ISSN 2278-3091

Volume 9, No.1.1, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse3991.12020.pdf

https://doi.org/10.30534/ijatcse/2020/3991.12020

Extracting Time-Aligned Topic Segments and Labels from the Speech Transcripts of Video Lectures

Melbert R. Bonotan[1], Rio Anthony Apego[2], Daniel Jonell Gorne[3], Joanna Victoria Saga[4]

[1]Caraga State University, Philippines, mrbonotan@carsu.edu.ph

[2] Caraga State University, Philippines, joannavicky26@gmailcom

[3] Caraga State University, Philippines, apegorioanthony@gmail.com

[4] Caraga State University, Philippines, daniel.gorne@carsu.edu.ph

ABSTRACT

The advent of the internet opens the learners with a pool of resources online with all the information, knowledge and data they need for their study. Different platforms are existing where they can access video lectures for them to study in advance, to further their study or cope with their behind topics. However, with vast video lectures available online, students need to watch or check the whole video just for them to identify its relevance, if it fits what they need. Thus, taking much of their time during their gathering of possible and helpful video. This research develops a platform that utilized Natural Language Processing (NLP) and Artificial Intelligence that will analyze video lectures and generate a table of contents. The table of Contents generated to act as an aid to student learning by helping them navigate to the content of the video. It provides the list of topics discussed in the video and link to a specific time frame the topic was discussed. Thus, outcomes to a more convenient and efficient in reviewing the video content.

Key words: Topic modeling, Automatic speech recognition, segmentation, Natural Language Processing.
.

1. INTRODUCTION

The use of technology, particularly the internet in learning environments has increased over the past decades. According to a study, a survey conducted by the Pew International and American Life Project found that approximately 78% of youth use the internet for learning [1]. The internet is filled with learning materials such as video lectures. The video lecture is one of the most popular learning tools for campus students as well as distance learners. These video lectures enable the learners to catch up with missed classes. Another survey conducted last 2011 states that 86.3% of learners find video lectures to be useful, whilst 63.7% felt that it improved their performance [2]. However, with the vast video lectures available online and with multiple topics discussed throughout the video, students need to check or watch the whole content to identify its relevance. This will take much of their time browsing and checking multiple videos for their reference. In this paper, the researcher will develop a framework utilizing Natural Language Processing (NLP) and Artificial Intelligence in analyzing video lectures and generate a textual representation of video content. This will provide an alternative way of browsing video with the use of the table of contents generated, where each content will be linked to the specific time frame of the video where the topic will be discussed, rather than common video data such as video title.

2. METHODOLOGY

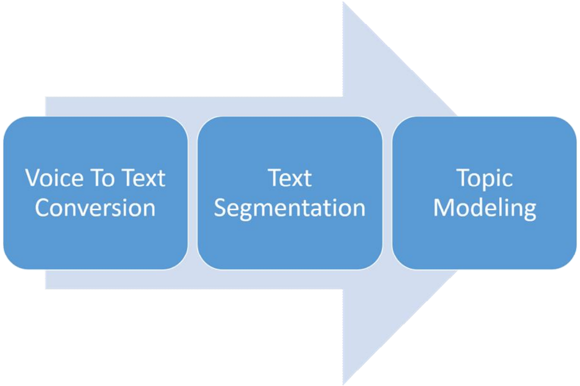


Figure 1: Methodological Framework

2.1 Voice to text conversion

Figure 1 depicts the methodological framework of the study. First, the video lecture file is converted to a lossless audio format with one audio channel. The audio is saved to Google Storage and a speech recognition request is sent to Google Speech API. This will convert the audio file to text format. Speech recognition does not automatically detect punctuations, therefore the text was sent to Punctuator2 (http://bark.phon.ioc.ee/punctuator) a Bidirectional Neural Network with Attention Mechanism for Punctuation Restoration.

2.2 Topic Segmentation and Labeling

The punctuated text is then separated into sentences using the Python Natural Language Toolkit. The topic segmentation algorithm Text Tiling groups the sentences based on word similarities between sentences [3]. After segments have been discovered, a topic model is trained on the entire English Wikipedia articles using Latent Dirichlet Allocation [4]. The goal of the training was to discover 100 topics and give the model the capability to predict which among the 100 topics the segment or word is likely to belong.

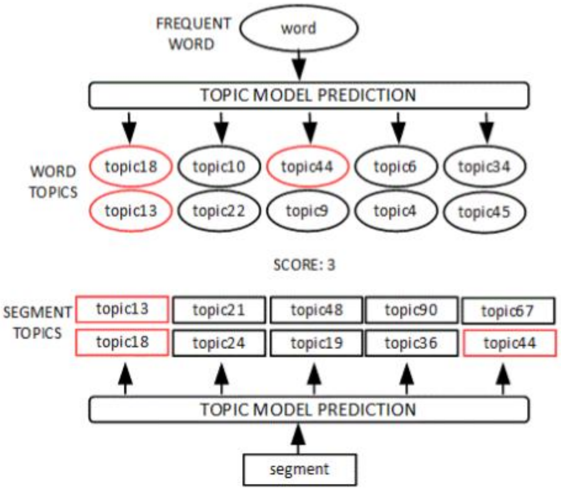


Figure 2: Scoring based on word-to-segment topic similarity.

Labeling each segment involves the use of the topic model. A label consists of the five most frequent words of the segment where each word was given a word-to segment topic similarity score predicted by the topic model. Selecting the five most frequent words on the segment requires two filters to be performed on the segment. First, the segment must not include stopping words that are frequently occurring but have no inherent meaning [5]. Second, the segment must keep nouns only to have the best possibility to represent a topic [6]. The ten most probable topics of each of the five-word are then compared to the ten most probable topics for the segment as predicted by the topic model. Words with no scores are dropped while words with at least one score are considered part of a label. Figure 2 depicts the comparison of ten topics predicted from a word against ten topics predicted from a segment having 3 similar topics, the score is then 3.

3. RESULTS AND DISCUSSION

3.1 Automatic Speech Recognition

A test is conducted on 20 video lectures from MIT Open Courseware (https://ocw.mit.edu/courses/audiovideo-courses/) five each from the fields Biology, Physics, Literature, and Economics. Figure 3 shows the comparison of the video length against the time taken for the Google Speech-to-Text conversion. Results show that it takes an average of 20 minutes for Google to convert 1 hour of speech to text or 25% of the total length of the video. Factors such as the file format of the video were excluded because, before speech recognition, the file format was converted to a uniform audio format. Another factor such as the client's internet speed was also excluded because the file was already stored in Google Storage as part of Google's cloud servers. Therefore, the time taken for the conversion process is proportional to the length of the video and independent of the client's internet speed.

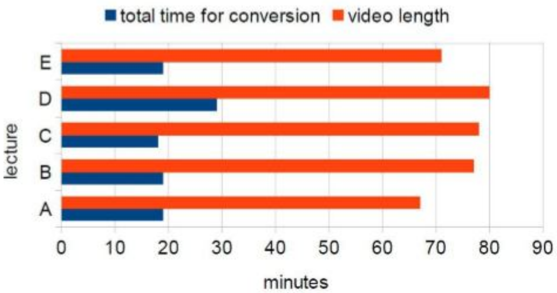


Figure 3: Comparison of video length against time taken for speech-to-text conversion on Biology video lectures

3.2 Topic Segmentation

Figure 4 shows the results of applying the TextTiling algorithm in the sentences. A pattern shows that the segment length increases as the video length increases. There was no segment overlap and each consists of an average of four sentences. There are cases that segment length only consisted of one sentence. This can be attributed to audience participation and math expressions. Video lectures from the field of Biology and Physics are rich in mathematical discussions. Non-alphabet characters were not considered in the TextTiling algorithm therefore segments were formed solely on the words, not math expressions. Varying duration serves as evidence of the unstructured style of the lecturer's discussion where changes in topics happen spontaneously [7].

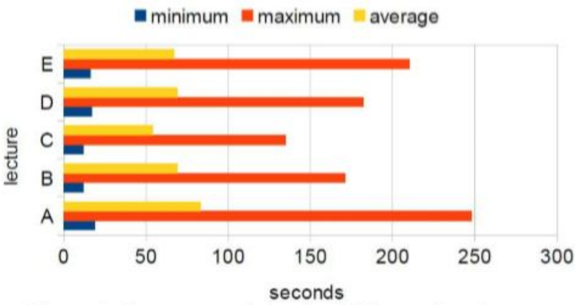
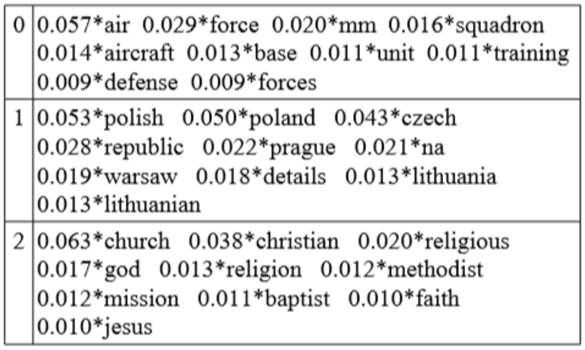


Figure 4: Segment statistics for Biology video lectures

3.3 Segment Labeling

The topic model is created by training a machine learning algorithm Latent Dirichlet Allocation to discover 100 topics using the entire English Wikipedia articles. The training of the model lasted approximately 13 hours on our machine from a 14.5 GB Wikipedia article dump. Table 1 shows the first three topics and words with a corresponding probability. If the topic model was asked to predict the topics of the word "aircraft" then it returns topic 0 with a probability of 0.014. However, a problem comes when the model was asked to predict the topics of "aircraft123" results in an error because it does not belong to its known vocabulary. The model also predicts the topics of multiple words such as a segment but its accuracy is negatively impacted because of words rarely seen in the same context. An example of this was a sentence containing the words "butterfly" and "presidential elections"

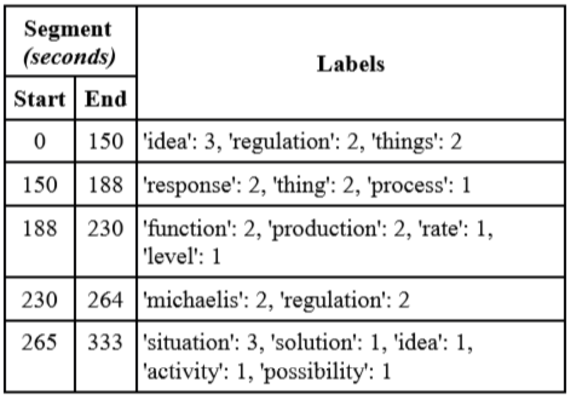
Table 1: First three topics and ten most contributing words



0	0.057*air 0.029*force 0.020*mm 0.016*squadron 0.014*aircraft 0.013*base 0.011*unit 0.011*training 0.009*defense 0.009*forces
1	0.053*polish 0.050*poland 0.043*czech 0.028*republic 0.022*prague 0.021*na 0.019*warsaw 0.018*details 0.013*lithuania 0.013*lithuanian
2	0.063*church 0.038*christian 0.020*religious 0.017*god 0.013*religion 0.012*methodist 0.012*mission 0.011*baptist 0.010*faith 0.010*jesus

Table 2 shows the results of the topic similarities of each of the five most frequent words compared to the topics of the segment itself. Out of the five most frequent words, only an average of 3 words has at least one similar topic. Some cases where a frequent word does not match any topics of the segment results in no label. This was a great issue that researchers have discovered and possibly the limit of the topic modeling algorithm. This issue results in segments with no labels and probably one of the weaknesses of the Latent Dirichlet Allocation.

Table 2: First five segments and labels of the first video lecture of Biology



Segment *(seconds)*		**Labels**
Start	**End**	
0	150	'idea': 3, 'regulation': 2, 'things': 2
150	188	'response': 2, 'thing': 2, 'process': 1
188	230	'function': 2, 'production': 2, 'rate': 1, 'level': 1
230	264	'michaelis': 2, 'regulation': 2
265	333	'situation': 3, 'solution': 1, 'idea': 1, 'activity': 1, 'possibility': 1

Figure 5 shows a web-based prototype in navigating the produced segment and label. The user can skip to the specific timestamps that interest them using the list of labels. The prototype also offers a way to search for a specific label or a specific segment. It must be noted that searching using the segment itself was costly. The reason was each video has 10 segments and 100 words each, then searching the segments of the video requires a comparison of 1000 words. The time it took for the prototype to process a video lecture varies depending on the internet speed of the user. Uploading the video lecture is dependent on the upload speed of the user and the size of the video file. On calculating time, the upload process of the video to the prototype was not taken into account. The speech recognition process takes the longest to complete averaging 20 minutes to convert 1 hour of speech to text. The time it took for the actual segmentation and labeling process was in averages 5 minutes. Therefore, a user should expect to wait 25 minutes after uploading the video to generate the segments and labels.

4. CONCLUSIONS AND RECOMMENDATION

The study approached video segmentation and labeling without relying on the video itself and prosodic features such as pauses on speech and loudness of voice. An advantage of this approach was its robustness against noises such as crowd noise or bad video and audio quality by taking advantage of the powerful automatic speech recognition system Google Speech-to-Text. The framework can be applied to a learning platform where students can browse video lectures relevant to their needs and skip to specific time segments with the aid of labels. Complementing our approach with prosodic features would enhance segmentation such as using pauses in speech as clues for a topic transition. The researchers have discovered that Google Speech-to-Text takes minutes for the conversion process. Since the trade-off between its speed and accuracy against other automatic speech recognition systems was not considered in the study. It might be better to use an automatic speech recognition system that is less accurate but faster. Given the time it takes for speech-to-text conversion, our framework is not fit for real-time segmentation and labeling. Instead, users must wait for minutes before getting results. The segmentation algorithm TextTiling the researchers used for segmentation relies on sentences. However, the speech was originally not punctuated. Therefore, any segmentation algorithm not relying on sentences would be promising research. Lastly, the machine used to train the topic model was not capable of discovering topics apart from the initial 100 topics. The researcher highly suggests that the next studies should aim to discover a minimum of at least 500 topics using a more powerful machine and using another algorithm apart from Latent Dirichlet Allocation.

REFERENCES

[1] O. Tilk, and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," Interspeech, pp. 3047-3051, 2016.

[2] B. Fralinger and R. Owens, “Youtube as a Learning Tool,” Journal of College Teaching and Learning, vol. 6, no. 8, pp. 15, 2009.
https://doi.org/10.19030/tlc.v6i8.1110

[3] D. M. Blei, “Probabilistic Topic Models”, Communications of the ACM, vol. 55, no.2, pp. 77-84, 2012.
https://doi.org/10.1145/2133806.2133826

[4] M. A. Hearst, “TextTiling: Segmenting text into multiparagraph subtopic passages,” Computational linguistics, vol. 23, no. 1, pp. 33-64, 1997.

[5] R. Lo, B. He, and I. Ounis, “Automatically Building a Stopword List for an Information Retrieval System,” Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop, vol. 5, pp. 1724, 2005.

[6] K. Chen. “Topic Identification in Discourse,” Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics, pp. 267-271

[7] Ballantine, J. (2004). Topic segmentation in spoken dialogue. Bachelor Thesis, Department of Computing, Division of ICS, Macquarie University, Sydney Australia.