INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

Exploring collections of multimedia archives through innovative interfaces in the context of Digital Humanities

Géraldine Damnati, Delphine Charlet, Marc Denjean
Orange Labs, Lannion, France
{geraldine.damnati,delphine.charlet,marc.denjean}@orange.com

Abstract

STIK is a platform that gathers Speech, Texts and Images of Knowledge. It allows browsing and navigating through collections of multimedia, facilitating access to archives in the domain of Knowledge resources. STIK includes a back-end with a specific automatic metadata extraction pipeline, a front-end with innovative interfaces for navigating within a document and a specific implementation of a search engine with dedicated key-word search functionality. It gathers multimedia contents from Canal-U, a French institution that exploits audiovisual archives produced by Higher Education and Research, with various formats and various academic disciplines. STIK is a contribution to the emerging domain of Digital Humanities.

Index Terms: multimedia document collections, navigation interface, automatic metadata extraction

1. Introduction

As part of the global current efforts in the domain of Digital Humanities, facilitating access to audiovisual archives remains an important issue. While official instances such as the French INA (National Audiovisual Institute) address this issue for Broadcast archives [1], we are interested in proposing new ways of retrieving and navigating through archives linked to Education and Knowledge. Through a bipartite partnership between Orange Labs and the FMSH (Fondation Maison des Sciences de l'Homme), which holds the Canal-U audiovisual documentary data-base of archives produced by French Higher Education and Research instances (https://www.canal-u.tv/), we propose a platform that gathers several Spoken Language Processing tools for automatic metadata extraction, along with an innovative exploration and navigation interface.

A selection of collections of documents has been processed by the platform. They have various formats (conferences, keynotes, scientific documentaries produced by scientists, short MOOC sequences, interviews of researchers, etc…) and cover all the academic disciplines. If other Educational projects propose access to lectures through innovative instrumentation, our purpose is to favor access to various type of Educational material, beyond lectures.

2. Automatic metadata extraction

Automatic Speech Recognition and LM Adaptation

ASR is performed with the VoxSigma software based on LIMSI technology [2] (http://www.vocapia.com/). Despite variable recording conditions, the main challenge remains the lexicon and Language Model (LM) issue. In fact each document is highly specialized and necessarily contains a high Out-of-Vocabulary rate or occurrences of words that can be present in the vocabulary but not necessarily in the same semantic context as in the LM training data. In order to alleviate this issue we provide the ASR engine with adaptation data along with each audio document to be transcribed. Adaptation data are collected from existing editorial metadata (title, summary, speakers' names) and through specific adaptation data enhancement methods.

Speaker Diarization and Identification

The audio content is segmented into speech/non-speech, and processed with a speaker diarization (SD) tool [3]. Recording conditions are usually far less favorable than professional conditions encountered with Broadcast contents. The number and roles of speakers depend on the type of content: some contents are interviews of a guest by an anchor, some are reports and some are conferences with one or two speakers and questions asked by the audience,. Even if the audience questions cover less than 1% of the total speech duration, they play an important role from an applicative point of view and must be correctly detected. Finally, a speaker identification step is applied, on the basis of identities provided in editorial meta-data and speaker role classification. Thanks to available editorial information, it is enough to automatically detect speaker role in order to find his/her identity. Role classification relies on rules on speaker time distribution (e.g. the speaker who speaks the most regularly in a content with multiple guests is the moderator).

Key-word extraction and weighting

Additionnaly to Named Entity extraction (Person, Place and Organisation) we have implemented an unsupervised, lexicon-free, term extraction approach that relies on syntactic analysis (POS tagging and chunk partitioning using the `lia_tagg` software). We apply rules on the syntactic categories of chunks to extract nominal groups and variable span sequences of such groups. For instance in the following utterance: *"a right temporal atrophy appears for adults after three weeks of sensorial **isolation**."*, we extract three levels of key-words (KW) around *isolation*: [*isolation*], [*sensorial isolation*] (immediate context) and [*three weeks of sensorial isolation*] (extended context). The extended context generates more meaningful expressions but increases the number of rare KWs, lowering their statistical significance. Keeping three extraction levels in a nested representation provides an interesting compromise as the user can chose the span of KWs to observe. The relevance is obtained by computing $\text{TF-IDF}_{\text{BM25}}$ weights [4] at two levels. The *inter-document* level reflects the relevance of a KW in the document relatively to the rest of the collection; the *intra-document* level reflects the relevance of a KW in a segment relatively to the document's other segments.

Topic Segmentation

An approach derived from previous work on TV Broadcast News topic segmentation [5] has been adapted to this particular context and submitted as a regular paper in this Interspeech conference. Contrarily to Broadcast News, topic definition is not straightforward: documents are usually globally monothematic and unlike lectures, we cannot rely on slides synchronization to help segmentation. Hence, consecutive sub-topics are not necessarily independent, but the objective is to propose some relevant eyemarks on a potentially long document, as a way for users to quickly browse back and forth through topically relevant segments.

3. Interface main functionalities

The front-end interface, implemented as a webapp, can be used both from a computer and a touch-screen device.

3.1. Finding a document

Documents can be retrieved by directly consulting the available collections tree diagram or by key-word search. Based on the available editorial and automatic metadata, we have built an index on the basis of the Elastic Search technology. Static editorial metadata are indexed as well as automatically extracted intra-content temporal meta-data (full transcription, key-words, speaker segments, topic segments). A completion strategy has been designed on the basis of the nested multi-span KWs. A scoring function has been designed in order to combine reliable editorial meta-data and automatic meta-data associated to confidence scores.

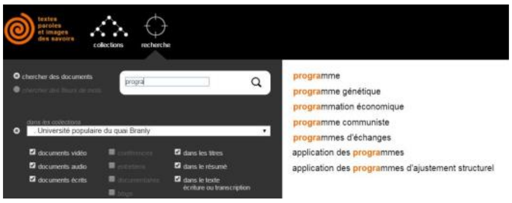

Figure 1: *search interface with key-word completion*

3.2. Navigating into a document

Document overview

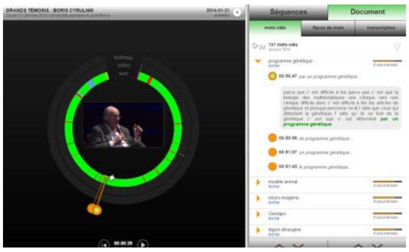

Figure 2: *Document overview with a circular player*

One of the originalities of this interface is the circular player surrounding the video. The wheel is segmented according to speaker diarization, each speaker being assigned a given color. In Figure 2, the document is a general public conference given at the Popular University of the Quai Branly Museum, where a journalist is conversing for 80 minutes with a famous psychiatrist (Boris Cyrulnik) followed by an interaction with the audience. A quick look at the wheel provides the structure of the conference, with the green segments corresponding to the guest and the other short segments corresponding to the mediator or to the audience. Eventually two additional concentric wheels can be used to see the structure of the video channel (not yet implemented) and the thematic segmentation. The panel on the right side provides the list of extracted KWs, sorted by their *inter-document* relevance weight. Selecting a KW reveals pins around the wheel, corresponding to its occurrences in the document. Each occurrence can be visualized in the right side panel in the context of its breath group, with the full transcription of the excerpt. Selecting a KW streams the audio to the beginning of the corresponding *breath group*. We define *breath groups* as sequences of words between two pauses (more than 60ms of silence between two consecutive words). This choice provides a relevant *audio feedback*, with sufficiently coherent context to understand the meaning while guaranteeing the KW will be uttered shortly.

Segment navigation

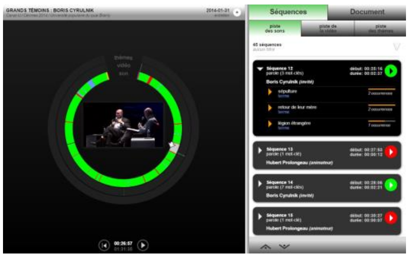

Figure 3: *Segment navigation*

The audio can be streamed both by clicking on the wheel or on one segment on the right side panel. Selecting a segment reveals its own metadata: the speaker's name and role, and the list of key-words in the segment, sorted by their relevance weight in the segment relatively to the other segments. It provides a quick access to a segment from a long conference, allowing to focus the attention on a given part or to quickly come back to a document after a first exhaustive listening.

4. Conclusion

This demonstration proposes new ways to access and explore multimedia archives in the domain of Digital Humanities. It exploits state-of-the-art Spoken Language Processing tools to extract metadata that are exploited within an innovative interface. The platform is under experimentation with several hundreds of hours of contents.

5. Acknowledgements

The authors thank the FMSH (Fondation Maison Sciences de l'Homme) for providing access to its multimedia archives.

6. References

[1] M. L. Viaud, O. Buisson, O., A. Saulnier, C. Guenais,. "Video exploration: from multimedia content analysis to interactive visualization", in *Proc. of ACM Multimedia*, 2010.

[2] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[3] D. Charlet, C. Barras, J-S. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates", in *Proc. ICASSP 2013*, Vancouver, Canada, 2013.

[4] K. S., Jones, S. Walker, S. E., Robertson, (2000). "A probabilistic model of information retrieval: development and comparative experiments", Information Processing & Management, 36(6), 2000.

[5] A. Bouchekif, G. Damnati, Y. Estève, D. Charlet, N. Camelin, "Diachronic Semantic Cohesion for Topic Segmentation of TV Broadcast News" ", in *INTERSPEECH 2015*, Germany, 2015.