Exploration de collections d'archives multimédia dans le contexte des Humanités Numériques : revisiter TALN'2015 ?

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

RÉSUMÉ

Cette démonstration présente un prototype d'exploration de contenus multimédias développé dans le but de faciliter l'accès aux contenus de la Connaissance. Après une extraction automatique de métadonnées, les contenus sont indexés et accessibles via un moteur de recherche spécifique. Des fonctionnalités innovantes de navigation à l'intérieur des contenus sont également présentées. La collection des enregistrements vidéo de TALN'2015 sert de support privilégié à cette démonstration.

ABSTRACT

Exploring multimedia archives in the context of Digital Humanities: browsing TALN'2015?
This demonstration presents an exploration prototype through multimedia contents developed in order to enhance access to Knowledge contents. After automatic metadata extraction, contents are indexed and become accessible through a dedicated search engine. Innovative navigation functionalities among contents are also presented. The collection of TALN'2015 video recordings is shown in this demonstration.

MOTS-CLÉS : collections multimédia, navigation, extraction automatique de métadonnées.

KEYWORDS: multimedia collections, navigation interface, automatic metadata extraction

1 Introduction

Parmi les enjeux soulevés par le domaine des Humanités Numériques, faciliter l'accès aux archives multimédia demeure un défi important. Si des travaux ont été menés dans le cadre des archives audiovisuelles, notamment à l'INA (Viaud et al., 2010), nous nous intéressons ici à proposer de nouveaux modes d'exploration dans le domaine de l'Education et de la Connaissance au sens large. A travers un partenariat entre Orange Labs et la FMSH (Fondation Maison des Sciences de l'Homme) qui gère entre autres le fond de documents audiovisuels produits par l'Enseignement Supérieur et la Recherche, via la webTV CanalU (https://www.canal-u.tv/), nous avons développé une plateforme pour l'exploration de ces contenus. Elle agrège plusieurs outils de TAL pour l'extraction automatique de métadonnées, ainsi qu'une méthodologie d'indexation dédiée et une interface innovante d'exploration et de navigation. Différents types de contenus sont disponibles (conférences, documentaires scientifiques, documents courts pour des MOOC, entretiens avec un chercheur, etc…). Les enregistrements vidéo des sessions orales de TALN'2015 ayant été intégrés au catalogue de CanalU par le CEMU (Centre d'Enseignement Multimédia Universitaire de l'Université de Caen), nous proposons dans cette démo un prototype d'exploration parmi les conférences de TALN'2015.

2 Extraction automatique de métadonnées

La **transcription de la parole** est réalisée à l'aide du logiciel Voxsigma (http://www.vocapia.com/) (Gauvain et al., 2002). La principale difficulté, dès lors que l'on traite des contenus spécialisés, demeure la couverture lexicale et l'adéquation du modèle de langage. Nous effectuons une **adaptation dédiée** pour chaque contenu. Les données d'adaptation sont établies *a minima* à partir des métadonnées éditoriales (titre, résumé, nom des intervenants) mais également à l'aide de mécanismes d'enrichissements automatiques. En outre pour le cas particulier des présentations orales de TALN'2015, nous avons exploité le texte de l'article associé.

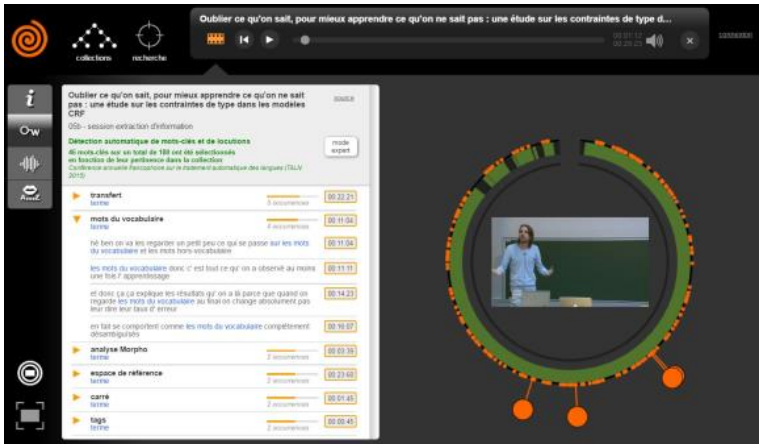
La ***diarization* en locuteurs** est réalisée à l'aide de l'outil détaillé dans (Charlet et al., 2013). L'**identification des locuteurs** est réalisée en s'appuyant conjointement sur les identités renseignées dans les métadonnées éditoriales et sur une analyse du contenu en rôle des locuteurs.

Outre l'extraction d'Entités Nommées (Personne, Lieu et Organisation), nous avons implémenté une approche non supervisée **d'extraction de mots clés** (KW pour *key-word*), ne s'appuyant pas sur un lexique mais sur une analyse morpho-syntaxique (étiquetage en POS et *chunking* à l'aide du logiciel `lia_tagg`). Un ensemble de règles sur les enchainements de *chunks* permet d'extraire des groupes nominaux ainsi que des séquences de groupes nominaux. Par exemple dans la séquence « appliquer le travail fait à la *simplification* lexicale de textes médicaux », nous extrayons trois niveaux de KW autour du terme « simplification » : [`simplification`], [`simplification lexicale`] (contexte immédiat) et [`simplification lexicale de textes médicaux`] (contexte étendu). Le contexte étendu permet d'obtenir des expressions sémantiquement riches, au détriment de la significativité statistique. En guise de compromis, nous avons adopté une représentation imbriquée qui conserve les trois niveaux, laissant à l'utilisateur le soin de choisir l'étendue du contexte qu'il souhaite observer. La pertinence des KW extraits est calculée à l'aide du coefficient TF-IDF$_{BM25}$.

3 Principales fonctionnalités de l'interface

L'**accès aux documents** se fait directement via l'arborescence des documents ou via un moteur de recherche. L'ensemble des métadonnées (éditoriales et automatiques) sont indexées. Une stratégie de complétion a été mise en place sur la base des KW à contexte variable. Ainsi la saisie « `similari` » dans la collection TALN'2015 propose la complétion ci-contre.

Plusieurs modalités de **navigation** sont proposées. L'une des originalités de cette interface est son *player* circulaire. La roue est segmentée par défaut selon la segmentation en locuteurs, ce qui permet de repérer facilement par exemple les questions finales. Le panel de gauche fournit la liste des KW triés par pertinence dans le document. La sélection d'un KW provoque l'ouverture d'un volet donnant le contexte de ses occurrences (transcription du groupe de souffle associé), et provoque également l'apparition de picots qui permettent de les visualiser dans la roue. La sélection d'une occurrence lance l'écoute au début du groupe de souffle associé, ce qui permet d'avoir un feedback sonore pertinent. D'autres fonctionnalités non détaillées dans cette description sont disponibles, comme la possibilité d'avoir accès à plusieurs roues de segmentation (comme par exemple la segmentation thématique (Bouchekif et al., 2015) ou une segmentation vidéo) ou la navigation par segment.



Références

VIAUD M.L., BUISSON O., SAULNIER A., GUENAIS C.,. Video exploration: from multimedia content analysis to interactive visualization, Actes de *ACM Multimedia*, 2010.

GAUVAIN J. L., LAMEL L., ADDA G., The LIMSI Broadcast News Transcription System, *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

CHARLET D., BARRAS C., LIENARD J-S., Impact of overlapping speech detection on speaker diarization for broadcast news and debates, Actes de *ICASSP 2013*, Vancouver, Canada, 2013.

BOUCHEKIF A., DAMNATI G., ESTÈVE Y., CHARLET D., CAMELIN N., Diachronic Semantic Cohesion for Topic Segmentation of TV Broadcast News, Actes de *INTERSPEECH 2015*, Allemagne, 2015.