(Bidirectional Long Short-Term Memory,双向长短时记忆网络)的关键信息提取模块,将抽取出的关键词与媒体文本作为输入,得到媒体文本的关键信息表示;最后将关键信息表示融入seq2seq模型的注意力机制中来生成媒体文本的摘要。该方法主要包括四个步骤,具体如下:

[0006] 步骤1:通过互联网采集多篇媒体文本,积累样本数据集;所述数据集中的一个样本包括一篇媒体文本以及该媒体文本的标准摘要;

[0007] 步骤2:对数据集中每一个样本构造三元组数据,一个三元组数据包括基准实例、正实例和负实例,基准实例为媒体文本的标准摘要、正实例为媒体文本原文、负实例为与正实例不同的另一篇媒体文本原文;

[0008] 步骤3:训练基于关键信息融合的seq2seq摘要生成模型。首先利用TextRank方法提取样本中媒体文本的关键词,然后基于BiLSTM的关键信息抽取模块抽取出文本的全局主题信息与关键词的局部要素信息并联合为关键信息表示,seq2seq摘要生成模块通过融合关键信息的注意力机制引导摘要生成过程,最后利用三元组损失与交叉熵损失联合训练所

述模型。

[0009] 步骤4:对待进行摘要的媒体文本生成摘要。对于待进行摘要的媒体文本,首先用TextRank方法提取关键词,将媒体文本原文与文本关键词输入到步骤(2)中训练好的seq2seq摘要生成模型中,生成媒体文本摘要。该方案能够从多个维度提取文本的关键信息,克服传统文本摘要方法主题不够一致、信息不够完整的问题,可应用于媒体文本关键信息的精确提取,提升媒体文本摘要的效果。

[0010] 相对于现有技术,本发明的优点如下:1)本发明采用的关键信息抽取模块,能够抽取出文本的全局主题信息与局部要素信息,对文本关键信息进行多层次的语义语境表示,补充了摘要生成过程缺失的关键特征,有效提高了摘要结果的主题一致性与信息覆盖度;2)本发明采用融合关键信息的注意力机制,能够有效融合多层次的关键信息并多角度地对摘要生成过程进行引导,减少了无关信息的干扰,有效提高了摘要结果的准确性。

附图说明

[0011] 图1为本发明实施例的处理流程图。

[0012] 图2为基于关键信息融合的seq2seq摘要生成模型的训练流程图。

具体实施方式

[0013] 为了加深对本发明的认识和理解,下面结合具体实施例,进一步阐明本发明。

[0014] 实施例1:参见图1、图2,一种融合关键信息的摘要生成方法,具体实施步骤如下:

[0015] 步骤1,积累样本数据集,不失一般性,本实施例首先从互联网上搜集大量媒体文本,并按人工规则生成标准摘要,它们共同构成样本数据集D。标准摘要的人工生成规则为:40%的媒体文本将标题作为标准摘要、40%的媒体文本使用前三句话作为标准摘要、20%的媒体文本采用人工编写的摘要作为标准摘要。

[0016] 步骤2,数据预处理,本实施例首先对数据集D的每篇媒体文本进行TextRank来提取关键词,选择TextRank打分最高的M个关键词作为媒体文本最终的关键词,并按原文中出现的位置顺序组织成关键词序列,本实施例中M取值为8。对数据集D的每份样本构建三元组数据(A、P、N),其中基准实例A为所属样本的标准摘要词序列、正例P为所属样本的媒体文本原文词序列、N为另一份样本的媒体文本原文词序列。

[0017] 步骤3,利用步骤2处理后的数据集D对关键信息融合的seq2seq摘要生成模型进行训练,该步骤的实施可以分为以下子步骤:

[0018] 子步骤3-1,构建输入层,输入层接收关键词序列与三元组数据作为输入,利用预训练的word2vec模型将每个词序列转化为词向量序列,分别得到映射后的关键词向量序列E_K、基准实例词向量序列E_A、正例词向量序列E_P与负例词向量序列E_N。

[0019] 子步骤3-2,构建文本编码层,本实施例采用一个两层BiLSTM循环神经网络对正例词向量序列E_P进行语义编码提取,得到词向量序列E_P的隐层状态向量BiLSTM(E_P)。

[0020] 子步骤3-3,构建关键信息提取层,关键信息提取层分为全局主题信息提取子层与局部要素信息提取子层,前者采用一个双层BiLSTM分别提取三元组词序列(E_A、E_P、E_N)的主题信息,本实施例将最后一层BiLSTM中前向LSTM与后向LSTM各自最后一个时刻的输出隐状态向量进行拼接后作为词序列的全局主题信息表示(T_A、T_P、T_N);后者采用一个单层BiLSTM

对关键词向量序列E_K进行消岐,得到要素词向量序列Elim(E_K)。

[0021] 子步骤3-4,构建摘要解码层。本实施例采用一个两层LSTM循环神经网络与注意力机制进行摘要的解码,首先利用两层LSTM得到当前摘要词的隐状态向量H,并将其作为查询向量Query与要素词向量序列Elim(E_K)进行注意力计算,得到局部要素信息向量表示K,然后将局部要素信息向量表示K、全局主题信息表示T_P、解码层隐状态向量H进行维度拼接后与子步骤3-1得到的隐层状态向量BiLSTM(E_P)进行注意力计算得到上下文向量c,公式如下:

[0022] $$c=\text{Attention}(H\oplus K\oplus T_P,\text{BiLSTM}(E_P)) \quad (1)$$

[0023] 其中⊕代表维度拼接运算。

[0024] 子步骤3-5,构建摘要概率化层,使用一个线性映射函数fc与softmax激活函数,将上下文向量c与解码层隐状态向量H转化为摘要词的预测概率分布P,计算公式如下所示:

[0025] $$P=\text{softmax}(fc(H,c)) \quad (2)$$

[0026] $$fc(H,c)=W_H H+W_c c+b \quad (3)$$

[0027] 其中,W_H、W_c和b是模型待训练的参数。

[0028] 子步骤3-6,构建损失函数层,本层联合主题信息表示的三元组损失与摘要词的交叉熵损失作为所述模型的训练损失函数。按如下损失函数计算公式得到本组样本的训练损失:

[0029] $$L_T=\max\{d(T_A,T_P)-d(T_A,T_N)+\text{Margin},0\} \quad (4)$$

[0030] $$d(T_A,T_P)=1-\cos(T_A,T_P) \quad (5)$$

[0031] $$d(T_A,T_N)=1-\cos(T_A,T_N) \quad (6)$$

[0032] $$L_{total}=\alpha L_S+\beta L_T \quad (7)$$

[0033] 其中L_T为三元组损失,Margin为边界距离,本实施例取值为1,以保证正实例与负实例在主题语义上存在差异性;$d(T_A,T_P)$代表基准实例A与正实例P的主题语义距离,$d(T_A,T_N)$代表基准实例A与负实例N的主题语义距离;cos函数用于计算两个主题向量夹角的余弦值,用以衡量主题向量间的语义相似度;α与β为超参数,代表两个损失各自的权重系数,本实施例中分别取值1与2;L_S为摘要词预测的交叉熵损失;L_{total}为本组样本的总体训练损失。

[0034] 子步骤3-7,训练所述模型。本实施例采用随机初始化的方式初始化所有待训练参数,在训练过程中采用Adam优化器进行梯度反向传播来更新模型参数,初始学习率设置为0.001。当训练损失不再下降或训练轮数超过50轮时,模型训练结束。

[0035] 步骤4,利用训练完毕的模型构建输出模型生成摘要。输出模型不需要事先构建三元组数据,只需要待进行摘要的媒体文本以及提取的关键词作为输入,然后在摘要解码层每一时刻的输入词为上一时刻生成的摘要词,初始摘要词为一个特殊的开始标记“<START>”,每一时刻的摘要词为摘要概率化层输出的概率最大的词,当输出结束标记“<END>”时,停止摘要生成,输出已生成的摘要词作为输入媒体文本的预测摘要。

[0036] 基于相同的发明构思,本发明实施例还提供一种融合关键信息的摘要生成装置,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,该计算机程序被加载至处理器时实现上述的融合关键信息的摘要生成方法。

[0037] 应理解这些实施例仅用于说明本发明而不用于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限

定的范围。

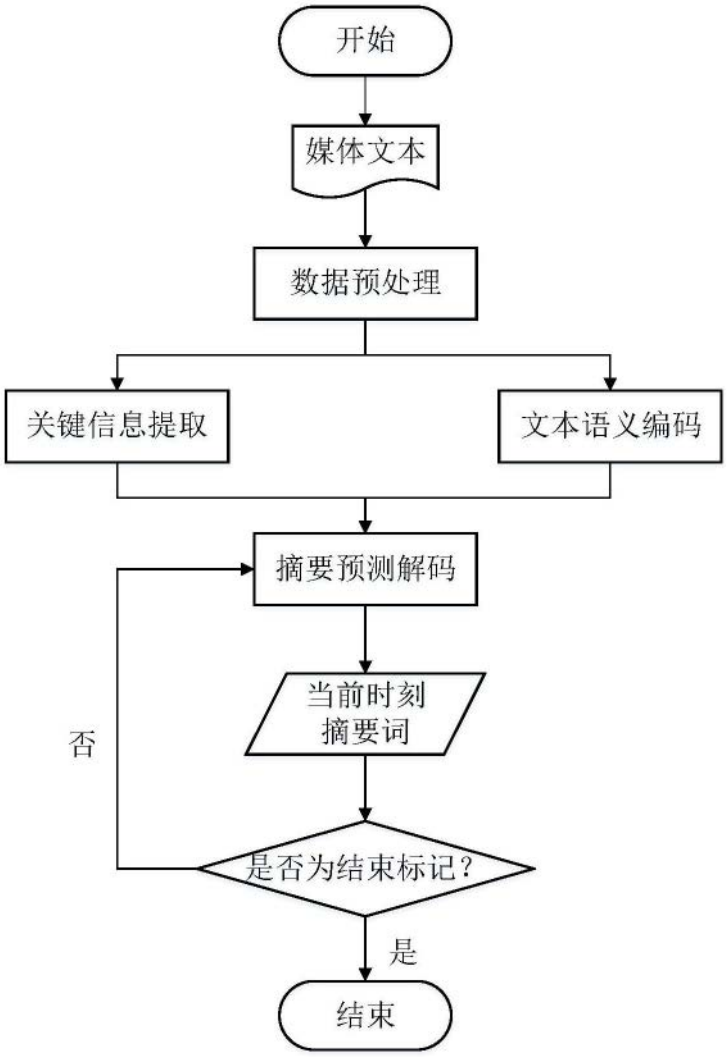

开始
媒体文本
数据预处理
关键信息提取
文本语义编码
摘要预测解码
当前时刻
摘要词
否
是否为结束标记?
是
结束

图1

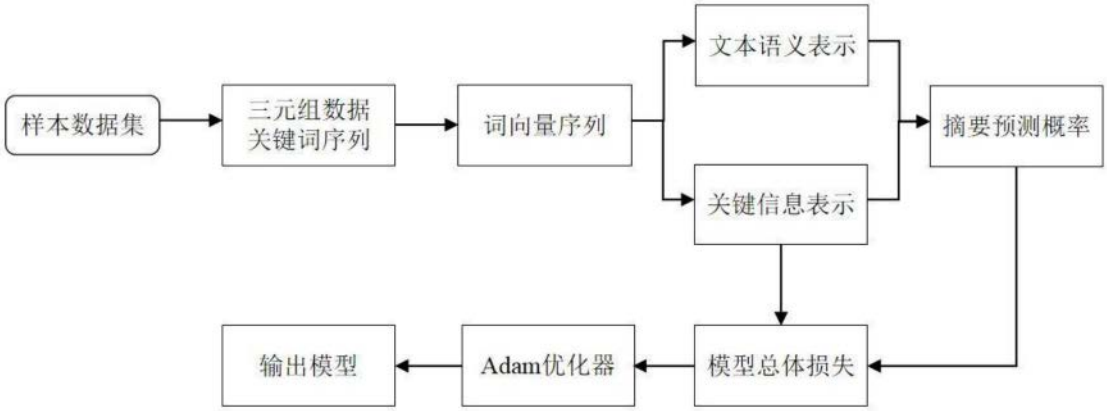

样本数据集
三元组数据
关键词序列
词向量序列
文本语义表示
关键信息表示
摘要预测概率
输出模型
Adam优化器
模型总体损失

图2