TALN 2005, Dourdan, 6–10 juin 2005

Segmentation thématique par chaînes lexicales pondérées

Laurianne Sitbon, Patrice Bellot
Laboratoire d'Informatique d'Avignon - Université d'Avignon
339, chemin des Meinajaries - Agroparc BP 1228
84911 AVIGNON Cedex 9 - FRANCE
Tél : +33 (0) 4 90 84 35 09
{laurianne.sitbon, patrice.bellot}@univ-avignon.fr

Mots-clefs : segmentation thématique, chaînes lexicales, entités nommées

Keywords: topic segmentation, lexical chains, named entities

Résumé Cet article propose une méthode innovante et efficace pour segmenter un texte en parties thématiquement cohérentes, en utilisant des chaînes lexicales pondérées. Les chaînes lexicales sont construites en fonction de hiatus variables, ou bien sans hiatus, ou encore pondérées en fonction de la densité des occurrences du terme dans la chaîne. D'autre part, nous avons constaté que la prise en compte du repérage d'entités nommées dans la chaîne de traitement, du moins sans résolution des anaphores, n'améliore pas significativement les performances. Enfin, la qualité de la segmentation proposée est stable sur différentes thématiques, ce qui montre une indépendance par rapport au type de document.

Abstract This paper presents an innovative and efficient topic segmentation method based on weighted lexical chains. This method comes from a study of different existing tools, and experiments where we considered the influence of a term at each precise place in the text. We build lexical chains with different kinds of hiatus (varying, none or density weighted). We demonstrate good results on a manually built french news corpus. We show that using named entities does not improve results. Finally, we show that our method tends to be domain-independent because results are similar on various topics.

1 Introduction

La segmentation thématique intervient dans différents domaines en organisation de l'information, tels que déterminer les limites entre des dépêches dans un flux d'informations (*broadcast news*, TDT (*Topic Detection and Tracking*)), ou créer un résumé automatique de textes pour lequel la segmentation sert à isoler les thématiques et les parties les plus représentatives (McDonald &

Chen, 2002). Enfin, du point de vue d'un utilisateur, la segmentation thématique a également des avantages en terme de facilités de lecture.

Beaucoup de moyens ont été imaginés pour segmenter un texte en thèmes cohérents. La principale différence entre ces méthodes tient au fait qu'elles sont ou non supervisées. Parmi les méthodes supervisées on trouve par exemple PLSA (Brants *et al.*, 2002) qui apprend des probabilités d'appartenance des termes à des classes sémantiques. D'autres méthodes se basent sur un apprentissage comme (Amini *et al.*, 2000) qui s'appuient sur des modèles de Markov cachés, ou bien (Caillet *et al.*, 2004) qui propose une classification des termes de même que (Chuang & Chien, 2004) et (Mekhaldi *et al.*, 2004). Nous avons développé une méthode non supervisée, à base de matrice de similarités et de chaînes lexicales du type de celles utilisées par (Utiyama & Isahara, 2001) ou (Galley *et al.*, 2003). Ce choix est lié à leurs capacités naturelles d'adaptation à de nouvelles thématiques, et à leur relative indépendance vis à vis de la langue des textes.

2 Méthodologie

Avant de concevoir une nouvelle méthode, nous avons fait une évaluation de l'état de l'art pour des textes en français (Sitbon & Bellot, 2004). Cette étude préliminaire a notamment permis d'analyser les différentes mesures d'évaluation de la segmentation, et de créer un corpus de test en français. Les outils que nous avons étudiés ont été comparées pour l'anglais par (Choi, 2000). Nos expériences ont montré que dans les conditions où le texte à segmenter est une suite d'articles bien distincts, et où la qualité des outils est évaluée automatiquement en fonction de la distance entre les frontières trouvées et celles à trouver, l'outil le plus efficace est C99 (Choi, 2000), qui ordonne localement les similarités entre chaque paire de phrases, puis fait des regroupements par maximum de densité. Nous avons montré que le type de document que l'on segmente, son thème, la taille et la variation de taille des segments à repérer, sont autant de caractéristiques influençant le travail des segmenteurs.

2.1 Construction et pondération des chaînes lexicales

Une chaîne lexicale relie des termes de manière linéaire dans un texte. Les méthodes actuelles de segmentation les utilisent pour relier les occurrences d'un même terme (ou lemme) qui sont "proches". Une chaîne est rompue lorsque le nombre de termes qui séparent deux occurrences dépasse une valeur fixée appelée hiatus. On peut alors recenser pour chaque phrase les chaînes actives.

Les applications des chaînes lexicales utilisent actuellement des hiatus définis de manière empirique, et la notion d'activité d'une chaîne est binaire (elle est active ou non active). Notre premier objectif est d'éliminer le caractère empirique du hiatus, afin que notre outil puisse s'adapter à n'importe quel type de texte sans intervention de l'utilisateur. Pour cela on peut

imaginer tout simplement la **suppression du hiatus**, ce qui revient à relier toutes les répétitions de termes. Nous avons également envisagé l'utilisation de **hiatus locaux** : le hiatus moyen est calculé pour chaque lemme. Ainsi si un mot est fortement répété à deux endroits distincts du texte, il y aura automatiquement la création de deux chaines. De plus, s'il est répété trois fois en début de texte, puis une fois à la fin, il n'y aura qu'une seule chaîne comprenant les occurrences de début de texte.

Ces techniques créent un déséquilibre dans la significativité de l'activité des chaînes en jeu dans le calcul des frontières. Il faut alors pondérer les chaînes, en fonction de leur compacité (ratio entre leur taille et le nombre d'occurrences). (Galley *et al.*, 2003) propose une pondération des chaînes en fonction de la compacité et de la fréquence du terme considéré, et obtient de bons résultats, malgré un hiatus déterminé de manière empirique. La catégorie des lemmes a été intégrée à cette pondération. Le poids d'une chaîne associée à un terme m est défini par :

$$score(Chaine, m) = max(Chaine, cat(m) \times freq(Chaine, m) \times log(\frac{L_{texte}}{L_{chaine}}) \qquad (1)$$

où freq(Chaine, m) est le nombre d'occurrences du terme m dans la chaîne, L_{texte} la longueur du texte, L_{chaine} la longueur de la chaîne (on est alors indépendant de la taille des textes à segmenter), et max(Chaine,cat(m)) le poids de la forme grammaticale la plus importante parmi les occurrences du terme dans la chaîne.

Puis on calcule les similarités à chaque fin de phrase, qui est une rupture thématique potentielle. La similarité est calculée avec :

$$sim(A, B) = \frac{\sum_m score(A, m) \times score(B, m)}{\sqrt{\sum_m score(A, m) \times \sum_m score(B, m)}} \qquad (2)$$

où A et B sont les ensembles de vecteurs représentant les poids des chaînes lexicales actives dans les n phrases avant et après (nous avons choisi n=2), score(X, m) étant le poids maximal du terme m dans l'ensemble des vecteurs X.

Les frontières retenues sont alors celles pour lesquelles la similarité est en dessous d'un seuil déterminé par $sim_{limit} = \mu + \frac{\sigma}{2}$ où μ et σ sont la moyenne et la variance de toutes les similarités calculées (Galley *et al.*, 2003).

2.2 Evaluation

Afin de constituer un corpus de grande taille aisément, on compose un corpus de test à partir d'articles journalistiques de thème globalement éloignés car classés manuellement dans différentes rubriques. Le corpus de test est composé de 4 séries de 100 documents, chaque série

correspondant à une taille moyenne pré-définie des segments. Chaque document est composé de 10 segments qui sont autant d'extraits d'articles du journal Le Monde, choisis aléatoirement. Ce journal proposant des thématiques très variées, on suppose alors les segments thématiquement cohérents et différents.

Pour évaluer l'efficacité de nos nouveaux algorithmes, nous avons utilisé la mesure Window Diff proposée par (Pevzner & Hearst, 2002), présentée et analysée dans (Sitbon & Bellot, 2004).

Nous avons testé les différentes approches pour le calcul des chaînes lexicales. L'utilisation d'un hiatus fixe de 11 est le paramètre préconisé par (Galley *et al.*, 2003) avec l'outil $LCseg$. Les résultats montrés dans cet article et rappelés ici dans le tableau 1 affichent de meilleures performances pour cette approche que C99 sur le Brown corpus, ainsi que sur le corpus TDT.

	Brown Corpus	TDT Corpus
$LCseg$	0,1137	0,0909
$C99$	0,1457	0,1272

Table 1: comparaison de $LCseg$ et C99 pour un nombre de segments inconnu, selon la mesure WindowDiff

Taille	LCseg	hiatus 120	hiatus locaux
9-11	0,3272	**0,3187**	0,3454
3-11	0,3837	**0,3685**	0,4016
3-5	0,4344	0,4309	**0,4204**

Table 2: Comparaison de $LCseg$ et de notre méthode pour des segments de différentes tailles (en nombre de phrases)

Nous avons donc décidé de comparer notre approche à celle de (Galley *et al.*, 2003). Les résultats sont présentés dans le tableau 2. Etant donné que les textes ont tous moins de 110 phrases (le maximum étant 10 segments de 11 phrases chacun), le hiatus 120 correspond à une absence de hiatus. Pour ces tests, les lemmes n'ont pas été pondérés en fonction de leur catégorie.

Pour les segments de grande taille (9-11), ou de tailles variables (3-11) , la meilleure méthode est finalement celle qui ne coupe pas les chaînes lexicales (hiatus 120). Pour des segments de petite taille, on observe une très faible amélioration lorsqu'on utilise des hiatus différents pour chaque lemme (hiatus locaux).

3 Exploitation d'une détection d'entités nommées

Si nous avons pris jusqu'ici les termes et leur fonction syntaxique comme seuls critères, nous pensons par ailleurs que la variation des noms propres est un indice intéressant.

On appelle entité nommée dans un texte tout ce qui fait référence à un identifiant unique (Chinchor, 1997). Il peut s'agit un mot ou d'un groupe nominal. Nous avons utilisé trois types d'entités nommées, repérées à partir d'un lexique fermé : listes de noms de personne, noms de lieux et noms d'organisations. Etant un identifiant unique, une entité nommée a tendance à moins se répéter d'un thème à l'autre. De plus, il est moins malvenu en français de les répéter que les noms communs ou adjectifs, même si le problème des anaphores reste à résoudre comme

on le verra.

Nous avons conduit plusieurs types d'expériences (table 3), en fonction de poids plus ou moins élevés attribués aux entités, ou en n'utilisant que les entités. Dans un premier temps nous avons multiplié le poids des chaînes contenant des entités nommées, par deux (ENx2) puis par dix (ENx10). Ensuite nous avons testé la méthode en n'utilisant que les chaînes lexicales correspondant à des entités nommées (EN seules).

Segments 9-11		
Méthode	hiatus 120	hiatus locaux
classique	**0,3187**	0,3454
ENx2	0,3211	0,3536
ENx10	0,3521	0,3888
EN seules	0,4235	0,4975

Segments 3-5		
Méthode	hiatus 120	hiatus locaux
classique	0,4309	0,4204
ENx2	0,4291	**0,4128**
ENx10	0,4315	0,4202
EN seules	0,4228	0,4291

Table 3: Evaluation sur un corpus journalistique avec différentes pondérations des entités nommées pour 2 tailles moyennes des segments

Les résultats présentés dans la table 3 montrent que l'amélioration avec une utilisation des entités nommées est très peu significative d'une part, et qu'il faut doser cet usage d'autre part. En effet on observe une perte de qualité lorsqu'on leur accorde un poids trop important ou lorsque on ne considère qu'elles.

Nous avons ensuite refait les mêmes tests sur un corpus journalistique composé uniquement d'articles traitant de sport, et sur lequel les méthodes testées dans (Sitbon & Bellot, 2004) donnaient les résultats les plus médiocres.

Segments 9-11		
Méthode	hiatus 120	hiatus locaux
classique	**0,3202**	0,3463
ENx2	0,3255	0,3321
ENx10	0,3561	0,3695
EN seules	0,3976	0,4621

Segments 3-5		
Méthode	hiatus 120	hiatus locaux
classique	0,4375	**0,4179**
ENx2	0,4359	0,4183
ENx10	0,4393	0,4265
EN seules	0,4430	0,4634

Table 4: Evaluation sur un corpus journalistique avec différentes pondérations des entités nommées pour 2 tailles moyennes des segments

Les résultats présentés sur la table 4 montrent que l'on n'obtient pas l'amélioration attendue par l'utilisation des entités nommées. Cela peut s'expliquer par une trop fréquente utilisation des anaphores qui limite la répétition des entités. De plus la reconnaissance à l'aide de listes limite le nombre d'entités utilisées, et il faudra recommencer cette étude avec un outil de détection automatique des entités nommées, afin de pouvoir en utiliser un plus grand nombre. On peut également utiliser des cooccurrences d'entités pour créer des chaînes "multi-lexicales".

On constate que les résultats pour un corpus thématiquement cohérent (sport), sont du même ordre que ceux pour un corpus généraliste. Cela tend à montrer que ces méthodes sont indépendantes du type de lexique utilisé dans les documents segmentés, ce qui apporte une forme d'indépendance dans le type de document, et qui était un de nos objectifs initiaux.

4 Conclusion

Les techniques que nous avons imaginées pour s'affranchir du paramètre du hiatus dans l'emploi des chaînes lexicales pour la segmentation thématique sont efficaces. Nous pensons pouvoir encore améliorer la qualité de la segmentation en calculant les probabilités de rupture thématique à partir des similarités, en utilisant des ordonnancement locaux à chaque frontière candidate, comme cela est fait dans C99. Le développement sous forme d'API est en cours, l'outil sera distribué prochainement dans le cadre du projet technolangue AGILE/OURAL (`http://www.technolangue.net/article79.html`).

Références

AMINI M., ZARAGOZA H. & GALLINARI P. (2000). Learning for sequence extraction tasks. In *Proceedings RIAO'2000*, Paris, France.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic anaysis. In *Proceedings of CIKM'02*, McLean, Virginia, USA.

CAILLET M., PESSIOT J.-F., AMINI M. & GALLINARI P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *Proceedings RIAO'04*, Avignon, France.

CHINCHOR N. (1997). Muc-7 named entity task definition. in http://www.itl.nist.gov/iaui/894.02/related projects/muc/proceedings/ne task.html.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, USA.

CHUANG S.-L. & CHIEN L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 127–136, Washington, D.C, USA.

GALLEY M., MCKEOWN K., FOLSER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of ACL'03*, Sapporo, Japan.

MCDONALD D. & CHEN H. (2002). Using sentence selection heuristics to rank text segments in txtractor. In *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*, p. 28–35.

MEKHALDI D., LALANNE D. & INGOLD R. (2004). Using bi-modal alignment and clustering techniques for documents and speech thematic segmentations. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 69–77, Washington, D.C, USA.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, p. 19–36.

SITBON L. & BELLOT P. (2004). Adapting and comparig linear segmentation methods for french. In *Proceedings RIAO'04*, Avignon, France.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *in Meeting of the Association for Computational Linguistics*, p. 491–498.