Textbook-Informed Lecture Video Segmentation for Fine-Grained OER Reuse

Zhaoyi Shi[1], Lubna Ali[2] and Ulrik Schroeder[2]

[1]RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
[2]Learning Technologies Research Group, RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany

Abstract

Lecture segmentation is an important technique for enabling fine-grained reuse of Open Educational Resource (OER) videos. However, existing lecture segmentation methods primarily model segmentation as flat boundary prediction and rely on human-annotated chapter boundaries that are often subjective in granularity. As a result, current approaches struggle to capture the hierarchical pedagogical structure naturally present in educational lectures. In this paper, we investigate the limitations of existing lecture segmentation paradigms through literature analysis and case studies. Based on these observations, we propose a novel textbook-informed training task that reformulates lecture segmentation as semantic relation prediction between lecture transcript chunks and textbook topic units. By introducing relations such as equivalence, containment, and overlap, the proposed task aims to help embedding-based segmentation models learn hierarchical topic structure beyond flat topic transitions, providing a conceptual foundation for future hierarchy-aware lecture segmentation systems.

Keywords

Lecture Segmentation, OER Reuse, Hierarchical Modeling, iTextbooks

1. Introduction

Open Educational Resources (OER) have become increasingly important in contemporary education because they improve accessibility, flexibility, and equity in learning [1]. Among different OER formats, educational video has become particularly prominent following the expansion of online and asynchronous learning [2]. However, despite the 5R (Reuse, Retain, Revise, Remix, and Redistribute) potential of OER, the reuse and adaptation of OER videos remain difficult in practice [3]. Building upon the existing OER conversion tool (convOERter) [4], this paper focuses on enabling fine-grained reuse of educational videos through lecture segmentation.

Although recent lecture segmentation methods have achieved promising performance, they largely overlook the hierarchical nature of educational lectures [5] and do not explicitly model hierarchical educational semantics. Moreover, many existing approaches rely on creator-provided chapter annotations, which are often subjective in granularity, further reinforcing flat segmentation biases. Recent LLM-based segmentation methods have demonstrated surprisingly strong performance [6, 7, 8], further suggesting that lecture segmentation is fundamentally a reasoning- and structure-intensive task.

One possible way to introduce hierarchical semantic understanding into lecture segmentation is through textbook enhancement. Modern textbooks commonly follow a hierarchical organizational paradigm consisting of chapters, sections, subsections, and subsubsections, where concepts are progressively decomposed into increasingly specific topics [9]. In addition, textbooks in domains such as mathematics often impose even more explicit semantic structures through elements such as definitions, lemmas, claims, proofs, exercises, and solutions. These structures reflect pedagogically meaningful relationships between concepts and levels of abstraction. Moreover, educational resources such as MIT OpenCourseWare frequently provide direct alignment between lectures and textbook chapters, making textbook-enhanced segmentation feasible in practice.

iTextbooks'26: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea
zhaoyi.shi@rwth-aachen.de (Z. Shi); ali@informatik.rwth-aachen.de (L. Ali); schroeder@informatik.rwth-aachen.de (U. Schroeder)
0009-0009-2296-0727 (Z. Shi); 0000-0003-2780-4319 (L. Ali); 0000-0002-5178-8497 (U. Schroeder)

This paper presents an early-stage conceptual investigation into hierarchical educational semantics for lecture transcript segmentation. We propose our research question: "How can textbook-informed supervision enable segmentation models to learn hierarchical educational semantics despite the limitations of flat and subjective segmentation annotations?" In this paper, we investigate the limitations of current segmentation paradigms and explore how structured educational resources such as textbooks may provide hierarchical semantic supervision beyond flat boundary annotations. Based on this observation, we further propose a novel training task designed to incorporate hierarchical pedagogical information into embedding-based lecture segmentation models.

2. Related Work

Existing lecture segmentation research largely inherits the flat boundary-oriented paradigm established by early methods. Early approaches mainly relied on single modalities [10, 11, 12], such as visual scene changes, slide transitions, pauses in speech, or transcript topic shifts, to detect segment boundaries. Transcript-based approaches [13, 14] further adapted classical text segmentation techniques such as TextTiling [15]. In these methods, non-transcript modalities primarily provided local temporal cues. Regarding transcript, although certain textual features, such as cue phrases or noun phrase transitions, may implicitly reflect hierarchical educational structures, the dominant modeling paradigm remained focused on detecting local discontinuities rather than understanding the global pedagogical organization of lectures.

Later work shifted toward deep learning and embedding-based approaches, yet hierarchical modeling remains limited due to existing segmentation annotations and evaluation datasets. State-of-the-art lecture segmentation methods mainly fall into supervised and unsupervised categories [6]. Supervised approaches such as AHMN [16, 17] rely on chapter annotations manually created by video authors. However, these annotations are often subjective, flat, and inconsistent in granularity, reflecting different interpretations of pedagogical structure. Unsupervised methods [18] similarly cluster transcript embeddings into flat segment groups without explicitly modeling semantic hierarchy. Moreover, current evaluation protocols are largely based on the same creator-provided chapter boundaries, causing existing methods to inherit and reinforce these flat segmentation assumptions. As a result, current approaches still struggle to capture the multi-level pedagogical organization naturally present in educational lectures.

Recent work [6, 7, 8] further highlights the potential of LLM-based approaches, which have achieved promising performance over SOA embedding-based methods, through their long-context understanding and stronger global semantic reasoning capabilities. Their strong performance suggests that effective lecture segmentation may depend not only on boundary detection, but also on modeling latent educational logic and hierarchical conceptual organization within lecture content.

3. Limitations of Flat Lecture Segmentation

In this section, we conduct exploratory case studies and adapt a previous state-of-the-art experiment to examine the hierarchical structure of lecture content.

3.1. Subjective Granularity in Human Annotation

To illustrate the subjective nature of lecture segmentation, we compare two representative lecture examples. As shown in Figure 1, we reference two lectures[1][2] from MIT OpenCourseWare[3] that exhibit notably different segmentation styles. Lecture A is divided into 23 segments with an average segment length of approximately 3 minutes, whereas Lecture B contains only 9 segments with an average length

[1] https://www.youtube.com/watch?v=t4K6lney7Zw
[2] https://www.youtube.com/watch?v=4sTKcvYMNxk
[3] https://ocw.mit.edu/

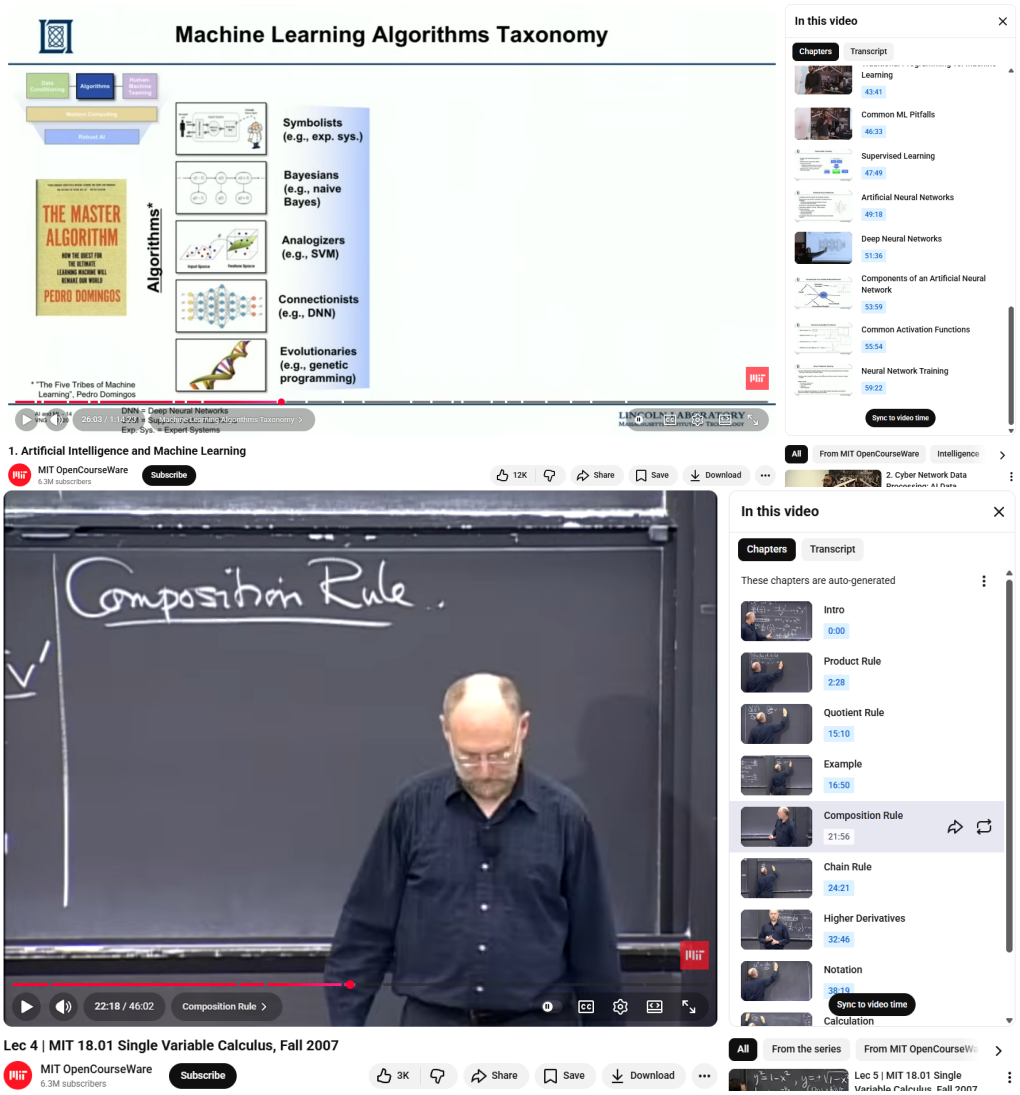

Figure 1: Screenshots for two MIT OpenCourseWare lectures, both under Creative Commons BY-NC-SA license, segmentation annotated by creator. Above: Lecture A, Mathematics for Big Data and Machine Learning (Lecture 1) [19]. Below: Lecture B, Single Variable Calculus (Lecture 1) [20].

of around 5 minutes. Furthermore, the chapter descriptions reveal different annotation granularities and naming conventions. Lecture A primarily emphasizes topic-oriented questions, while Lecture B focuses more on conceptual explanations and illustrative examples. Moreover, as shown in lecture A of Figure 1, segmentation granularity can also vary substantially within a single lecture. The largest segment of video simultaneously covers multiple subtopics, like distributed training and inference mechanisms, both of which are comparable in granularity to independently segmented topics in earlier examples.

3.2. Evidence of Hierarchical Boundaries in SOA Predictions

This subsection revisits a case study from prior work [6] to examine how the absence of hierarchical modeling and the data's subjectivity of granularity can negatively affect lecture segmentation performance. The experiment was conducted on a randomly selected MIT OCW lecture from the MITFLD dataset using the strongest model reported in the paper, namely a BiLSTM built on jointly trained visual-text embeddings (JVTE). In their analysis, prediction results where the ground truth (GT) is negative but the model predicts a boundary are divided into two categories: actual false positives and reasonable finer-grained sub-topic boundaries that are not annotated in the GT.

Table 1
Case Study of Video_107 from MITFLD with BiLSTM (JVTE) [6]

Time	GT	Prediction	Remarks
00:02	No	Yes	Sub-topic (Opening Animation)
02:41	Yes	Yes	True positive
06:53	No	Yes	Sub-topic
08:25	No	Yes	Sub-topic
10:07	Yes	Yes	True positive
11:39	No	Yes	Sub-topic
14:30	No	Yes	False positive
17:41	Yes	Yes	True positive
17:59	No	Yes	False positive
26:59	Yes	No	False negative
30:28	No	Yes	Sub-topic
32:17	Yes	Yes	True positive
32:22	No	Yes	Sub-topic
32:24	No	Yes	False positive
33:42	Yes	No	False negative
35:48	No	Yes	Sub-topic
36:36	No	Yes	Sub-topic (Copyright Page)

As shown in Table 1, the model predicts 15 boundary points, while the ground truth only contains 6 annotated boundaries. Among the 6 GT boundaries, 4 are correctly identified by the model, indicating that the model is capable of capturing major topic transitions. More importantly, 8 out of the 15 predicted boundaries are regarded by the original authors as reasonable finer-grained sub-topic transitions rather than clear errors. This suggests that the model tends to segment lectures at a finer granularity than the annotations. One possible explanation, also discussed in the original paper, is the distribution bias of the dataset, where the average number of segments per lecture is around 15, whereas this example only contains 6 annotated boundaries.

However, the predicted boundaries also reveal model's inconsistencies of granularity. For example, timestamp 06:53 marks the beginning of a new article discussion and timestamp 08:25 corresponds to the third bullet point under that article, both of which are treated as sub-topic boundaries. Yet the second bullet point is not segmented, resulting in inconsistent granularity within the same discourse structure. In addition, 3 out of the 15 predictions are genuine false positives, while 2 ground truth boundaries are missed entirely.

4. Proposed Training Task

We propose a pretraining paradigm that leverages aligned textbook–lecture pairs to improve hierarchical semantic understanding prior to downstream lecture segmentation. Unlike prior lecture segmentation approaches that model segmentation as flat boundary prediction, we formulate hierarchical pedagogical relation prediction between lecture segments and textbook-derived topic units, enabling models to learn asymmetric semantic granularity and topic containment relations. The core intuition behind the proposed task is that a model capable of understanding the hierarchical organization of educational semantics should also be able to infer structured semantic relations between a lecture segment and another topic unit, including equivalence, hierarchical containment, and partial semantic overlap.

To formalize this idea, we introduce a relation set inspired by set relations and operations. Let A and B denote two topic-related text segments. The semantic relation between them is defined as shown in Table 2.

The training dataset is constructed using existing lecture chapter annotations and textbook-video alignment information. Although creator-provided chapters may exhibit subjective granularity, they still represent meaningful pedagogical topic units and can therefore serve as anchors for relation

Table 2
Semantic relations between topic segments

Relation	Mathematical Expression
Equivalent	$A = B$
Parent	$A \supset B$
Child	$A \subset B$
Overlap	$A \cap B \neq \emptyset,\ A \neq B$
Unrelated	$A \cap B = \emptyset$

construction. For each lecture chapter, the fully aligned textbook section forms an *equivalent* pair, while the hierarchical structure of the textbook naturally provides *parent* and *child* relations through ancestor and descendant topic units. *Overlap* relations can be derived from partially aligned textbook sections, whereas strong *unrelated* relations can be generated through sampling from different topics within the same course. During training, the model is required to predict the semantic relation between paired inputs.

Such supervision encourages the model to reason about the semantic role and granularity of pedagogical units rather than relying solely on flat topic transitions. To correctly infer the relation between two segments, the model must understand both the underlying topic semantics and the hierarchical organization of concepts, exercises, and explanations. In particular, semantically similar pairs labeled as *parent*, *child*, or *overlap* provide informative supervision signals analogous to hard negative examples in contrastive learning. Combined with further training on human-annotated segmentation boundaries, this objective may help the model gradually reconstruct the hierarchical pedagogical structure underlying lecture content.

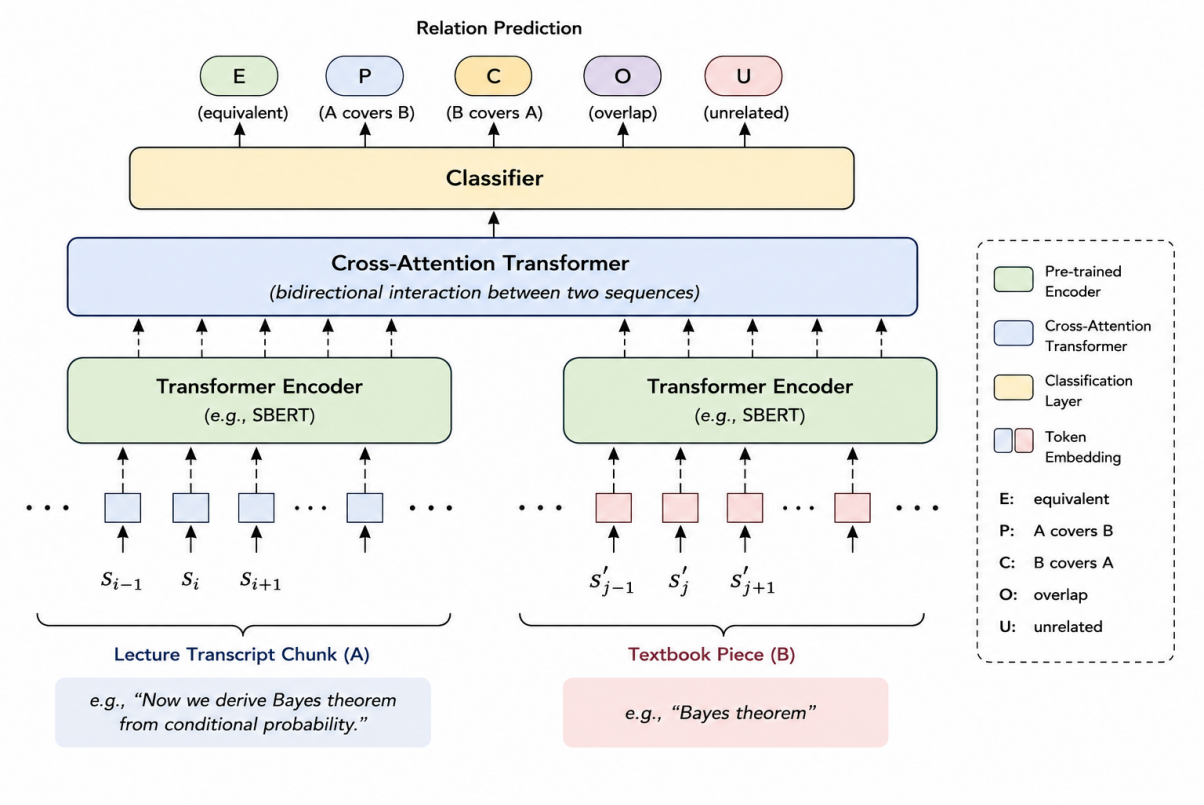

Figure 2: Cross-attention architecture for hierarchical semantic relation prediction between lecture transcript chunks and textbook pieces.

We further present an example implementation of the proposed hierarchical relation prediction paradigm within a transcript-based lecture segmentation model. As shown in Figure 2, the model adopts an S-BERT encoder as the underlying sentence embedding backbone, followed by a global Transformer layer to produce context-aware embeddings over the entire lecture segment. On top of these representations, a shallow cross-encoder Transformer is introduced to enable deep semantic interaction between the lecture segment and the textbook topic segment. Through cross-attention, the model can jointly reason over the semantic structures of both inputs and infer their hierarchical relation. By training on such objectives, the underlying chunk encoder is encouraged to develop a deeper understanding of pedagogical topics and their hierarchical organization within lecture content.

5. Future Work and Conclusion

5.1. Future Work

Future work should address both the evaluation of hierarchical semantic understanding and its contribution to hierarchy-aware lecture segmentation. Regarding the evaluation of hierarchical semantic understanding, beyond standard classification metrics on the proposed semantic relation labels, future studies may investigate retrieval-based evaluation settings, such as retrieving relevant lecture transcript chunks from textbook topics and vice versa. Such evaluations could provide additional evidence that the learned representations capture meaningful pedagogical relationships between educational resources.

A key challenge concerns the evaluation of downstream segmentation performance. Existing lecture segmentation benchmarks primarily rely on creator-provided chapter boundaries, which our analysis suggests may reflect subjective and inconsistent granularity choices. Future evaluation protocols should therefore combine traditional segmentation metrics with human-centered measures that directly assess content reuse, retrieval effectiveness, and educational utility.

More broadly, this work aims to provide an initial step toward more objective representations of educational content structure. The proposed framework is motivated by the observation that effective OER reuse requires reusable and objective segment metadata that can support retrieval and adaptation across educators and learning contexts. By incorporating textbook-informed pedagogical supervision, we hope to facilitate finer-grained organization of lecture content beyond subjective chapter annotations. In our future work, we plan to implement and evaluate the proposed framework on lecture segmentation tasks and explore its integration into the convOERter ecosystem, ultimately supporting fine-grained OER video reuse and granularity-aware educational content retrieval.

5.2. Conclusion

Motivated by the need for granularity-consistent segment metadata to support OER video retrieval and reuse, this paper presents a conceptual investigation into the limitations of current lecture segmentation methods, particularly their flat boundary modeling and reliance on subjective segmentation annotations that fail to capture the hierarchical nature of educational content. Through literature review, case studies, and adaptation of previous experimental results, we analyze how these limitations emerge and explore the potential role of textbooks as a source of hierarchical pedagogical supervision. Based on this observation, we further propose a novel textbook-informed training paradigm for modeling hierarchical educational semantics in lecture segmentation. By introducing semantic relation types such as semantic containment and partial overlap, the proposed task may provide a foundation for future work on hierarchy-aware educational representation learning and lecture understanding.

Declaration of Generative AI Use

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Olivier, A. Rambow (Eds.), Open Educational Resources: A Catalyst for Social Justice and Equality, Springer Nature Singapore, Singapore, 2023, pp. 15–33. URL: https://doi.org/10.1007/978-981-19-8590-4_2. doi:10.1007/978-981-19-8590-4_2.

[2] M. Noetel, S. Griffith, O. Delaney, T. Sanders, P. Parker, P. Cruz, C. Lonsdale, Video improves learning in higher education: A systematic review, Review of Educational Research (2021). URL: https://doi.org/10.3102/0034654321990713. doi:10.3102/0034654321990713.

[3] P. H. D. Valle, R. Capilla, V. dos Santos, D. Feitosa, E. Y. Nakagawa, Open educational resources: Barriers and open issues, 2026. URL: https://arxiv.org/abs/2603.10013. arXiv:2603.10013.

[4] L. K. N. Ali, convOERter: a technical assistance tool to support semi-automatic conversion of images in educational materials as OER, Ph.D. thesis, RWTH Aachen University, 2024. URL: https://publications.rwth-aachen.de/record/978511. doi:10.18154/RWTH-2024-01352, artwork Size: pages 1 Online-Ressource : Illustrationen Pages: pages 1 Online-Ressource : Illustrationen Publication Title: Dissertation Volume: RWTH Aachen University.

[5] S. A. Ambrose, M. W. Bridges, M. DiPietro, M. C. Lovett, M. K. Norman, How learning works: Seven research-based principles for smart teaching, John Wiley & Sons, 2010. URL: https://books.google.com/books?hl=zh-CN&lr=&id=6nGaDwAAQBAJ&oi=fnd&pg=PR13&ots=Ki-XR1ZvU1&sig=tliC83gApm1XyKN0yuUN7wyKmfY.

[6] J. Wang, R. Y.-K. Kwok, E. C. H. Ngai, Towards key point identification (kpi) for lecture videos: Approaches and performance evaluation, ACM Trans. Multimedia Comput. Commun. Appl. 21 (2025). URL: https://doi.org/10.1145/3746640. doi:10.1145/3746640.

[7] A. Krassovitskiy, R. Mussabayev, K. Yakunin, Llm-enhanced semantic text segmentation, Applied Sciences 15 (2025). URL: https://www.mdpi.com/2076-3417/15/19/10849. doi:10.3390/app151910849.

[8] F. Retkowski, A. Waibel, Paragraph segmentation revisited: Towards a standard task for structuring speech, 2026. URL: https://arxiv.org/abs/2512.24517. arXiv:2512.24517.

[9] C. M. Reigeluth, The elaboration theory: Guidance for scope and sequence decisions, in: Instructional-design theories and models, Routledge, 2013, pp. 425–453.

[10] H. J. Jeong, T.-E. Kim, M. H. Kim, An accurate lecture video segmentation method by using sift and adaptive threshold, in: Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia, MoMM '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 285–288. URL: https://doi.org/10.1145/2428955.2429011. doi:10.1145/2428955.2429011.

[11] X. Che, H. Yang, C. Meinel, Lecture video segmentation by automatically analyzing the synchronized slides, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 345–348. URL: https://doi.org/10.1145/2502081.2508115. doi:10.1145/2502081.2508115.

[12] R. R. Shah, Y. Yu, A. D. Shaikh, S. Tang, R. Zimmermann, Atlas: Automatic temporal segmentation and annotation of lecture videos based on modelling transition time, in: Proceedings of the 22nd ACM International Conference on Multimedia, MM '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 209–212. URL: https://doi.org/10.1145/2647868.2656407. doi:10.1145/2647868.2656407.

[13] D. Galanopoulos, V. Mezaris, Temporal lecture video fragmentation using word embeddings, in: I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, S. Vrochidis (Eds.), MultiMedia Modeling, Springer International Publishing, Cham, 2019, pp. 254–265.

[14] R. R. Shah, Y. Yu, A. D. Shaikh, R. Zimmermann, Trace: Linguistic-based approach for automatic lecture video segmentation leveraging wikipedia texts, in: 2015 IEEE International Symposium on Multimedia (ISM), 2015, pp. 217–220. doi:10.1109/ISM.2015.18.

[15] M. A. Hearst, Multi-paragraph segmentation expository text, in: 32nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Las Cruces, New Mexico, USA, 1994, pp. 9–16. URL: https://aclanthology.org/P94-1002/. doi:10.3115/981732.981734.

[16] J. Wu, Y. Sun, Y. Kong, H. Shu, L. Senhadji, AHMN: A multi-modal network for long MOOC videos chapter segmentation, Multimedia Tools and Applications 83 (2024) 88523–88541. URL: https://doi.org/10.1007/s11042-023-17654-2. doi:10.1007/s11042-023-17654-2.

[17] F. Retkowski, A. Waibel, From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 406–419. URL: https://aclanthology.org/2024.eacl-long.25/. doi:10.18653/v1/2024.eacl-long.25.

[18] D. Singh S, A. Gupta, C. V. Jawahar, M. Tapaswi, Unsupervised audio-visual lecture segmentation, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5221–5230. doi:10.1109/WACV56688.2023.00520.

[19] MIT OpenCourseWare, Artificial intelligence and machine learning, https://www.youtube.com/watch?v=t4K6lney7Zw, 2020. YouTube video, accessed 2026-05-18.

[20] MIT OpenCourseWare, Single variable calculus lecture, https://www.youtube.com/watch?v=4sTKcvYMNxk, 2006. YouTube video, accessed August 18, 2026.