RECITAL 2001, Tours, 2-5 juillet 2001

Extraction d'information de documents textuels associés à des contenus audiovisuels

Estelle Le Roux
INA
4 avenue de l'Europe, 94366 Bry sur Marne Cedex France
eleroux@ina.fr
LIMSI
Bt 508, Université Paris-Sud, 91403 Orsay Cedex
Estelle.Le.Roux@limsi.fr

Résumé - Abstract

L'indexation audiovisuelle, indispensable pour l'archivage et l'exploitation des documents, se révèle être un processus délicat, notamment à cause de la multiplicité de significations qui peuvent être attachées aux images. Nous proposons dans cette communication une méthode d'instanciation de "patrons d'indexation" à partir d'un corpus d'articles de journaux écrits. Cette méthode repose sur un processus "d'amorçage hiérachisé", qui permet de trouver de nouveaux termes à partir de termes connus dans leur voisinage et de leurs relations taxinomiques sous forme d'ontologie.

Mots-clés Amorce - Extraction d'information - Ontologie - Patron d'indexation

Audiovisual indexation, essential for filing and using documents, is a difficult process notably because of the multiplicity of meanings which can be associated to the pictures. We propose a method of instanciation of "indexation patterns" from a corpus of articles from newspapers. This method is based on a "hierarchical bootstrapping" which can find new terms from known terms in their neighbourhood and from an ontology.

Key words Bootstrapping - Indexation patterns - Information extraction - Ontology

1 L'indexation audiovisuelle

1.1 L'INA : les documents audiovisuels et leur exploitation

L'INA [1] (*Institut National de l'Audiovisuel*) a pour vocation de constituer le patrimoine audiovisuel en sauvegardant, numérisant et restaurant, en un mot en *archivant*, les émissions de radio

[1] http://www.ina.fr

et de télévision publiques. L'institut a, actuellement, en sa possession plus d'un million et demi d'heures de radio et de télévision, auxquels viennent s'ajouter plus d'un million de documents photographiques. 70 000 heures de programmes par an sont documentées, conservées et recopiées. Les fonds d'archives de l'INA sont donc une ressource importante, pour tous les professionnels de l'audiovisuel mais aussi pour les chercheurs, enseignants et étudiants.

Pour que tous ces documents audiosivuels soient exploitables, nous devons les indexer. Or indexer signifie interpréter mais il va sans dire que l'interprétation de documents est une opération coûteuse, aussi bien en temps qu'en personnes. Un des moyens qui s'offre à nous pour pouvoir réduire ce coût est de pouvoir instrumenter l'interprétation le plus possible. A ce niveau, nous nous trouvons confronter au fait que l'image ne possède pas d'interprétation particulière (Metz, 1968). Comment, alors, arriver à interpréter, à indexer un document audiovisuel ?

Dans cette communication, nous proposons, après avoir restreint le domaine télévisuel aux journaux et aux magazines, d'extraire des informations dans des documents textuels en relation avec les documents audiovisuels à indexer : des articles de presse écrite qui relatent le même événement. Ainsi, par exemple, les journaux nous ont raconté l'histoire d'un enfant retrouvé vivant plusieurs jours après le tremblement de terre qui a eu lieu en Inde au début de l'année ; la télévision, quant à elle, a montré les images du sauvetage du jeune rescapé. La partie qui permettra de relier les informations textuelles aux documents audiovisuels sera faite par le département de la recherche à l'INA.

Nous présenterons dans un premier temps les problèmes intervenant dans l'indexation audiovisuelle, puis nous préciserons pourquoi les articles de presse écrite semblent pertinents pour l'interprétation des documents télévisuels. Nous parlerons ensuite d'un système basé sur une stratégie "d'amorçage hiérarchisé" qui permet d'extraire de nouveaux termes dans les articles à partir "d'amorces" déjà connues et organisées sous forme taxinomique dans une ontologie. Nous discuterons enfin des différentes perspectives et conclusions que soulèvent cette recherche.

1.2 L'audiovisuel et l'écrit

1.2.1 La problématique de l'indexation audiovisuelle

Indexer un document consiste à lui attribuer des *descripteurs*[2]. L'un des principaux enjeux de l'indexation est donc de disposer d'index permettant d'une part de refléter adéquatement le contenu et d'autre part de se prêter à une manipulation aisée. Or il est difficile d'obtenir un système possédant en même temps ces deux aspects.

De plus, pour l'indexation audiovisuelle, les unités[3] de localisation permettant de définir la partie du document qui sera associée à un index n'est pas simple à trouver car il n'existe pas d'éléments tels que les blancs dans un texte écrit.

Enfin, il est également nécessaire d'obtenir des unités de caractérisation pour interpréter et proposer un index permettant de reformuler le contenu d'une partie du document mais cette étape pose aussi des problèmes car une même image peut avoir différentes significations.

[2] Un *descripteur* est une forme symbolique qui permet de caractériser un document lors de l'indexation (Auffret, 2000).

[3] Par *unité*, nous entendons *objet signifiant repéré*.

Voyant que l'interprétation des documents audiovisuels à partir de ces mêmes documents ne semble pas pouvoir se faire à l'aide d'une technique aisée et peu coûteuse, nous devons trouver un moyen d'obtenir une interprétation explicite. La piste alors envisagée est de partir de textes écrits produits par la presse.

1.2.2 Des journaux nationaux comme aide à l'indexation des documents audiovisuels

Jusqu'à aujourd'hui, le document de base à l'INA pour pouvoir exploiter les documents audiovisuels est un document écrit, appelé *notice*, créé par les documentalistes. Les notices contiennent différents champs tels que le type, l'auteur, le résumé d'une émission indexée. Sur ce document, il est possible de faire des recherches en texte plein. Cela suppose alors d'avoir des mots orthographiés correctement, notamment les noms propres. Si nous reprenons l'exemple du tremblement de terre, les documentalistes peuvent avoir des problèmes pour écrire le nom de certaines villes tel que celui de la ville de Buhj. La presse écrite leur vient donc en aide pour vérifier rapidement l'orthographe d'un mot.

Le deuxième intérêt, l'un des plus importants, pour utiliser les journaux réside dans le fait que les documentalistes peuvent se faire une meilleure idée des usages futurs. Il ne faut pas, en effet, oublier que l'indexation des documents n'est utile que si les documentalistes peuvent indexer les informations qui seront pertinentes dans un futur plus ou moins proche.

Enfin, il est important de rappeler que l'utilisation des journaux dans ce cadre est motivée par le fait que les documents audiovisuels montrent un événement tandis que les journaux écrits décrivent un événement. Il est donc intéressant de voir quel genre de lien existe entre ces deux media. Nous trouvons dans le quotidien *Le Monde* du 26 juillet 1999 les phrases suivantes : *Les Français n'ont encore emporté aucune victoire dans ce Tour de France. Et se profile le spectre de 1926, dernière édition où la nation hôte fut déclarée fanny*. Si nous lisons la notice correspondant au journal télévisé de 13 heures du 26 juillet 1999, nous trouvons dans le champ *Résumé*, la phrase suivante : *Pour la première fois depuis 1926, les coureurs français ont terminé le Tour de France sans la moindre victoire d'étape*. Nous pouvons voir clairement le lien qui existe : ces deux media font référence au fait qu'aucun cycliste français n'a remporté une étape lors du Tour de France 1999, tout comme en 1926. Partant de là nous pouvons extraire des informations dans la presse écrite, informations qui auront des chances d'intervenir dans le cadre de l'indexation des documents audiovisuels. Ces informations textuelles seront alors utilisées comme des métadonnées.

2 L'instanciation des patrons du Tour de France

2.1 Des informations intéressantes issues du Tour de France

Notre travail consiste à venir en aide aux documentalistes en leur proposant des informations, extraites des journaux écrits à l'aide de patrons d'indexation, qu'ils seront susceptibles de retrouver dans les documents audiovisuels.

Nous avons constitué un corpus journalistique issu du *Monde*, du *Parisien*, de *Libération*, de *L'Equipe* et de l'*AFP* et ayant pour thème le Tour de France cycliste 1999. La figure 1 nous

montre un extrait de notre corpus [4].

Le Monde,

7 juillet 1999, page 20

TOUR DE FRANCE 1999 Sur la route de Saint-Nazaire, Casino a raflé la mise Une chute collective lors de la deuxième étape a permis au peloton de distancer certains prétendants au maillot jaune . Alex Zulle (Banesto), Ivan Gotti (Polti), Michael Boorgerd (Rabobank) . mais pas de décourager le robuste Estonien Jaan Kirsipuu (Casino), nouveau leader de la course

BORDENAVE YVES

TOUR DE FRANCE 1999 L'Estonien Jaan Kirsipuu (Casino) a revêtu, lundi 5 juillet, pour la première fois le maillot jaune à l'issue de la deuxième étape Challans-Saint-Nazaire (176 km), gagnée au sprint par le Belge Tom Steels. CASINO, qui ne veut plus associer son nom au cyclisme, va se désengager à la fin de l'année. PLUSIEURS FAVORIS, parmi lesquels l'Italien Ivan Gotti (Polti), le Suisse Alex Zulle (Banesto) et le Néerlandais Michael Boogerd (Rabobank) ont été pris dans une chute collective qui a provoqué la cassure du peloton au passage du Gois. L'EQUIPE ONCE du controversé directeur sportif Manolo Saiz a, contrairement aux usages en début d'épreuve, roulé à fond pour creuser un écart de six minutes avec la centaine d'attardés. (...)
Maillot jaune, lundi 5 juin, à Saint-Nazaire (Loire-Atlantique), victorieux à l'étape de Challans (Vendée) la veille, Jaan Kirsipuu est un coureur comblé. (...)
Profitant du jeu des bonifications glanées sur les 176 km de cette deuxième étape disputée entre deux averses de Challans à Saint- Nazaire, remporté au sprint par le Belge Tom Steels (Mapei), il a ravi la première place à l'Américain Lance Amstrong (US Postal), qui le suit désormais à 14 secondes. (...)
YVES BORDENAVE

FICHE DOCUMENTAIRE

Titre complémentaire: 2E ETAPE CHALLANS - SAINT-NAZAIRE, 5 JUILLET 1999; DANS ENSEMBLE DE 2 PAGES Sujets - France: 1999; CLUB SPORTIF; COMPETITION SPORTIVE; CYCLISME; DOPING; SPORTIF Sujets - International: CYCLISME Noms propres: CASINO; KIRSIPUU JAAN; LAVENU VINCENT Taille: MOYEN 990707LM654460

Figure 1: Extrait du journal *Le Monde*

Nous voyons dans cet exemple que de nombreuses informations sont pertinentes dans le cadre de l'indexation des documents : nous apprenons quelle équipe a gagné l'étape du 07 juillet 1999

[4]La partie se trouvant dans la fiche documentaire renvoie aux différents champs utilisés par les journalistes du *Monde*.

(*7 juillet 1999 ; Sur la route de Saint-Nazaire, Casino a raflé la mise*), qui était le vainqueur et qui était le maillot jaune le 05 juillet 1999 (*L'Estonien Jaan Kirsipuu (Casino) a revêtu, lundi 5 juillet, pour la première fois le maillot jaune [...]*). Dans cet article, nous apprenons également qu'il y a eu une chute au passage du Gois (*Plusieurs favoris [...] ont été pris dans une chute collective qui a provoqué la cassure du peloton au passage du Gois.*).

Ces informations importantes vont être extraites en utilisant des patrons d'indexation. Par *patrons d'indexation* nous entendons des structures génériques définissant un niveau et un type de description sur les objets montrés, les paroles entendues, les concepts évoqués et ils seront liés à une ontologie.

Pour instancier ces patrons, nous allons faire une analyse textuelle sur les articles à l'aide d'un système d'extraction possédant des amorces et devant être hiérarchisé. Nous allons également coupler ce système à une analyse syntaxique afin d'obtenir de meilleurs résultats.

2.2 Une système d'armorçage hiérarchisé

Les amorces Dans le but d'instancier des patrons d'indexation, il est nécessaire dans un premier temps, d'établir des liens sémantiques entre les différents termes de notre corpus. Pour ce faire, nous allons créer un système utilisant des *amorces*, système qui servira de dictionnaire sémantique. Les amorces sont des termes appartenant à une catégorie définie qui vont nous permettre d'identifier d'autres termes appartenant à cette même catégorie. Comme nous possédons un corpus relativement cohérent et homogène, nous devrions trouver des amorces qui nous permettront de définir des catégories telles que EQUIPE, NATIONALITE. Une fois les différentes amorces et catégories définies, nous passerons notre système sur notre corpus pour qu'il puisse extraire toutes les phrases possédant une amorce. Si, par exemple, nous avons une catégorie EQUIPE, comportant les amorces suivantes : *Casino, Polti, US Postal, Rabobank*, notre système sera alors en mesure de pouvoir extraire la phrase suivante de l'article donné en 3.1 : *Alex Zulle (Banesto), Ivan Gotti (Polti), Michael Boorgerd (Rabobank)*. Nous pourrons ainsi définir quel cycliste appartient à quelle équipe, élément intéressant pour toutes les recherches concernant les équipes. Nous reprenons ainsi la démarche de E. Riloff qui a développé les systèmes AutoSlog puis AutoSlog-TS (Riloff, 1996), (Riloff, Lorenzen, 1999). Leur système s'applique bien à tout ce qui concerne les listes, les appositions mais nous ne sommes pas sûrs de pouvoir appliquer cette méthode à notre corpus. Nous pensons que nous avons intérêt à faire intervenir plus de syntaxe dans le but d'éviter de mauvaises appartenances catégorielles par exemple. Sur ce dernier point, nous rejoignons Roark et Charniak (Roark, Charniak, 1998).

En n'extrayant que des phrases contenant des amorces, nous allons ainsi éviter de passer les patrons sur l'ensemble du corpus. Enfin, le fait de définir des catégories, avec les amorces, va nous servir dans le processus de hiérarchisation.

Un système hiérarchisé De simples amorces ne suffisent effectivement pas pour pouvoir instancier des patrons d'indexation corrects. Si nous définissons, par exemple, de manière grossière, le patron d'indexation suivant : CYCLISTE – VAINQUEUR, avec comme catégorie CYCLISTE comprenant le nom des différents coureurs du Tour de France tel que *Armstrong, Virenque, Jalabert* et la catégorie VAINQUEUR comportant les amorces *remporter, gagner, rafler*, nous n'allons pas pouvoir extraire la phrase *Sur la route de Saint-Nazaire, Casino a raflé la mise*. Il n'y aura pas, en effet, de correspondance entre l'amorce et le terme *Casino*. Si,

en revanche, nous avons un système hiérarchisé tel que Cycliste EST-UN Coureur, Coureur APPARTIENT-A une Equipe, nous pourrons alors établir des correspondances entre Casino et Coureur et par la même à Cycliste. Nous envisageons donc d'utiliser une ontologie qui va nous permettre d'avoir une représentation structurée des différentes catégories sémantiques utilisables dans le cadre de l'extraction d'information.

3 Conclusions

L'indexation audiovisuelle est un processus complexe et coûteux car il est difficile de trouver des unités représentant correctement le contenu : les images possèdent trop de significations et il est difficile de pouvoir localiser une unité significative. Pour faciliter l'indexation audiovisuelle, nous proposons d'extraire des informations pertinentes d'articles issus de la presse écrite portant sur les mêmes événements que les images à indexer, à l'aide de patrons d'indexation et d'une ontologie.

Références

Auffret G. (2000), *Structuration de documents audiovisuels et publication électronique - Constitution d'une chaine éditoriale numérique pour la mise en ligne de collections audiovisuelles*, Thèse de Doctorat, Université de Technologie de Compiègne.

Habert B., Fabre C. (1999), Elementary Dependency Trees for Identifying Corpus-specific Semantic Classes, *Computers and the Humanities*, Vol. 33, n°3, 207-219.

Metz C. (1968), *Essais sur la signification au cinéma*, Paris, Klincksieck.

Riloff E. (1996), Using Learned Extraction Patterns for Text Classification, Connectionist. In Wermter S., Riloff E., Scheler G. (eds.), *Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Berlin, Springer-Verlag, 75-289.

Riloff E., Lorenzen J. (1999), Extraction-based Text Categorisation: Generating Domain-specific Role Relationships Automaticaly. In Strzalkowki (ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, 167-196, .

Riloff E., Shepherd J. (1997), A Corpus-Based Approach for Building Semantic Lexicons, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Brown University, Providence, Rhode Island, USA 127-132.

Roark B., Charniak, E. (1998), Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *the 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Canada, 1110-1116.