Recherche d'information textuelle et phonétique pour le contrôle de l'étiquetage automatique d'émissions dans un flux télévisuel

Camille Guinaudeau

IRISA/INRIA
Campus de Beaulieu
35042 RENNES Cedex
France
Camille.Guinaudeau@irisa.fr

RÉSUMÉ. *En 2007, Naturel (Naturel, 2007) a proposé un système qui associe automatiquement une étiquette, c'est-à-dire un titre, à des émissions issues du découpage d'un flux TV. Cependant, ce système ne permet pas de vérifier la correction des associations étiquette-émission. Nous proposons dans cet article de contrôler cet étiquetage en nous basant sur les transcriptions textuelle et phonétique de la bande sonore contenue dans le flux. Nous montrons que des méthodes de recherche d'information permettent d'associer à chaque émission une description, issue d'un guide de programmes TV, description qui est ensuite comparée avec l'étiquette originale de l'émission. La technique proposée permet de contrôler un peu plus de 45% des émissions étudiées et de diminuer de nombre d'erreurs de l'étiquetage original de 3,5%.*

ABSTRACT. *In 2007, Naturel (Naturel, 2007) developed a method which, given a segmented video stream, associated a label with each segment. However, this method did not automatically check the accuracy of the results obtained. In this paper we propose to control these results, by taking each segment, and associating the corresponding phonetic or textual transcription of the soundtrack with descriptions extracted from a TV guide. Using techniques inspired from information retrieval methods, a description is linked to each segment, which can then be compared with the label associated by Naturel's method. This new method allows us to make a decision for 45% of the segments, and to lower the original labeling error rate by 3.5%.*

MOTS-CLÉS : *transcription automatique de la parole, recherche d'information textuelle, recherche d'information phonétique, multimédia, étiquetage des segments de flux TV*

KEYWORDS: *automatic speech recognition, textual information retrieval, phonetic information retrieval, multimedia, TV stream segments labeling*

Camille Guinaudeau

1. Introduction

Le nombre toujours grandissant de collections de documents télévisuels disponibles (et de documents multimédia en général) pose désormais le problème de la navigation dans ces flux, de leur interrogation, *etc.*, ce qui nécessite un accès à leur contenu sémantique, pouvant impliquer, dans un premier temps, leur structuration. Une des étapes de structuration consiste à découper un flux TV en émissions successives. Un tel flux est composé de deux types de segments : les inter-programmes (publicités, *jingles*, bandes-annonces), également notés IP, et les programmes (émissions de variétés, reportages, films, *etc.*). En 2007, Xavier Naturel (Naturel, 2007) a proposé une méthode permettant de découper automatiquement en programmes et inter-programmes les flux TV correspondant à plusieurs semaines d'enregistrement, et d'associer à chacun des segments obtenus une étiquette, c'est-à-dire une information sémantique qui se limite, dans ce travail, au titre de l'émission. Les IP étant généralement des programmes répétés, l'une des idées directrices de (Naturel, 2007) est d'utiliser ces répétitions pour réaliser la distinction entre les deux types de programmes. L'association d'une étiquette à chaque segment est ensuite faite grâce à un alignement entre un guide de programmes électronique, de type Télé Magazine –contenant les titres des émissions accompagnées de leurs heures de diffusion –et la segmentation obtenue. Cet appariement est mis en place par un algorithme d'alignement dynamique temporel (*Dynamic Time Warping* –DTW) qui calcule la distance entre la segmentation et le guide en prenant en compte la similarité de la longueur des segments ainsi que celle des horaires de diffusion. Cette tâche est complexe car un guide de programmes n'est pas complet. En effet, les IP n'y sont pas mentionnés, ainsi que certains programmes tels que la météo. De plus, des retards et des modifications de dernières minutes opérées par les chaînes par rapport au guide de programmes compliquent la tâche d'étiquetage. Les résultats de la segmentation sont globalement bons, mais, en ce qui concerne l'étiquetage, le processus associe parfois à plusieurs segments consécutifs une même étiquette. En étudiant ces cas de plus près, on constate qu'il ne s'agit pas d'un problème de sur-segmentation (où un programme serait découpé en plusieurs segments) mais bien d'un problème d'étiquetage. De plus, le processus d'étiquetage ne peut pas gérer les changements dans la grille de diffusion car il ne considère pas les informations sémantiques contenues dans les émissions mais uniquement leurs horaires de diffusion. Il est donc nécessaire de vérifier que les étiquettes liées aux segments correspondent aux programmes effectivement diffusés.

Dans cet article, nous proposons une méthodologie permettant de contrôler automatiquement, voire d'améliorer, l'étiquetage des segments proposé par (Naturel, 2007). Lors de notre contrôle, nous prenons en compte les informations de sens portées par les segments d'émissions. En effet, nous cherchons à caractériser le contenu des segments en travaillant sur les transcriptions –textuelles et phonétiques –de la bande sonore de ces segments, les paroles prononcées dans les programmes étant fortement représentatives de ce qu'ils renferment. Une telle caractérisation peut être mise en oeuvre de différentes façons. Certaines études, s'appuyant sur le fait qu'on utilise un certain vocabulaire pour parler d'un sujet puis que l'on change d'unités lexicales pour passer à un autre thème, se fondent sur la notion de cohésion lexicale. Par exemple, dans (Ferret *et al.*, 2001), les auteurs proposent un module qui extrait des signatures thématiques à partir de segments thématiquement homogènes obtenus préalablement. Ces signatures sont constituées de l'ensemble des mots maintenant élevé le niveau de cohésion lexicale et correspondent de ce fait à une représentation des thèmes abordés dans les textes. D'autres travaux mettent en place des méthodes issues de la recherche

d'information (RI). Dans (Lecorvé *et al.*, 2008) par exemple, les auteurs caractérisent les documents grâce aux mots dont le poids *tf * idf* est le plus important. Ce sont des méthodes similaires à ces derniers travaux que nous allons utiliser dans cet article.

La méthodologie proposée ici consiste à associer automatiquement à la transcription des segments résultant du travail présenté dans (Naturel, 2007) une description textuelle extraite d'un guide télévisuel grâce à des méthodes de RI. Le guide de programmes utilisé contient exactement les mêmes émissions que celui ayant servi lors de l'étape d'étiquetage dans le travail de (Naturel, 2007) ; il est cependant plus complet dans la mesure où chacun des programmes est associé à une description composée du titre de l'émission et, sauf exception, d'un résumé de son contenu qui peut aller jusqu'à 250 mots pour les plus longs. La description liée grâce à notre technique à chaque segment transcrit est ensuite comparée à l'étiquette fournie par Naturel afin de décider si cet étiquetage semble correct ou non. Lier transcription et description est toutefois difficile. En effet, les descriptions des programmes sont parfois très courtes et peu informatives (si limitées au seul titre de programme ou à un très bref résumé) et les transcriptions de la bande sonore des segments sont parfois très éloignées de ce qui a été prononcé dans la réalité. Le but de cet article est tout d'abord de montrer que des techniques de RI sont utilisables dans le contexte de la transcription de la télévision ; en effet, il n'existe pas, à notre connaissance, de travaux appliquant ces méthodes sur un tel matériau. Nous démontrons également qu'en adaptant des techniques de RI, il est possible de travailler sur du texte dégradé –les transcriptions de certains programmes télévisuels ayant un taux d'erreurs pouvant aller jusqu'à 80% –tout en obtenant des résultats encourageants. Nous réussissons en effet à contrôler l'étiquetage de près de la moitié des segments.

Nous décrivons dans un premier temps la méthode que nous avons mise en place pour contrôler l'étiquetage proposé par Naturel. Nous présentons ensuite les premiers résultats, obtenus sur les deux journées-test des 10 et 11 mai 2005, avant de conclure sur les travaux à mener afin de les améliorer.

2. Méthodologie

Le principe de la méthode que nous proposons pour contrôler l'étiquetage se décompose en deux étapes principales (*cf figure* 1), étapes présentées dans la suite de cette section. La première consiste à attacher à chaque segment reconnu dans le flux TV par Naturel une description issue du guide de programmes. L'association segment/description se base sur la similarité entre le contenu du segment, obtenu grâce aux transcriptions textuelle et phonétique de sa bande sonore, et celui de la description, et prend donc en compte des informations lexicales et sémantiques. Dans la seconde étape, la description associée par notre méthode est comparée, pour chacun des segments, avec l'étiquette proposée par (Naturel, 2007) afin de valider ou non cette dernière et éventuellement de la remplacer.

2.1. *Association segment/description*

L'association entre les segments de programmes et les descriptions est mise en place grâce à deux méthodes de recherche d'information appliquées respectivement sur les transcriptions textuelles et phonétiques de la bande sonore contenues dans les segments. Les résultats de ces deux recherches sont combinés *a posteriori*.

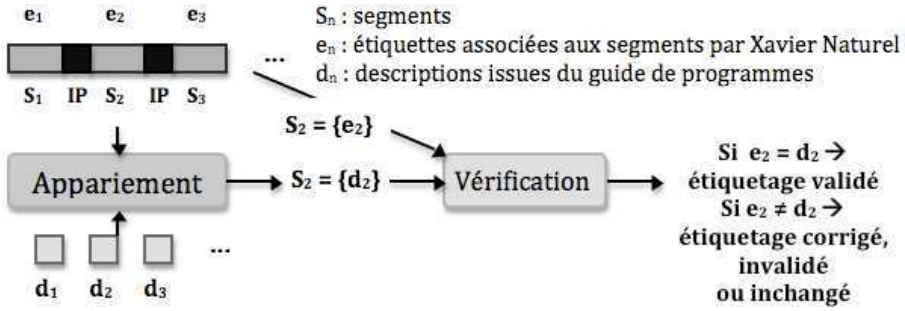

Figure 1. *Architecture globale de la méthode de contrôle de l'étiquetage*

2.1.1. *Recherche textuelle*

Pour lier descriptions et transcriptions textuelles des segments, ces deux types d'informations sont représentés, après lemmatisation, par des vecteurs de mots pondérés par un score *tf * idf*. La fréquence documentaire inverse est calculée à partir d'articles journalistiques extraits du journal *Le Monde* entre 1987 et 2003, soit 800 000 articles[1]. Une similarité entre les vecteurs représentant les descriptions et ceux représentant les transcriptions est calculée grâce à une mesure angulaire de type cosinus. La valeur de cette mesure est ensuite utilisée pour calculer, pour chaque couple segment transcrit/description, un score final qui prend en compte d'autres éléments que nous décrivons dans la suite de ce paragraphe. Notre méthode présente deux originalités : la première est la prise en compte, dans le calcul du poids *tf * idf* des mots contenus dans la transcription, de l'indice de confiance qui leur est associé. À chaque mot qu'il produit, le système de transcription automatique de la parole joint, en effet, une valeur qui traduit la confiance qu'il porte en sa transcription. Ainsi, dans notre technique, les mots ayant un indice de confiance faible sont discrédités par le biais d'un système de malus appliqué à leur poids suivant l'idée de (Lecorvé *et al.*, 2008). La seconde originalité est que l'ensemble des transcriptions des programmes est tour à tour considéré comme un ensemble de documents et un ensemble de requêtes. De la même manière, les descriptions de programmes sont vues soit comme des documents, soit comme des requêtes. Nous obtenons ainsi pour un segment transcrit s la liste classée, par ordre décroissant de similarité, de toutes les descriptions possibles, et pour une description d la liste classée, par ordre décroissant de similarité, de tous les segments possibles. Ces deux « passes » se justifie dans le calcul du score final: nous souhaitons, en effet, que pour chaque couple segment/description, la description soit celle qui corresponde le mieux au segment et *vice versa* ; or le score associé aux couples s/d et d/s n'est pas symétrique car il prend en compte la valeur du cosinus mais également la position du couple dans la liste résultat et la différence entre le score de ce couple et celui du couple suivant (*cf.* (Guinaudeau, 2008) pour plus de détails). Pour finir pour chaque segment s, l'association retenue –c'est-à-dire celle pour laquelle la description d cor-

1. Bien que le vocabulaire de ce corpus soit parfois assez éloigné de celui prononcé dans les émissions télévisées –ce qui influence sans doute la valeur des scores *tf * idf* –le choix de son utilisation s'explique par le fait qu'il a été employé lors de l'apprentissage du modèle de langue du système de transcription. Les mots qu'il contient sont donc reconnus par le système.

respond le mieux au programme contenu dans le segment –est celle pour laquelle le score final, égal à la somme des scores des couples s/d et d/s, est le plus élevé.

2.1.2. *Recherche phonétique*

Cependant, comme le montrent (Cardillo *et al.*, 2002, Logan *et al.*, 2002), appliquer une recherche textuelle seule pour explorer des transcriptions de documents sonores se heurte au problème des mots hors vocabulaire. En effet, les systèmes de reconnaissance automatique de la parole utilisent, en règle générale, pour produire des transcriptions, un dictionnaire phonétique qui permet d'associer des mots aux sons entendus. Ils ne peuvent donc pas retourner des termes n'appartenant pas au dictionnaire. Or ces mots dits hors vocabulaire peuvent posséder un fort contenu sémantique. C'est en particulier le cas des noms propres, souvent absents des dictionnaires. Nous choisissons donc de traiter différemment les descriptions issues du guide télévisuel contenant des noms propres, repérés grâce à l'outil LIA_PHON (Béchet, 2001), de celles n'en possédant pas, qui ne subissent, elles, que le seul dispositif explicité ci-dessus.

Pour chaque description d'émissions contenant des noms propres, nous mettons tout d'abord en place une recherche phonétique permettant de chercher à repérer, à l'intérieur de la bande sonore d'un segment, les noms propres présents dans cette description. Le système de reconnaissance de la parole produit, dans son processus de transcription automatique de la bande sonore, une transcription phonétique qui traduit sous forme de phonèmes les sons prononcés dans l'émission. Les noms propres, quant à eux, sont phonétisés à partir des descriptions grâce à l'outil LIA_PHON qui transforme chaque mot de la description en suite de phonèmes par un système de règles. La recherche phonétique permet de retrouver une petite suite de phonèmes –ici un nom propre –dans une séquence de phonèmes plus grande –la transcription phonétique de la bande sonore contenue dans le segment. Nous ne cherchons toutefois pas à repérer dans la transcription phonétique exactement la suite de phonèmes qui constitue le nom propre. En effet, les noms propres n'apparaissant pas dans le dictionnaire du système de reconnaissance de la parole, ils font l'objet d'erreurs de transcription. De plus, la transcription phonétique de la bande sonore est perturbée par les marques d'hésitation ainsi que les accents des locuteurs, ce qui modifie la prononciation des noms propres recherchés. Notre but est donc de rechercher dans la transcription phonétique la séquence de phonèmes qui minimise le plus la distance avec la phonétisation du nom propre. Cette distance est calculée grâce à une adaptation de la distance d'édition qui autorise le nom propre phonétisé à être situé n'importe où dans la transcription phonétique (*cf.* (Muscariello *et al.*, 2009) pour plus de détails sur la distance d'édition utilisée). La méthodologie que nous proposons est la suivante : pour chacun des noms propres contenus dans une description, nous récupérons, pour tous les segments de programmes, le coût minimal engendré par la transformation de la chaîne de phonèmes correspondant au nom propre phonétisé en une partie de la chaîne de phonèmes correspondant à la transcription phonétique de la bande sonore du segment. Ce score minimal est ensuite additionné avec les scores obtenus pour chacun des noms propres de la description. Cependant, les descriptions qui contiennent des noms propres possèdent également de nombreux autres mots tout aussi porteurs de sens. C'est pourquoi nous appliquons aussi sur ces descriptions, parallèlement à la recherche phonétique, une recherche textuelle identique à celle employée pour les descriptions ne contenant pas de noms propres.

2.1.3. *Combinaison des deux types de recherche*

Finalement, pour chacun des couples segment/description, les scores de ces deux recherches sont combinés *a posteriori* –celui issu de la recherche phonétique et additionné à celui obtenu par les deux « passes » de la recherche textuelle multiplié par un facteur 200 –et les paires segment/description retenues sont celles dont le score global est le plus élevé.

2.2. *Vérification de l'étiquetage de Xavier Naturel*

La phase d'appariement expliquée ci-dessus, appliquée à un corpus de deux jours d'enregistrement de télévision, nous permet d'associer une description à 60 segments sur 133. Nous souhaitons, dans cette seconde étape, contrôler automatiquement la qualité de l'étiquetage proposé dans (Naturel, 2007). Pour cela, nous comparons pour chaque segment du flux télévisé l'étiquette fournie par Naturel avec la description liée à ce segment par notre technique. Si le titre contenu dans notre description et l'étiquette correspondent, l'étiquetage est considéré comme correct. Si, au contraire, la description que nous avons attachée à un segment est différente de l'étiquette de Naturel, nous comparons alors les horaires de début de diffusion correspondant aux deux programmes désignés respectivement par l'étiquette et la description avec l'heure du début du segment. Si l'heure de début du programme désigné par l'étiquette est la plus proche de celle de début de diffusion du segment, l'étiquetage est considéré correct, sinon il est dit faux et on remplace l'étiquette par notre description. Cependant, si la différence entre l'heure de début du programme désigné par l'étiquette ou par la description et l'heure de début de diffusion du segment est supérieure à une demi-heure, on considère que l'étiquetage est faux sans proposer de nouvelle étiquette.

3. Résultats

Les résultats que nous fournissons ici concernent globalement tant la méthode de recherche textuelle « pure » que celle « hybride », combinant recherche phonétique et textuelle pour les descriptions contenant des noms propres. Notre technique de vérification nous permet de contrôler automatiquement l'étiquetage d'un peu plus de 45% des segments repérés par Naturel dans les deux journées-test. Ce pourcentage peut s'expliquer par le fait qu'environ 30% des segments de notre corpus contiennent des programmes –tels que la météo ou les programmes interstitiels [2] –qui ne possèdent pas de description dans le guide des programmes. Nos résultats se répartissent comme présentés dans la figure 2.

Si l'étiquetage de seulement 45% des segments a pu être contrôlé par notre méthode, ces résultats sont tout de même prometteurs. En effet, des expériences successives nous ont montré que les paires segment/description fournies par la combinaison étaient un peu meilleures que celles fournies par une recherche textuelle seule. D'une part, le nombre de couples pertinents retournés augmente légèrement (plus 2 pour les deux journées-test). D'autre part, le nombre de fausses alarmes, c'est-à-dire d'asso-

2. Les programmes interstitiels sont des programmes d'une minute environ tels que « Un jour, un arbre » ou « Conso mag ».

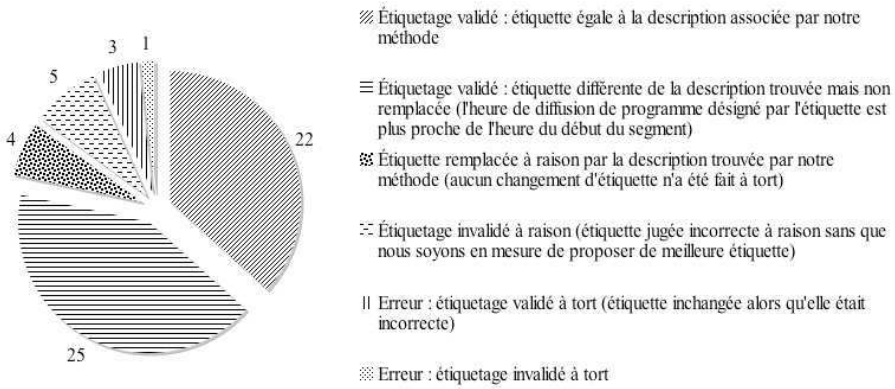

Figure 2. *Résultats de la vérification des étiquettes fournies par Xavier Naturel*

ciations considérées à tort comme pertinentes, par rapport au nombre d'associations pertinentes trouvées diminue pour les deux journées. Cependant ces résultats sont encore perfectibles. Nous étudions dans la partie suivante les différentes techniques que nous pourrions mettre en place afin d'accroître les performances obtenues.

4. Discussions et perspectives

La méthodologie de contrôle que nous avons proposée et qui s'avère conduire à des résultats encourageants repose sur une combinaison originale de techniques de RI textuelle et phonétique, et tire d'ailleurs profit de cet assemblage. Il reste cependant à étendre la couverture de notre technique à tous les programmes puisqu'elle ne contrôle actuellement qu'environ la moitié des segments produits par (Naturel, 2007). Pour ce faire, diverses pistes sont envisageables.

Une première perspective, à court terme, consiste à faire fonctionner notre méthode sur un corpus plus étendu. En effet, nous constatons de grandes disparités au niveau des résultats sur les deux journées des 10 et 11 mai 2005. Travailler sur un corpus plus important nous permettrait d'avoir une vision plus complète des performances de notre système et de traiter un ensemble plus large de types d'émissions. En plus du corpus de mai 2005 constitué de trois semaines d'enregistrement segmentées et étiquetées, un nouveau corpus de 6 mois d'enregistrements en continu des chaînes TF1 et France 2 vient d'être construit à l'IRISA.

Nous devons également travailler sur une catégorisation automatique des programmes télévisuels. Nous avons, dans cet article, proposé deux modes de recherche différents en fonction de la présence ou non de noms propres dans les descriptions mais nous n'avons pas pris en compte les disparités au niveau des émissions. Or, l'utilisation de méthodes différentes, adaptées à chacun des types d'émissions, nous permettrait sans doute de contrôler l'étiquetage d'un plus grand nombre de segments.

Certains travaux, (Fischer *et al.*, 1995) par exemple, proposent des techniques permettant de décider de la classe d'un programme – météo, journal télévisé ou sport

–en se basant sur la longueur des scènes, la couleur dominante du fond d'écran, la pureté des silences, *etc*. En utilisant des indices vidéo et audio, nous pourrions ainsi améliorer les résultats du système en utilisant, pour chaque programme, une méthode plus adéquate.

Enfin, un dernier élément qui explique peut être partiellement nos résultats est que nous supposons dès le départ que la segmentation de Xavier Naturel est correcte, ce qui se révèle assez juste de façon générale ; cependant certains segments contiennent plusieurs programmes et donc généralement plusieurs thèmes, ce qui gêne le calcul de similarité entre les descriptions et la transcription textuelle de ces segments, car les divers thèmes se parasitent dans la représentation à l'aide de vecteurs de mots pondérés. Nous pourrions donc réfléchir à l'utilisation d'une méthode étudiant la cohésion lexicale du segment afin de le découper en sous-parties si cette cohésion passe sous un seuil critique et de caractériser son thème en utilisant les termes qui ont permis de maintenir cette cohésion. Ceci suppose toutefois que les sous-parties soient suffisamment longues pour que l'on puisse effectivement distinguer une cohérence au niveau du vocabulaire.

Le découpage en émissions est une étape essentielle à la structuration de flux télévisés mais n'est cependant pas suffisant à la navigation à l'intérieur de ceux-ci. Notre objectif, à plus long terme, est de regrouper les thèmes similaires à travers les émissions et les chaînes, dans le but de pouvoir faire une recherche dans le flux vidéo sur un sujet particulier, ou de mener une étude comparative du traitement d'un même sujet entre différentes chaînes. Un prolongement de notre travail consistera donc à mettre en place une segmentation et un étiquetage des émissions en sous-sections abordant plusieurs sujets –les différents reportages d'une émission d'investigation par exemple.

5. Bibliographie

Béchet F., « LIA_PHON : un système complet de phonétisation de textes », *Traitement automatique des langues*, vol. 42, n° 1, p. 47-67, 2001.

Cardillo P. S., Clements M., Miller M. S., « Phonetic Searching vs. LVCSR : How to Find What You Really Want in Audio Archives », *International Journal of Speech Technology*, vol. 5, n° 1, p. 9-22, 2002.

Ferret O., Grau B., « Utiliser des corpus pour amorcer une analyse thématique », *Traitement automatique des langues*, vol. 42, n° 2, p. 517-545, 2001.

Fischer S., Lienhart R., Effelsberg W., « Automatic Recognition of Film Genres », *3rd International ACM Conference on Multimedia*, 1995.

Guinaudeau C., « Contrôle automatisé de contenu télévisuel », 2008. Rapport de stage de Master 2 recherche de l'université de Caen Basse Normandie, Caen, France.

Lecorvé G., Gravier G., Sébillot P., « On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems », *6th International Language Resources and Evaluation*, 2008.

Logan B., Moreno P., Deshmukh O., « Word and Sub-word Indexing Approaches for Reducing the Effects of OOV Queries on Spoken Audio », *2nd International Conference on Human Language Technology Research*, 2002.

Muscariello A., Gravier G., Bimbot F., « Variability Tolerant Audio Motif Discovery », *International Multimedia Model Conference*, 2009.

Naturel X., Structuration automatique de flux vidéos de télévision, PhD thesis, Université de Rennes 1, France, 2007.