Segmentation thématique : apport de la vectorisation

Vincent Claveau[*] — **Sébastien Lefèvre**[**]

** IRISA-CNRS*
Campus de Beaulieu
F-35042 Rennes cedex
Vincent.Claveau@irisa.fr

*** VALORIA, Université de Bretagne-Sud*
Campus de Tohannic
F-56017 Vannes cedex
Sebastien.Lefevre@univ-ubs.fr

RÉSUMÉ. *Dans cet article, nous nous intéressons à la segmentation thématique d'émissions télévisées à partir de la transcription automatique de leur bande-son. La segmentation thématique de textes a fait l'objet de travaux depuis de nombreuses années, et les techniques mises en œuvre reposent souvent sur des descriptions de contenu et des calculs de similarité utilisés en recherche d'information. Dans cet article, nous proposons une technique s'inspirant des travaux de morphologie mathématique utilisés en segmentation d'image. Nous montrons de plus que la technique d'appariement par vectorisation proposée par (Claveau* et al., *2010) peut être utilisée dans l'étape clef de calcul de similarité entre les segments. Nous évaluons cette approche sur deux corpus d'émissions de télévision. Les résultats obtenus au travers de ces expériences dépassent nettement ceux des approches existantes et montrent le bien-fondé de notre démarche.*

ABSTRACT. *This paper deals with topic segmentation of TV broadcasts using their transcription obtained by automatic speech recognition. Topic segmentation has been studied for several years, and most often the techniques proposed rely on information retrieval techniques to compute similarities between segments. In this paper, we propose a new segmentation approach inspired by mathematical morphology studies developed in the framework of image segmentation. We also show that using the similarity technique called vectorization and first developed for IR (Claveau* et al., *2010) can be efficiently used in this context. This new topic segmentation technique is evaluated on two corpora of TV broadcasts on which it outperforms other existing approaches.*

MOTS-CLÉS : *Segmentation thématique, vectorisation, ligne de partage des eaux, calcul de similarité, flux TV, plongement*

KEYWORDS: *Topic segmentation, vectorization, watershed, similarity measure, TV streams, embedding*

1. Introduction

La première étape essentielle à tout traitement d'un flux vidéo comme la télévision est sa délinéarisation, c'est-à-dire son découpage en documents. Dans cet article, nous nous intéressons à la segmentation thématique, c'est-à-dire le découpage d'une émission en éléments plus fins traitant d'un même sujet. Pour effectuer cette segmentation thématique, nous nous appuyons sur le contenu oral des bandes-son qui sont transcrites automatiquement.

Différentes hypothèses et contraintes sous-tendent notre travail. Tout d'abord, nous ne supposons aucune connaissance sur le contenu à segmenter, que ce soit en terme de sujets abordés ou en terme de format des émissions. D'autre part, la segmentation que nous cherchons à effectuer est une segmentation linéaire ; nous supposons donc que les différents sujets sont abordés successivement au cours des émissions traitées.

Dans cet article, nous proposons une méthode efficace reposant sur deux apports. Tout d'abord, nous montrons notamment que la vectorisation (Claveau *et al.,* 2010), technique de calcul de similarité initialement développée pour la RI, est particulièrement adaptée à cette tâche. Ensuite, nous montrons qu'en dressant une analogie entre cette tâche de segmentation de texte et ce qui est fait en segmentation d'image, il est possible de bénéficier des apports en terme de techniques et de formalisations développées dans ce domaine. L'approche mise en œuvre dans cet article s'appuie ainsi sur une technique commune en image appelée *ligne de partage des eaux*.

Dans les sections suivantes, nous présentons successivement ce parallèle entre segmentation d'image et segmentation thématique, le principe de ligne de partage des eaux, les approches existantes en segmentation thématique, et le calcul de similarité par vectorisation, étape clef de notre système de segmentation. Des expériences menées sur deux corpus d'émissions télévisées sont ensuite rapportées.

2. De la segmentation morphologique d'image à la segmentation thématique de texte

L'analyse et le traitement des images ont motivé depuis plusieurs décennies de nombreuses recherches. Parmi les cadres théoriques qui ont amené à l'élaboration de ces techniques, la morphologie mathématique a suscité depuis une quarantaine d'année un grand intérêt de par sa capacité à gérer implicitement les informations de forme dans les images (Soille, 2003) et donc pour effectuer de la segmentation d'image. Celle-ci consiste à partitionner une image initiale pour former un ensemble de régions homogènes au sens d'un certain critère. De nombreuses méthodes de segmentation d'image ont été proposées dans la littérature (Gonzalez *et al.,* 2008), et nous nous intéressons ici à l'approche la plus connue, la Ligne de Partage des Eaux (LPE).

La méthode LPE de segmentation morphologique procède comme suit. En représentant l'image à segmenter I comme un relief (ou surface topographique), elle identifie dans ce dernier les lignes de partage des eaux. Celles-ci sont alors associées aux

frontières des régions ainsi segmentées. Ce principe relativement simple a donné lieu à différents paradigmes pour sa mise en œuvre ; nous avons choisi ici d'utiliser l'approche la plus connue, celle de Vincent et Soille (Vincent *et al.,* 1991) qui consiste à simuler l'inondation progressive du relief par ses minima, et à séparer les différents bassins associés à chaque minimum par des digues. A l'issue du processus, ces digues représentent les lignes de partage des eaux, ou autrement dit les frontières des régions. La figure 1 illustre ce processus dans le cas d'une image à une dimension.

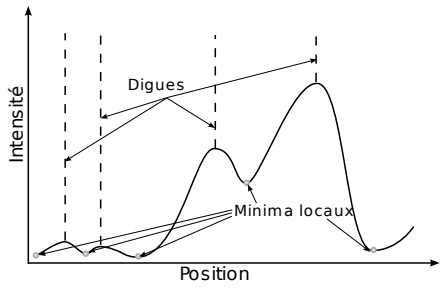

Figure 1. *Illustration du principe de segmentation par Ligne de Partage des Eaux*

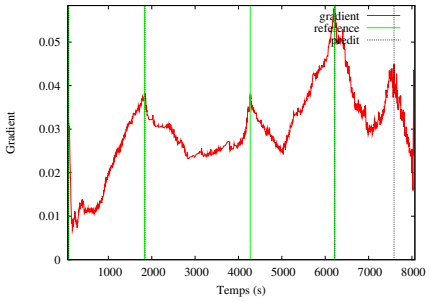

Figure 2. *Gradient en fonction du temps de début des groupes de souffle*

Cependant, l'approche précédente n'est que rarement appliquée directement sur l'image. La segmentation est plutôt précédée d'une transformation de l'image d'origine, afin d'en rehausser les valeurs des pixels de frontières tout en limitant les valeurs des pixels dans les zones homogènes. On calcule donc généralement un gradient de l'image (ou dérivée première du signal) noté ∇I pour mettre en exergue les zones de transition, généralement représentatives des frontières des objets.

L'analogie entre la segmentation d'image et celle de texte se fait très simplement. Dans notre cadre de segmentation de vidéos, nos textes sont issus d'un système de transcription automatique. Ces transcriptions ne sont pas composées de phrases mais de groupes de souffle (suite de mots prononcés entre des respirations ou des pauses) repérés par un *timestamp*. Ce sont donc ces groupes de souffle qui constituent l'unité minimale manipulée, équivalente au pixel, entre lesquels sont recherchées les frontières thématiques. Ce flux de groupes de souffle a, contrairement à une image, une seule dimension, mais comme nous l'avons vu en figure 1, la technique de ligne de partage des eaux s'applique parfaitement dans ce cas.

Notre approche repose donc sur le calcul d'un gradient sur la suite des groupes de souffle, et les ruptures thématiques sont détectées par LPE. La calcul de ce gradient, étape clef du processus, est détaillé dans la section 4. La technique LPE que nous utilisons est celle décrite précédemment ; elle est simplement précédée d'un lissage du gradient permettant d'éliminer les minima locaux non significatifs.

3. Travaux connexes

Différentes approches ont déjà été développées pour la segmentation thématique. Parmi celles-ci, il faut distinguer plusieurs grandes familles. Il y a d'une part des approches s'appuyant sur des propriétés de formatage des documents, ou sur la détection de marqueurs discursifs (Christensen *et al.,* 2005). L'autre grande famille d'approches s'appuie sur le contenu des documents pour détecter les changements de thème. C'est dans cette famille que s'inscrit notre approche et beaucoup des systèmes existants. Parmi ceux-ci, il est intéressant de noter les similarités qu'elles partagent avec les méthodes de segmentation d'image. Ainsi, Utiyama et Isahara (Utiyama *et al.,* 2001) proposent une méthode de segmentation statistique basée sur un modèle de Markov caché. Ce type d'outil est également utilisés en image (Salzenstein *et al.,* 2006), notamment pour leurs capacités à gérer le bruit.
Dans l'approche SEGMENTER (Kan *et al.,* 1998), la segmentation s'appuie sur une représentation des chaînes lexicales proche d'un graphe qu'il faut partitionner. Dans le domaine de l'analyse d'image, des approches similaires sont également utilisées (Shi *et al.,* 2000).
Les approches DOTPLOTTING (Reynar, 2000) et *C99* (Choi, 2000), même si elles utilisent deux représentations différentes des données, s'appuient toutes deux sur une méthode de clustering pour regrouper ensuite les segments cohérents. En segmentation d'image, cette démarche est très classique (Gonzalez *et al.,* 2008).
Les frontières produites par l'approche TEXT-TILING (Hearst, 1997) correspondent aux zones où la cohésion lexicale avec les blocs de texte précédents et les blocs de texte suivants est associée à un minimum local significatif. Les minima qui sont retenus sont associés aux zones où la cohésion diffère fortement des blocs voisins. Cette approche est celle qui présente le plus de similitudes avec la segmentation morphologique par LPE, et donc avec l'approche que nous présentons ci-après.

Ces différentes approches ont été comparées, que ce soit sur l'anglais (Choi, 2000) ou le français (Sitbon *et al.,* 2004). Malheureusement, elles l'ont été sur des données construites artificiellement dans des contextes applicatifs très différents du nôtre. Par ailleurs, ces approches étant basées sur la répétition des mots, un problème handicapant est le manque de termes partagés entre les segments à comparer quand ceux-ci sont petits. Pour minimiser l'impact de ce problème, Guinaudeau et collègues (Guinaudeau *et al.,* 2010) s'appuient sur le modèle de (Utiyama *et al.,* 2001) mais y intègrent des termes reliés sémantiquement pour enrichir la description de chaque candidat-segment comme pour les systèmes d'extension de requêtes en RI. Notre approche basée sur la vectorisation est pressentie comme mieux adaptée à ce genre de situations.

4. Calcul du gradient par vectorisation

4.1. *Principes de la vectorisation*

La vectorisation est une technique de plongement (*embedding*) qui permet de projeter n'importe quel calcul de similarité entre deux documents (ou un document et une requête pour la RI) dans un espace vectoriel. Elle a été proposée et testée dans un cadre de RI standard (Claveau *et al.,* 2010) dans lequel elle a montré son intérêt en terme de complexité et de qualité des résultats.

Son principe est relativement simple : pour chaque document de la collection considérée, il consiste à calculer avec la mesure de similarité initiale, quelle qu'elle soit, des scores de proximité avec m documents-pivots. Les m scores obtenus forment ainsi un vecteur de m dimensions représentant le document. Dès lors, la comparaison de deux documents (ou d'un document et d'une requête) peut donc s'effectuer de manière standard dans cet espace vectoriel, par exemple en calculant une distance L2. De nombreux algorithmes permettent de calculer de telles distances très rapidement.

Plus formellement, on note Vect$(D, \mathcal{P}, Sim)$ le vecteur représentant le document D construit à partir de la mesure de similarité initiale Sim sur les documents pivots $\mathcal{P}$. Par exemple, Vect$(D, [P_1, P_2, P_3], Okapi)$ est un vecteur de dimension 3, avec pour première composante le score fourni par un système Okapi entre le document D, considéré comme une requête, et le document pivot P_1, etc.

Il est important de noter que la vectorisation change l'espace de représentation. Il ne s'agit donc pas seulement d'une réduction de l'espace ou d'une approximation de la distance originelle comme proposée par exemple dans les travaux de (Bourgain, 1985). C'est ce changement d'espace qui est à l'origine de deux propriétés intéressantes.

D'une part, cet *embedding* permet de réduire la complexité quand le calcul de similarité initial est très coûteux (Claveau *et al.,* 2010). Cette propriété, potentiellement intéressante pour la RI, n'est pas spécifiquement utile pour notre tâche de segmentation. D'autre part, la vectorisation permet que deux documents soient considérés comme proches s'ils sont proches des mêmes documents-pivots. Cette comparaison indirecte, ou affinité du second ordre, permet de mettre par exemple en relation des documents textuels qui ne contiennent pourtant aucun mot en commun. C'est cette propriété qui va nous intéresser dans notre tâche de segmentation. Elle doit en effet nous permettre de pallier le problème de manque de répétition entre groupes de souffle expliqué précédemment.

4.2. *Utilisation*

Un gradient est calculé entre chaque groupe de souffle, c'est-à-dire que l'on calcule la similarité par vectorisation entre les groupes de souffle précédents et les suivants. Il faut noter qu'on ne compare pas seulement le groupe de souffle précédent au groupe de souffle suivant, mais on considère, comme pour TEXT-TILING, les n précédents aux n

suivants. On note $Struct_{pred}(i)$ (respectivement $Struct_{succ}(i)$) l'union des groupes de souffle i et ceux le précédant (resp. le suivant). En s'inspirant de ce qui se fait en segmentation d'image, on donne par ailleurs une importance décroissante aux groupes de souffle au fur et à mesure que l'on s'éloigne de la frontière-candidate.

Dans les expériences rapportées en section suivante, les documents-pivots que nous utilisons sont simplement des suites de groupes de souffle issues de découpages aléatoires de l'émission considérée, et la mesure de similarité initiale utilisée dans la vectorisation est une distance L_2 associée à une pondération des groupes de souffle en $\sqrt{TF}$. Formellement, le gradient se définit donc par :

$$\nabla(i) = L_2(\text{Vect}(Struct_{pred}(i-1), \mathcal{P}, \sqrt{TF}/L_2), \text{Vect}(Struct_{succ}(i), \mathcal{P}, \sqrt{TF}/L_2))$$

La figure 2 montre un exemple de gradient calculé par vectorisation sur une émission d'un de nos corpus (*cf. infra*) contenant 4 segments (bornes indiquées en trait plein vert ; les ruptures détectées par LPE dans notre système sont en pointillés). Les groupes de souffles sont représentés par leur temps de début. À un point donné, plus le gradient est élevé, plus il indique une dissimilarité entre ce qui précède et ce qui suit, c'est-à-dire une rupture thématique.

5. Expérimentations

5.1. *Données expérimentales*

Nos expériences sont menées sur deux corpus d'émissions en français développés par (Guinaudeau *et al.,* 2010). Le premier corpus est un ensemble de 60 journaux télévisés de France 2 (*JT*, par la suite) d'une durée d'environ 40 minutes chacun. Le second corpus appelé *Reportages* est constitué d'émissions de reportages : 12 *Envoyé spécial*, de 2 heures chacun, et 16 *Sept à huit*, durant 1 heure. Ces deux corpus ont des caractéristiques différentes : le corpus *JT* a 1180 segments contre environ 140 pour le corpus *Reportages*. La segmentation de référence a été effectuée indépendamment en considérant qu'un changement de thème a lieu à chaque changement de reportage. Cette définition de la rupture thématique a l'avantage de correspondre à un besoin applicatif réel et bien défini. Les bandes-son de ces deux corpus ont été transcrites automatiquement par le système de reconnaissance de la parole IRENE (Huet *et al.,* 2010). Ces transcriptions ont ensuite été étiquetées en parties-du-discours avec TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`) ; seuls les noms, verbes et adjectifs sont conservés et racinisés.

5.2. *Résultats*

Notre système a donc été évalué sur les corpus *JT* et *Reportages*. Comme pour (Guinaudeau *et al.,* 2010), les performances sont mesurées en terme de rappel, pré-

cision et F-mesure, en considérant comme correcte une frontière de segment placée à moins de 10 secondes d'une de référence. À des fins de comparaison, nous indiquons les résultats obtenus, lorsque disponibles, sur les mêmes corpus par le système de Utiyama et Isahara (Utiyama *et al.,* 2001) tel qu'implémenté dans (Guinaudeau *et al.,* 2010), les meilleurs résultats obtenus par les différentes variantes du système de Guinaudeau et collègues (Guinaudeau *et al.,* 2010), et également les résultats obtenus par une implémentation inspirée de TEXT-TILING (Hearst, 1997). Pour cette dernière, nous utilisons la même préparation des données et dont nous faisons bénéficier de la même technique de détection de rupture par LPE que notre système. La seule différence avec notre technique par vectorisation réside donc dans le calcul du gradient qui se note : $\nabla(i) = cosinus(Struct_{pred}(i-1), Struct_{succ}(i))$.

Le tableau 1 présente les résultats sur nos deux corpus. On y constate dans les deux cas un gain important de notre système par rapport aux systèmes existants. La comparaison avec notre implémentation de TEXT-TILING est particulièrement intéressante puisqu'elle met bien en valeur l'intérêt du calcul de gradient par vectorisation plutôt qu'avec un cosinus associé à une représentation TF-IDF. Cet intérêt est encore plus marqué sur le corpus *JT* ; cela s'explique par le fait que les segments y sont très courts, ce qui rend le calcul de gradient direct tel que fait dans TEXT-TILING peu fiable.

	Corpus JT			Corpus Reportages		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
(Utiyama *et al.,* 2001)	-	-	59.44	-	-	51.09
(Guinaudeau *et al.,* 2010)	-	-	61.41	-	-	62.92
TEXT-TILING	44.17	41.97	43.04	59.32	60.93	60.12
vectorisation	67.47	61.6	64.4	77.38	69.65	73.31

Tableau 1. *Performances des systèmes de segmentation sur les corpus JT et Reportages*

6. Conclusions

Le parallèle que nous avons dressé dans cet article entre la segmentation d'image et la segmentation thématique nous a permis de développer une technique très performante reposant sur la technique de ligne de partage des eaux. Notre approche bénéficie aussi d'un modèle de calcul de similarité, la vectorisation, plus évolué que ceux habituellement utilisés pour cette tâche. Les évaluations rapportées montrent que l'utilisation de cette technique de vectorisation permet de pallier le manque de répétition entre des groupes de souffle, qui s'avère un problème important lorsque les segments thématiques recherchés sont courts. Cette conclusion renforce donc celle de Guinaudeau et collègues, qui avait tenté de résoudre ce problème en utilisant des données externes. À ce titre, notre approche offre un cadre plus efficace pour traiter ce problème, et ne nécessite aucune connaissance externe.

7. Bibliographie

Bourgain J., « On Lipschitz embedding of finite metric spaces in Hilbert space », *Israel Journal of Mathematics*, 1985.

Choi F. Y. Y., « Advances in domain independent linear text segmentation », *Proc. of the 1st metting of the North American Chapter of the Association for Computational Linguistics*, États-Unis, 2000.

Christensen H., Kolluru B., Gotoh Y., Renals S., « Maximum entropy segmentation of broadcast news », *Proc. of the 30th IEEE ICASSP*, 2005.

Claveau V., Tavenard R., Amsaleg L., « Vectorisation des processus d'appariement document-requête », *7e conférence en recherche d'informations et applications, CORIA'10*, Sousse, Tunisie, p. 313-324, March, 2010.

Gonzalez R., Woods R., *Digital Image Processing*, 3ème edn, Prentice Hall, 2008.

Guinaudeau C., Gravier G., Sébillot P., « Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels », *Actes de la conférence Traitement automatique des langues*, Montréal, Canada, 2010.

Hearst M., « Text-tiling : segmenting text into multi-paragraph subtopic passages », *Computational Linguistics*, vol. 23, n° 1, p. 33-64, 1997.

Huet S., Gravier G., Sébillot P., « Morpho-Syntactic Post-Processing with N-best Lists for Improved French Automatic Speech Recognition », *Computer Speech and Language*, vol. 24, n° 4, p. 663-684, October, 2010.

Kan M.-Y., Klavans J. L., McKeown K. R., « Linear segmentation and segment significance », *Proc. of the 6th International Workshop of Very Large Corpora (WVLC-6)*, 1998.

Reynar J. C., Topic Segmentation : Algorithms and applications, PhD thesis, University of Pennsylvania, 2000.

Salzenstein F., Collet C., « Fuzzy Markov Random Fields versus Chains for Multispectral Image Segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n° 11, p. 1753-1767, 2006.

Shi J., Malik J., « Normalized cuts and image segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 8, p. 888-905, 2000.

Sitbon L., Bellot P., « Évaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au fran cais », *Actes de la conférence Traitement automatique des langues*, Fez, Tunisie, 2004.

Soille P., *Morphological Image Analysis : Principles and Applications*, Springer-Verlag, Berlin, 2003.

Utiyama M., Isahara H., « A statistical model for domain-independent text segmentation », *Proc. of the 9th conference of the ACL*, 2001.

Vincent L., Soille P., « Watersheds in Digital Spaces : An Efficient Algorithm Based on Immersion Simulations », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, n° 6, p. 583-598, 1991.