Journal of Al-Azhar University Engineering Sector

Vol. 18, No. 69, October 2023, 900 -913

SEMANTIC SEGMENTATION OF EDUCATIONAL VIDEOS FOR MICRO LEARNING OBJECTS IN ADAPTIVE E-LEARNING

Abdelrahman S. Halawa*, Shehab Gamalel-Din, Abdurrahman A. Nasr
Systems & Computers Department, Faculty of Engineering, Al-Azhar Unversity, Cairo, Egypt.

*** Correspondence**: ahalawa@azhar.edu.eg

Citation:

A.S. Halawa, S. Gamalel-Din, A. A. Nasr "Semantic Segmentation of Educational Videos for Micro Learning Objects in Adaptive E-Learning", Journal of Al-Azhar University Engineering Sector, vol. 18, pp. 900-913, 2023.

Received: 7 August 2023

Accepted: 21 September

DoI:10.21608/auej.2023.234873.1416

ABSTRACT

E-Learning is gaining prominence, especially in lifelong learning, primarily through lecture videos. However, these videos often encompass multiple topics or serve various instructional roles within a single subject. In adaptive e-Learning, the smaller and granular the units, the more versatile presentations and personalized lectures are composed. Such units are known as Micro Learning Objects (MLOs). Consequently, the necessity emerges to segment these lecture videos into multiple MLOs, each fulfilling a distinct instructional role in a lecture. This article presents an automatic model leveraging advanced language models to segment lecture videos semantically into Micro Learning Objects (MLOs). Additionally, a new well-segmented dataset of educational videos (YT-EV) was introduced, in which the video is segmented according to a pre-defined timestamped agenda. The model is trained on general text datasets to understand LO segments and subsequently fine-tuned using transfer learning on video datasets to achieve better segmentation results. The experimental results showed an F1-score of value 0.657, which is considered promising and emphasizes the significance of text transcript-based video segmentation for enhancing adaptive e-Learning.

KEYWORDS: language model, NLP, Learning Object, video segmentation, Transfer Learning.

التقسيم الدلالي لمقاطع الفيديو التعليمية لكائنات التعلم الصغيرة في التعلم الإلكتروني التكيفي

عبد الرحمن سمير حلاوة*، شهاب جمال الدين، عبد الرحمن نصر
قسم هندسة النظم والحاسبات، جامعة الأزهر، القاهرة، مصر

***البريد الاليكتروني للباحث الرئيسي:** ahalawa@azhar.edu.eg

الملخص العربي: -

التعلم الإلكتروني يكتسب شهرة، خاصة في التعلم طويل المدى، وذلك بشكل رئيسي من خلال المحاضرات التي تحتوي مقاطع الفيديو. ومع ذلك، تغطي هذه المقاطع في كثير من الأحيان مواضيع متعددة أو تخدم أدوارًا تعليمية متنوعة ضمن موضوع واحد. في التعلم الإلكتروني التكيفي، كلما كانت الوحدات أصغر وأكثر تفصيلًا، كلما تم إعداد عروض ومحاضرات مخصصة أكثر مرونة. تُعرف مثل هذه الوحدات بأسماء كائنات التعلم الصغيرة Micro Learning Objects (MLOs). وبناءً على ذلك، يظهر الحاجة إلى تقسيم مقاطع فيديو المحاضرات هذه إلى MLOs متعددة، حيث يقوم كل منها بأداء دور تعليمي مميز في المحاضرة. يقدم هذا المقال نموذجًا تلقائيًا يستفيد من نماذج اللغة المتقدمة لتقسيم مقاطع فيديو المحاضرات بشكل دلالي إلى كائنات تعلم صغيرة (MLOs). بالإضافة إلى ذلك، تم تقديم مجموعة بيانات جديدة

مجزأة جيدًا لمقاطع الفيديو التعليمية (YT-EV)، حيث يتم تقسيم الفيديو وفقًا لجدول زمني محدد مسبقًا. تم تدريب النموذج على مجموعات بيانات نصية عامة لفهم أقسام كائنات التعلم ومن ثم يتم ضبطه بشكل أفضل باستخدام التعلم عبر النقل على مجموعات بيانات الفيديو لتحقيق نتائج تقسيم أفضل. أظهرت النتائج التجريبية F1-score بقيمة 0.657، والتي تعتبر واعدة وتؤكد على أهمية تجزئة الفيديو المستندة إلى النص لتعزيز التعلم الإلكتروني التكيفي.

الكلمات المفتاحية: نموذج لغة، معالجة اللغة الطبيعية، كائن التعلم، تقسيم الفيديو، التعلم عبر النقل.

1. INTRODUCTION

E-Learning is offering a flexible alternative to traditional educational methods and, hence, is gaining traction and expectedly will play a major role in the future of Long-Life Learning. However, concerns arise about the one-size-fits-all approach in asynchronous e-Learning where the instructor is absent. Adaptive teaching, resembling personalized tutoring, tailors learning experiences to each individual student [1]. While private tutoring is effective, scalability is a challenge. Adaptive/personalized online learning provides scalable personalization, while adjusting content and teaching strategies for each individual learner separately [2], hence, mirrors one-to-one tutoring in enhancing learning.

Adaptive e-Learning strives for efficient, effective, and engaging knowledge acquisition. It relies on student models, granular Learning Objects (LOs), and concept ontologies [3]. The smaller the granularity of the LOs the effectiveness in shaping and personalizing new lectures. Atomic LOs are called Micro LOs (MLOs). MLOs in adaptive learning systems play an analogous role of the small cubes of the Lego™ puzzle game in which the same small cube can be reused in constructing multiple shapes as needed; similarly, MLOs can be reused in the composition of multiple adaptive lessons. The focus in this research is on segmenting large granular lecture's videos into smaller sized MLOs.

Text segmentation is a common NLP activity that divides text into components based on established criteria. It may be applied to documents or video transcripts, with the goal of generating logically cohesive units. Dividing a text into segments of a single mode can enhance and expedite subsequent applications. As an example, text segmentation has proven effective in enhancing information retrieval by indexing subdocument components instead of entire documents [4]. Similarly, other functions like summarization and information extraction can reap advantages from text [5]. These same advantages can also be realized in video transcripts.

This study was built upon our pervious [6] which focused on segmenting well-structured documents such as articles and theses. In contrast, this research extends their efforts to the realm of video segmentation, which is a prevalent form of multimedia in asynchronous e-Learning. Noteworthy, in video lectures are commonly unstructured unlike articles that are well structured by authors. In addition, video segmentation for extracting MLOs is different from other types of video segmentation in terms of its specific educational focus and purpose.

We introduce three techniques for segmenting semi/unstructured text, aiming to handle fully unstructured text, such as video transcripts. To achieve this, we harness the capabilities of language models like pre-trained BERT [7] further enhancing their performance through transfer learning on new datasets. Due to the absence of a suitably segmented video dataset, we undertook the creation of a new dataset by collecting educational videos from platforms like YouTube™, as elaborated in

Sections 4 and 5. The segmentation outcomes are intended for use in generating Learning Objects (LOs) through segment merging or splitting, a topic discussed in Section 3, albeit beyond the scope of this article.

In this article, Section 2 examines similar studies and highlights a few common segmentation features especially for videos, while Section 3 explains the concept of Learning Objects in the context of adaptive e-Learning systems. Section 4 describes how the segmentation model and methodology evolved via a series of experiments with the target of auto-detecting segmentation in semi-structured text. Section 5 applies this evolved model on video transcript. Finally, Section 6 summarizes the findings and suggests a few future directions.

2. RELATED WORK

Video segmentation plays a critical role in a broad range of practical applications, especially in Multimedia Learning Object domains. Extensive research has been conducted in this field over the years. Video segmentation approaches can be categorized into two main types: only text-based algorithms and text/audio/image-based algorithms, as discussed below.

2.1 Video segmentation based on text algorithms.

One approach within text-based methods involves the integration of content-based and boundary-based techniques [8]. In this approach, the initial segmentation is computed in the first pass by analyzing the temporal distribution and arrival rate of features. Subsequently, the second pass involves the detection of changes in content-bearing words using the content-bearing features. [9] have introduced a method that enhances accuracy by combining various segmentation features, including noun phrases, topic noun phrases, verb classes, word stems, combined features, cue phrases, and pronouns. In contrast, [10] utilize natural language processing techniques such as noun phrase extraction and leverage lexical knowledge sources like WordNet to segment lecture videos. Additionally, [11] propose an approach for identifying segment boundaries by matching blocks of SRT (subtitle resource tracks) with Wikipedia texts related to the lecture video's topics. They begin by generating feature vectors based on noun phrases within both the Wikipedia text blocks (each corresponding to a Wikipedia topic) and SRT blocks (consisting of 120 words each). Subsequently, cosine similarity is employed to measure the similarity between a Wikipedia block and an SRT block. Finally, a segment boundary is defined as an SRT block that exhibits both the maximum cosine similarity and surpasses a predefined similarity threshold.

Furthermore, [12] introduces a cross-segment attention mechanism designed to identify significant boundaries within text through the capture of inter-segment relationships. This method, which considers context and connections between segments, shows promising outcomes in enhancing the accuracy and effectiveness of text segmentation. On the other hand, [13] present an unsupervised approach that utilizes a novel similarity score based on BERT embeddings [14], distinguishing itself from similarity score heuristics unrelated to neural models. Additionally, [15] proposes a transformer-over-transformer system, named transformer2, for neural text segmentation. It comprises bottom-level sentence encoders employing pre-trained transformers and an upper-level transformer-based segmentation model utilizing sentence embeddings. In contrast, [16] suggests a method leveraging a pre-labeled text corpus in conjunction with an enhanced neural deep learning

model. BERT serves as a robust sentence encoder, demonstrating that state-of-the-art results can be achieved with minimal training through the use of a text segmentation-focused data augmentation strategy.

2.2 Video segmentation based on audio/image/text algorithms.

A considerable body of research leveraging video features, including audio and frames, has been documented. For instance, in the work of [17], an approach was presented for the migration of legacy video lectures into digital learning objects. This method identifies slide transitions, extracts information from a presentation document (such as author name, title, and date of creation), acquires slide images, populates learning object metadata, and captures the presentation's table of contents. Another study by [18] focuses on the segmentation of videos into distinct scenes. This is achieved by detecting frame transitions through the analysis of color histograms within lecture video frames. [19] introduced a framework for the development of effective multimedia learning objects, which can be seamlessly integrated with Learning Management Systems (LMSs). This framework facilitates the creation, storage, distribution, and evaluation of learning objects automatically extracted from digital media.

Furthermore, in the study conducted by [20], a supervised method is proposed that incorporates visual features along with transcripts. This approach involves training a Support Vector Machine (SVM) on lecture videos to detect changes in events, such as "speaker writing on the blackboard" or "slide presentation," from which fragment boundaries are extracted. In another work by [21], an innovative solution is introduced that employs boosted margin maximizing neural networks to efficiently index educational videos. Through the utilization of neural networks and boosted margin maximization, this method exhibits promise in accurately identifying and organizing crucial segments within lecture videos, thereby enhancing content retrieval and accessibility. [22] focuses on the development of a segmentation method for lecture videos based on speech patterns, including pitch, volume, pause rates, and the initial time of each audio chunk, as well as content cues. Leveraging speech recognition techniques, the proposed approach aims to accurately identify and segment lecture videos into meaningful segments.

All of these researches focused on segmentation for different purposes, not including MLOs. To explain, an MLO explains one single tiny instructional role (e.g., example, overview, introduction …etc.) for a single concept. Therefore, the main focus of this research is to identify MLOs in video lectures.

3. THE CONCEPT OF LEARNING OBJECTS (LO) IN THE CONTEXT OF E-LEARNING

Learning materials in e-Learning are made up of multimedia learning objects known as Learning Objects (LOs). LOs are reusable, modular instructional content units developed to improve digital learning experiences. They are self-contained learning materials that, when associated with particular learning objectives, provide flexibility in curriculum design, enabling varied teaching styles as personalization necessitates [23]. They come in a variety of sizes, ranging from little multimedia pieces to large modules. Indeed, LOs encourage resource efficiency, scalability, and embedded assessments, making them essential tools in current educational technology and e-

learning, allowing for the production of flexible, adjustable, and pedagogically successful digital learning experiences.

Noteworthy, the smaller the granularity of the LOs the effectiveness in shaping and personalizing new lectures. Atomic LOs are called Micro LOs (MLOs). MLOs in adaptive learning systems play an analogous role of the small cubes of the Lego™ puzzle game in which the same small cube can share in constructing multiple shapes as needed. The focus in this research is on segmenting large granular lecture's videos into smaller sized MLOs [3]. Every MLO must be properly specified by a set of metadata attributes through which the object selection criteria are determined [19]. Noteworthy, the metadata determination is outside the scope of this article, the focus is only on LOs identification.

Video segmentation for extracting MLOs is different from other types of video segmentation in terms of its specific educational focus and purpose. While general video segmentation aims to divide videos into segments based on various criteria like scene changes or content shifts, video segmentation for LO extraction is primarily driven by educational objectives. In MLO extraction, the goal is to identify and separate video segments that contain distinct and meaningful educational content.

4. DEFINING A SEGMENT IN MACHINE TRAINING

This research foresees a video as a multidimensional data element, three dimensions for accuracy, namely, voice, images, and transcript. All dimensions cooperate to identify accurate segmentation. This research foresees the transcript as the base segmentation element, while the other two dimensions aid accuracy enhancements. Therefore, this section discusses how the video transcript is segmented.

A core step in machine learning to build a model for a certain concept is machine training. The goal was to train the machine to be able to identify the proper segmentation for a video transcript. Therefore, the first research problem was "how to define a text segment and train the machine to identify it?" This section discusses the experimental work that this research went through to answer this research question.

To answer this research question, two approaches were experimented. The first approach tests the use of the full context of the input text, which takes care of each sentence in the context and thus affects the final constructed model. while the second one focuses on the context around the expected separators, in which the context is divided into pieces of sentences and then labeled based on their positions. Three experiments were conducted; two of them employed the first approach, while the third employed the second approach.

The aim of this research is to find the best method to train the machine on how to predict the place separating two segments. This section exploits the power of a few recent language models to detect such separators in a text in general, while the next chapter applies the results of the experimental work of this section on video transcripts.

The three experiments used a pre-trained small BERT English uncased version [7] (it uses 4 hidden layers, a hidden size of 512 dimensional embeddings, and 8 attention heads) model, The pre-trained model is then fine-tuned, involving training the weights of the pre-trained model on new data to

enhance its performance in segmentation tasks. The three approaches use the same model architecture with different preprocessing and labeling techniques. The architecture of the used model consists of the following main five layers (as depicted in **Fig. 1**):

1. Input Layer
2. Pre-processing Layer
3. Encoder Layer, then use the pooled output from the BERT output map
4. Dropout layer
5. Dense layer with sigmoid activation

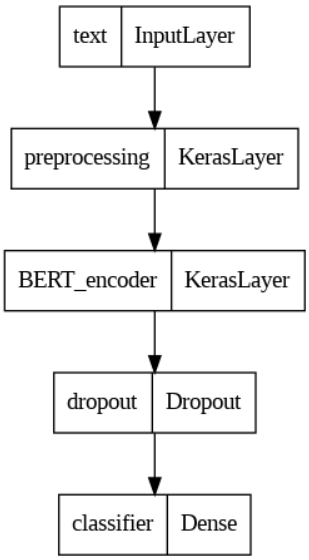


Fig. 1: The architecture of the used model.

4.1 The Dataset

Halawa, et al. [6] tried to segment well-structured text (articles and theses), while in this article we continue our research to handle unstructured text data (video transcripts). This section steps over semi-structured data, leaving the fully unstructured text data (video transcripts) to the next section. Two datasets, Choi's dataset [24] and Manifesto [25] dataset, were selected for the three experiments on semi-structured text data, especially because they met the following criteria:

1. Unstructured/semi-structured text dataset.
2. Published and commonly used in segmentation research.
3. Already segmented or can be segmented easily by code without human efforts.

Choi's dataset [24] is an early dataset containing 920 synthetic documents made of concatenated extracts of news articles. Each document is made of 10 segments, where each segment was created

by sampling a document from the Brown corpus and then sampling a random segment length up to 11 sentences. For our planned experimentation, this data set was randomly sampled as follows: 25% of documents as a test set, 10% of documents as a validation set, and the remaining 65 % of documents are used for training.

Manifesto Project[1] dataset [25] created a real-world dataset of political texts from the Manifesto Project, which were manually labeled by domain experts with segments of seven different topics (e.g., economy and welfare, quality of life, foreign affairs, etc.). The selected manifestos contain between 1000 and 2500 sentences, with segments ranging in length from 1 to 78 sentences each.

In each of the three experiments, both datasets, namely, Choi's [24] and Manifesto [25], were used to ensure that the results do not depend on the dataset itself, as discussed below. Before using the datasets, text inputs had been preprocessed by NLP techniques, such as text normalization, removal of stop words, elimination of Unicode characters, etc. In fact, each of the three experiments had its own data preparation to make the data ready for training, as discussed below.

4.2 Experiment 1: Labeling Data by Zeros and Ones.

In this method, and before applying the model as shown in **Fig. 1**, the data were preprocessed where each segment was split into a sequence of sentences, the last sentence of which (the separator sentence) was labeled by "1", while all other sentences were labeled by "0". **Table 1** and **Table 2** show a random sample from Choi and Manifesto dataset respectively, while **Table 3** represents the results of the experiment.

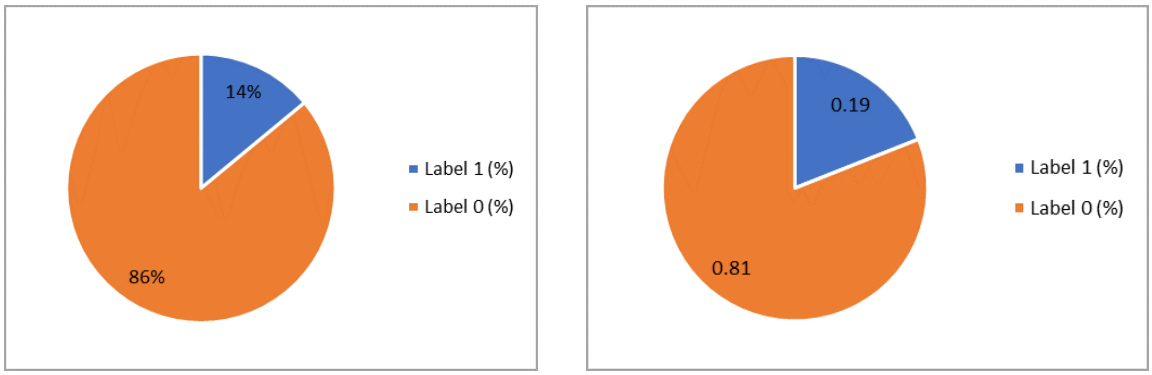
Table 1: Random sample from Choi dataset with labeling [24]**.**

Index	Sentence	Label
12159	Another marked difference is noted here.	1
17839	I could talk to you for three hours and still not be able to give you all of our plans ! !	1
16682	Guilford-Martin personality inventories.	0
41733	Modifications of the last technique have been applied by several groups of investigators.	0
32985	Two major types of fully distributed cost analysis 1.	0
14380	"He has given only the one pass in his 27 innings, an unusual characteristic for a southpaw. "	0
16429	"The religions of the people include Christianity, Mohammedanism, paganism, ancestor worship and animism. "	0
24591	At each step of the calculation the operating variables of only one stage need be varied.	1

[1] https://manifestoproject.wzb.eu/

Table 2: Random sample from Manifesto dataset with labeling [25].

Index	Sentence	Label
204	will reward those who create private-sector jobs here in America,	0
383	and that the country urgently needs comprehensive immigration reform that brings undocumented immigrants out of the shadows	0
3116	Global competitiveness will increasingly require an entrepreneurial culture of cooperation and teamwork.	1
1686	We believe in an America where every child comes to school ready to learn.	0
820	In part, this is in recognition that the United States has been, and always will be, a Pacific power.	0
3090	All those gains are jeopardized if Democrats gain unfettered power once again.	1
5564	Every teacher and every student deserves a safe classroom in which to work and learn.	0
1663	We will secure more funding for aggressive biomedical research seeking affordable and effective therapies based on real science.	0



a) Choi's dataset **b)** Manifesto's dataset

Fig. 2: Labels distribution in both datasets.

Table 3: The Results of Experiment 1.

Dataset	precision	recall	F1-score
Choi	0.688	0.215	0.328
Manifesto	0.500	0.225	0.311

This approach suffers from a few defects. The number of sentences labeled "0" was very big compared to those labeled "1" as shown in **Fig. 2**. In addition, the results (as shown in **Table 3**) are not good enough, e.g., the harmonic mean of precision and recall (F1-score) is only 0.328 and 0.311 for Choi and Manifesto datasets respectively. Therefore, Experiment 2 tried to tackle such defects, but unfortunately faced another set of defects as discussed in the following subsection.

4.3 Experiment 2: Labeling Data by a percentage between 0 and 1.

In this approach, the data was augmented, in which each segment was split into sentences, each of which was labeled with a calculated ratio of the current sentence index to the total number of

sentences (rounded to one decimal place). To explain, let us consider a segment of 5 sentences, the preprocessed labeling will take place in 5 iterations as follows (as depicted in **Fig. 3**):

Iteration 1: the first sentence is labeled by 0.2 (the index of the first sentence is 1 and the total number of sentences is 5, then the label will be 1/5 = 0.2),

Iteration 2: the first sentence and the second sentence were concatenated into one string that was separated by a space and labeled by 0.4.

Iteration N: in the last iteration, all the sentences were combined together and labeled by 1.

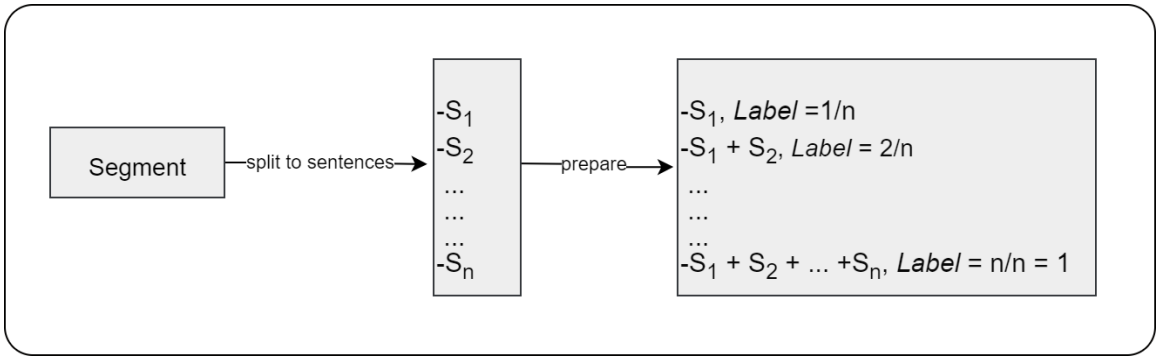


Fig. 3: Data Labeling in Experiment 2.

Table 4 shows the result of this experiment. Notice that the recall is improved, but all other metrics declined. This is due to the long input string after concatenation, which indicated the need for making a tradeoff between the lengths of the input strings and keeping attention to the sequence of sentences during training.

Table 4: The results of Experiment 2.

Dataset	precision	recall	F1-score
Choi	0.214	0.326	0.237
Manifesto	0.201	0.230	0.200

4.4 Experiment 3: Cross segments labeling

In this experiment, each segment was divided into pieces, each of which was then given a label. Each piece was composed of a fixed number of sentences (in practice, the number of sentences per piece was set to 2). Each piece was labeled according to its position, e.g., If the two pieces were in the same segment (in the middle and not on a boundary), it gets a label of "0", otherwise, when a piece crosses segments, it gets a label of "1" (e.g., when the first piece of sentences is at the end of the first segment and the second piece is at the start of the second segment). See **Fig. 4** for details. The main motivation for this model is its simplicity in defining the target separators. The final form of the input pieces is *[CLS_TOKEN] + first piece+ [SEP_TOKEN] + second piece + [PAD_TOKENs]*.

Table 5 shows the results of this experiment, which indicate improvements in all parameters, with very promising results. The Choi dataset exhibits remarkably minor errors, largely due to its high

standard deviation. It's essential to emphasize that the Choi dataset is relatively small and synthetic in nature, consequently having limitations. Given that each document is a combination of excerpts from randomly selected news articles, the task is artificially simplified, which is evident from the previously established low error rate. In the next section, cross segmentation labeling is applied to video transcript datasets (a complete unstructured text).

Table 5: The results of Experiment 3.

Dataset	precision	recall	F1-score
Choi	0.990	0.992	0.991
Manifesto	0.650	0.586	0.616

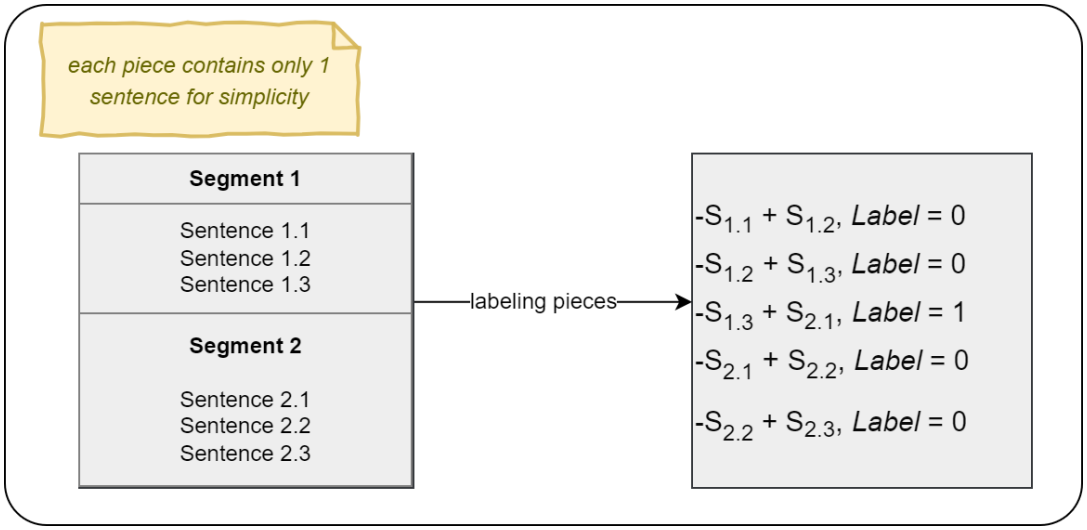


Fig. 4: Cross segmentation labeling.

5. VIDEO SEGMENTATION

Video segmentation is very challenging and important, especially in adaptive e-learning domains. A video has several integrated components, namely, transcript/text, audio, and image frames. In e-Learning, the transcript reveals much about the learning message to transfer. Transcripts can be obtained by applying Automatic Speech Recognition (ASR) and Speech-to-Text (STT) algorithms. This section discusses this part of our research on segmenting videos via segmenting its text transcript.

The results obtained on segmenting semi-structured texts, as explained by the experiments of Section 4, were to be tested on the unstructured video transcript texts. Two video datasets were used in this experimental part of the research. The first dataset was VT-SSum [26], while the second dataset was specially prepared by this research from **Y**ou**T**ube's **E**ducational **V**ideos (YT-EV)[2].

[2] This new dataset is available at https://github.com/abdelrahmansamir/YT-EV-dataset.

VT-SSum is a benchmark dataset [26] with spoken language for video transcript segmentation and summarization. It includes 125K transcript-summary pairs extracted from 9,616 videos. VT-SSum takes advantage of the videos from VideoLectures.NET by leveraging the slides content as the weak supervision to generate the extractive summary and segmentation for video transcripts. The segmentation assumptions of this dataset do not match those for Learning Objects (LO).

To train the experimental models, AdamW optimizer is used [27] with a 15% dropout rate as well as a linear warmup procedure. Learning rates are set between 1e-5 and 3e-5. Since this is a binary classification problem and the model outputs are probabilities, Binary Crossentropy loss function was used. Binary Crossentropy is mathematically formulated to calculate the average cross-entropy loss between the predicted probabilities and the true binary labels, providing a measure of how well the model's predictions align with the actual data.

5.1 Experiment 4: Semi-Structured Dataset Segmentation (VT-SSum-based dataset):

This experiment used VT-SSum data set as a set of semi-unstructured text. The results of this experiment are listed in **Table 6**, which reveals precision, recall, and F1-score as 0.524, 0.295, and 0.377, respectively. These results indicate that the VT-SSum dataset segments are not accurate in most cases, which aligns with the segmentation assumption used. This raised the need to work with a different dataset that is more related to the data of a general lecture video, as discussed in Experiment 5.

Table 6: Test set results of VT-SSum segmentation experiment.

Dataset	precision	recall	F1-score
VT-SSum	0.524	0.295	0.377

5.2 Experiment 5: Fully unstructured text segmentation (YT-EV-based dataset):

The VT-SSum dataset assumes that the videos are based on presentation slides and that each slide is a segment on its own, which does not necessarily guarantee an LO segment. Therefore, this research prepared its own dataset based on YT-EV videos, a dataset that is closer to the real-world common lecture videos.

YT-EV is a dataset collected from YouTube. The made-up dataset was based on educational videos in different fields, and it included more than 700 videos with 12.5K segments in total. This data set was automatically collected as follows:

1. Prepared is a list of search keywords (Educational course names) and another list of some YouTube playlists.
2. Using the YouTube Data API[3], the two prepared lists were searched in YouTube.com.
3. From the search results, only videos having timestamps were selected. Timestamps are set by the video authors to link a segment description to a specific moment in the video; hence, they are well-defined segments that match the definition of an LO.
4. By using the YouTube-transcript-api [28] , transcripts were generated.

[3] https://developers.google.com/youtube/v3/docs

5. Finally, the timestamps (as determined in Step 3) were used to mark the segments on the extracted transcripts.

Table 7 reports the results of the experiment on YT-EV dataset. In general, the results clearly indicate that YT-EV dataset surpasses VT-SSum dataset in terms of precision, recall, and F1-score, this is due to the nature of the datasets themselves. The VT-SSum dataset segments were constructed based on strict assumptions as explained above, while the YT-EV dataset is more properly segmented since the segmentation was made by the authors themselves.

Noteworthy, due to computational limitations on the available resources, this experiment created a relatively limited-size YT-EV dataset and used a compact version of the BERT model, but surprisingly the revealed promising outcomes.

Table 7: Test set results of YT-EV segmentation experiment.

Dataset	precision	recall	F1-score
YT-EV	0.722	0.603	0.657

CONCLUSION

This article presented three models designed for tasks involving the segmentation of text and video content. Among these models, the "cross-segment labeling" approach, which leverages contextual information near expected separators, yielded the most favorable outcomes. Our evaluation encompassed the use of four distinct datasets, with metrics including Precision, Recall, and F1-score systematically computed for each. The obtained results instill confidence in the viability of relying on transcripts as a primary component for video segmentation, particularly in the context of Learning Object (LO) identification. We presume that the remaining components, namely voice and image frames, could further enhance performance. Additionally, even more promising results could be attained by employing larger datasets and expanding computational resources in future endeavors.

REFERENCES

[1] L. W. Anderson and D. R. Krathwohl, A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Addison Wesley Longman, Inc., 2001.

[2] S. Gamalel-Din, "Smart e-Learning: A greater perspective; from the fourth to the fifth generation e-learning," Egyptian Informatics Journal, vol. 11, no. 1, pp. 39--48, 2010.

[3] S. Gamalel-Din, "An intelligent etutor-student adaptive interaction framework," Proceedings of the 13th International Conference on Interacción Persona-Ordenador, pp. 1--8, 2012.

[4] F. Llopis, A. Ferrández and J. L. Vicedo, "Text segmentation for efficient information retrieval," International Conference on Intelligent Text Processing and Computational Linguistics, pp. 373--380, 2002.

[5] G. Shtekh, P. Kazakova, N. Nikitinsky and N. Skachkov, "Applying topic segmentation to document-level information retrieval," Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia, pp. 1--6, 2018.

[6] A. Halawa, S. Gamalel-Din and A. Nasr, "Exploiting bert for malformed segmentation detection to improve scientific writings," Applied Computer Science, vol. 19, no. 2, pp. 126-141, 2023.

[7] I. Turc, M.-W. Chang, K. Lee and K. Toutanova, "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models," arXiv preprint arXiv:1908.08962v2, 2019.

[8] D. Ponceleon and S. Srinivasan, "Automatic discovery of salient segments in imperfect speech transcripts," Proceedings of the tenth international conference on Information and knowledge management, pp. 490--497, 2001.

[9] M. Lin, J. F. Nunamaker, M. Chau and H. Chen, "Segmentation of lecture videos based on text: a method combining multiple linguistic features," 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the, pp. 9--pp, 2004.

[10] M. Lin, M. Chau, J. Cao and J. F. Nunamaker Jr, "Automated video segmentation for lecture videos: A linguistics-based approach," International Journal of Technology and Human Interaction (IJTHI), vol. 1, no. 2, pp. 27--45, 2005.

[11] R. R. Shah, Y. Yu, A. D. Shaikh and R. Zimmermann, "TRACE: linguistic-based approach for automatic lecture video segmentation leveraging Wikipedia texts," in 2015 IEEE International Symposium on Multimedia (ISM), IEEE, 2015, pp. 217--220.

[12] M. Lukasik, B. Dadachev, G. Simoes and K. Papineni, "Text segmentation by cross segment attention," arXiv preprint arXiv:2004.14535, 2020.

[13] A. Solbiati, K. Heffernan, G. Damaskinos, S. Poddar, S. Modi and J. Cali, "Unsupervised topic segmentation of meetings with BERT embeddings," arXiv preprint arXiv:2106.12978, 2021.

[14] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[15] K. Lo, Y. Jin, W. Tan, M. Liu, L. Du and W. Buntine, "Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence," arXiv preprint arXiv:2110.07160, 2021.

[16] A. Maraj, M. V. Martin and M. Makrehchi, "A More Effective Sentence-Wise Text Segmentation Approach Using BERT," Document Analysis and Recognition--ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5--10, 2021, Proceedings, Part IV 16, pp. 236--250, 2021.

[17] A. De Lucia, R. Francese, I. Passero and G. Tortora, "Migrating legacy video lectures to multimedia learning objects," Software: Practice and Experience, vol. 38, no. 14, pp. 1499--1530, 2008.

[18] D. Ma and G. Agam, "Lecture video segmentation and indexing," Document Recognition and Retrieval XIX, vol. 8297, pp. 238--245, 2012.

[19] A. S. Imran and F. A. Cheikh, "Multimedia learning objects framework for e-learning," 2012 International Conference on E-Learning and E-Technologies in Education (ICEEE), pp. 105--109, 2012.

[20] C. A. Bhatt, A. Popescu-Belis, M. Habibi, S. Ingram, S. Masneri, F. McInnes, N. Pappas and O. Schreer, "Multi-factor segmentation for topic visualization and recommendation: the must-vis system," Proceedings of the 21st ACM international conference on Multimedia, pp. 365--368, 2013.

[21] D. Ma, X. Zhang, X. Ouyang and G. Agam, "Lecture vdeo indexing using boosted margin maximizing neural networks," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 221--227, 2017.

[22] D. Chand, "Lecture video segmentation using speech content (Master's thesis)," 2020.

[23] S. Gamalel-Din, "Smart E-Learning School of the Future: Project Report," the Proceedings of the 6th E-learning Applications Conference, Cairo, pp. 10--12, 2009.

[24] F. Y. Choi, "Advances in domain independent linear text segmentation," arXiv preprint cs/0003083, 2000.

[25] G. Glavaš, F. Nanni and S. P. Ponzetto, "Unsupervised text segmentation using semantic relatedness graphs," Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, pp. 125--130, 2016.

[26] T. Lv, L. Cui, M. Vasilijevic and F. Wei, "Vt-ssum: A benchmark dataset for video transcript segmentation and summarization," arXiv preprint arXiv:2106.05606, 2021.

[27] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," arXiv:1711.05101v2, 2018.

[28] J. Depoix, "youtube-transcript-api," jun 16 2023. [Online]. Available: https://pypi.org/project/youtube-transcript-api/.

[29] O. Bohl, J. Scheuhase, R. Sengler and U. Winand, "The sharable content object reference model (SCORM)-a critical review," International Conference on Computers in Education, 2002. Proceedings., pp. 950--951, 2002.

[30] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez and J. Garcia-Rodriguez, A review on deep learning techniques applied to semantic segmentation, arXiv preprint arXiv:1704.06857, 2017.