# Discriminative segmentation of long speech recognition transcripts.

## Introduction

Online learning has shown significant growth in the past decade[[1]](#footnote-1). Moreover, the Covid-19 outbreak has forced schools, universities, and companies to leverage high quality online videos to create and spread knowledge resources. However, despite the abundance of online resources there exist several open challenges in video search, indexing, and segmentation that hinder users from specifically looking up segments of video lectures relevant to their query. In this project we aim to address the issue of topic segmentation in video lectures. Extraction of such semantically coherent clips from within a video improves the video search through better semantic indexing[[2]](#footnote-2). Through a discriminative model we aim to segment video transcripts into lexically cohesive topics using bidirectional representations.

## Task Description, Challenges and Previous work

Unlike text documents, topic segmentation on transcriptions cannot readily exploit sentence boundaries and other graphical features in text. Additionally, topic segmentation of transcripts is also posed with the challenge of maintaining temporal coherence in topics. University and online video lectures can be seen as a monologue discourse and face similar challenges with transcription, segmentation and retrieval[[3]](#footnote-3). Lectures have the unique advantage of being (usually) well-structured: we might expect the breaks between topics and units, for example, to be fairly clear. Initial efforts to segment topics in monologue transcripts can be traced back to as early as 1997[[4]](#footnote-4)-98[[5]](#footnote-5). Many of these initial systems were based on the idea of lexical cohesion[[6]](#footnote-6): the tendency of well-formed segments to induce a compact and consistent lexical distribution. Consequently, some these approaches such as the TopicTiling[[7]](#footnote-7) algorithm successfully explored marking boundaries in areas of low lexical cohesion, while others clustering based approaches such as the prominent C99[[8]](#footnote-8) uses neighbourhood of similarity measures to identify topics. Additionally, a brand of discriminatively trained models based on Hidden Markov Models (HMM)[[9]](#footnote-9), Conditional Random Fields (CRF)[[10]](#footnote-10), neural networks[[11]](#footnote-11) and Support Vector Machines[[12]](#footnote-12) have been used to tackle the problem.

Most of these approaches relied on hand-crafted cohesion features and metrics to describe boundaries and predefined (pseudo) topics. Long documents are often modelled using sliding windows and text-tiling algorithms with the aim to capture long-term dependencies in monolog transcripts. However, It’s often difficult to build labelled datasets with predefined topic boundaries and hand crafted features. This project investigates a classification based discriminative approach using embeddings from transformer models and bidirectional representations to identify topic transitions in lecture transcripts. At the core of our methodology we rely on simulating topic transitions in transcripts by combining smaller lectures transcripts and identifying topic transitions as a classification task. This is by no means a novel approach, for instance, Reynar[[13]](#footnote-13) concatenates Wall Street Journal articles with the intention to simulate monologue discourse. However, the approach relies on parts of speech tags and close class words to detect boundaries. In contrast, our approach solely relies on the text input from the transcript to identity topic transitions.

## Related work

The closets work to the best of our knowledge is presented by TCS Innovation Labs[[14]](#footnote-14) where they use bi-directional representations from RNNs to perform topic segmentation of French newswire transcripts named EuroNews. However, due to the use of LSTMs, their methodologies still fails to model long sequences as they are often struck by the vanishing gradient problem[[15]](#footnote-15). We plan to overcome these challenges in our work through the use embeddings from the Transformer XL[[16]](#footnote-16) architecture that is specifically developed to tackle long sequences.

While all approaches have been show to benefit from inclusion of non-lexical features such as audio, prosody and video segmentation features this work focuses solely on the text transcripts. This is largely motivated by recent advancements in modelling long text sequences[[17]](#footnote-17) using attention mechanism[[18]](#footnote-18) and the availability of transcripts in most online learning platforms.

## Dataset

To perform the task of topic segmentation we collect approximately 10,000 video transcripts for lectures from the Khan Academy API[[19]](#footnote-19). We leverage the Khan Academy Topic Tree[[20]](#footnote-20) to find lectures transcripts organized under various topics, subjects and units. The topics are split on a hierarchy with varying degrees of granularity.

An example Record from the raw dataset:

|  |
| --- |
| **{**  **"title":** "Number grid"**,**  **"translated\_title":** "Number grid"**,**  **"translated\_description\_html":** "Sal goes through all the numbers from 0 to 100 and shows some interesting patterns."**,**  **"id":** "xb19b2406"**,**  **"description\_html":** "Sal goes through all the numbers from 0 to 100 and shows some interesting patterns."**,**  **"kind":** "Video"**,**  **"translated\_description":** "Sal goes through all the numbers from 0 to 100 and shows some interesting patterns."**,**  **"youtube\_id":** "9XZypM2Z3Ro"**,**  **"keywords":** ""**,**  **"transcript": [**  **{**  **"text":** "Voiceover: The goal of this video \nis to essentially write down"**,**  **"endTime":** 3384**,**  **"startTime":** 660**,**  **"kaIsValid":** true  **},**  **{**  **"text":** "all the numbers in order\nfrom zero to a 100."**,**  **"endTime":** 7726**,**  **"startTime":** 3384**,**  **"kaIsValid":** true  **},**  **"text":** "That was pretty neat."**,**  **"endTime":** 297470**,**  **"startTime":** 295123**,**  **"kaIsValid":** true  **} …]}** |

We additionally pre-process the above data deduplicate and store the transcripts as a single string separated by new-line character. The resulting record for a single record looks as follows:

|  |
| --- |
| **{**  **"title":** "Number grid"**,**  **"id":** "xb19b2406"**,**  **"description\_html":** "Sal goes through all the numbers from 0 to 100 and shows some interesting patterns."**,**  **"transcript\_text":** "Voiceover: The goal of this video is to essentially write down all the numbers in order from zero to a 100.\nBut I’m going to do it in an interesting way, a way that maybe will allow us to see some patterns in the numbers themselves.\nSo let me just start, so I’m gonna start at zero, one, two, three, four, five, six, seven … now we got to 10 followed by a zero.\nThat was pretty neat."**,**  **}** |

Finally, to create the training, validation and test dataset we randomly concatenate two or more transcripts from different topics and mark the topic boundaries. We additionally explore sampling strategies that take into account course and fine-grained differences in the topics in the topic tree.

## Exploratory Data Analysis

Through the exploratory data analysis phase of the dataset we wish to identify the following:

1. Number of total topics in the dataset. – This aids in developing better sampling strategies when creating the model training dataset.
2. The distribution of transcripts across the various topic granularities.
3. Identify the right length of training samples that approximately works on at least the median length of the transcripts.
4. Explore and address the impact of the chosen metrics in four basic conditions of boundary detection i.e. hypothesizing no boundaries, boundaries everywhere, evenly-spaced boundaries or randomly-spaced boundaries.

## Baseline and Experimental Setup

For the experimentation we randomly split the dataset into train, validation and test splits in a ratio of 70/15/15 and perform all model development and tuning on the train and validation sets and report metrics on the test set. For training and test, the punctuation marks are removed and all words are converted to lower case, as it would be with ASR transcripts. Due to the novelty of the dataset we have no previous baselines on the dataset. Therefore, we report the baseline performance of the bi-directional models on input representations using word2vec[[21]](#footnote-21). We test the performance of the various configurations of the transformer models against this baseline model trained on word2vec representations.

## Metrics and Evaluation

For comparison of the proposed and baseline methods, we use the standard topic segmentation evaluation measures Pk[[22]](#footnote-22). Pk indicates the probability of segmentation error. It is the average probability, given two points in the dataset, that the hypothesized segment is incorrect as to whether they are separated by a boundary or not. A lower value indicating a better performance.

We also calculate the Precision (P) and Recall (R) on the segmentation results, and report the F1 score which was calculated as F1 = 2(P ×R)/(P +R). A window size(k) of 20 words on each side of a true segmentation point will be used to label a hypothesised segmentation point as true positive or false positive.

## Model and Experimental Setup

As mentioned earlier we develop our baseline method using Word2Vec and RNNs. However, to address the drawbacks related to modelling long sequences using RNNs we use embedding from the transformer XL model that have proven to be effective on sequences of length up to 8906 tokens long. We feed the output of the transformer XL model to an bi-directional LSTM layer and model the similarity of a token at time step t to itself in the forward and backward direction. Concretely, the forward representation of a token at time step tf corresponds to the hidden activation of the forward LSTM and the backward representation at time step tb corresponds to the hidden activation of the backward LSTM. The dot product of these representations capture the similarity between the forward that backward representations. We hypothesize that lower similarity values should indicate topic transitions in the sequence. This can be modelled with the binary cross entropy loss with the sigmoid activation function in the final layer to indicate the transition in topics. We intend to experiment with various model configurations including testing activations such as inverted tanh for the final activation layer and report the results of our experiments.

## Deployment Strategy

The model is largely developed in python using tensorflow. As such, we intend to deploy the model as an API accessible via POST request. Since this work involves modelling large sequences we intend to investigate the appropriateness of a large model deployment using tf-serving[[23]](#footnote-23) and FastAPI [[24]](#footnote-24)for the use.

The following is the expected schema for the request and response of the API.

|  |
| --- |
| **Request:**  **{**  **"text":** "very .... long .. transcript"  **}**  **Response:**  **{**  **"segments": [**  "seg\_1"**,**  "seg\_2",…  **]**  **}** |

1. Garrett, R., & Legon, R. (2019). CHLOE 3 Behind the Numbers: The Changing Landscape of Online Education 2019. [↑](#footnote-ref-1)
2. James Allan, Topic Detection and Tracking: EventBased Information Organization, Kluwer Academic Publishers, Norwell, MA, USA, 2002. [↑](#footnote-ref-2)
3. James Glass, Timothy J. Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay, 2019, Recent Progress in the MIT Spoken Lecture Processing Project, MIT, Computer Science and Artificial Intelligence Laboratory Cambridge, Massachusetts, USA [↑](#footnote-ref-3)
4. Passonneau RJ and Litman DJL 1996 Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence and linguistic devices In Interdisciplinary Perspectives on Discourse (ed. Hovy E and Scott D) Springer-Verlag. [↑](#footnote-ref-4)
5. Allan J, Carbonell J, Doddington G, Yamron J and Yang Y 1998 Topic detection and tracking pilot study: Final report Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop. [↑](#footnote-ref-5)
6. Matthew Purver, Spoken Language Understanding, chapter Topic Segmentation, pp. 291–317, John Wiley & Sons, Ltd, 2011. [↑](#footnote-ref-6)
7. Riedl, Martin and Biemann, Chris, ACL 2012, TopicTiling: A Text Segmentation Algorithm Based on LDA, Proceedings of ACL 2012 Student Research Workshop, pp, 37–42. [↑](#footnote-ref-7)
8. Freddy Y. Y. Choi, “Advances in domain independent linear text segmentation,” in Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, 2000, pp. 26–33. [↑](#footnote-ref-8)
9. . David M. Blei and Pedro J. Moreno, “Topic segmentation with an aspect hidden markov model,” in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 343–348. [↑](#footnote-ref-9)
10. Xiaoxuan Wang, Lei Xie, Mimi Lu, Bin Ma, Engsiong Chng, and Haizhou Li, “Broadcast news story segmentation using conditional random fields and multimoda features,” IEICE Transactions, vol. 95-D, no. 5, pp. 1206–1215, 2012 [↑](#footnote-ref-10)
11. Jia Yu, Xiong Xiao, Lei Xie, Eng Siong Chng, and Haizhou Li, “A dnn-hmm approach to story segmentation,” in INTERSPEECH, 2016, pp. 1527–1531. [↑](#footnote-ref-11)
12. Mehryar Mohri, Pedro Moreno, and Eugene Weinstein, “Discriminative topic segmentation of text and speech,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 533–540. [↑](#footnote-ref-12)
13. Reynar C Jeffrey, 1994, An Automatic Method of Finding Topic Boundaries [↑](#footnote-ref-13)
14. Imran Sheikh, Dominique Fohr, Irina Illina. Topic segmentation in ASR transcripts using bidirectional rnns for change detection. ASRU 2017 - IEEE Automatic Speech Recognition and Understanding Workshop, Dec 2017, Okinawa, Japan. [↑](#footnote-ref-14)
15. http://people.idsia.ch/~juergen/fundamentaldeeplearningproblem.html [↑](#footnote-ref-15)
16. Zihang Dai and Zhilin Yang and Yiming Yang and Jaime Carbonell and Quoc V. Le and Ruslan Salakhutdinov, Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, 2019. [↑](#footnote-ref-16)
17. Beltagy, Iz & Peters, Matthew & Cohan, Arman. (2020). Longformer: The Long-Document Transformer. [↑](#footnote-ref-17)
18. Nikita Kitaev and Łukasz Kaiser and Anselm Levskaya. (2020). Reformer: The Efficient Transformer. [↑](#footnote-ref-18)
19. http://api.khanacademy.org/api/internal/videos/{}/transcript [↑](#footnote-ref-19)
20. http://api.khanacademy.org/api/v1/topictree [↑](#footnote-ref-20)
21. T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, NIPS 2013, Distributed representations of words and phrases and their compositionality. [↑](#footnote-ref-21)
22. Doug Beeferman, Adam Berger, John Lafferty, Machine Learning 34, 177–210 (1999), Statistical Models for Text Segmentation. [↑](#footnote-ref-22)
23. https://www.tensorflow.org/tfx/guide/serving [↑](#footnote-ref-23)
24. https://fastapi.tiangolo.com/ [↑](#footnote-ref-24)