



(12) 发明专利申请

(10) 申请公布号 CN 114051154 A

(43) 申请公布日 2022. 02. 15

(21) 申请号 202111305567.6

H04N 21/8547 (2011.01)

(22) 申请日 2021.11.05

G06V 30/414 (2022.01)

(71) 申请人 新华智云科技有限公司

地址 310012 浙江省杭州市西湖区文一西路460号文娱中心430室

(72) 发明人 刘潇婧

(74) 专利代理机构 杭州裕阳联合专利代理有限公司 33289

代理人 吴文杰

(51) Int. Cl.

H04N 21/233 (2011.01)

H04N 21/234 (2011.01)

H04N 21/439 (2011.01)

H04N 21/44 (2011.01)

H04N 21/845 (2011.01)

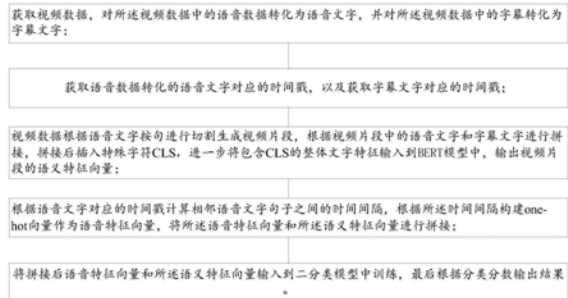
权利要求书2页 说明书5页 附图1页

(54) 发明名称

一种新闻视频拆条方法和系统

(57) 摘要

本发明公开了一种新闻视频拆条方法和系统,所述方法包括:获取视频数据,对所述视频数据中的语音数据转化为语音文字,并对所述视频数据中的字幕转化为字幕文字;获取语音数据转化的语音文字对应的时间戳,以及获取字幕文字对应的时间戳;将视频数据根据语音文字按句进行切割生成视频片段,根据视频片段中的语音文字和字幕文字进行拼接,拼接后插入特殊字符CLS,进一步将包含CLS的整体文字特征输入到BERT模型中,输出视频片段的语义特征向量;根据语音文字对应的时间戳计算相邻语音文字句子之间的时间间隔,根据所述时间间隔构建one-hot向量作为语音特征向量,将所述语音特征向量和所述语义特征向量拼接并输入到二分类模型中根据分类分数输出结果。



1. 一种新闻视频拆条方法,其特征在于,所述方法包括:

获取视频数据,对所述视频数据中的语音数据转化为语音文字,并对所述视频数据中的字幕化为字幕文字;

获取语音数据转化的语音文字对应的时间戳,以及获取字幕文字对应的时间戳;

视频数据根据语音文字按句进行切割生成视频片段,根据视频片段中的语音文字和字幕文字进行拼接,拼接后插入特殊字符CLS,进一步将包含CLS的整体文字特征输入到BERT模型中,输出视频片段的语义特征向量;

根据语音文字对应的时间戳计算相邻语音文字句子之间的时间间隔,根据所述时间间隔构建one-hot向量作为语音特征向量,将所述语音特征向量和所述语义特征向量进行拼接;

将拼接后语音特征向量和所述语义特征向量输入到二分类模型中训练,最后根据分类分数输出结果。

2. 根据权利要求1所述的一种新闻视频拆条方法,其特征在于,采用ASR语音识别技术将视频数据中的语音数据转化为语音文字,并获取对应语音文字的时间戳,采用OCR文字识别技术识别视频字幕文字,并获取对应文字的时间戳。

3. 根据权利要求1所述的一种新闻视频拆条方法,其特征在于,所述拆条方法还包括:将获取的语音文字按句进行切割,并根据切割的语音文字将对应的视频数据进行切割,生成对应的视频片段,获取切割后的视频片段的字幕文字,并将切割后的视频片段的字幕文字进行合并拼接。

4. 根据权利要求1所述的一种新闻视频拆条方法,其特征在于,所述拆条方法还包括:对获取的语音文字的句子进行标注,设置结尾句标签字符和非结尾句标签字符,并建立语音文字句子的标签特征向量。

5. 根据权利要求1所述的一种新闻视频拆条方法,其特征在于,所述拆条方法包括:以128个视频片段为长度将视频分为连续不重复的子块,且每个子块作为独立的视频作为输入数据。

6. 根据权利要求1所述的一种新闻视频拆条方法,其特征在于,所述语音特征向量构建方法包括:根据视频片段中每个句子之间时间间隔,将时间间隔分段赋值,其中时间间隔为0/s的赋值为0, (0s, 5s]为1, (5s, 10s]为2, (10s, +∞)为3,并将上述0、1、2、3转化为one-hot向量作为语音特征向量。

7. 根据权利要求6所述的一种新闻视频拆条方法,其特征在于,所述拆条方法还包括:将拼接后语音特征向量和所述语义特征向量输入到预训练模型BERT进行特征提取后,将提取后的特征输入到全连接层,并接入由sigmoid函数构建的二分类模型中对当前片段是否为结尾句进行分类判断。

8. 根据权利要求7所述的一种新闻视频拆条方法,其特征在于,在二分类模型的训练过程中,计算由多个视频片段构成的视频子块的熵交叉误差,用于计算结尾句的概率:

$$J = \sum_{i=1}^{n-1} [-y_i \log p_i - (1 - y_i) \log(1 - p_i)]$$

其中J为熵交叉误差, y_i 为标签、 p_i 是结尾句的概率,采用梯度下降方法计算所述熵交叉

误差的最小值作为训练完毕指标,并用验证集对所述二分类模型训练结果进行验证。

9.一种新闻视频拆条系统,其特征在于,所述系统执行上述权利要求1-8中任意一项所述的一种新闻视频拆条方法。

10.一种计算机可读存储介质,其特征在于,计算机可读存储介质存储有计算机程序,所述计算机程序可被处理器执行上述权利要求1-8中任意一项所述的一种新闻视频拆条方法。

一种新闻视频拆条方法和系统

技术领域

[0001] 本发明涉及新闻媒体技术领域,特别涉及一种新闻视频拆条方法和系统。

背景技术

[0002] 新闻拆条的主要任务是针对某个新闻视频(如新闻联播、新闻30分、地方新闻播报等),根据一定的业务逻辑对视频内容进行片段拆分,从而为后续的素材整理、内容分发提供数据基础。目前主要有两种技术方案:1)基于图像:根据镜头场景的转化进行视频拆分,如通过主持人静坐和播报新闻实况的镜头场景不同,进行切分判断。2)基于规则:根据固定字幕出现的位置、大小和时间等特征,对新闻切分点进行判断。现有技术存在如下缺陷:1、根据镜头场景的转化进行新闻片段切分,没有考虑新闻的语义信息,无法覆盖主持人一直静坐或画面连续切换等场景的新闻视频。2、利用规则进行新闻片段切分,通用性和可复用性差,人工成本高。

发明内容

[0003] 本发明其中一个发明目的在于提供一种新闻视频拆条方法和系统,所述方法和系统同时利用自动语音识别技术ASR和文字识别技术OCR分别获取语音播报和视频字幕的对应文字和文字对应的时间戳,通过两种识别手段对新闻视频进行视频切分点的判断,从而可以有效地提高视频切分点准确率。

[0004] 本发明其中一个发明目的在于提供一种新闻视频拆条方法和系统,所述方法和系统通过语音识别技术获取的文字和视频字幕获取的文字进行拼接,输入到预训练模型BERT进行训练,生成具有联合特征的语义特征向量,所述语义特征向量可以避免单独视频切分因为主持人静坐或连续切换造成视频切条不准确现象。

[0005] 本发明其中一个发明目的在于提供一种新闻视频拆条方法和系统,所述方法和系统通过将自动语音识别技术ASR时间差特征和具有联合特征的语义特征进行拼接,通过二分类模型判断是否存在新闻的结尾句,进一步执行新闻拆条,因此本发明涉及的新闻拆条不用考虑规则问题,适用性更好。

[0006] 为了实现至少一个上述发明目的,本发明进一步提供一种新闻视频拆条方法,所述方法包括:

[0007] 获取视频数据,对所述视频数据中的语音数据转化为语音文字,并对所述视频数据中的字幕转化为字幕文字;

[0008] 获取语音数据转化的语音文字对应的时间戳,以及获取字幕文字对应的时间戳;

[0009] 视频数据根据语音文字按句进行切割生成视频片段,根据视频片段中的语音文字和字幕文字进行拼接,拼接后插入特殊字符CLS,进一步将包含CLS的整体文字特征输入到BERT模型中,输出视频片段的语义特征向量;

[0010] 根据语音文字对应的时间戳计算相邻语音文字句子之间的时间间隔,根据所述时间间隔构建one-hot向量作为语音特征向量,将所述语音特征向量和所述语义特征向量进

行拼接；

[0011] 将拼接后语音特征向量和所述语义特征向量输入到二分类模型中训练,最后根据分类分数输出结果。

[0012] 根据本发明其中一个较佳实施例,采用ASR语音识别技术将视频数据中的语音数据转化为语音文字,并获取对应语音文字的时间戳,采用OCR文字识别技术识别视频字幕文字,并获取对应文字的时间戳。

[0013] 根据本发明另一个较佳实施例,所述拆条方法还包括:将获取的语音文字按句进行切割,并根据切割的语音文字将对应的视频数据进行切割,生成对应的视频片段,获取切割后的视频片段的字幕文字,并将切割后的视频片段的字幕文字进行合并拼接。

[0014] 根据本发明另一个较佳实施例,所述拆条方法还包括:对获取的语音文字的句子进行标注,设置结尾句标签字符和非结尾句标签字符,并建立语音文字句子的标签特征向量。

[0015] 根据本发明另一个较佳实施例,所述拆条方法包括:以128个视频片段为长度将视频分为连续不重复的子块,且每个子块作为独立的视频作为输入数据。

[0016] 根据本发明另一个较佳实施例,所述语音特征向量构建方法包括:根据视频片段中每个句子之间时间间隔,将时间间隔分段赋值,其中时间间隔为0/s的赋值为0, (0s, 5s]为1, (5s, 10s]为2, (10s, +∞)为3,并将上述0、1、2、3转化为one-hot向量作为语音特征向量。

[0017] 根据本发明另一个较佳实施例,所述拆条方法还包括:将拼接后语音特征向量和所述语义特征向量输入到预训练模型BERT进行特征提取后,将提取后的特征输入到全连接层,并接入由sigmoid函数构建的二分类模型中对当前片段是否为结尾句进行分类判断。

[0018] 根据本发明另一个较佳实施例,在二分类模型的训练过程中,计算由多个视频片段构成的视频子块的熵交叉误差,用于计算结尾句的概率:

$$[0019] \quad J = \sum_{i=1}^{n-1} [-y_i \log p_i - (1 - y_i) \log(1 - p_i)]$$

[0020] 其中J为熵交叉误差, y_i 为标签、 p_i 是结尾句的概率,采用梯度下降方法计算所述熵交叉误差的最小值作为训练完毕指标,并用验证集对所述二分类模型训练结果进行验证。

[0021] 为了实现至少一个上述发明目的,本发明进一步提供一种新闻视频拆条系统,所述系统执行上述一种新闻视频拆条方法。

[0022] 本发明进一步提供一种计算机可读存储介质,计算机可读存储介质存储有计算机程序,所述计算机程序可被处理器执行上述一种新闻视频拆条方法。

附图说明

[0023] 图1显示的是本发明一种新闻视频拆条方法的流程示意图;

[0024] 图2显示的是本发明一种新闻视频拆条系统的模型示意图。

具体实施方式

[0025] 以下描述用于揭露本发明以使本领域技术人员能够实现本发明。以下描述中的优

选实施例只作为举例,本领域技术人员可以想到其他显而易见的变型。在以下描述中界定的本发明的基本原理可以应用于其他实施方案、变形方案、改进方案、等同方案以及没有背离本发明的精神和范围的其他技术方案。

[0026] 可以理解的是,术语“一”应理解为“至少一”或“一个或多个”,即在一个实施例中,一个元件的数量可以为一个,而在另外的实施例中,该元件的数量可以为多个,术语“一”不能理解为对数量的限制。

[0027] 请结合图1-图2,本发明公开了一种新闻视频拆条方法和系统示意图,其中所述方法包括如下步骤:首先需要收集视频数据,其中所述视频数据可以采用爬虫技术从网络上获取,比如通过爬虫技术获取网络上的新闻视频数据1000条,将1000条所述新闻视频数据的80%作为训练集,将所述新闻视频数据的20%作为验证集。在完成新闻视频数据的收集后,对所述新闻视频数据进行预处理,所述预处理的方法包括:采样现有的语音识别技术(Automatic Speech Recognition,ASR)将所述新闻视频数据中的语音数据转化为语音文字,所述语音文字为文本形式的文字,获取每一语音文字对应的时间戳;进一步将获取的新闻视频进行解帧,将每一新闻视频转化为图片帧,进一步采用OCR文字识别技术(optical character recognition)获取每一帧图片的字幕文字和对应的时间戳。需要说明的是上述语音识别技术(Automatic Speech Recognition,ASR)和OCR文字识别技术(optical character recognition)均为现有技术,本发明对识别过程不再赘述。

[0028] 进一步的,在对所述新闻视频数据进行预处理后,根据获取的语音文字进行完整新闻视频的片段切割,所述切割方法包括:将视频中的语音识别的语音文字按照单句对视频进行分割,比如语音文字按句可以识别为: $S = (s_1, s_2, s_3, \dots, s_n)$,根据所述语音文字获取的时间戳, s_i 表示集合S中的任意一个句子,将对应句的视频片段分割为 $V = (v_1, v_2, v_3, \dots, v_n)$,其中 v_i 表示对应句语音文字 s_i 的视频片段。进一步的,需要将每个被切割的视频片段 v_i 中的字幕拼接为对应视频片段 v_i 的字幕文字 c_i 。

[0029] 进一步的,需要对所述语音识别获取的语音文字分割后的句子 $S = (s_1, s_2, s_3, \dots, s_n)$ 进行人工标注,根据新闻对应的题材、内容将所有分割后的句子 $S = (s_1, s_2, s_3, \dots, s_n)$ 判断是否是结尾句,若当前句子 s_i 为结尾句,则将当前句子 s_i 人工标注为1,构成当前句子的结尾句标签,若当前句子 s_i 不是结尾句,则将当前句子 s_i 人工标注为0,构成当前句子的非结尾句标签,因此针对所有分割后的句子其是否为结尾句构成0和1的组合: $y = (y_1, y_2, y_3, \dots, y_n)$,其中 y_i 表示对应切割后句子 s_i 对应的结尾句判断标签,其中 $y_i \in \{0, 1\}$ 。举例来说:xxx出席xxx大会。**【END】**北京市举办一年一度的庙会活动。很多人都兴致勃勃地来参加。活动包括xxx。**【END】**。其中**【END】**表示结尾句,其映射的标签为1,其他句号后映射的标签皆为0。也就是说,在特定新闻语境下,句号为结尾的句子并非真实的结尾句,因此通过人工标注的方式识别出特定新闻语境下的结尾句的形式,便于后续的模式训练。

[0030] 值得一提的是,在完成所述判断结尾句标签的人工标注后,需要进行语义特征提取,其中所述语义特征提取方法包括如下步骤:以上述切割的句子对应的视频片段大小为粒度,将每段视频片段的语音文字和对应的字幕文字进行拼接,其中拼接方式包括: $s_i = (w_{i1}, w_{i2}, \dots, w_{im})$ [SEP] $c_i = (t_{i1}, t_{i2}, \dots, t_{ik})$,其中 w_{im} 为切割后语音句子的单个文字字符,而 t_{ik} 为上述对应字幕文字句子中的单个文字字符,[SEP]为拼接符。并且在上述拼接的过程中同时在拼接句的句首插入特殊字符[CLS],使得形成完整的拼接后的特征:[CLS] $s_i =$

$(w_{i1}, w_{i2}, \dots, w_{im})$ [SEP] $c_i = (t_{i1}, t_{i2}, \dots, t_{ik})$, 将所述拼接后的特征输入到预先训练好的BERT模型中进行语义特征提取, 利用所述特殊字符[CLS]输出的向量可以表示每个视频片段的联合语义特征向量。

[0031] 本发明在建立每个视频联合语义特征向量构建后, 进一步构建语音特征向量, 所述语音特征向量构建方法包括如下步骤:

[0032] 以128个视频片段为长度将视频分成连续不重复的视频子块, 也就是说每个视频子块包含128个视频片段, 其中所述视频子块的视频片段数量并非特定, 本发明仅做举例说明, 其中将每个子块都作为独立的视频作为分类模型的输入数据, 进一步根据所述语音文字的相邻两个句子之间的时间间隔构建语音特征向量, 其中将相邻两个句子之间的时间间隔为0s的句子的值定义为0, 并定义相邻两个句子时间间隔(0s, 5s]为1, (5s, 10s]为2, (10s, +∞)为3, 且将上述定义的值0、1、2、3转化为one-hot向量作为当前句的语音特征向量, 且最后一个视频片段的语音特征向量的值为3, 其中所述语音特征向量可以句末的时间间隔定义, 比如第二句和第一句之间的时间间隔为3s, 则对应的第一句的语音特征向量的值为1。

[0033] 进一步的, 本发明将上述语音特征向量和每个视频片段对应的语义特征向量进行拼接, 其中拼接方式为两向量直接拼接, 所述向量直接拼接的方式使得向量的维度相加, 并将语音特征向量和每个视频片段对应的语义特征向量拼接的结果再次输入到所述预先训练好的BERT模型中进一步提取特征, 将所述BERT模型提取的特征向量输入到全连接层, 并接入由sigmoid函数构建的二分类模型中对当前片段是否为结尾句进行分类判断。对于一个由n个视频片段构成的视频子块, 定义cross-entropy errors (熵交叉误差) 为:

$$[0034] \quad J = \sum_{i=1}^{n-1} [-y_i \log p_i - (1 - y_i) \log(1 - p_i)]$$

[0035] 其中 y_i 为上述判断结尾句的标签, p_i 为是结尾句的概率, 在训练集数据上利用梯度下降的方式进行训练使上述熵交叉误差公式最小, 并在验证集上进行效果验证, 取验证集上效果最好的一轮作为最后的模型保存。

[0036] 在完成所述二分类模型的训练后, 将视频按照上述步骤进行识别, 其中所述二分类模型的识别结果可以为:0010001, 则将所述二分类模型的识别的结果进行合并, 其中上述二分类模型的识别结果可以知道第三句和第七句为结束句, 因此进一步将前三句进行合并, 同时合并第四句到第七句, 从而完成最后新闻拆条的视频片段结果。

[0037] 特别地, 根据本发明公开的实施例, 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, 该计算机程序可以通过通信部分从网络上被下载和安装, 和/或从可拆卸介质被安装。在该计算机程序被中央处理单元(CPU)执行时, 执行本申请的方法中限定的上述功能。需要说明的是, 本申请上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。计算机可读存储介质例如可以是但不限于电、磁、光、电磁、红外线段、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段的电连接、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)、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。在本申请中,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。而在本申请中,计算机可读的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。计算机可读的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输,包括但不限于:无线段、电线段、光缆、RF等等,或者上述的任意合适的组合。

[0038] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、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,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。

[0039] 本领域的技术人员应理解,上述描述及附图中所示的本发明的实施例只作为举例而并不限制本发明,本发明的目的已经完整并有效地实现,本发明的功能及结构原理已在实施例中展示和说明,在没有背离所述原理下,本发明的实施方式可以有任何变形或修改。

获取视频数据，对所述视频数据中的语音数据转化为语音文字，并对所述视频数据中的字幕转化为字幕文字；

获取语音数据转化的语音文字对应的时间戳，以及获取字幕文字对应的时间戳；

视频数据根据语音文字按句进行切割生成视频片段，根据视频片段中的语音文字和字幕文字进行拼接，拼接后插入特殊字符CLS，进一步将包含CLS的整体文字特征输入到BERT模型中，输出视频片段的语义特征向量；

根据语音文字对应的时间戳计算相邻语音文字句子之间的时间间隔，根据所述时间间隔构建one-hot向量作为语音特征向量，将所述语音特征向量和所述语义特征向量进行拼接；

将拼接后语音特征向量和所述语义特征向量输入到二分类模型中训练，最后根据分类分数输出结果。

图1

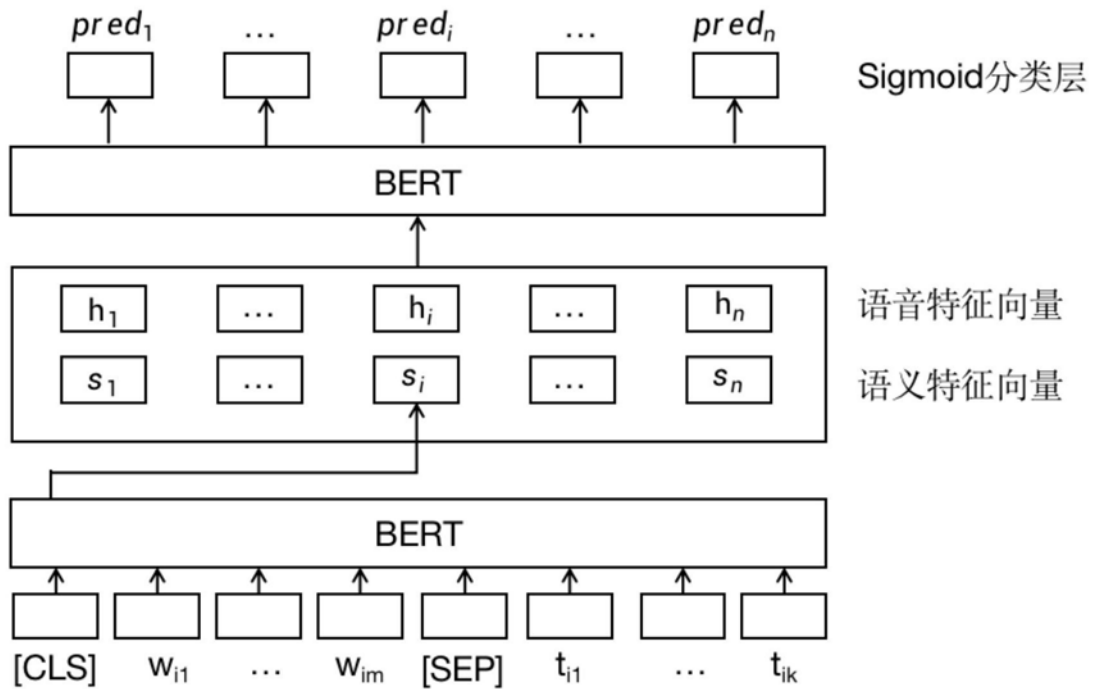


图2