多模态智慧课堂主题分割

郝玉泽[1],周斌[1*],胡波[2]
(1 中南民族大学 计算机学院,武汉 430074;2 武汉市东信同邦信息技术有限公司,武汉 430074)

摘 要 智慧课堂环境下,教学视频的主题分割对于提升教学分析和内容组织效率具有重要意义.然而,现有主题分割方法在智慧课堂场景中面临自动语音识别(ASR)转录错误率高、课堂内容不连贯等挑战,导致分割效果不佳.针对上述挑战,提出了一种多模态智慧课堂主题分割模型.该模型首先使用ASR获取语音转录文本,同时设计了一种全新的动态OCR识别与相似度检测去重算法从视频帧中提取文本信息,有效减少冗余帧处理;随后将ASR文本和OCR文本分别送入文本编码器进行编码,采用多模态注意力机制融合两种模态特征;最后通过BiLSTM和MLP网络进行主题边界的预测.实验结果表明:文中提出的解决方法在智慧课堂领域内的准确性和扩展性优于基线方法,同时,提出的动态OCR识别与相似度检测去重算法减少了模型的处理时间,可以满足现实应用的要求.

关键词 智慧课堂;多模态;主题分割;去重
中图分类号 TP391.1 **文献标志码** A **文章编号** 1672-4321(2026)02-0180-11
doi:10.20056/j.cnki.ZNMDZK.20250829

Multimodal smart classroom topic segmentation

HAO Yuze[1], ZHOU Bin[1*], HU Bo[2]
(1 College of Computer Science, South-Central Minzu University, Wuhan 430074, China; 2 Wuhan Dongxin Tongbang Information Technology Co. Ltd, Wuhan 430074, China)

Abstract In smart classroom environments, topic segmentation of instructional videos is of great significance for improving teaching analysis and content organization efficiency. However, existing topic segmentation methods face challenges in smart classroom scenarios, including high error rates in Automatic Speech Recognition (ASR) transcription and incoherent classroom content, leading to poor segmentation performance. To address these challenges, a multimodal smart classroom topic segmentation model is proposed. The model firstly uses ASR to obtain speech transcription text, while designing a novel dynamic OCR recognition and similarity detection de-duplication algorithm to extract textual information from video frames, effectively reducing redundant frame processing. Subsequently, ASR text and OCR text are fed into text encoders for encoding, and a multimodal attention mechanism is employed to fuse features from both modalities. Finally, topic boundary prediction is performed through BiLSTM and MLP networks. Experimental results demonstrate that the proposed solution method outperforms baseline methods in terms of accuracy and scalability in the smart classroom domain. Meanwhile, the proposed dynamic OCR recognition and similarity detection de-duplication algorithm reduces the model's processing time, meeting the requirements of real-world applications.

Keywords smart classroom; multimodal; topic segmentation; de-duplication

近年来,随着现代科学技术的不断发展与进步,智慧课堂的新模式不断走入各级教育机构,特别是高校.随着新模式的普及,通过智慧课堂的录制产生了丰富的教学资源,其中通过自动语音识别(Automatic Speech Recognition,ASR)系统生成的口语文本记录缺乏结构性的注释,如段落.而段落或章节是语义上连贯的子文件单位,通常与主题或子主题的边界相呼[1].由于语音转录文本整段内容缺

收稿日期 2024-12-08
* **通信作者** 周斌(1971-),男,教授,博士,研究方向:大数据处理,E-mail:binzhou@mail.scuec.edu.cn
基金项目 湖北省技术创新专项基金资助项目(2019ADC071);中央高校基本科研业务费专项资金资助项目(CZY23006)

乏段落会给信息的提取和利用带来困难，因此如何在智慧课堂主题分割领域充分利用口语文本这一资源成为语言理解领域的一项新挑战.

智慧课堂主题分割，旨在将数小时的教学视频分割成更小的片段，这些片段由一系列讲述相同或相似主题内容的视频帧和相对应的语音转录文本组成，如图 1，它显示了主题分割后的视频时间线，中间是章节中的 ppt 内容，右侧显示了教师授课时的视频文字记录.智慧课堂主题分割完成了对课堂内容的分割工作，将长视频转换成一系列主题连贯的片段，这些片段揭示了智慧课堂背后的粗粒度语义结构.

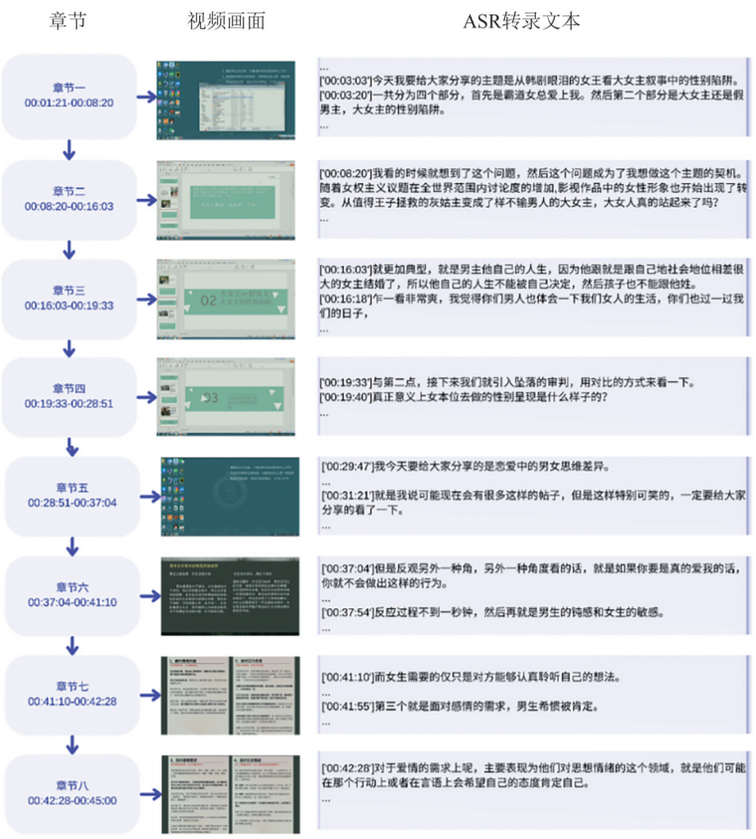


图 1　智慧课堂的示例

Fig. 1　Example of a smart classroom

智慧课堂主题分割在现代教育中具有重要意义，它产生的章节段落结构不仅简化了知识理解，还帮助教师和学生能够轻松找到感兴趣的课堂内容.更重要的是，它还影响着非常多的下游任务.例如，生成章节摘要，对于一节一个半小时的课程，转录的文本会超过 20000 字，无法通过直接输入到大模型中生成课程摘要.相反，通过使用智慧课堂主题分割模型将长转录文本分割成连贯的，主题单一的章节，然后分别对相应章节执行文本摘要，这可对整节课堂内容生成更精确更有效的内容简介.此外，还为教育研究和教学质量评估提供了重要的数据支持，有助于改进教学方法和课程设计、构建知识图谱[2]等，同时有助于辅助智慧课堂中的下游任务（话语解析）[3].

主题分割问题已成为自然语言处理（Natural Language Processing，NLP）领域里的热点问题，主要包括文本主题分割和视频主题分割.在文本主题分割的研究中，现在的无监督模型主要通过计算文本相似[4]或探索文本的主题[5]表示来推断主题边界.与无监督方法精心设计和使用的浅层特征不同，有监督神经模型可以对更深层的语义信息进行建模，并从标注数据中探索话题转移的线索[6].有监督模型通过预训练语言模型（Pre-trained Language Models，PLMs）和在大规模有监督数据集上进行微调，在话题分离方面取得了很大的进步[7-10].

近些年，研究人员[11-14]通过对文本连贯性进行外部建模，提高了主题分割性能，但这些要么忽略相邻句子的上下文建模，要么需要依赖高质量的长文档，或者需要为辅助任务构建增强数据，在长文档的主题分割上将消耗更多的训练时间.因此，基

于单一文本特征进行主题段落分割的现有方法存在诸多局限.

在早期的视频分割计算模型中,主要针对镜头或场景检测,仅利用时空方面或帧颜色等表面视觉特征[15-18].这些方法通常测量视频时间线上的时间相似性,以预测镜头/场景边界.尽管与镜头/场景分割的定义不同,智慧课堂主题分割的任务更多地关注视频中与主题相关的语义,这不一定与视觉变化相一致.如图1所示,尽管主题发生了巨大变化,但在智慧课堂的视觉背景中,视频画面有相当长的一段时间内仍然保持相似.此外,之前的视频分割方法大多集中在视觉变化清晰、模式简单的短视频[19-22].短视频的这些独特特征可以在模型设计中强调,也可以在监督设置中学习,但同时使这些模型对更长、更细微的视频内容(如纪录片或教学直播,智慧课堂录播)的适应性呈下降趋势.

针对现有问题,本文提出了一种新颖的多模态智慧课堂主题分割模型,该模型可以将对齐的语音转录文本和视频帧经过光学字符识别(Optical Character Recognition, OCR)处理后得到的相应文本作为输入,这种丰富的输入大大提高了模型的性能,因为文本和视频帧的互补可以更全面地反映输入视频的主题相关语义[23];同时模型加入了动态OCR识别与相似度检测去重算法,从而去除冗余帧,保留关键帧,提高模型处理速度,使得多模态融合处理更加高效.模型中也引入了多模态注意力层,使模型对视频信息和语音转录文本信息不再是简单的融合,而是以互补的方式有效地整合文本和视觉信号.在新收集的大学智慧课堂录播数据集上训练和测试本文提出的分割器.

1 多模态智慧课堂主题分割模型

1.1 问题定义

受到基于神经网络的文本主题分割监督方法的启发,本文将智慧课堂主题分割作为序列标记任务,以通过ASR转录得到的文本作为标记单位.一个智慧课堂的教学视频包含:(1)ASR转录得到的一系列带时间范围的文本,(2)一系列有时间戳的视频帧.与传统文档主题分割不同的是,传统文档主题分割的最小单元是独立的句子,本文中提出的模型则使用ASR转录独立的文本作为最小单元,这个最小单元往往包含一个或多个句子,单元的大小取决于语音端点检测技术(Voice Activity Detection, VAD)的切割结果,它能从带有噪声的语音中准确地定位出语音的开始点和结束点.为了提高模型的稳健性,本文将智慧课堂中的ASR(自动语音识别)转录文本划分为独立单元进行处理.在智慧课堂中,讲话者通常会连贯地表达相同主题的内容,这些内容不应参与主题分割的概率计算.同时,由于ASR转录后的独立单元文本通常属于同一讲话者,在面对多方对话的智慧课堂场景中,采用ASR转录后的独立文本作为最小处理单元,有助于模型保持稳健性.

智慧课堂的主题分割旨在将一段长视频及其对应的ASR语音转录文本划分为若干个相对独立的片段,每个片段包含若干句转录文本及相应的视频内容.为了实现主题分割功能,本文将该任务转化为针对文本及其对应视频的二分类问题.具体而言,为每个转录文本预测一个二进制标签,表示该文本是否指示主题段的边界,即是否为段落的最后一句.通过这种方法,模型能够有效识别主题边界,提高主题分割的准确性和鲁棒性.

1.1.1 模型输入

模型输入包括视频V及ASR转录文本$T_V=\{s_1,s_2,s_3,\cdots,s_n\}$以及每个独立文本的开始时间和结束时间序列$\{(b_1,e_1),(b_2,e_2),(b_3,e_3),\cdots,(b_n,e_n)\}$,同时还包括视频帧$X_v=\{x_1,x_2,x_3,\cdots,x_n\}$和单个帧$x_i$具有的时间戳$t_i$,单位为ms.

1.1.2 模型输出

模型输出为一系列标签$Y=\{y_1,y_2,y_3,\cdots,y_n\}$,其中$y_i$是一个二进制值,用于表示文本$s_i$是否为一个新主题章节的开始.对于转录文本序列,1表示相应的文本与视频主题片段边界重叠,否则为0.模型不需要预测最后一句话的标签y_n,因为根据定义,它必然是最后一段的最后一句,因此它等于1,即最后一段的末尾.

1.2 模型框架

本文提出的多模态主题分割模型基于分层编码-解码架构(图2),其框架与TextSeg[24]相似.它由两个分层链接的编码层组成:即上下文编码层(Contextual Encoder)和上下文解码层(Contextual Decoder).同时为了使文本和视觉模式对模型的预测起到互补作用,模型中添加了视频帧到文字转换的OCR方法和多模态注意力机制(图2中的交叉注意力层).这种设计可以将之前单模态的框架扩展到

多模态，旨在融合语音转录文本与视频帧的互补语义信息，并通过跨模态注意力机制增强主题边界的识别能力；同时还可以解决现有的多模态框架中传统方法直接处理原始视频帧，未针对长视频的重复内容（如静态PPT页面）进行去重，导致计算效率低下、冗余帧处理不足的问题、降低无关背景对主题边界预测的干扰.

本文模型接受语音转录和视频帧作为输入，并基于它们预测主题片段边界.模型框架的核心组件及流程见图2.

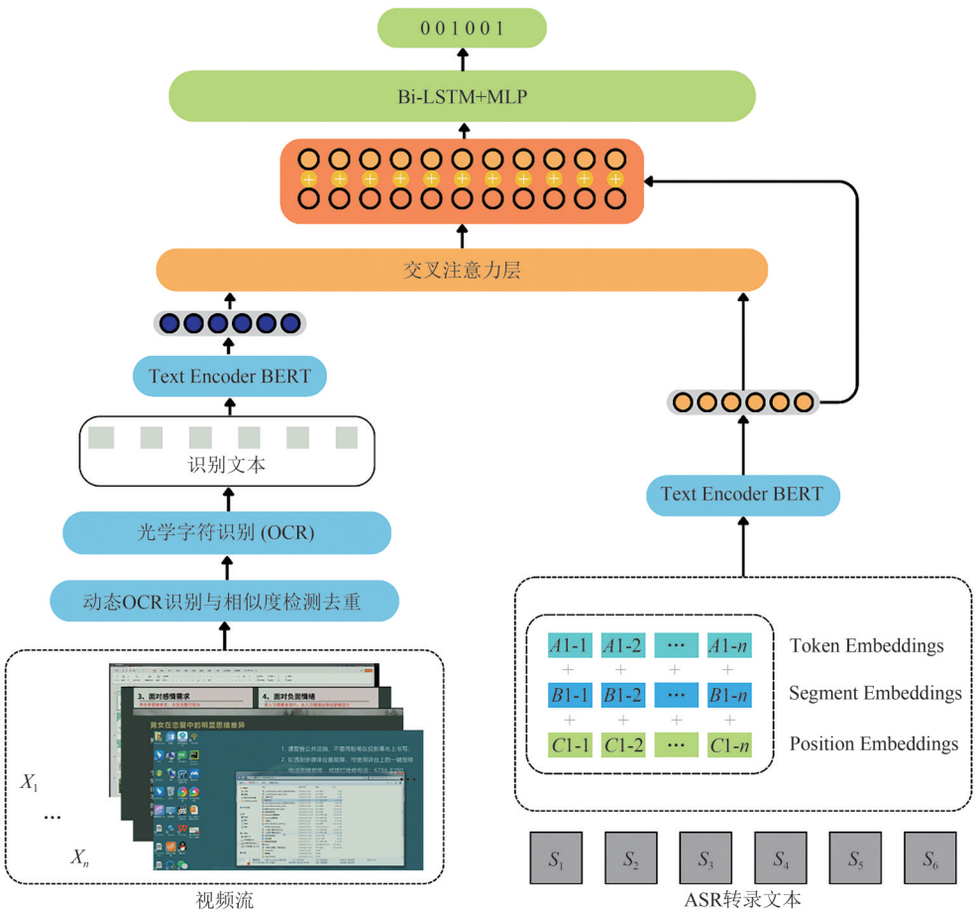


图2　多模态智慧课堂主题分割模型框架

Fig. 2　Multimodal smart classroom topic segmentation model

1.2.1　文本编码器

文本编码器为语音转录文本及视频帧OCR提取的文本进行深度语义建模，生成低级特征，实现上下文编码（Contextual Encoding）.令 $T_V=\{s_1,s_2,s_3,\cdots,s_n\}$ 表示一篇具有 n 个句子的转录文本，并且令每一个句子都是由若干数量的词组成，即 $s_n=\{t_1,t_2,t_3,\cdots,t_m\}$ 表示一个具有 m 个分词的句子.鉴于预训练模型BERT[25]在文本分割方面的优势[26]，本模型使用中文BERT-base模型来提取文本特征，BERT-base模型具有12层网络，使用768维嵌入和12个注意力头.其中，使用[PAD]特殊分词来对句子进行填充.通过对最后一层隐藏状态的平均池化操作，可以将变长句子编码为固定维度向量 $T^i\in\mathbb{R}^{768}$，从而实现将不定长的文本用统一大小的向量来表示，其计算公式如下：

$$\left\{T^n\right\}_{n=0}^{N}=\text{mean}-\text{pooling}\left(\text{BERT}_{\text{base}}\left(\left(\left\{t_n^m\right\}_{m=0}^{M}\right)\right)\right),\quad(1)$$

式中：t_n^m 表示 S_n 句的第 m 个分词；T^n 表示对相应句子编码后得到的句子嵌入.

1.2.2　动态OCR识别与相似度检测去重算法

图片去重算法的主要目的是识别并删除数据集中相同或非常相似的图像，从而优化数据管理，提高存储和处理效率.目前常见的图片去重算法主要有：(1)基于帧差分的去重算法：通过比较连续帧之间的差异来识别重复帧，如计算两帧图像对应像素点的灰度差值或计算结构相似性指数（SSIM）；(2)基于哈希的去重算法：对每一帧图像进行哈希处理，比较哈希值以识别重复图像，如PHash、dhash、aHash；(3)基于特征点匹配的去重算法：提取图像中的特征点，通过特征点匹配来识别重复图

像,如使用SIFT算法等.但这些方法普遍存在一些问题:方式单一、去重效果受阈值设定的影响、计算复杂度较高、耗时长等.

为了提高多模态融合处理的效率,本文提出了一种全新的动态OCR识别与相似度检测去重算法(其流程见图3),该算法通过分析视频帧中的OCR识别框数量变化和图像相似度,去除冗余帧并保留关键帧,从而优化处理流程.在动态OCR识别过程中,视频帧中提取的文本可能包含大量无效信息,如桌面图标、系统时间、噪点等.为确保提取的文本与教学内容相关,本研究采用了一套多层次的过滤策略.首先,通过空间过滤剔除无效区域,仅保留高度≥20像素、宽度≥50像素的文本区域,并排除位于视频边缘(底部5%区域)的文本,这些区域通常包含无关字幕或噪点;其次,利用正则表达式匹配过滤非教学相关文本,如时间戳"12:30"或桌面图标文字"回收站",同时丢弃字符数≤3的短文本片段(如"是"、"的"等无意义短词).

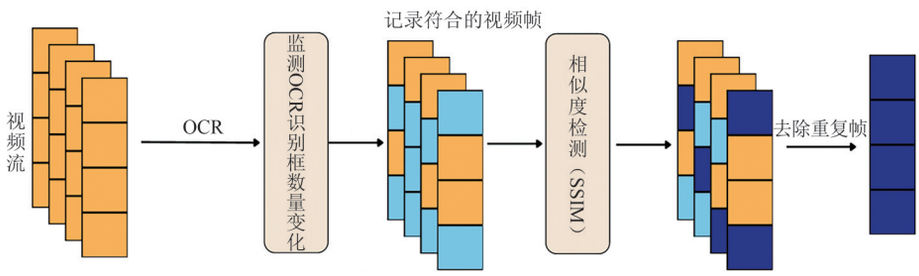


图3 动态OCR识别与相似度检测去重算法

Fig. 3 Dynamic OCR recognition and similarity detection de-duplication algorithm

首先,对于每一帧视频图像,进行OCR识别并记录识别框的数量.监测识别框的变化,当识别框数量出现显著下降或突变时,认为该帧可能包含新的信息,并记录当前帧,对识别框数量顺序递增的视频帧,智慧课堂中的老师还处在同一章节下的教学任务中,其包含的视频信息与按策略保留的视频帧信息存在重合,可以舍弃,接下来,通过计算相邻帧之间的图像相似度,进一步去除重复.在相似度检测过程中,本文采用了结构相似性指数(SSIM)作为度量标准.当两帧的相似度超过设定的阈值(0.85)时,将其标记为重复帧并删除.

在智慧课堂数据集上的实验表明,动态OCR识别与相似度检测去重算法对于智慧课堂视频(0.3帧/s)的去重率在85%以上,同时去重率受课堂授课中视频播放在整节课时中占比多少的影响,若课堂中播放视频占比越小,去重率越高.该算法显著减少了视频数据中的冗余帧数量,从而提升了处理效率.实验结果表明,在保证视频质量不受损的前提下,该方法有效地去除了重复帧,使得多模态融合处理更加高效.

1.2.3 生成视频帧表示

智慧课堂因其场景的特殊性,具有时间长、视频内容背景变化不明显、教师普遍依赖ppt作为教学工具等特点,传统方式中依靠对视频帧画面进行编码的操作很难充分利用视频画面的信息,同时也会显著增加模型处理的时间,阻碍了其在高校等场景下的大规模应用.

因此,本文创新性地通过提取视频帧中的文本信息来生成帧表示(具体流程见图4).由于授课老师的语音转录文本持续时间在秒级别,本文模型对视频进行每3 s采样一次,得到一系列视频帧X_v.这种采样方式既能保留足够的视频信息,又能避免因过高的帧率导致模型处理延迟过长.采用动态OCR识别和相似度检测去重算法,去除重复帧,最终得到一组新的视频帧$X_v'=\{x_1,x_2,x_3,\cdots,x_n\}$,每帧$x_i$具有对应的时间戳$t_i$,时间戳单位为ms.

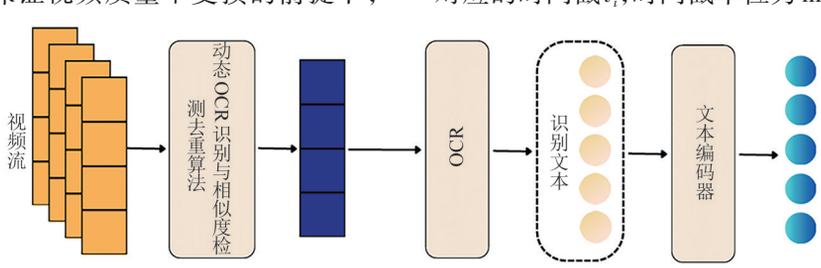


图4 生成视频帧表示

Fig. 4 Generate a representation of the video frame

在文本信息提取方面，选用PP-OCR[27]模型对视频帧中的文字进行识别，作为视频帧信息的表达方式．随后，将每帧画面的识别文本输入文本编码器，提取固定大小的文本嵌入，获得对应帧的帧表示．

1.2.4　文本感知的视觉信息表示（多模态注意力机制）

如果给出一个ASR转录文本i及其时间区间$[b_i, e_i]$，以及这个时间区间内的视频帧，则可以通过文本编码器和生成视频帧表示两个模块，得到ASR识别文本表示和文本对应的视频帧的帧表示．

随后，为了实现文本与视觉信息的动态融合，模型引入了跨模态注意力机制，为每个自动语音识别（ASR）识别的内容生成文本感知的视觉表示，如图5所示．这一过程并非简单地通过均值池化一组覆盖句子区间的帧来获得视觉表示．设计这一机制的初衷是为了让与文本共享更多语义信息的帧，在生成视觉表示时被赋予更高的关注权重；相反，那些与文本无关或几乎无关的帧可能会对传递给后续模块的融合多模态表示的质量产生负面影响[28]．在实际应用中，多模态注意力模块采用标准的标量点积注意力函数[29]来实现．对于每个ASR句子嵌入，通过前馈网络生成查询向量q_i；对于视频帧嵌入，则生成键K_i和值V_i．

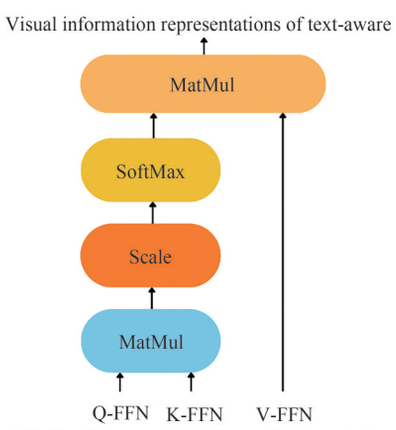


图5　文本感知的视觉信息表示

Fig. 5　Visual information representations of text-aware

通过键值，计算文本感知的视觉表示vr_i如下：

$$vr_i = A_i V_i, \tag{2}$$

$$A_i = \mathrm{softmax}\left(\frac{q_i K_i^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{3}$$

其中$q_i \in \mathbb{R}^{1\times d_k}$、$K_i \in \mathbb{R}^{1\times d_k}$、$V_i \in \mathbb{R}^{1\times d_k}$表示通过将句子表示和帧表示分别传递3个并行前馈层（即Q-FFN、K-FFN和V-FFN）生成的查询向量、键和值矩阵，其中K-FFN和V-FNN共享相同的参数．

1.2.5　上下文解码与边界预测

最后模型将所有获得的文本感知视觉表示与它们对应的句子表示拼接成联合特征$H_i = [T^i;\ vr_i]$，并馈入双向LSTM（BiLSTM）捕捉跨句子依赖关系，该层执行上下文并返回隐藏状态．接下来，多层感知机（MLP）和Softmax充当主题边界预测器，根据阈值τ对输入隐藏状态进行二进制预测．如果转录句子的输出概率超过τ，则标记为1，表示分段边界，整个模型使用二分类交叉熵损失进行训练．

$$\mathrm{Loss} = -\left(y \times \log\left(\hat{y}\right) + \left(1 - \mathrm{y}\right) \times \log\left(1 - \hat{y}\right)\right), \tag{4}$$

其中$\hat{y}$是模型预测样本是正例的概率，即该文本为主题段落的最后一句，y则是样本标签，其中如果样本属于正例，取值为1，否则取值为0．

2　实验设置和实验结果

为了评估本文提出的智慧课堂主题分割模型的有效性和通用性，本文在自建的数据集上进行了评估．

2.1　域内数据集

域内数据集（In-domain Dataset）是指与研究目标或任务在特征、数据分布和上下文等方面高度相关和一致的数据集．使用域内数据集进行模型训练和测试，可以确保模型在实际应用中的有效性和可靠性，因为这些数据集与预期的应用场景具有相似的属性和特点．

对于域内分段推理，本文使用来自同一域（语料库）的数据，训练和测试模型．由于视频分割领域缺乏大规模的人工注释数据集，本文收集了一个新的语料库，由采集自多所国内高校的智慧课堂中的录播视频组成，涵盖多类学科（政治、计算机、心理等），同时包含多种课堂教学形式（如教师授课，翻转课堂等），视频分辨率统一为1920 × 1080，帧率25 fps，音频采样率32 kHz．在视频采集过程中，采用严格的筛选标准以确保数据集质量．这些标准包括对视频长度（45 min左右）、字数（平均>0.5字/s）、章节持续时间（平均时间>5 min）和句子长度的限制（长度<60个标记），以确保数据分布符合真实课堂场景．该语料库中的每个视频都与一系列章节时间戳相关联，这些章节时间戳指示了该视频中包含的每个主题章节的开始．因此，本文使用这些可用的

章节开始时间戳作为基本事实主题段边界,因为主题章节从定义上讲是视频中的主要主题划分.语音转录结果则是由阿里ASR模型生成,带有令牌级偏移.如表1所示,本文数据集中视频时长总计5048 min,ASR识别文本约110万字.同时以6:2:2的比例划分数据集.与其他现有的视频理解语料库相比,本文构建的智慧课堂语料库具有如下优点:(1)规模更大;(2)涵盖更多不同的课堂主题,包括政治,历史,数学,计算机,心理等相关主题;(3)包含多种课堂教学形式;(4)具有可靠分割.

表1 智慧课堂语料库数据集的统计信息

Tab. 1 Statistical information on the Smart Classroom Corpus dataset

数据集指标	值
句子数	56346
字数(含标点符号)	1099485
句子平均长度	19.5
段落数	855
视频总时长/min	5048
语速/(字数/min)	217

2.2 评估指标

模型性能通过3项指标进行评估,分别是精确率(Precision)、$F1$分数、P_k,其中Precision和$F1$分数是流行于各领域的常用指标,P_k[30]是主题分割领域的专用评估指标.

精确率(Precision)描述了正确预测的边界数占所有预测边界数的比例,召回率(Recall)描述了正确预测的边界数占所有真实边界数的比率,$F1$则对精确率和召回率进行了折中处理.P_k则是一种基于统计技术的指标,指标越小,算法越优,取值为0时表示完美分割.各指标的公式如下:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

准确率用于衡量预测边界与所有边界的准确匹配程度.

由于主题边界的判断相对主观,通常没有绝对的标准.尽管准确率和$F1$分数能够描述分割点的匹配情况,但它们难以体现各段文本主题的一致性.比如,现有的分割算法A和B,对同一序列进行了如图6所示的分割.

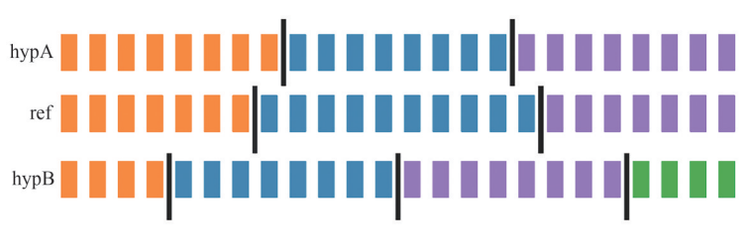


图6 算法A和算法B对主题边界的预测与真实分段位置的对比

Fig. 6 Comparison of predictions of topic boundaries with true segment locations by algorithm A and algorithm B

在图6中,ref代表真实边界,hypA和hypB分别是算法A和B的预测边界.同一颜色的方框表示这些句子属于同一主题,间隔则是划分的潜在边界,竖线表示真实边界或预测边界.

在提供的示例中,A和B的准确率分别为82.60和78.26,但$F1$分数均为0.尽管A的准确率更高,显示出其算法性能更优,但在多数情况下,稀疏的边界分布会使算法比较变得复杂.对于$F1$分数,尽管A的预测接近真实情况,并且在主题一致性方面表现良好,但$F1$分数无法区分两者的差异[31].

BEEFERMAN在1999年引入了一种新的评估指标:

$$P_k = \frac{1}{N - k + 1} \sum_{i=1}^{N-k+1} \delta\left(\text{ref}(i, i + k - 1) \overline{\oplus} \text{hyp}(i, i + k - 1)\right), \tag{8}$$

其中,P_k是评估指标值,N是文本中的句子总数,k是评估窗口的大小,通常取语料库内平均主题段落长度的一半,$\text{ref}(i, i + k - 1)$、$\text{hyp}(i, i + k - 1)$分别表示参考分割和预测分割在位置i和$i + k$之间是否存在边界(同一段为0,不同段为1),$\delta(x, y)$表示阶跃函数,当两者相同(同属一个段落或都不属于某一段落)时返回0,不同则返回1,$\overline{\oplus}$表示同或运算.该指标用于计算每个评估窗口内,参考分割和预测分割是否一致.指标值越低,表示分割结果与参考分割越接近,分割效果越好.

它就是使用一个窗口大小为k的滑动窗口,判断窗口的2个边缘的节点是否属于同一个主题,然后再看标准的判断和预测的判断是否一致即可,最后将一致的数量除以滑动次数就得出P_k值.

2.3 评估实验

2.3.1 模型处理时间

为了验证多模态模型在实际应用中的可行性,

本文进行了4组对比实验，使用智慧课堂数据集，具体实验设计如下：1）在不加入对视频的处理时，仅对长文段文本进行分段；2）使用本文提出的多模态智慧课堂主题分割模型进行主题分割，但不加入动态OCR识别与相似度检测去重算法，OCR模型已提前转换为onnx格式，以模拟工程实践中的环境；3）加入动态OCR识别与相似度检测去重算法后的时间，评估动态OCR识别与相似度检测去重算法对处理效率的提升；4）使用研究人员[32]应用于YouTube短视频的视频主题分割方法，其目的是与其他主流的多模态模型的运行时间进行对比.

所有实验均在以下硬件配置上进行：处理器为Intel Core i7-13700K，内存为64GB DDR5，显卡为NVIDIA GeForce RTX 4090，存储为2TB NVMe SSD，操作系统为Ubuntu 22.04 LTS.通过以上设置对每种方法的运行时间进行了模拟，并得到表2中的实验结果.

表2　处理时间对比

Tab. 2　Processing time comparison

实验编号	实验内容	运行时间（ms/节课）
1	仅文本	3178
2	文本+视频	294140
3	文本+视频（去重算法）	108452
4	文本+视频（YouTube视频分割模型）	324492

通过实验结果可以看出，不同数据处理方法在文本与视频处理中的运行效率差异显著.实验内容包括仅处理文本、同时处理文本与视频、应用去重算法及使用YouTube视频分割模型.每种方法的运行时间分别为3178 ms、294140 ms、108452 ms和324492 ms.

仅处理文本的实验运行时间最短，仅为3178 ms，这是由于文本处理相对简单，不涉及复杂的计算和多媒体数据处理.处理文本与视频的实验，运行时间显著增加至294140 ms，相对于仅处理文本增加了约9倍，这是因为视频数据处理需耗费较多的计算资源，包括视频解码、帧提取以及特征提取等过程.在处理文本与视频的基础上引入去重算法后，运行时间减少至108452 ms，节约了大约63.1%的时间，表明去重算法能够有效地减少重复数据的处理，从而降低了总体运行时间.这种优化在处理大规模数据时尤为显著.使用YouTube短视频分割模型的实验运行时间最高，达到324492 ms，相对于仅处理文本增加了约10倍，这主要是由于视频分割模型的复杂性和高计算资源需求.虽然该模型能够提供更精细的分割效果，但也增加了处理时间.

结合大学中智慧课堂的应用情况，由于处理视频并没有实时性的要求，且虽然智慧课堂主题分割模型的运行时间即使比单文本的运行时间要长，但在整个智慧课堂分析流程的用时中依然只占很少的比重，本文提出的智慧课堂主题模型能够更好地处理不同类型的数据，提升课堂教学视频在下游任务中的处理效果，尤其是良好的分段对于生成章节摘要和评价教师的教学质量更为重要.通过动态OCR识别与相似度检测去重算法，可以有效减少冗余数据，在保证实验结果精度变化不大的前提下大幅提高处理速度，适用于大规模数据分析和实时处理需求.通过这些实验，验证了多模态模型的实际可行性和高效性，为智慧课堂的应用提供了强有力的技术支持.

2.3.2　对比实验

目前，有研究人员[32]提出了双对比域自适应多模态主题分割模型，其在所有视觉和文本分割模型比较中表现最佳，针对研究中的私有YouTube数据集分割表现良好，Pr分数为30.61，可见视频与文本的结合可以提高主题分割的效果.

本文与该模型都为多模态视频分割模型，但本研究采用了智慧课堂的教学视频作为数据集，与研究人员[32]使用的YouTube视频存在显著差异.在引用该研究的方法时，有必要对数据集的差异进行详细描述和分析.在内容差异上，YouTube视频内容多样，涵盖娱乐、教育、新闻等多个领域，而智慧课堂视频专注于教育内容，具有更强的专业性和结构化；在视频质量上，YouTube视频的分辨率和质量参差不齐，而智慧课堂视频通常具有较高且一致的质量，以确保教学效果；在场景与背景上，YouTube视频的场景和背景变化较大，而智慧课堂视频通常在教室或固定的教学环境中录制，背景相对单一；在视频长度上，YouTube视频的长度不定，从几分钟到数小时不等，而智慧课堂视频往往有固定时长，如一节课时；在多模态信息上，智慧课堂视频通常包含教师的讲解、板书和PPT展示等多模态信息，而YouTube视频中，多模态信息形式更加多样.

因数据集存在较大的特征差异，所以本研究仅将双对比域自适应多模态主题分割模型引入运行时间的对比实验中.

实验细节上，文本编码器采用中文BERT-base模型（隐藏层维度768，12层Transformer），通过最后一层隐藏状态的平均池化生成句子嵌入.视频帧文

本信息通过PP-OCRv3模型提取,并经过动态OCR识别与相似度检测去重算法处理,保留关键帧后生成文本感知的视觉表示.跨模态注意力模块中,查询向量(Q-FFN)、键值矩阵(K-FFN/V-FFN)的维度设置为768,其中K-FFN与V-FFN共享参数.上下文解码部分采用双向LSTM(2层,隐藏单元256),后接MLP和Sigmoid激活函数进行边界预测.模型训练使用Adam优化器,初始学习率为0.00003.

实验中的参考指标主要选用精确率P、$F1$分数、P_k三个指标,其中精确率P和$F1$分数数值越大效果越好,P_k指标正好相反,数值越小,则模型效果越优.不使用准确率是因为在智慧课堂的主题分割中,智慧课堂具有视频时间长、分割段落点少的特点,导致分割点稀疏,而模型真正应该关注的是那些预测为1的分割点,而不是非分割点.

本文将与以下代表性基线模型进行比较:textspilt[33],TextSeg[24],Cross-segment BERT[9],Seq-BERT-Base[7].在智慧课堂环境中,由于ASR模型转录文本带来的识别错误,以及教师在授课内容间穿插的与教学内容无关的话语(如维持课堂纪律),导致长文本章节内容缺乏句子间的强连贯性;同时由于之前的方法大多基于单纯的文本进行分割,导致以往的模型都严重依赖文本的质量和文本之间内容的连贯性.而在引入视频信息之后,本文提出的模型在智慧课堂主题分割应用中有更好的鲁棒性,其性能相比于对比的基线模型都有提高,见表3.

表3 智慧课堂主题分割方法实验对比结果

Tab. 3 Experimental comparison results of smart classroom topic segmentation methods

算法	智慧课堂数据集		
	$P\uparrow$	$F1\uparrow$	$P_k\downarrow$
textspilt	6.3	9.4	51.81
TextSeg	24.12	37.59	34.89
Cross-segment BERT	27.88	43.45	31.06
Seq-BERT-Base	33.88	50.28	27.54
ours	42.96	59.42	23.22

通过实验结果可以看出,不同算法在智慧课堂数据集上的表现差异显著.其中,textspilt的表现最差,精确率和$F1$值最低,分割错误率最高.这表明该模型在智慧课堂数据集上的效果不理想.TextSeg模型的结果表明其在分割准确度上有所改进.Cross-segment BERT和Seq-BERT-Base进一步提高了各项指标,其中Seq-BERT-Base相比表现较好,其P值33.88,$F1$值为50.28,P_k值为27.54.

而本文提出的算法在所有指标上均表现最佳,其中P值为42.96,$F1$值为59.42,P_k值最低,为23.22.这验证了本文算法的有效性和优越性,突显了其多模态融合在处理智慧课堂视频分割任务中的潜力和优势.

仅依赖传统算法和单一模型,难以有效规避或消除这些噪音对分割结果的影响,从而限制了在具体的智慧课堂实践中的应用效果.然而,本文提出的新算法在这方面取得了显著进步.通过多模态智慧课堂主题分割模型,能更好地处理长视频中的稀疏分割点问题,从而提高了精确率P和$F1$值,降低了P_k值,这一改进为智慧课堂视频分割提供了更为精确和可靠的解决方案,在智慧课堂场景中的实际应用价值更高,能够为下游任务提供更加准确和可靠的分段结果.

3 结论

本研究以智慧课堂为背景,针对ASR系统的转录错误较高、课堂主题段落较为分散、段落内授课内容不连贯等问题,提出了一种多模态视频主题分割模型,该模型接受视频转录文本和帧作为输入.在自建的域内数据集上的实验结果表明:本文提出的模型在新收集的大学智慧课堂语料库上可以实现对课堂内容的整体段落分割,能够有效地应用在智慧课堂下游任务中,同时提出的动态OCR识别与相似度检测去重算法可以有效去除长视频中的重复帧,极大的提高了模型的运行速度.

参 考 文 献

[1] ALAM T, KHAN A, ALAM F. Punctuation restoration using transformer models for high-and low-resource languages [C]//Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). Stroudsburg: ACL, 2020: 132-142.

[2] XU Z, CRUZ M J, GUEVARA M, et al. Retrieval-augmented generation with knowledge graphs for customer service question answering [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC: ACM, 2024: 2905-2909.

[3] JIANG F, LIU W, CHU X, et al. Advancing Topic segmentation and outline generation in Chinese texts: The paragraph-level topic representation, corpus, and benchmark [C]//The 2024 Joint International Conference

on Computational Linguistics, Language Resources and Evaluation.Torino: ELRA and ICCL,2024:495-506.

[4] BARALDI L, GRANA C, CUCCHIARA R. A deep Siamese network for scene detection in broadcast videos [C]//Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane:ACM, 2015: 1199-1202.

[5] LAN D, BUNTINE W, JOHNSON M. Topic segmentation with a structured topic model[C]//The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: ACL, 2013:190-200.

[6] BADJATIYA P, KURISINKEL L J, GUPTA M, et al. Attention-based neural text segmentation [C]//Advances in Information Retrieval. Cham: Springer, 2018: 180-193.

[7] ZHANG Q, CHEN Q, LI Y, et al. Sequence model with self-adaptive sliding window for efficient spoken document segmentation[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Cartagena:IEEE, 2021: 411-418.

[8] INAN H, RUNGTA R, MEHDAD Y. Structured summarization: Unified text segmentation and segment labeling as a generation task[J]. arXiv:2022,2209.13759.

[9] LUKASIK M, DADACHEV B, PAPINENI K, et al. Text segmentation by cross segment attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2020: 4707-4716.

[10] 赵怡博, 蒋峰, 李培峰. 一种基于BERT的多级连贯性文本分割方法[J]. 计算机应用与软件, 2024, 41(10): 262-268, 324.

[11] GLAVA Š G, SOMASUNDARAN S. Two-level transformer and auxiliary coherence modeling for improved text segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7797-7804.

[12] LO K, JIN Y, TAN W, et al. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: ACL, 2021: 3334-3340.

[13] WANG L, LI S, LV Y, et al. Learning to rank semantic coherence for topic segmentation[C]//Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen. Stroudsburg, : ACL, 2017: 1340-1344.

[14] ARNOLD S, SCHNEIDER R, CUDRÉ-MAUROUX P, et al. SECTOR: A neural model for coherent topic segmentation and classification[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 169-184.

[15] RAO A, XU L, XIONG Y, et al. A local-to-global approach to multi-modal movie scene segmentation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10143-10152.

[16] RASHEED Z, SHAH M. Scene detection in Hollywood movies and TV shows[C]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.Madison:IEEE, 2003: II-343.

[17] RUI Y, HUANG T S, MEHROTRA S. Exploring video structure beyond the shots [C]//Proceedings of IEEE International Conference on Multimedia Computing and Systems.Austin:IEEE, 2002: 237-240.

[18] CHEN S, NIE X, FAN D, et al. Shot contrastive self-supervised learning for scene boundary detection [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 9791-9800.

[19] GYGLI M, GRABNER H, RIEMENSCHNEIDER H, et al. Creating summaries from user videos[C]//Computer Vision-ECCV 2014. Cham: Springer, 2014: 505-520.

[20] JADON S, JASIM M. Unsupervised video summarization framework using keyframe extraction and video skimming [C]//2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA). Greater Noida:IEEE, 2020: 140-145.

[21] JAYARAMAN D, GRAUMAN K. Slow and steady feature analysis: Higher order temporal coherence in video[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 3852-3861.

[22] ZHU W, PANG B, THAPLIYAL A V, et al. End-to-end dense video captioning as sequence generation[C]//The 29th International Conference on Computational Linguistics (COLING 2022). Gyeongju: ICCL, 2022: 5651-5665.

[23] JAMES N, TODOROV K, HUDELOT C. Combining visual and textual modalities for multimedia ontology matching [C]//Semantic Multimedia. Berlin: Springer, 2011: 95-110.

[24] KOSHOREK O, COHEN A, MOR N, et al. Text segmentation as a supervised learning task[C]//Proceedings of the 2018 Conference of the North American Chapter Ofthe Association for Computational Linguistics: Human LanguageTechnologies, Volume 2 (Short Papers). Stroudsburg:ACL, 2018: 469-473.

[25] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language

understanding[C]//The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL,2019: 4171-4186.

[26] XING L, HACKINEN B, CARENINI G, et al. Improving context modeling in neural topic segmentation [C]//The 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). Virtual Event: ACL,2020: 626-636.

[27] DU Y, LI C, GUO R, et al. PP-OCR: A practical ultra lightweight OCR system[J]. arXiv: 2020,2009.09941.

[28] WANG Z, ZHONG Y, MIAO Y, et al. Contrastive video-language learning with fine-grained frame sampling[C]// Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: ACL, 2022: 694-705.

[29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st Conference on Neural Information Processing Systems. Long Beach: Curran Associates,2017: 6000-6010.

[30] BEEFERMAN D, BERGER A, LAFFERTY J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1): 177-210.

[31] PEVZNER L, HEARST M A. A critique and improvement of an evaluation metric for text segmentation [J]. Computational Linguistics, 2002, 28(1): 19-36.

[32] XING L, TRAN Q, CABA F, et al. Multi-modal video topic segmentation with dual-contrastive domain adaptation [M]//MultiMedia Modeling. Cham: Springer Nature Switzerland, 2024: 410-424.

[33] ALEMI A A, GINSPARG P. Text segmentation based on semantic word embeddings[J]. arXiv:2015,1503.05543.

(责编 曹东,校对 雷建云)