An interactive system to automatically generate video summaries and perform subtitles synchronization for persons with hearing loss

I. Cuzco-Calle, P. Ingavélez-Guerra, V. Robles-Bykbaev and D. Calle-López
GI-IATa, Cátedra UNESCO Tecnologías de Apoyo para la Inclusión Educativa
Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador
Email: {icuzco, pcingavelez, vrobles, dcallel}@ups.edu.ec

***Abstract*—According to World Health Organization (WHO), approximately 328 million of adults and 32 million of children present hearing loss in the world. Likewise, the number of people with such impairment increased from 42 million in 1985 to about 360 million in 2011. However, the most of multimedia and web contents, whether they are educational or leisure, are not accessible for persons with hearing loss. In this line, this paper presents an interactive system aimed at automatically generating video summaries and performing subtitles synchronization for persons with hearing loss. Our proposal relies on an educational platform (MOODLE) and Natural Language Processing (NLP) to provide an environment fully configurable for these persons. The module that generates the video summaries uses techniques such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), whereas the synchronization module is based on forced alignment between audio streams and text. With the aim of validating our environment, we have tested our approach on 15 videos, obtaining a score of 80% for three criteria related to the summary content: understandability, concordance, and context appropriateness.**

***Keywords*—Video summaries; Hearing loss; Disabilities; Latent Semantic Analysis.**

I. Introduction

The optimal use of multimedia tools to promote accessibility and adaptability in the interaction of students with disabilities is a constant challenge. The videos are currently considered as pedagogical strategies to achieve meaningful learning, since they interact several elements in their access modes such as reading, capturing visual, textual, auditory information, description of environments, among others [1]. In the area of education and disabilities, several advances have been made, but it is still difficult to evaluate the successful application of accessible tools that strengthen the student's interaction with their virtual educational environment. In many cases the virtual environment is particularized to a specific disability, generating new accessibility barriers for other pathologies [2].

Research on learning technologies for students with disabilities in Ecuador, demand constantly updated research finding that go in line with a changing reality. On the issue of disability it is important to consider the figures worldwide trend, since about 1000 million, or 15% of the world population, have some form of disability, and its incidence is higher in developing countries [3], this is not very encouraging, given that currently the structures for health care, rehabilitation and special education do not reach their full development, which makes us reflect on the need to have methodological proposals favoring inclusion.

978-1-5386-5491-0/18/$31.00

Currently, countries face the challenge of providing quality education for all, strengthening the inclusion approach that is gradually gaining ground in the educational and social fields, facing the high rates of exclusion, discrimination and educational inequality. The creation of the conditions for the development of education for everyone, which guarantees quality with equity, implies transformations in the educational system, in its cultures, policies and practices, involving in an active and participative way evaluative processes that feed back the efforts made. Focusing on people with disabilities is to open a diverse range of studies, efforts, perceptions and finding solutions.

Tanking into account that the inclusion of students with disabilities depends on a process of constant monitoring and the search for a universal learning that allows a true inclusion, support and graduation of the student, the present project explains the development of an application to generate summaries of videos accordingly to the needs of the user, thereby providing a tool that contributes to the systematization and capture of explanatory videos that may be lacking in students with hearing, intellectual or limited time limitations in order to listen to the entire video, requiring focusing on the most relevant aspects.

The rest of the paper is organized as follows. In Section 2 we present a brief overview contributions related with the text summarization. The general architecture of the proposed system is described in Section 3. The pilot experiment carried out with the aim of determining the real feasibility of the system is presented in Section 4. Finally, Section 5 presents the conclusions and some ideas for future work.

II. Related work

A. A novel clustering method for static video summarization

Static video summarization is recognized as an effective way for users to quickly browse and comprehend large numbers of videos. Inspired by the idea from high density peaks search clustering algorithm, they propose an efficient clustering algorithm integrating important video properties to gather similar frame into clusters. The input is a number of video frames, and the output is a storyboard composed of representative frames, the proposed method includes four steps [4]:

- Pre-sampling
- Video frame representation
- Clustering
- Video summarization result generation

B. Important Objects for Egocentric Video Summarization

In this work, we are interested in creating object-driven summaries for videos captured from a wearable camera, they first train a regression model from labeled training videos that scores any regions likelihood of belonging an important person or object, for the input variables, they develop a set of high-level cues to capture egocentric importance, such as frequency, proximity to the camera wearers hand, and object-like appearance and motion, as well as a set of low-level cues to capture region properties such as size, width, and height [5].

C. Video Summarization with Long Short-Term Memory

They propose a novel supervised learning technique for summarizing videos by selecting keyframes or key subshots. Casting the task as a structured prediction problem, their main idea is to use Long Short-Term Memory (LSTM) to model the variable-range temporal dependency among video frames, so as to derive both representative and compact video summaries. The proposed model successfully accounts for the sequential structure crucial to generating meaningful video summaries, leading to state-of-the-art results on two benchmark datasets [6].

D. An iteratively re-weighting algorithm for dynamic video summarization

In recent years, it has been recognized that the summary of condensed and meaningful video for the viewers is a beneficial and attractive research in the multimedia community.

This paper proposes an iteratively reweighting dynamic video summarization (IRDVS) algorithm based on the joint and adaptive use of the visual modality and accompanying subtitles. The proposed algorithm takes advantage of our developed SEmantic inDicator of videO seGment (SEDOG) feature for exploring the most representative concepts for describing the video [7].

III. General system architecture

With the aim of providing accessible educational multimedia materials for persons with hearing loss (and other specific needs), we propose an interactive system that implements several functionalities for both persons with special needs and developers. In Figure 1 we describe the main stages that are conducted by the system with the aim of generating accessible contents:

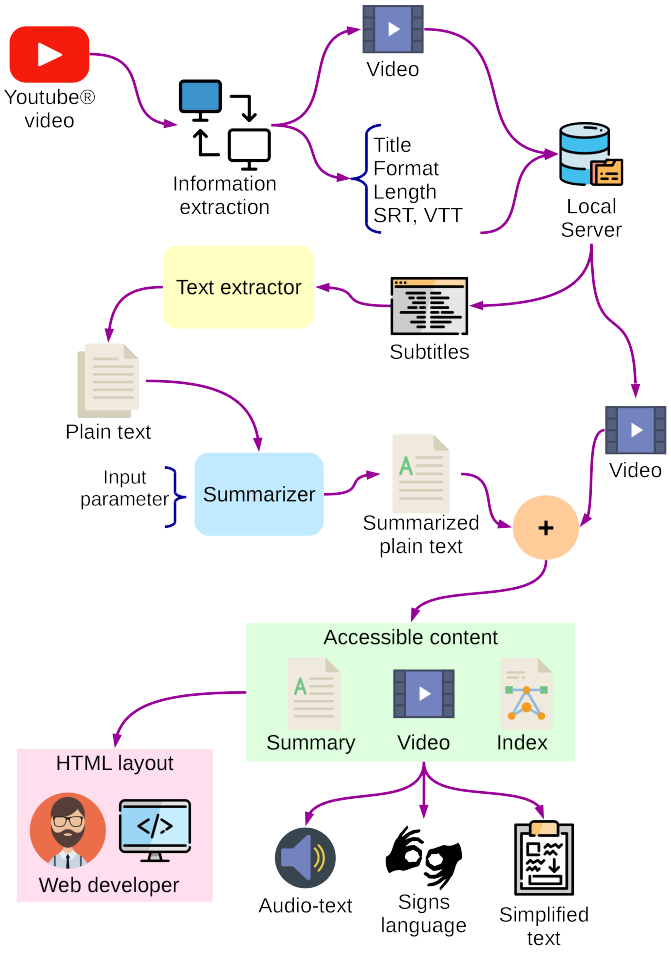

Fig. 1. The general system architecture.

- *Stage 1:* in this stage, the system retrieves the educational video from Youtube and the following information related with it: title, format (MP4, FLV, etc.), length, and subtitles (SubRip, Web Video Text Tracks, etc.). This information is indexed according to the educational category and stored on a local server. The subtitles can be used later to extract keywords that will be employed to establish times where exists a topic change in the video. These keywords are especially useful to create navigation menus according to the addressed topics in the video.
- *Stage 2:* this stage consists of obtaining the text summarization (through LDA and LSA techniques) [8] according to a specific level that can be set on the system (input parameter).

TABLE I
MATCHING PERCENTAGE

ID	Tool keywords	Online tool keywords	Matching percentage
1	Mam, Cleta, Teo	Cleta, Teo, Pablo	0.66
2	Onenote,onenote_class_notebooks, onenote_class_notebook, Onenote	Onenote,onenote_class_notebooks, onenote_class_notebook, Onenote	1
3	Nio, ordenador, colegio	Aplausos, Risas, Megafan	0
4	LED, resistencia, voltio	LED, miliamperio, voltio	0.66
5	Planeta, Facundo, movimiento	Facundo, Urano, Marte	0.33
6	Ciudad, Mesopotamia, imperio	Mesopotamia, Babilonia, Mediterraneo	0.33
		Average	**0.4966**

For the application of the LDA technology, a corpus is specified as input, and we generate a list of words that define a set of topics in a process composed of two stages:

- Randomly choose a distribution on the specified topics
- We relate the topic of the distribution chosen in step 1 with the input parameter and we show the words on the selected topic, this list of words are considered as keywords.

For LSA technology the text is presented as a matrix where terms are converted into rows and each unit of text as a column, resulting in a matrix with X_{ij} elements. Each element of this matrix can be interpreted as the number of times it presents a term i within a text unit j. Once the matrix is defined, the decomposition of its elements into singular values is performed, and then proceed to calculate the Cosine distances over them and then extrapolate the smallest distances between the elements. With this, we will have obtained the elements with greater similarity in their contexts.

- *Stage 3:* in this stage, the system combines the summarized text with the video to assemble the accessible educational content. This new content contains the following features that allow supporting the requirements of a relevant spectrum of persons with disabilities or special needs:
  - A simplified text that describes the main aspects of the educational contents addressed in the video. In this resource, the system substitutes words that do not appear commonly in a text with their synonyms or with words with the same meaning but more colloquial/ straightforward. For example, this resource can be used by older adults, persons with intellectual disability (mild or moderate), etc.
  - A version of the summary translated to Signs Language. With this resource, a deaf person can quickly analyze the video with the aim of extracting the most relevant of its aspects as well as to determine if is an educational material of her/his interest.
  - An audio transcription of the summarized text. This result can is useful for blind persons or persons with low-vision. The system can generate an audio transcription using the summary made or the simplified text of the summary.
  - With the aim of supporting the web developers tasks, the system can generate a set of suggestions to embed the video considering the WCAG 2.0 guidelines [9].

On the other hand, the problem of synchronizing texts with audiovisual elements involves extracting audio from the multimedia file and working on their data sets expressed as waves. The objective of doing this is to be able to recognize within the audio element, the phonemes related to the language so that, based on the most probable coincidences, words and sentences that can be added within the video are generated.

In order to make sense within each moment of the video, it is necessary to assign time labels on the textual transcription in such a way that it can be reproduced in a synchronous manner with the source audio element.

IV. EXPERIMENTS AND RESULTS

In order to sustain the objective of the developed system, two stages of experimentation were proposed, the first one consists of a comparative analysis of the keywords extracted from 6 random videos, where it was measured the percentage of agreement between our system and an on-line summary tool, thus obtaining an average of 0.4966 (Table I).

In Figure 2, we can see the words most frequently found at the time that the analysis of the plain text of the video processed by our intelligent system is performed; for the extraction of these words we used freeling [10], that allows us to obtain tuples $< lemma, label, probability >$.

On the other hand, the aim of the second experimentation stage was to determine the real precision of the system to obtain the video-summaries. In this line, a group of experts has selected 15 random videos of the total of 40 stored in the system's database. The group of experts saw each video, and after that, they read the 15 summaries generated by the platform. As a final step, the experts evaluated each summary according to the following criteria:

- *Content understandability (Cu):* determines how understandable is the summary of the video considering the total content of it.
- *Content concordance (Co):* specifies whether the summary sentences are meaningful with respect to the total content of the video.
- *Content appropriate context (Cc):* determines whether the words and sentences in the summary maintain the same context as the content of the entire video.

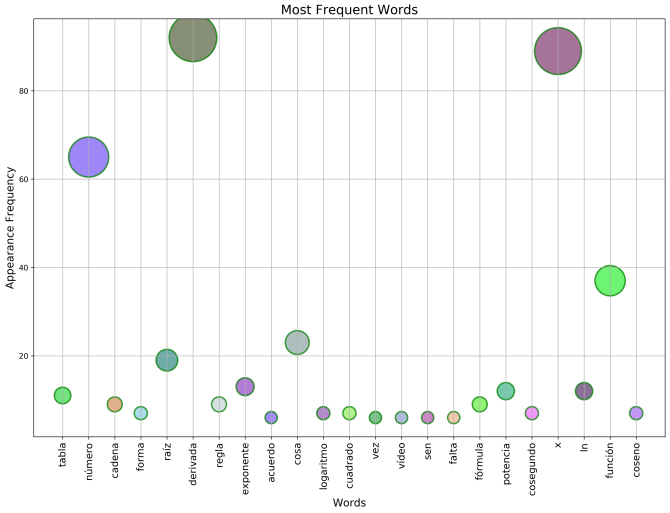

Fig. 2. A scatter plot of the most frequent words found in 6 videos randomly chosen.

Each of the above-described criteria was evaluated in the Likert scale (5=strongly agree, 4=agree, 3=undecided, 2=disagree, and 1=strongly disagree). As it can be seen in Figure 3, the achieved results are encouraging, given that the scores obtained are 4.13, 4.13, and 4.33, respectively.

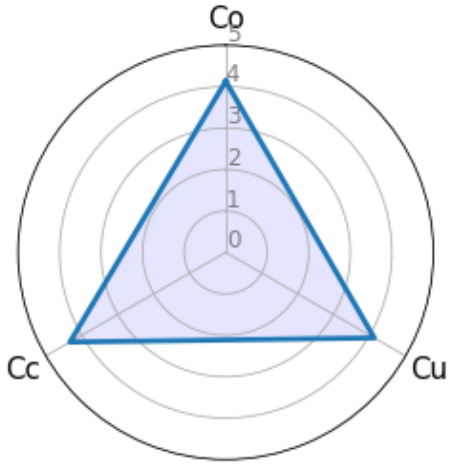

Fig. 3. Scores reached by the system in the summaries generation, according to the criteria of the group of experts.

V. Conclusion

Focusing on people with disabilities is to open a diverse range of studies, efforts, perceptions and search for solutions. There are several considerations to take into account to make viable or prevent the access of a person with disabilities to virtual environments. In many cases, oral and/or spoken expression generates problems in understanding the environment. A multimedia auditory element with long duration, could lower interest and lack importance, that is why we have proposed this prototype with the intention that the user does not lose interest in educational issues, for it is necessary to develop tools that branch out alternatives of access to educational resources, strengthening their accessibility and universal design.

As lines of future work we propose the following ones:

- To develop a module based on image analysis with the aim of determining the semantic content of different groups of video frames.
- To develop an intelligent module to perform the transcription of texts (captions) to Sign Language.

Acknowledgment

This work has been funded by the "Sistemas Inteligentes de Soporte a la Educación (v5)" research project, the Cátedra UNESCO "Tecnologías de apoyo para la Inclusión Educativa" initiative, and the Research Group on Artificial Intelligence and Assistive Technologies (GI-IATa) of the Universidad Politécnica Salesiana, Campus Cuenca.

References

[1] M. Á. H. Batista, "Consideraciones para el diseño didáctico de ambientes virtuales de aprendizaje: una propuesta basada en las funciones cognitivas del aprendizaje," *Revista Iberoamericana de Educación*, vol. 38, no. 5, p. 2, 2006.

[2] M. V. de Castro, M. A. S. Bissaco, B. M. Panccioni, S. C. M. Rodrigues, and A. M. Domingues, "Effect of a virtual environment on the development of mathematical skills in children with dyscalculia," *PloS one*, vol. 9, no. 7, p. e103354, 2014.

[3] J. Bickenbach, "The world report on disability," *Disability & Society*, vol. 26, no. 5, pp. 655–658, 2011.

[4] J. Wu, S.-h. Zhong, J. Jiang, and Y. Yang, "A novel clustering method for static video summarization," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9625–9641, 2017.

[5] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 38–55, 2015.

[6] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[7] P. Dong, Y. Xia, S. Wang, L. Zhuo, and D. D. Feng, "An iteratively reweighting algorithm for dynamic video summarization," *Multimedia Tools and Applications*, vol. 74, no. 21, pp. 9449–9473, 2015.

[8] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[9] M. Debevc, I. Kožuh, S. Hauptman, A. Klembas, J. B. Lapuh, and A. Holzinger, "Using wcag 2.0 and heuristic evaluation to evaluate accessibility in educational web based pages," in *International Workshop on Learning Technology for Education in Cloud*. Springer, 2015, pp. 197–207.

[10] X. Carreras, I. Chao, L. Padró, and M. Padró, "Freeling: An open-source suite of language analyzers." in *LREC*, 2004, pp. 239–242.