White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/195705/

Version: Accepted Version

Proceedings Paper:
Mohammed, A (2022) Video Segmentation and Characterisation to Support Learning. In: Lecture Notes in Computer Science. EC-TEL 2022, 12-16 Sep 2022, France. Springer Nature, pp. 229-242. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-031-16290-9_17

This is an author produced version of a conference paper published in the Lecture Notes in Computer Science book series. Uploaded in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Video Segmentation and Characterisation to Support Learning

Abrar Mohammed[1] and Vania Dimitrova[1]

School of Computing, University of Leeds, UK

Abstract. The predominance of using videos for learning has become a phenomenon for generations to come. This leads to a prevalence of videos generating and using open learning platforms. However, learners may not be able to detect the main points in the video and relate them to the domain for their study. This can hinder the effectiveness of using videos for learning. To address these challenges, our research aims to develop automatic ways to segment videos, characterise them and finalise the segmentation work by aggregating adjacent segments within a video with the same focus of domain topic(s) or topic-concept(s). We present a framework for automated video segmenting and characterising to support learning (VISC-L). We assume that the domain we are processing videos from has been computationally presented (via ontology). We are using the Deep learning BERT-BASE-Uncased model with a binary classifier to identify the focus topic of each segment. Then we use a semantic tagging algorithm to identify the focus concept within the topic. The adjacent segments within a video with the same focus topic/concept are aggregated to generate the final characterised video segments. We have evaluated the usefulness of watching the identified segments and characterisations compared with video segmentation provided by Google.

Keywords: Video-based learning · Video transcript · Text analytics · Domain ontology · Video characterisation · Video aggregation.

1 Introduction and Related work

The use of videos for learning has increased rapidly. It offers the flexibility of having visual and auditory channels that make it easier for learners to get the information and to support their learning [11, 14]. There is a massive amount of freely available videos that learners have access to. Not only does learning from such videos take up a significant amount of time for watching, but, crucially, it can be hard for the learners to identify key points in the videos and link these points to the study domain. [3, 21].

Manual video segmentation and characterisation. A widely used approach for video segmentation and characterisation is manual annotation. A common technique when using videos for learning is note-taking which makes a reference of important points mentioned within a segment in a video [10]. This allows learners to identify relevant video segments and to indicate key points in

these segments. However, such video annotation requires manual effort. In [5], segmenting videos has been done by teachers who provide the sections in the videos related to specific courses. In order to improve learner engagement and to aid the digesting of the learning material, teachers in [22] characterise learning videos by highlighting the contents with a phrase or a keyword or by adding questions. While such characterisation is closely linked to the learning goals set by the teachers, it is subjective and does not scale across different domains.

Characterising videos can also be done by learners. For instance, teachers have asked learners to annotate videos and test the effect of it on their learning [16]. Though learner annotation can improve engagement with videos, it is dependent on learner engagement (e.g. high self-efficacy learners engage better) and their prior knowledge (e.g. learners may not be able to see key points).

Automatic segmentation and characterisation. Recent works have developed approaches for automating the process of video segmentation. This falls into two categories - using learner interactions and using video content. In [17] learners' comments while watching videos are aggregated to identify "high attention intervals" which refer to key points noted by learners. These intervals are used to facilitate interaction with videos by offering an interactive visualisation interface. While using learner interaction data to segment and characterise videos which can give the learners reactions and perspective, the segmentation depends on the learner engagement and learners may not appropriately capture the key points in the videos. Alternative approaches focusing on the video content are proposed. To detect teaching practices (presenting, guiding, administration) in recorded lessons of trainee teachers, acoustic features from the audio and text feature from the transcripts are used in [21]. Machine learning models, trained by using annotations by expert observers, are dependent on the availability of previously annotated segments which may not always be practical.

Recent video segmentation approaches based on video content utilise state-of-the-art tools in natural language processing and tap into the availability of knowledge models. MOOC video lectures were automatically segmented in [6] by using a neural network over adjacent sentences; the neural network was trained on Wikipedia pages. To characterise the video segments, topics are extracted from slide titles. In [7], topical segmentation of lecture videos is performed by using a domain knowledge graph. A BERT model is used to compute the semantic similarity between different concepts in the video. [4] uses different text sources (transcript, slide text, hand written text on whiteboards) to segment and annotate videos. The segmentation is based on the transition between slides, while the annotation uses Wikidata and DBPedia to find the entity type and to compute semantic similarity between tokens in the video segment's text. All existing approaches have evaluated only the technical performance of their segmentation algorithms; their usefulness to aid learning has not been assessed.

In this paper, we address the following research question:

How to automatically segment and characterise videos to support learning?

We present a generic ontology-underpinned framework, called VISC-L, which uses video transcripts to segment, characterise and link videos to the domain

knowledge covered in the segments. Similarly to the last approaches, we use existing knowledge models, in the form of a domain ontology, to identify the domain concepts as well as use the ontology hierarchy and a language model based on BERT to identify focus topics and concepts for each video segment.Our work has a key difference from previous approaches. While none of them assesses the effect on learning, we provide here an evaluation study with users to examine the effect of the segmentation and characterisation in a learning context. We compare with a state-of-the-art video segmentation and characterisation interface that is available for YouTube videos[1].

The main contributions of the work presented here is: (a) a novel framework for segmenting and characterising videos by using video transcripts and linking them to domain concepts; (b) application of the framework in a representative learning domain (presentation skills); (c) evaluating the usefulness of video segmentation and characterisation for learning and drawing wider implications for adoption. The work is part of broader research that explores how to generate video narratives to support learning by linking video segments to help learners to identify and link key points in videos.

The paper is organised as follows. Section 2 outlines the VISC-L framework, and Section 3 presents how VISC-L is applied in the Presentation Skill domain. A user evaluation study is presented in Section 4, and Section 5 is a conclusion.

2 Framework for Video Segmentation and Characterisation for Learning (VISC-L)

The proposed framework for Video Segmentation and Characterisation for Learning (VISC-L) is presented in Fig. 1. It includes three main steps: selecting initial segments, characterising those segments, and aggregating the segments based on common domain topics.

Input. VISC-L is based on two assumptions. Firstly, it is assumed that the **video transcript** relates to the domain which will be learned (e.g. the videos can be lectures/tutorials/conversations linked to a specific topic). Hence, the text in the video transcripts is taken as an input. The second assumption is that there is a **domain ontology** $\Omega = \{C, H\}$ which includes the relevant domain concepts C $\neq \phi$ linked in a concept hierarchy H. Available ontologies - The Linked Open Data Cloud: [2] can be used or the ontology can be developed with domain experts. The later is used in this work. We use $c_i \subset c_j$ to denote that c_i is a subclass of c_j. The top level concepts in the concept hierarchy define the main domain topics $\{T_1, ..., T_m\}$. In order to identify the main topics in the video transcripts, as part of the characterisation step (see below), **training data** with domain topics as labels are needed. This can either be created with expert annotators or collected from past user interactions (the later is followed in the application of VISC-L presented in the next section).

[1] Offered by Google, produced by Google Video AI https://cloud.google.com/video-intelligence/

[2] The Linked Open Data Cloud https://lod-cloud.net/

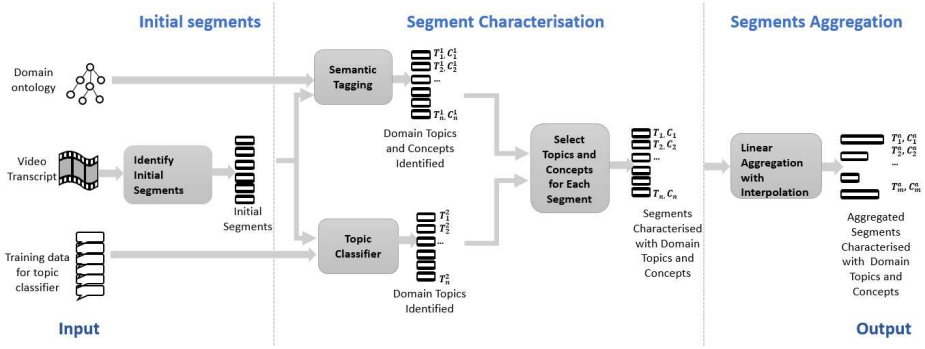


Fig. 1. Video segments characterisation and aggregation framework.Notice that T means a set of focus topics and C means a set of focus concepts within the focus topics, e.g. T_1C_1 means there is a focus concept C_1 in the focus topic T_1.

Output. The output of VISC-L is a set of aggregated video segments with a start and end time in the corresponding video. Each aggregated video segment i is characterised with a set of domain focus topics T_i^a (top concepts in Ω) and a set of concepts C_i^a from the focus topics which are mentioned in the transcript of the video segment (for every c_i from C_i^a, $c_i \subset T_i^a$).

Initial Segments. Our video segmentation approach is inspired by text-tilling in text summarisation [12] - starting with smaller units (e.g. sentences) and aggregating them to get larger coherent units (e.g. paragraphs). Hence, we include an initial segmentation step where the video transcripts are cut into small segments that are used as a starting point for aggregation. Initial segments can be done by using a certain number of text lines (e.g. the approach presented in the next section) or by using pre-defined segments (e.g. high attention intervals from past interactions [18]).

Segment Characterisation. In order to aggregate the initial segments, we need to identify what domain content is presented in each segment. This is done during the segment characterisation step which links each video segment i with a set of focus topics T_i and a set of concepts C_i. To do so, we propose using two algorithms: semantic tagging and topic classification.

The first algorithm is **semantic tagging** which was developed in our previous work [18]. This algorithm links each video segment to focus topics and concepts by mapping the terms from the ontology to the text in the video transcript. It first pre-processes the transcript[3], including: tokenise the transcript, clean it from stop words and punctuation, select nouns and noun phrases and match the ontology terms to the noun phrases. If there is a match, the ontology concept c_i will be identified (tagged to the text), noting also the path to reach a top level concept (i.e. linking to a focus topic T; $c_i \subset T$). As a result, each segment i is linked to a set of focus topics and their corresponding concepts; we

[3] We have used Natural Language Tool Kit (NLTK) https://www.nltk.org/

denote this as $< T_i^1, C_i^1 >$ (where ^{1} indicates that this is an output from the first segment characterisation algorithm). A key challenge for this algorithm is word sense disambiguation. This is not that prominent with carefully selected videos. However, if the videos are selected automatically from open social spaces, there will be a high risk of word sense ambiguity. Hence, we need to disambiguate the topics based on the context, which is done with the second algorithm.

The second algorithm is a **topic classifier** which identifies a domain topic based on the context of that topic. Following the latest development in natural language processing, we use Bidirectional Encoder Representations from Transformers (BERT)[8] as a topic classifier. BERT embeds pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Accordingly, it can be fine tuned with just one additional output layer to create state-of-the-art models for different language tasks, topic classification in this case. First, the BERT model is fine-tuned using training data with domain topic labels (which is part of the input for VISC-L). Then, the fine-tuned model is used as a classifier to link each segment i to its domain topics T_i^2 (where ^{2} indicates that this is an output from the second segment characterisation algorithm).

The last step in segment characterisation is to **combine the outputs from both algorithms**. For each segment i, the outcomes from the two algorithms $< T_i^1, C_i^1 >$ and T_i^2 are combined by intersecting the focus topics $T_i = T_i^1 \cap T_i^2$ and selecting the concepts C_i from C_i^1 that belong to T_i. Hence, each segment is characterised by $< T_i, C_i >$ (a set of focus topics T_i and their concepts C_i).

Segments Aggregation. Following the text-tilling approach [12], small segments will be aggregated into larger segments. To maintain the flow of information within adjacent segments, we have developed an aggregation algorithm based on thematic progression theory [2]. It states that a good written text should have a relation between theme, (the main clause), and rheme, ("the remainder of the message where the Theme is developed") [2]. Three patterns for coherent text are suggested: *Constant theme* (when the first theme in one sentence is carried on and used at the beginning of the second sentence); *Linear theme* (the important message in a rheme of one sentence is carried on into following clause as a theme in the second sentence), and *Split theme* (a development of a rheme with important information to be used as themes in subsequent clauses in the following sentence). Relating to video segments and using the characterisation, we associate the focus topic with the segment's theme and the focus concepts with the segment's rheme. We propose a **linear aggregation with an interpolation algorithm** (see Algorithm 1). The linear theme pattern was selected as the most appropriate, as it allows keeping a continuous focus topic and at the same time take into account the specific concepts within that topic. Some segments can be without characterisation which can be because the speaker is silent or is digressing from the domain. If we look strictly for adjacent segments, these *gap segments* which break the topic flow will lead to starting a new aggregate. To smoothen the aggregation, we use interpolation. If the segments before and after a *gap segment* have common focus concepts, it is assumed that the

common concepts are spread across the three segments. Hence, the *gap segment* will be interpolated in the aggregated segment.

3 Application of VISC-L in a Presentation Skill Domain

3.1 Input and Initial Segments

Domain and ontology. To apply VISC-L we have selected the Presentation Skill domain which represents a transferable skill that can assist learners in transmitting their message or to convince others with their ideas [9, 15]. This domain is supported by a domain ontology designed by [1]. The main topics T in this ontology are Delivery, PresentationAttribute, Structure and Visual Aids. Each topic has its own concepts C and this domain ontology has 302 concepts.

Video selection and initial segmentation. If there is no pre-selected set of videos related to the domain, the ontology can be used to collect videos from available social platforms (like YouTube). Following the concept hierarchy, a search schema can be developed, similarly to [19]. For example, using a combination of $< Domain, T_i, C_i, "tutorial" >$ as search terms, videos with tutorials related to the topic T_i in the domain can be collected. We have implemented this search schema using the library youtube-search-python [4]. We have applied a time filter so that each video duration should be > 3 minutes and selected the videos that associate with the YouTube generated transcript. As a result, we have collected 63 videos that have a corpus of 110594 tokens. Then, we applied the Initial segmentation step from VISC-L on the 63 collected video transcripts and generated 2382 segments.

Training data for the topic classifier. To fine tune the BERT-BASE-Uncased model, we used the training data which we obtained from the six studies conducted on the [17] learning platform. The domain of the videos used in [17] is the Presentation skill domain where the students can write comments or rate other students' comments. The total number of participants who watched the videos was 38 and they wrote 2038 comments. These comments had been labelled by other students with the domain topics: Delivery, Structure and VisualAids; notice that the topic PresentationAttribute is missed from the labels- we have solved this issue in the semantic tagging step below.

Segment Characterisation The characterisation of video segments includes two steps: semantic tagging and a topic classifier. With **Semantic tagging** we have applied the semantic tagging algorithm, as demonstrated in section 2, which has two inputs: the transcript of the 2382 segments generated from the Initial segmentation step in VISC-L, and the ontology [1] of the Presentation skill domain. The transcript of each segment has been tokenized, cleaned and POSTAGed to get the resulted nouns and noun phrases to be semantically tagged to the ontology terms to decide the focus topic/concept $< T_i^1, C_i^1 >$. This algorithm also succeeds in noticing the topic, PresentationAttribute, and collect its concept. To overcome the issue of word sense ambiguity mentioned in section 2

[4] https://pypi.org/project/youtube-search-python/

Algorithm 1: VISC-L algorithm. The input is the list of the segments from the videos with their focus topic FT and concepts FC. We aggregate the segments from the same video

Data: $S = \{s_1, s_2, .., s_n\}$; $n \geq 0$; $s =< FT, FC >$ where $FT, FC \neq \phi$,CurrentSeg=ϕ,NextSeg=ϕ,PreviousAgg=False,AggList=[],Gap=ϕ, $i = 0$

Result: $AllAgg = \{S_1a, S_2a, .., S_ma\}$; $m <> 0$

while $i < n$ **do**
  CurrentSeg=S_i;
  if *PreviousAgg=False* **then**
    NextSeg=S_{i+1}; $FocusC \leftarrow CurrentSeg \cap NextSeg$
    if $FocusC <> \phi$ **then**
      if $Gap \neq []$ **then**
        $AggList \leftarrow CurrentSeg, Gap, NextSeg$
      else
      $AggList \leftarrow CurrentSeg, NextSeg$;$i \leftarrow i+2$; $PreviousAgg = True$
    else
    if $Gap = []$ **then**
      $Gap \leftarrow NextSeg$; $i \leftarrow i+1$; Go to Step 3 ;
    else
      $AllAgg \leftarrow CurrentSeg$;$AggList, FocusC \leftarrow []$;CurrentSeg= Gap; $i \leftarrow i+1$; Go to step 3
    end
  else
    if *PreviousAgg=True* **then**
      $FocusC \leftarrow CurrentSeg \cap FocusC$;
      if $FocusC <> \phi$ **then**
        if $Gap = []$ **then**
          $AggList \leftarrow CurrentSeg$; $i \leftarrow i+1$;
        else
          $AllAgg \leftarrow CurrentSeg, Gap$; $Gap \leftarrow []$; $i \leftarrow i+1$;
        end
      else
      $AllAgg \leftarrow AggList, CurrentSeg$; $AggList \leftarrow []$; CurrentSeg= Gap;
      $Gap = []$;$FocusC = []$; $i \leftarrow i+1$; PreviousAgg=False;
    else
    end
  end
end
if $i = n$ **then**
  $CurrentSeg = S_i$;$FocusC \leftarrow CurrentSeg \cap FocusC$;
  if $FocusC <> \phi$ **then**
    if $Gap = []$ **then**
      $AggList \leftarrow CurrentSeg$; $AllAgg \leftarrow AggList$; $AggList = []$
    else
      $AggList \leftarrow CurrentSeg, Gap$; $Gap \leftarrow []$; $AllAgg \leftarrow AggList$; $AggList = []$
    end
  else
    if $Gap = []$ **then**
      $AllAgg \leftarrow AggList, CurrentSeg$; $AggList \leftarrow []$
    else
      $AllAgg \leftarrow AggList, Gap, CurrentSeg$; $AggList, Gap, FocusC \leftarrow []$; PreviousAgg=False;
    end
  end
end

by identifying the tokens in the transcript that are contextually related to the domain, we need the topic classifier model.

The second step is the **Topic classifier**. To select the best Deep learning model to be considered as the VISC-L Topic classifier, we have conducted two experiments. First, we compared different pre-trained BERT models (Roberta Base [13], Distill Bert [20] and BERT-BASE-Uncased [8]) as topic classifiers. These models are widely used for topic modelling and semantic analysing tasks. After we fine-tuned the models with our training data, we passed the 2382 video segments generated from the initial segmentation step to the models to be classified with the domain topics. To choose the best model, we compared between their precision, recall and F1-score values. Hence, BERT-BASE-Uncased has been selected as it gives higher (precision, recall and F1-Score) results and is better to be used as a binary classifier as shown in Table 1.

To get the final segment characterisation, we run the step of **Combining the characterisation results** identified from the semantic tagging and the topic classifier. For instance, a segment i has two characterisations, one from the semantic tagging algorithm $< T_i^1, C_i^1 >$ and one from the topic classifier model $< T_i^2 >$. The final characterisation of the segment i is the result of combining the two characterisations: $< T_i, C_i >=< T_i^1, C_i^1 > \cap < T_i^2 >$. This means, the focus topic is the one identified in both characterisations. Notice that the topic *PresentationAttribute* can only be recognised by the semantic tagging algorithm as mentioned in Section 2.

Table 1. BERT-BASE-Uncased model as multiple and binary classifier result.

Topic	Multiple Classifier			Binary Classifier		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Delivery	0.89	0.94	0.92	0.91	0.93	0.92
Structure	0.76	0.74	0.75	0.69	0.71	0.70
Visual Aids	0.97	0.85	0.91	1.00	0.87	0.93

Characterisation Outcome. The final characterisation result revealed that 1877 segments have been characterised with a focus topic(s) and concept(s). However, there are 505 segments with no characterisation. The average segment duration is 14 seconds with (STDV = 6) and the average number of focus topic(s)/concept(s) per segment is (1 and 2) respectively. According to the characterisation results, 62% of all segments focus only on one topic while 30% focus on two topics and 7% of all segments focus on three topics. The number and type of the focus topic(s) within the segments can inform the next step in our framework which is the aggregation of video segments (see section 4.3). Additionally, the segments' characterisation can inform their usage for learning (useful for creating video-segment-narratives). For example, to get an in-depth focus of the concepts within a topic, the segments with one focus topic can be used. Whereas, to find the relationship between two topics, the segments with two focus topics can be helpful, e.g. 10% of segments (the higher percentile) focus on the topics

Delivery and Structure together. The segments that focus on three topics could be used as introductory segments to the domain by mentioning most of its topics.

The characterisation results showed that more segments have the same focus topic which can be aggregated together to get longer segments with the same focus topic and concept. Before we commenced the aggregation step, we first evaluated the characterisation of the single segments.

Characterisation Evaluation. In order to evaluate the characterisation of the single segments, we have asked an external expert to asses the accuracy of the characterisation of 137 random segments taken from the 2382 segments that we have characterised from all the videos. The expert is a researcher who has some work done on the same domain of this work (Presentation skills). We provided the expert with the topics and concepts in the domain ontology with their hierarchy to be familiar with the nature of the video segments' characterisation. The expert has been asked to do the following: add new topics or concepts if it is thought to be missed from the characterisation, mark the assigned topic or concept to segment as wrong if they thought it was irrelevant or leave the characterisation if it was correct. The overall number of topics/concepts that were assessed was 345. By analysing the evaluation results, we found that the new suggested concepts from the evaluators either added to the ontology if it was relevant to the domain or ignored if it was irrelevant. To measure the agreement between the expert and our characterisation, we ran the Cohen Kappa formula. The agreement value we got from the formula was 0.91 which is 90%, proving a high agreement between our characterisation and the expert.

Linear Aggregation with Interpolation. We ran the third step of VISC-L on the characterised segments to aggregate them based on their focus topic/ concept. The input to the aggregation Algorithm 1 is the video segments with their final characterisation. The algorithm compares the adjacent segments $i, i{+}1$ within a video and checks their focus topic and concept $<T_i, C_i>, <T_{i+1}, C_{i+1}>$. It checks whether they intersect with each other and have some similar concept(s) $FocusC =< T_i, C_i > \cap < T_{i+1}, C_{i+1} >$. If $FocusC \neq \phi$ then these segments will be aggregated and their final characterisation will be $< T_i^a, C_i^a >$, which is the focus topic and concept of the aggregates. The duration of the aggregates starts from the beginning of the first segment and finishes at the end of the last segment in the aggregate. If $FocusC = \phi$, then the second segment $i + 1$ will be a gap segment and the algorithm will check the intersect between the first and third segment $FocusC =< T_i, C_i > \cap < T_{i+2}, C_{i+2} >$. If $FocusC \neq \phi$ then the segment $i, i + 1, i + 2$ will be aggregated - this is the reason we call it aggregation with interpolation. Otherwise, if $FocusC = \phi$, the first segment i will be saved as a single segment and a new aggregation will start from the segment $i + 1$ which will be considered as the first one. The aggregation result revealed that the number of the segments decreases to become 933 (where the original number was 2382). This showed that many adjacent segments have the same focus topic and concept. This is proved with the increase in the percentage of the segments (67.7%) that focus on one topic and concept. Subsequently, the size of the new aggregates has been increased with an average duration of 36 seconds.

Furthermore, the predominant topics, after aggregating the segments, are still the topics Delivery with (29.1% of the segments) and Structure with (20.6% of the segments). There is a decrease in the number of segments that focus on two topics (13%). Nevertheless, the topics Delivery and Structure stand out as the more correlated topics among other topics which highlighted that they are necessary to understand each other. On the other hand, the topic Presentation Attribute appears alongside other topics in the aggregates instead of being a unique focus topic. This indicates that this topic is better to be demonstrated by presenting its relationship with other topics.

4 User Study

To evaluate the usefulness of the characterised video segments to support learning, we have conducted a user study focusing on soft-skills (giving presentations).

4.1 Experimental Setup

Participants. 18 people (10 Male, 7 Female and 1 other) took place in the study; 16 PhD students from the University of Leeds and 2 from Industry. 13 participants were 18-29 years old and 5 participants were above 30 years. The training level is varied: 13 have some training, the rest either have a lot of training or received no training before. Their presentation experience is varied: 10 have a Medium level, 5 have either an experienced level or little experience. 12 participants are native English speakers. 10 participants watch YouTube videos every week for learning and every day for other purposes, whereas the rest use YouTube occasionally.

Materials and Procedure. 8 videos have been selected for the study, based on: their popularity, the duration of the video should be between > 4 and <6 minutes so the study will not last for more than one hour. A survey (using Google Forms) was prepared to assess the learning effect, perceived usefulness, cognitive demand and usability, comparing the VISC-L and Google algorithm. The participants went through the following steps in the survey: 1.Read and accept the consent form, 2.Complete a short pre-study section to collect their profiles, 3.Watch several suggested video-segments with characterisation generated using one of the algorithms (VISC-L or Google), 4.Give feedback on the video-segments and the provided characterisation, 5. Provide a short video summary, 6.Give feedback on the usability and usefulness of the recommended video-segments for learning about giving presentations, Repeat [3-6] with segments generated by the other algorithm (Google or VISC-L).

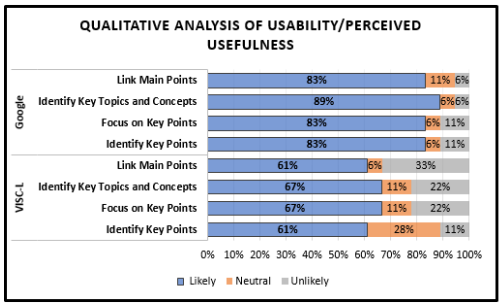
The study was approved by the ethical committee of of the Faculty of Engineering and Physical Sciences , University of Leeds.

Data analysis. Due to the limited number of participants, when comparing the VISC-L and the Google algorithm with regard to learning effect, perceived usefulness, cognitive demand and usability, we use the non-parametric Mann Whitney U-test, there was no statistical significance at $p<0.05$.

4.2 Results

To assess the **Learning effect** of watching the video segments, we compared the participants' domain terms mentioned in the pre-test with the new domain terms mentioned after watching the video segments generated by either VISC-L or the Google algorithms. During the pre-test, an average of 6 terms (STDV=5.9) were mentioned by the participants. After watching VISC-L segments, the participants named on average 7 (STDV=4.6) new terms, while after watching the Google segments, the participants named on average 7 (STDV=4.8) new terms. With both algorithms (no significant difference), the video segments with characterisation led to identifying new domain terms.

Perceived usefulness comparison between the characterisation of VISC-L and Google considered whether participants managed to identify and link main points in the videos to the topics in the domain and to identify key points and focus on them. The results presented in Fig. 2-**A** showed that in general, the characterised segments were LIKELY to meet their goals. The participants preferred the characterisation generated by Google more because the language used was extracted directly from the transcript and was easy to recognise in the video, while VISC-L was referring to key domain concepts related to the transcript.



A

		Mental	Effort	Frustration	Performance
Google	AVG	9.39	7.22	9.89	11.89
	STDV	2.93	2.94	5.07	6.04
VISC-L	AVG	7.89	7.39	8.18	10.65
	STDV	3.72	5.01	5.13	6.13

B

Fig. 2. A: Perceived usefulness of the characterised video segments for learning using VISC-L and Google. **B:**Cognitive demand results of the characterised video segments for learning using VISC-L and Google. The values range from 1(low) to 20(high).

Furthermore, participants were asked what they found positive or negative when watching characterised video segments generated by VISC-L or Google. For **both VISC-L and Google Positives** the participants found that the segments offered them a strategy for learning and were good to help them focus as these segments were short and with description. For **VISC-L Negatives** the participants noticed that the characterisation was scripted and not in a natural way while for **Google Negatives** they found that some descriptions did not

match with the video content, commented that there were too many segments within some videos, and pointed at inaccurate starting times.

Cognitive demand was assessed using the NASA-TLX questionnaire[5], including mental demand, effort, frustration, and performance - the results are showed in Fig. 2-**B**. Participants were asked to provide comments to justify their scores. For **Mental Demand** they reported that the video segments generated with both VISC-L and Google had low Mental demand and required low **Effort**. This is because the segments were short, easy to watch and the characterisation helped the participants to focus on a single topic. Meanwhile, high mental demand and effort was reported because some of the video content added little or no new knowledge or the description was not in-line with the video or did not specify the focus topics. With regard to **Frustration**, for VISC-L, 4 participants reported high frustration because they found some segments' characterisation did not align with the actual content. While with Google, 5 participants gave high frustration because they found some segments' start time was inaccurate or the characterisation was incomplete. Regarding **Performance**, there was similar feedback for both segments generated with VISC-L and Google. Participants gave high performance as they found the segments were very good at explaining the key terms and helped them to think of the domain topics. Whereas, few participants reported low performance as they did not enjoy some of the video content and did not feel they learned new things.

To assess **Usability**, we asked the participants to rank whether the segments were useful and the characterisation was helpful for learning presentation skills. The number of participants given as an average and standard deviation for each ranked factor. For VISC-L segments, an (avg=0.47, STDV=0.36) found that the segments were useful but an (avg=0.30, STDV=0.26) found that these segments were not useful. In comparison, for Google segments, an (avg=0.54, STDV=0.34) found these segments were useful but an (avg=0.24, STDV=0.23) found these segments were not useful. Furthermore, we tested whether the characterisation of the segments was helpful or not. For VISC-L characterisations, an (avg=0.46, STDV=0.29) found that the characterisation was helpful but an (avg=0.39, STDV=0.30) found that these characterisations were unhelpful. For Google characterisations, an (avg=0.60, STDV=0.26) found them helpful but an (avg=0.26, STDV=0.17) found that the characterisations unhelpful. These results indicate that the participants agreed that the using of characterised video segments for learning was helpful.

5 Conclusion and Future Work

We proposed the generic video segmentation and characterisation framework VISC-L to support learning. It was applied in a presentation skills domain. An evaluation study examined the usefulness and usability of video segmentation and characterisation, comparing VISC-L and Google. The results from the

[5] https://humansystems.arc.nasa.gov/groups/tlx/

study gave two indications. Firstly, they indicated that using characterised video segments could improve learners' domain knowledge, as the learners were able to identify new domain terms. Secondly, the results showed that there was no statistical significant difference between VISC-L and Google video segmentation and characterisation. With regards to learning effect, for both VISC-L and Google, there was improvement in learning because there were unique new terms mentioned in the summaries after watching the videos. Hence, the study provides support for using segmentation with characterisation to support learning. The perceived usefulness of segmentation and characterisation with Google was slightly better than VISC-L. Participants' feedback indicated that the format used to present the characterisation has influenced the usefulness - the natural language descriptions offered by Google were easier to follow than the list of concepts offered in the VISC-L interface. The usability with both VISC-L and Google shows that their generated characterised segments were helpful.

Having a characterisation in the form of terms linked to a domain ontology will allow us to develop algorithms for connecting video segments to create video narratives (combining several segments) to focus on specific domain concepts. We will combine VISC-L with the Google approach: VISC-L to extract the concepts and Google to create initial segments and to formulate titles. VISC-L is currently being applied in healthcare where we focus on awareness of patients' health-related quality of life needs, using online videos with patient stories.

Acknowledgments The authors wish to thank the participants in the user study. The authors thank Prof. Tanja Mitrovic and her colleagues at the University of Canterbury New Zealand for sharing the user interaction data, which was used to fine-tune the BERT model. The work on VISC-L is partially funded by the the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750 (InADVANCE project).

References

1. Abolkasim, E.N.A.: Semantic Approach to Model Diversity in a Social Cloud. Ph.D. thesis, University of Leeds (2019)
2. Bloor, M., Bloor, T.: The practice of critical discourse analysis: An introduction. Routledge (2013)
3. Bywater, J.P., Floryan, M., Chiu, J.L.: Discs: A new sequence segmentation method for open-ended learning environments. In: International Conference on Artificial Intelligence in Education. pp. 88–100. Springer (2021)
4. Cagliero, L., Canale, L., Farinetti, L.: Visa: A supervised approach to indexing video lectures with semantic annotations. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference. vol. 1, pp. 226–235. IEEE (2019)
5. Castro, M.D.B., Tumibay, G.M.: A literature review: efficacy of online learning courses for higher education institution using meta-analysis. Educ. and Info. Techn. **26**(2), 1367–1385 (2021)
6. Das, A., Das, P.P.: Automatic semantic segmentation and annotation of mooc lecture videos. In: Int. C. on Asian Digital Libraries. pp. 181–188. Springer (2019)

7. Das, A., Das, P.P.: Semantic segmentation of mooc lecture videos by analyzing concept change in domain knowledge graph. In: Int. C. on Asian Digital Libraries. pp. 55–70. Springer (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
9. Di Gangi, P.M., Goh, S.H., Lewis, C.C.: Using social media to support presentation skill development in traditional classroom environments. Journal of Organiz. and End User Comput. (JOEUC) **29**(3), 68–91 (2017)
10. Dodson, S., Roll, I., Harandi, N.M., Fels, S., Yoon, D.: Weaving together media, technologies and people: Students' information practices in flipped classrooms. Information and Learning Sciences (2019)
11. Haridakis, P., Hanson, G.: Social interaction and co-viewing with youtube: Blending mass communication reception and social connection. Journal of broadcasting & electronic media **53**(2), 317–335 (2009)
12. Hearst, M.A.: Texttiling: A quantitative approach to discourse segmentation. Computational Linguistics **23**(1), 33–64 (1997)
13. Huertas-Garcıia, Á., Huertas-Tato, J., Martín, A., Camacho, D.: Civic-upm at checkthat! 2021: integration of transformers in misinformation detection and topic classification. arXiv:2109.12987v1 (2021)
14. Khan, M.L.: Social media engagement: What motivates user participation and consumption on youtube? Computers in human behavior **66**, 236–247 (2017)
15. Maican, C., Cazan, A.M., Lixandroiu, R., Dovleac, L., Maican, M.A.: Exploring the factors influencing the use of communication and collaboration applications. Journal of Organizational and End User Computing (JOEUC) **33**(4), 94–124 (2021)
16. Mirriahi, N., Jovanović, J., Lim, L.A., Lodge, J.M.: Two sides of the same coin: video annotations and in-video questions for active learning. Educa. Techn. Research and Develop. **69**(5), 2571–2588 (2021)
17. Mitrovic, A., Dimitrova, V., Weerasinghe, A., Lau, L.: Reflective experiential learning: Using active video watching for soft skills training. In: Proceedings of the 24th international conference on computers in education. Asia-Pacific Society for Computers in Education (2016)
18. Mohammed, A., Dimitrova, V.: Characterising video segments to support learning. In: Proceedings of the 28th Inter. Conf. on Computers in Education. Asia-Pacific Society for Computers in Education (in print) (2020)
19. Pritoni, M., Paine, D., Fierro, G., Mosiman, C., Poplawski, M., Saha, A., Bender, J., Granderson, J.: Metadata schemas and ontologies for building energy applications: a critical review and use case analysis. Energies **14**(7), 2024 (2021)
20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
21. Schlotterbeck, D., Uribe, P., Jiménez, A., Araya, R., Molen Moris, J.v.d., Caballero, D.: Tarta: Teacher activity recognizer from transcriptions and audio. In: Int. Conf. on Artificial Intelligence in Education. pp. 369–380. Springer (2021)
22. Tseng, S.S.: The influence of teacher annotations on student learning engagement and video watching behaviors. Int. J. of Ed. Techn. in HE **18**(1), 1–17 (2021)