Extraction de mots-clefs dans des vidéos Web par Analyse Latente de Dirichlet

Mohamed Morchid[1] Georges Linarès[1]
(1) LIA-CERI, Université d'Avignon et des Pays de Vaucluse
mohamed.morchid@univ-avignon.fr, georges.linares@univ-avignon.fr

RÉSUMÉ

Cet article présente une méthode d'étiquetage de vidéos collectées sur une plate-forme de partage de vidéos. Cette méthode combine un système de reconnaissance de la parole, qui extrait les contenus parlés des vidéos, et un module d'extraction de mots-clefs opérant sur les transcriptions automatiques. La difficulté majeure, dans cette caractérisation de vidéos par un ensemble de mots-clefs, est liée aux performances du SRAP qui sont souvent très faibles sur des vidéos générées par les utilisateurs.Dans cet article, une méthode d'extraction de mots-clefs robuste aux erreurs de reconnaissance est proposée. Cette méthode repose sur la projection des contenus parlés dans un espace thématique obtenue par Analyse Latente de Dirichlet. Nos expériences sont réalisées sur un ensemble de vidéos collectées sur une plate-forme de partage communautaire. Elles montrent l'intérêt du modèle proposé, en particulier dans les situations d'échec du système de transcription automatique.

ABSTRACT

LDA-based tagging of Web videos

This article presents a method for the automatic tagging of youtube videos. The proposed method combines an automatic speech recognition system, that extracts the spoken contents, and a keyword extraction system that aims at finding a small set of tags representing the video. In order to improve the robustness of the tagging system to the recognition errors, a video transcription is represented in a semantic space obtained by Latent Dirichlet Allocation (LDA), in which each dimension is automatically characterized by a list of weighted terms and chuncks. Our experiments demonstrate the interest of such a model to improve the robustness of the tagging system, especially when speech recognition (ASR) system produce highly errorneous transcript of spoken contents.

MOTS-CLÉS : Reconnaissance de la parole, analyse des contenus, catégorisation audio, multimédia.

KEYWORDS: Speech recognition, content analysis, audio categorization, multimedia.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 273–280,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les plates-formes de partage de vidéos sur Internet se sont fortement développées ces dernières années. En 2011, YouTube augmentait d'une heure d'enregistrement déposée toutes les secondes. Malheureusement, l'utilisation de ces collections de vidéos, souffre de l'absence d'informations structurées et fiables. L'indexation réalisée par l'hébergeur repose essentiellement sur les mots-clefs fournis par les utilisateurs, éventuellement sur les résumés ou le titre des documents. Malheureusement, ces méta-données sont souvent incomplètes ou erronées, parfois volontairement : certains utilisateurs choisissent des mots-clefs qui favorisent le référencement jusqu'à s'éloigner significativement du contenu réel de la vidéo déposée. Ceci implique donc des tags non représentatifs du contenu même de la vidéo..

Cet article propose une méthode pour l'extraction automatique de mots-clefs des contenus parlés d'une vidéo. Cette méthode repose sur un processus en 2 étapes qui réalisent respectivement la transcription automatique de la parole puis l'extraction de mots-clefs.

Un des problèmes majeurs de cet enchaînement extraction/analyse des contenus est lié au composant de reconnaissance de la parole, qui est souvent peu performant sur des données Web, dont la diversité de forme et de fond est extrême et qui sont généralement éloignées des conditions d'entraînement des systèmes.

Deux pistes sont typiquement suivies pour exploiter des transcriptions bruitées par les erreurs de reconnaissance. La première consiste à améliorer la robustesse du reconnaisseur de parole, de façon à éviter les situations d'échec massif du système, qui rendraient la transcription inutilisable. Cette voie requière le plus souvent des données caractéristiques de la tâche, données qui sont difficiles à collecter et coûteuses à annoter. L'autre possibilité est d'améliorer la tolérance du système d'analyse aux erreurs de reconnaissance. Cet article présente une stratégie robuste pour l'extraction de mots-clefs issus d'une transcription automatique des segments parlés d'une vidéo.

Cette méthode repose sur l'idée que le niveau lexical est particulièrement sensible aux erreurs de reconnaissance et qu'une représentation de plus haut niveau pourrait permettre de limiter l'impact négatif de ces erreurs sur les modules d'analyse. Le document source est projeté dans un espace thématique dans lequel le document peut être vu comme une association de thèmes. Cette représentation intermédiaire est obtenue par une analyse Latente de Dirichlet appliquée à grand corpus de textes. Une méthode originale utilisant cette décomposition pour déterminer les mots-clefs caractéristiques du document source est ensuite proposée.

L'extraction de mots-clefs est un thème classique du traitement automatique du langage naturel. La section suivante dresse un panorama rapide des approches les plus courantes et discute de leur capacité à traiter des textes bruités par les erreurs de reconnaissance. Notre proposition est ensuite détaillée : la section 3 décrit l'architecture globale du système, la section 4 présente le processus de construction de l'espace thématique et les métriques utilisées. La méthode d'extraction des mots-clefs dans cet espace est présentée dans la section 5, puis évaluée dans la section 6. L'article se termine par une conclusion et quelques perspectives.

2 État de l'art

La recherche de mots-clefs dans des documents textuels est un problème classique du traitement automatique du langage naturel. L'approche la plus populaire consiste à extraire les mots de plus fort TFxIDF (**T***erm* **F***requency.***I***nverse* **D***ocument* **F***requency*), qui mesure la fréquence du mot dans le document, normalisée par la fréquence du mot dans un grand corpus. Cette mesure a été déclinée en différentes variantes utilisant des ressources externes, la position du mot candidat dans le document (Frank *et al.*, 1999; Deegan *et al.*, 2004; HaCohen-Kerner *et al.*, 2005) ou des connaissances linguistiques (Hulth, 2003).

La recherche de mots-clefs dans des documents parlés présente des difficultés particulières, dues aux spécificités de l'oral et à l'usage de systèmes de reconnaissance automatique de la parole pour l'extraction des contenus linguistiques. Quelques travaux utilisent des approches haut niveau, basées sur des ontologies ou des connaissances linguistiques explicites. (van der Plas *et al.*, 2004) utilise Wordnet et EDR, un dictionnaire électronique de noms propres, pour extraire les concepts dominants d'un texte annoté automatiquement.

D'autre approches reposent sur des modèles statistiques, utilisées initialement sur des bases purement textuelles ; par exemple, (Suzuki *et al.*, 1998) utilise LSA pour l'extraction de mots-clefs dans une base de données encyclopédique. Ce type de techniques a ensuite été appliqué avec succès à de très nombreux problèmes de traitement de la parole. Par exemple, (Bellegarda, 2000) utilise LSA (Latent Semantic Analysis) pour extraire les phrases les plus significatives d'un document parlé.

L'extraction de mots-clefs peut être vue comme une forme extrême de résumé. Dans (**?**), les auteurs utilisent le modèle CBC (Comitee Based CLustering) et l'analyse latente de Dirichlet (LDA) pour extraire un ensemble de mots supposés résumer le document. Les résultats obtenus démontrent l'efficacité de LDA et semble robuste aux erreurs de reconnaissance, qui est un des points critiques des systèmes d'analyse des contenus parlés.

Notre proposition est d'utiliser la décomposition en thèmes latents obtenus par LDA pour trouver les thèmes latents composants la retranscription issu de la vidéo. Ainsi, est extrait les mots-clefs caractéristiques de vidéos disponibles sur Internet, qui représentent des conditions peu contrôlées, particulièrement difficiles à traiter par un système de reconnaissance de la parole.

3 Méthode proposée - Architecture générale

La méthode d'étiquetage proposée repose d'abord sur la transcription automatique des contenus parlés du document. Les sorties du module de reconnaissance sont ensuite projetées dans un espace thématique obtenu par LDA.
Nous considérons que le concept principal du document, qui doit être caractérisé par les mots-clefs, est une combinaison de thèmes latents représentés par les thèmes LDA. Les mots-clefs sont donc extraits par un processus de sélection et de combinaison des thèmes les plus représentatifs du document à étiqueter (cf.*Figure 1*). Diverses règles de combinaisons, telles que l'union et l'intersection sont proposées et évaluées. Ces méthodes sont comparées à l'approche classique à base de TFxIDF.
Ce système d'extraction de tags se décompose en trois phases : (1) Création d'un espace de

thèmes, (2) projection de la retranscription dans cet espace pour (3) déterminer un ensemble de thèmes proche puis (4) en extraire un ensemble de tags représentatif de la vidéo (cf.*Figure 1*) :

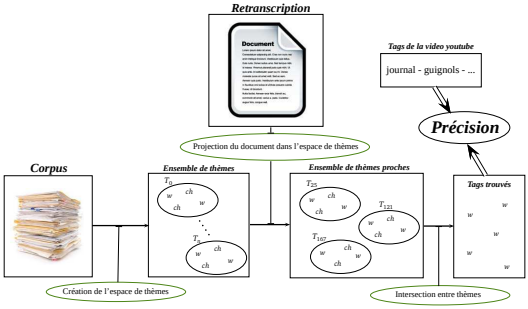

FIGURE 1 – **Extraction des mots-clefs par union ou intersection de thèmes.**

4 Espace thématique

LDA est un modèle génératif qui considère un document comme un *sac de mots* résultant d'une combinaison de thèmes latents. L'étude statistique des cooccurrences des mots dans une base permet d'extraire des classes de mots cooccurrents, qu'on assimile souvent à des thèmes, bien que rien ne permette d'établir explicitement un lien entre le modèle statistique et l'interprétation en thèmes qui pourrait en être faite. (Rigouste *et al.*, 2006) a clairement établi les avantages de LDA comparé à d'autres modèles génératifs du même type, largement utilisés en traitement automatique du langage naturel, par exemple LSI (Latent Semantic Indexing), équivalent à LSA (Kubota Ando et Lee, 2001)) ou sa variante probabiliste (Probabilistic Latent Semantic Indexing, PLSI, (Hofmann, 1999)).

Toutes ces méthodes requièrent des ensembles de données suffisants pour être une estimation correcte de leurs paramètres. Nous avons choisi d'utiliser Wikipédia et la collection des dépêches de l'agence France presse (AFP) de 2000 à 2006, qui représentent respectivement 1G et 4.5 Go pour un total d'environ 1 milliard de mots et 3 millions d'articles. Ces deux bases sont lemmatisées avec TreeTagger (Stein et Schmid, 1995) pour retirer l'ensemble de mots vides (articles, propositions, ...) avant d'estimer un modèle à 5000 thèmes de 30 mots, nombre choisi empiriquement. Nous obtenons un ensemble de mots en minuscule sans termes *vide*. Ceci permet de traiter des mots pouvant être trouver par le système de reconnaissance de la parole.

4.1 Représentation des thèmes LDA

Les thèmes LDA sont représentées par un vecteur composé du poids des mots du lexique dans le thème ($P(w_i|t)$). Le vecteur de poids d'un thème V_t est composé des probabilités des mots w_i

sachant le thème t, pondérées par l'IDF (*Définition dans la section 4.2*) du mot :

$$V_t[i] = P(w_i|t).idf(w_i)$$

4.2 Représentation vectorielle des documents

Un document peut être considéré comme un point dans un espace vectoriel R^k, chaque coordonnée i du vecteur W_d représentant un indicateur du poids du mot dans le document. Ce poids ($W_d[i] = tf(w_i) \times idf(w_i) \times rp(w_i)$) combine la mesure ($tf.idf$) et la position ($rp$) du mot w_i dans le document (Salton, 1989) :

$$tf(w) = \frac{n(w)(d)}{\sum_{i=0}^{k} n(w_i)(d)}, \quad idf(w) = log\frac{|D|}{|\{d : w \in d\}|}, \quad rp(w) = \frac{|d|}{fp_w}$$

où $n(w)(d)$ est le nombre d'occurrences du mot w dans d, et D est la collection complète de documents. Cette valeur est identique pour tous les mots d'un même document, la pondération liée à la première occurrence du mot dans le document est notée rp. Elle peut donc être simplifié par $rp(w) = \frac{1}{fp_w}$ car tout les mots sont issus d'un même document. fp étant la position de la première occurrence du mot dans le document.

4.3 Similarité document/thème

Nous avons vu que les documents étaient caractérisés par des vecteurs de TFxIDFxRP, et que les thèmes étaient représentés par des vecteurs de probabilités de mots conditionnées aux thèmes et normalisées par l'IDF. La mesure du cosinus est utilisée pour évaluer la similarité entre ces deux vecteurs :

$$Sim(d,t) = \frac{\sum_{w_i \in d} V_t[i].W_d[i]}{\sqrt{\sum_{w_i \in d} V_t[i]^2}.\sqrt{\sum_{w_i \in t} W_d[i]^2}}.$$

5 Extraction des mots-clefs

Il s'agit ici d'extraire les n mots-clefs de la projection du document dans l'espace thématique. La stratégie proposée est d'isoler les m thèmes principaux et de combiner leurs vecteurs caractéristiques. Dans nos expériences, m est fixé empiriquement à 100 thèmes. Deux approches peuvent être suivies ; la première consiste à chercher les éléments communs aux thèmes principaux du document ; l'autre consiste à extraire les mots de plus forts poids de l'ensemble des thèmes caractéristiques. La première approche va nous amener à chercher les mots-clefs dans l'intersection des thèmes, la seconde dans l'union. Pour les deux méthodes, les n mots de score $sc(w)$ les plus

élevé de chacun des thèmes sont sélectionnés pour commencer, avec n égal au nombre de tags associés à la vidéo :

$$sc(w) = \sum_{k=0}^{m} Sim(d, t_k).P(w|t_k) \quad (1)$$

où $P(w|t_k)$ représente la probabilité du mot w sachant la thème t_k et $Sim(d, t_k)$ la similarité t_k et d le document.

6 Expériences

Le corpus de tests est composé d'environ 100 vidéos françaises comportant 14 tags en moyenne hébergées sur la plate-forme YouTube. Le premier traitement consiste à appliquer le système de reconnaissance à la bande son de ces vidéos. Le système, dérivé de celui que le LIA a engagé dans la campagne d'évaluation ESTER 2008 (Linarès *et al.*, 2007), est utilisé. La segmentation est produite avec l'outil GPL du LIUM (Meignier et Merlin, 2010). Le moteur de reconnaissance utilise des modèles classiques à base de modèles de Markov et de statistiques 4-grammes, avec un algorithme de recherche A* appliqué à un treillis de phonèmes.

Les modèles acoustiques sont des modèles contextuels avec un partage d'états par arbres de décisions. Ces modèles sont appris sur les données produites par les 2 campagnes ESTER successives et par le projet EPAC, pour un total d'environ 250 heures de données annotées. Le jeu de modèles est composé de 20000 HMMs pour un peu plus de 5000 états partagés. Les modèles à mélange de gaussiennes associés à ces états sont des mixtures à 32 composantes. La dépendance au genre est introduite à la fin du processus d'estimation, par adaptation MAP de ce modèle. Les modèles de langage sont des 4-grammes classiques estimés, pour l'essentiel, sur environ 200M de mots du journal français Le Monde, le corpus ESTER (environ 2M de mots) et le corpus GIGAWORD, composé des dépêches d'informations sur la période de 2000 à 2006, pour environ 1 milliard de mots.

Le décodeur exécute deux passes. La première est un décodage rapide (3xRT) en 4-grammes, qui permet l'adaptation au locuteur des modèles acoustiques ; la seconde est effectuée avec une exploration plus complète de l'espace de recherche (les coupures sont moins strictes) et utilise ces modèles adaptés. Elle est exécutée en environ 5 fois le temps réel sur une machine de bureau standard. La transcription manuelle de la parole est une tâche lourde. De façon à estimer le niveau de performance du système sur des vidéos issues du Web, 10 des 100 vidéos de test, choisies aléatoirement, ont été transcrites manuellement, ce qui représente environ 35 minutes de parole. Sur cet échantillon assez réduit, le système obtient un taux d'erreur mot de 63.7%, évidemment très élevé mais conforme à ce qu'on pouvait attendre du décodage de documents Web, très divers sur le fond et enregistrés dans des conditions variables - mais le plus souvent difficiles et peu contrôlées. Ces vidéos sont le résultats de la requête "journal actualité" soumise à la plate-forme YouTube. Ainsi, l'ensemble des vidéos contient 6166 tags dont 903 absents du vocabulaire du modèle LDA. Ceci représente environ 14% de mots hors-vocabulaire (cf.*Table 1*).

L'extraction de mots-clefs est vue comme une tâche de détection des tag utilisateurs, qui sont considérés ici comme référence unique. Les performances sont mesurées de façon classique, en

Méthode	Tags trouvés	Précision
Tags utilisateurs	iranien atomique netanyahou livni intel arabe	1
TFxIDFxRP	vrai ehoud **iranien** jérusalem sécuritaire iran	0.16
Intersection	**iranien** étranger vrai iran **atomique** jérusalem	0.33
Union	chancelier gouvernement **iranien** ancien virtuel laisser	0.16

TABLE 1 – Exemple de tags trouvés par les méthodes d'extraction de mots-clefs par intersection, union des thèmes et TFxIDFxRP comparées au tags de l'utilisateur pour une vidéo.

terme de Précision.

Les résultats montrent que l'intersection est très clairement supérieure à l'union des thèmes, ce qui peut sembler assez contre-intuitif. Par ailleurs, cette méthode est très nettement supérieure à la classique TFxIDFxRP, ce qui valide l'idée, qui a motivé ce travail, que l'abstraction réalisée par la projection dans l'espace thématique limite l'effet négatif des erreurs de reconnaissance. Dans la (cf.*Table 2*) est présentée la précision pour chacune des méthodes.

	Intersection	Union	TFxIDFxRP
Précision	**4.8%**	0.9%	2.8%

TABLE 2 – Précision comparées de la méthode d'extraction de mots-clefs par intersection, union des thèmes thématiques, par TFxIDFxRP

La faiblesse de ces résultats est due au taux de tags n'apparaissant pas dans la retranscription (14%), et aux erreurs du système RAP.

7 Conclusions et perspectives

Dans cet article, une méthode d'étiquetage automatique de vidéos est proposée. Celle-ci repose sur la localisation du document dans un espace thématique issu d'une analyse latente de Dirichlet. Différentes méthodes d'extraction des mots clefs dans cet espace sont proposées. Nos expériences montrent que cette représentation thématique du document permet de réduire l'effet négatif des erreurs de reconnaissance : les méthodes proposées dépassent très significativement l'approche classique à base de TFxIDFxRP, qui opère dans une représentation de niveau lexical.

Dans l'absolu, les performances obtenues peuvent sembler relativement faibles (moins de 5% de Précision). Néanmoins, cette évaluation repose sur la comparaison des résultats du système et des tags qui sont effectivement donnés par les utilisateurs ; on peut considérer qu'il s'agit la d'une référence elle-même assez bruitée et des expériences complémentaires permettraient probablement de consolider les résultats obtenus ici (par exemple en proposant des annotations produites par différents utilisateurs plutôt que par l'uploader seul).

De façon plus générale, nos résultats confirment la robustesse supérieure apportée par la projection du document dans des représentations de relativement haut niveau. Valider ce principe sur d'autres tâches classiques liées à l'indexation ou à l'interprétation de la parole est une des pistes que nous développerons dans l'avenir.

Références

BELLEGARDA, J. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

DEEGAN, M., SHORT, H., ARCHER, D., BAKER, P., MCENERY, T. et RAYSON, P. (2004). Computational linguistics meets metadata, or the automatic extraction of key words from full text content. *RLG Diginews*, 8(2).

FRANK, E., PAYNTER, G., WITTEN, I., GUTWIN, C. et NEVILL-MANNING, C. (1999). Domain-specific keyphrase extraction. *In International joint conference on artificial intelligence*, volume 16, pages 668–673. Citeseer.

HACOHEN-KERNER, Y., GROSS, Z. et MASA, A. (2005). Automatic extraction and learning of keyphrases from scientific articles. *Computational Linguistics and Intelligent Text Processing*, pages 657–669.

HOFMANN, T. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

HULTH, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *In Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10*, pages 216–223. Association for Computational Linguistics.

KUBOTA ANDO, R. et LEE, L. (2001). Iterative residual rescaling : An analysis and generalization of lsi.

LINARÈS, G., NOCÉRA, P., MASSONIE, D. et MATROUF, D. (2007). The lia speech recognition system : from 10xrt to 1xrt. *In Proceedings of the 10th international conference on Text, speech and dialogue*, pages 302–308. Springer-Verlag.

MEIGNIER, S. et MERLIN, T. (2010). Lium spkdiarization : an open source toolkit for diarization. *In CMU SPUD Workshop*, volume 2010.

RIGOUSTE, L., CAPPÉ, O. et YVON, F. (2006). Quelques observations sur le modele lda. *Actes des IXe JADT*, pages 819–830.

SALTON, G. (1989). Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.

STEIN, A. et SCHMID, H. (1995). Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues*, 36(1-2):23–35.

SUZUKI, Y., FUKUMOTO, F. et SEKIGUCHI, Y. (1998). Keyword extraction using term-domain interdependence for dictation of radio news. *In Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1272–1276. Association for Computational Linguistics.

van der PLAS, L., PALLOTTA, V., RAJMAN, M. et GHORBEL, H. (2004). Automatic keyword extraction from spoken text. a comparison of two lexical resources : the edr and wordnet. *Arxiv preprint cs/0410062*.