

TALN 2008, Avignon, 9–13 juin 2008

Comparaison de méthodes lexicales et syntaxico-sémantiques dans la segmentation thématique de texte non supervisée

Alexandre Labadié¹ Violaine Prince¹

(1) LIRMM, 161 rue ADA, 34392 Montpellier cedex
labadie@lirmm.fr, prince@lirmm.fr

Résumé. Cet article présente une méthode basée sur des calculs de distance et une analyse sémantique et syntaxique pour la segmentation thématique de texte. Pour évaluer cette méthode nous la comparons à un un algorithme lexical très connu : c99. Nous testons les deux méthodes sur un corpus de discours politique français et comparons les résultats. Les deux conclusions qui ressortent de notre expérience sont que les approches sont complémentaires et que les protocoles d'évaluation actuels sont inadaptés.

Abstract. This paper present a semantic and syntactic distance based method in topic text segmentation and compare it to a very well known text segmentation algorithm : c99. To do so we ran the two algorithms on a corpus of twenty two French political discourses and compared their results. Our two conclusions are that the two approaches are complementary and that evaluation methods in this domain should be revised.

Mots-clés : Methodes d'évaluation, segmentation de texte, segmentation thématique.

Keywords: Evaluation methods, text segmentation, topic segmentation.

Introduction

Il existe beaucoup de tâches dites de «segmentation de texte». Par exemple, la recherche et l'extraction de textes dans des documents multimédia, où le texte est mélangé à de la vidéo et de l'image, sont des tâches assimilées à la «segmentation de texte» (Karatzas, 2003). Regrouper des mots en morphèmes, ou en unités linguistiques plus importantes, est aussi nommé segmentation de texte (par exemple, dans le traitement des langues asiatiques, qui utilisent des idéogrammes, les frontières des mots sont difficiles à déterminer (Wu & Tseng, 1993), (Yang & Li, 2005)). Dans cet article, nous nous intéressons à la «**segmentation thématique de texte**», c'est à dire à l'opération qui a pour but de trouver la structure thématique (Hearst & Plaunt, 1993) d'un texte et d'en proposer une décomposition par thème (Ponte & Croft, 1997). Si la plupart des textes traitent d'un sujet unique, ils abordent en général plusieurs thèmes en leur sein. Plus le texte est volumineux, plus il est probable que ses thèmes, ou sous-thèmes d'un sujet donné, soient nombreux. Fondamentalement, la segmentation thématique de texte recherche, au sein d'un texte, le début et la fin des thèmes. Pour des raisons pratiques, nous utiliserons pour le reste de cet article le terme «segmentation thématique» plutôt que segmentation thématique de texte.

Si l'on considère que la segmentation thématique doit diviser le document en plusieurs segments cohérents et distincts sur le plan thématique, alors chaque segment ne doit idéalement traiter que d'un seul thème. Mais un thème est une unité complexe sur le plan rhétorique, qui nécessite souvent des digressions, des exemples et des argumentations.

Ce qui nous amène à nous poser la question de la définition de la notion de thème. Dans la littérature, nous en trouvons plusieurs définitions. En général, un thème est : *le sujet d'une conversation ou d'une discussion* (définition du dictionnaire). En linguistique, on le définit comme : *l'élément d'un énoncé qui est réputé connu par les participants à la communication* (on l'oppose souvent au rhème qui est l'information nouvelle apportée par l'énoncé). Nous admettrons ici que le thème d'un segment de texte est *ce dont il parle*. La segmentation thématique doit donc diviser le texte en portions dont chacune des phrases «parle» de la même chose que les autres.

Dans cet article nous comparons Transeg, notre méthode de segmentation thématique basée sur des calculs de distance et une analyse sémantique et syntaxique a priori, à c99, un algorithme de référence à l'heure actuelle dans le domaine. Nous voulons déterminer l'importance des informations d'ordre sémantique et syntaxique portées par les phrases, qui sont les éléments constitutif du texte, dans le cadre de la segmentation thématique. Notre méthode tient compte de cette information, alors que c99 est essentiellement de granularité lexicale. Nous présenterons les deux approches testées ici dans une première partie, pour examiner en détail leur résultats respectifs sur notre corpus de discours politiques. Nous concluons sur la validité des méthodes d'évaluation de segmentation thématique et les possibilités d'évolution de Transeg.

1 Présentation des méthodes comparées.

1.1 C99

Développé par Choi (Choi, 2000), c99 est un algorithme de segmentation thématique s'appuyant fortement sur la notion de cohésion lexicale (Morris & Hirst, 1991). C'est un des algorithmes donnant les meilleurs résultats dans le domaine de la segmentation thématique (Bestgen & Pié-

rard, 2006).

C99 s'appuie sur une matrice de similarité entre les phrases du texte. D'abord projetées dans l'espace des mots du texte, les phrases sont ensuite comparées deux à deux à l'aide d'une mesure de similarité (en général le cosinus) pour former une matrice de similarité. Plus récemment, Choi a amélioré c99 en y adjoignant une analyse sémantique latente (LSA), afin de réduire l'espace des mots à un espace de «concepts» (Choi *et al.*, 2001).

L'originalité de Choi est de ne pas travailler directement sur la matrice de similarité, mais sur une matrice de classement locaux. Chaque case de la matrice est comparée aux cases environnantes (leur nombre dépendant de la taille du masque appliqué, qui est paramétrable), et obtient un score en fonction du nombre de cases ayant un score de similarité inférieur. Evidemment ces *rangs*, ainsi nommés par Choi, sont normalisés par le nombre de cases effectivement comprises dans le masque pour éviter les effets de bord.

La recherche des frontières de thème se fait en optimisant la densité de sous-matrices le long de la diagonale de la matrice de rang, de manière récursive. L'algorithme s'arrête lorsque la frontière idéale désignée par l'algorithme est la dernière phrase de la matrice courante ou, si l'utilisateur fournit un maximum en paramètre à l'algorithme, quand le nombre maximum de frontières est atteint.

La plupart des approches basées sur la cohésion lexicale n'utilisent que peu (voir pas) d'information syntaxique ou sémantique. C99 ne fait pas exception. L'ajout d'une LSA en prétraitement du texte introduit certes un peu de sémantique, mais cela reste au niveau lexical, et ne gomme que partiellement ce défaut d'informations sémantiques. Les informations syntaxiques sont ignorées.

1.2 Transeg

Nous avons développé une méthode de segmentation thématique, appelée Transeg, que nous avons voulu la plus sensible possible aux variations thématiques intra-textuelles. Elle part de l'hypothèse que le texte est constitué par des phrases, qui à leur tour, ne sont pas jetées en vrac, mais produites et ordonnées selon une intention discursive. Dès lors, les informations de position et de construction de ces dernières sont importantes pour la reconnaissance des frontières, aussi bien que de la structure des thèmes qui jalonnent le document.

Les travaux préliminaires qui ont servi à l'élaboration de Transeg ont été évalués lors de du défi DEFT'06 (Azé *et al.*, 2006). Bien qu'encore incomplète, la méthode Transeg a été classée en milieu de tableau face aux différentes autres approches évaluées.

1.2.1 Représentation du texte

La première étape de notre approche est de convertir chaque phrase du texte en un vecteur sémantique. Ce vecteur est obtenu grâce à l'analyseur morpho-syntaxique de la langue française SYGFRAN (Chauché, 1984). Ces vecteurs sont des vecteurs sémantiques à la Roget (Roget, 1852), mais se basant sur le thésaurus Larousse (Larousse, 1992) comme référence. Le vecteur de chaque phrase est calculé de manière récursive en combinant linéairement les vecteurs des constituants de la phrase, eux même obtenus par combinaison linéaire des vecteurs de mots. Le poids de chaque constituant dépend du résultat d'une analyse morpho-syntaxique en constituant et en dépendance¹.

¹La formule est donnée par (Chauché & Prince, 2007)

1.2.2 Segmentation du texte

En nous appuyant sur cette représentation de la phrase, nous avons cherché à identifier ce que nous nommons les «zones de transition» à l'intérieur du texte. La notion de zone de transition vient de l'hypothèse selon laquelle la frontière entre deux thèmes au sein d'un texte n'est pas une phrase unique, mais probablement une courte succession de phrases (2 ou 3). Pour retrouver ces zones de transition nous faisons glisser une fenêtre le long du texte. Cette fenêtre, d'une largeur fixe de 20 phrases (sa taille reste paramétrable), représente un enchaînement supposé de deux segments thématiques. Chaque moitié de la fenêtre est donc considérée comme un segment thématique potentiel. On calcule alors un centroïde pour chacun d'entre eux. Ce centroïde est un barycentre pondéré, ce qui nous permet d'incorporer un peu d'information stylistique. En effet, les premières phrases et les introductions comportent très souvent plus d'informations pertinentes que les autres phrases (Labadié & Chauché, 2006), (Lelu *et al.*, 2006). Les poids de chaque phrase sont donc calculés selon une régression linéaire donnant plus d'importance aux premières phrases d'un segment comparativement aux dernières. Finalement, nous calculons une distance (que nous nommons distance thématique) entre ces deux barycentres. Cette distance est attribuée comme score de transition à la phrase du milieu de la fenêtre (la frontière potentielle donc, voir figure 1). Le choix d'une fenêtre de 20 phrases se base sur l'observation empirique de la taille moyenne d'un segment sur les différents corpus fournis lors de l'évaluation DEFT'06 (Azé *et al.*, 2006), qui est d'environ 10 phrases (10, 12). On notera que le masque utilisé par Choi dans son algorithme est par défaut de 11 phrases, les deux méthodes envisagent donc une taille moyenne du segment thématique proche de 10 phrases.

Les zones de transition sont donc des phrases successives avec un score de transition supérieur à un seuil déterminé. Ce seuil est le résultat d'une observation détaillée du corpus de discours politiques fourni lors de l'évaluation DEFT'06 (Azé *et al.*, 2006). Nous avons calculé les distances qui séparent des segments thématique de bon nombre de discours (plus de 100000 phrases au total) et nous avons trouvé une distance moyenne de 0.45 pour un écart type de 0.08. Une fois les zones de transition identifiées, on sélectionne les phrases frontières en leur sein. Une description plus détaillée de notre approche est disponible dans (Prince & Labadié, 2007). Dans notre première mise en œuvre de la méthode, nous utilisons la distance angulaire pour le calcul du score de transition. Dans cet article, nous utilisons une version modifiée de la distance de concordance proposée par (Chauché *et al.*, 2003).

1.2.3 La distance de concordance

Les vecteurs sémantiques issus de l'analyse par SYGFRAN ont 873 composantes, et la grande majorité de ces dernières ne sont pas activées pour une phrase donnée. Avec autant de valeurs nulles, la distances angulaire n'est pas assez discriminante. La distance de concordance a pour objectif d'être plus discriminante en se concentrant sur les composantes les plus activées et leur classement relatif.

Considérons deux vecteurs \vec{A} et \vec{B} , nous classons leurs composantes de la plus activée à la moins activée et ne conservons que les premières valeurs ($\frac{1}{3}$ de la taille initiale du vecteur). \vec{A}_{tr} et \vec{B}_{tr} sont les versions triées et réduites de respectivement \vec{A} et \vec{B} . Comme nous ne conservons que les composantes les plus fortes de chaque vecteur, \vec{A}_{tr} et \vec{B}_{tr} peuvent très bien ne pas avoir de composantes en commun (dans ce cas la distance qui les sépare sera de 1). Dans le cas où \vec{A}_{tr} et \vec{B}_{tr} ont au moins une composante en commun nous pouvons calculer deux différences :

La différence de rang : Si i est le rang de C_t une composante de \vec{A}_{tr} et $\rho(i)$ le rang de la même

Comparaison de méthodes lexicale et syntaxico-sémantique dans la segmentation thématique non supervisée

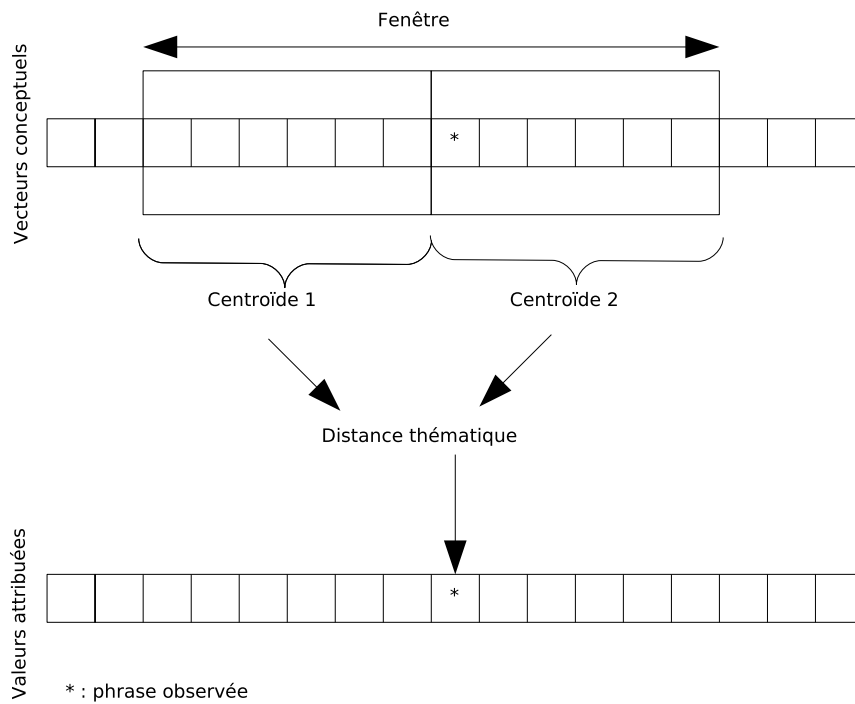


FIG. 1 – Attribution d'un score de transition

composante dans \vec{B}_{tr} , alors nous avons la différence de rang :

$$E_{i,\rho(i)} = \frac{(i - \rho(i))^2}{Nb^2 + (1 + \frac{i}{2})} \quad (1)$$

Où Nb est le nombre de composantes conservées.

La différence d'intensité : Il nous faut également comparer la différence d'intensité des différentes composantes communes. Pour cela nous considérons a_i l'intensité de la composante de rang i dans \vec{A}_{tr} et $b_{\rho(i)}$ l'intensité de la même composante dans \vec{B}_{tr} (et dont le rang est $\rho(i)$), alors nous avons la différence d'intensité :

$$I_{i,\rho(i)} = \frac{\|a_i - b_{\rho(i)}\|}{Nb^2 + (\frac{1+i}{2})} \quad (2)$$

Ces deux différences nous permettent de calculer la concordance :

$$P(\vec{A}_{tr}, \vec{B}_{tr}) = \left(\frac{\sum_{i=0}^{Nb-1} \frac{1}{1 + E_{i,\rho(i)} * I_{i,\rho(i)}}}{Nb} \right)^2 \quad (3)$$

Toutefois, la concordance P se concentre sur l'intensité et le rang des composantes et n'a pas la notion de direction que possède la distance angulaire. Nous introduisons donc la notion de direction en combinant la concordance avec la distance angulaire. Ainsi si $\delta(\vec{A}, \vec{B})$ est la distance angulaire entre \vec{A} et \vec{B} , nous avons :

$$\Delta(\vec{A}_{tr}, \vec{B}_{tr}) = \frac{P(\vec{A}_{tr}, \vec{B}_{tr}) * \delta(\vec{A}, \vec{B})}{\beta * P(\vec{A}_{tr}, \vec{B}_{tr}) + (1 - \beta) * \delta(\vec{A}, \vec{B})} \quad (4)$$

Où β est un coefficient donnant plus ou moins de poids à P . Il est aisé de prouver que $\Delta(\vec{A}_{tr}, \vec{B}_{tr})$ n'est pas symétrique.

$\Delta(\vec{A}_{tr}, \vec{B}_{tr})$ a été conçu au départ dans un contexte de classification, afin de comparer des vecteurs de textes à des vecteurs de classes. Comme seule la proximité entre le texte et la classe était importante la symétrie n'était pas nécessaire. Dans notre contexte de segmentation de texte, où un segment n'est pas plus important que celui qui le précède ou lui succède, la symétrie est indispensable. Ainsi la distance de concordance symétrique vaut :

$$D(\vec{A}, \vec{B}) = \frac{\Delta(\vec{A}_{tr}, \vec{B}_{tr}) + \Delta(\vec{B}_{tr}, \vec{A}_{tr})}{2} \quad (5)$$

2 Expérience : segmentation thématique de vingt deux discours politiques français.

Afin de mesurer l'efficacité de Transeg nous l'avons comparé à un algorithme reconnu et éprouvé, c99 de (Choi *et al.*, 2001). Les deux algorithmes sont non supervisés et donc indépendants des données (ils n'effectuent ni apprentissage, ni adaptation aux données et donc l'usage répété d'un même corpus n'aura pas d'influence sur les résultats). Les tests ont été effectués sur un corpus de vingt deux discours politiques extrait du corpus d'apprentissage fourni lors de l'atelier DEFT'06 (Azé *et al.*, 2006). Nous décrivons dans cette section la préparation des données, les conditions de l'expérience et nous commentons les résultats.

2.1 Présentation des données : Un corpus de vingt deux discours politiques français

Notre choix s'est porté sur le corpus de discours politiques fourni par l'atelier DEFT'06 pour deux raisons principales :

- Les frontières thématiques au sein des discours ont été identifiées par des personnes pouvant être considérées comme «expertes» dans le domaine (le personnel travaillant à la rédaction et à la mise en ligne des discours présidentiels). Ainsi ces frontières thématiques paraissent moins artificielles que des débuts de texte concaténés ou des débuts de paragraphe, comme c'est le cas en général quand on essaie d'évaluer des méthodes de segmentation.
- En tant que textes argumentatifs par excellence, les discours politiques offrent, en général, une structure thématique claire.

En d'autres termes, ce corpus est tout à fait approprié pour une réelle évaluation de la segmentation thématique de textes. Malheureusement, le corpus initial proposé par DEFT'06 était extrêmement bruité. Certains discours étaient uniquement en lettres capitales par exemple (ce qui est préjudiciable dans une langue comme le français qui utilise beaucoup d'accents et s'en sert pour la reconnaissance lexicale), d'autres sont en fait des interviews. Il a donc été nécessaire de sélectionner et de nettoyer manuellement vingt deux discours dans un ensemble de plusieurs centaines de textes concaténés.

Sur un corpus initial de plus de 300000 phrases (de qualité douteuse) nous avons donc extrait 22 discours totalisant 1895 phrases et 54551 mots. Nous ne disposons d'aucune information quant au début des discours au sein de cet ensemble (en dehors des début de segments thématiques

qui pouvaient être indifféremment des débuts de discours, comme de simples frontières thématiques). L'opération a donc pris beaucoup de temps et a considérablement réduit la masse de données sur laquelle nous avons mené l'expérience, mais c'était nécessaire afin de disposer d'un jeu de données propre, et d'éviter les biais expérimentaux qui pourraient entacher l'objectivité des résultats.

2.2 Présentation de l'expérience : Comparaison des méthodes

Nous avons donc comparé Transeg et c99 sur ce jeu de données. Afin d'être sûr de ne pas faire d'erreur de programmation, nous avons utilisé la version 1.3 de c99 fournie par Choi lui même (<http://www.lingware.co.uk/homepage/freddy.choi/software/software.htm>). Cette version de c99 bénéficie de l'amélioration LSA présentée dans (Choi *et al.*, 2001).

Nous avons lancé les deux algorithmes sur les 22 textes, sans jamais les concaténer. En effet, considérer la reconnaissance d'un texte comme la reconnaissance d'un thème ne nous paraît pas être de nature à rendre compte de la variation thématique intratextuelle. Or malheureusement, dans la majorité des campagnes d'évaluation, on ne fait pas de différence entre les deux cas. En séparant bien les textes, c'est vraiment la segmentation intra-textuelle que nous avons isolée et testée, évitant ainsi des biais expérimentaux.

Pour comparer les résultats, nous avons utilisé le rappel et la précision avec fenêtre de tolérance présentés dans (Azé *et al.*, 2006). Ils comptent comme correctes des phrases que les algorithmes ramèneraient et qui seraient juste avant ou juste après la phrase identifiée par l'expert comme étant la frontière. Dans le cadre de l'expérience, la fenêtre était de deux phrases avant ou après. L'équipe de DEFT'06 a constaté que l'usage de ces mesures ne changeait pas le classement des méthodes présentées par rapport à un rappel et une précision stricts. Cette tolérance permet de ne pas sanctionner un algorithme qui sélectionne comme frontière possible une phrase juste à coté de la frontière identifiée par l'expert.

A partir de ce rappel et de cette précision nous calculons un *FScore* selon la formule bien connue :

$$FScore = \frac{(\beta^2 + 1) * rappel * precision}{\beta^2 * precision + rappel} \quad (6)$$

Avec $\beta = 1$.

On notera tout de même que c99 comme Transeg considère toujours la première phrase d'un texte comme une frontière thématique, et que lors de notre évaluation nous considérons cette réponse comme correcte. Aussi pour chaque texte les deux méthodes ont au moins un retour correct.

2.3 Résultats : avantage Transeg

Le premier constat que nous faisons à la vue des résultats, est qu'ils sont plutôt décevants en terme de *FScore* quelle que soit la méthode utilisée. Le *FScore* est une mesure très stricte, et même en utilisant un rappel et une précision tolérante nous obtenons au mieux un *FScore* de 42.86 (pour Transeg sur le texte 9 ; nous donnons ici les valeurs sous forme de pourcentage, pour des raisons de lisibilité). Cela nous permet de constater que la marge de progression dans le domaine de la segmentation thématique demeure importante.

Plus en détail, nous constatons que Transeg a un meilleur *FScore* sur 16 des 22 textes considérés. Sur ces 16 textes, Transeg a toujours un rappel supérieur ou égal à celui de c99, et son *FScore* est au minimum supérieur de 20% à celui de c99 (texte 1) pour aller à plus de 4 fois supérieur (texte 9). On notera également que Transeg possède également le meilleur résultat sur

l'ensemble des textes.

C99 dépasse Transeg dans 6 textes, mais avec un *F Score* au maximum deux fois supérieur à celui de Transeg. Toutefois on notera que c99 est en général toujours performant en terme de précision comparativement à Transeg. Ce résultat s'explique facilement en étudiant les conditions de l'expérience. En effet les deux méthodes ramènent toujours au moins la première phrase, qui est toujours comptabilisée comme correcte. C99 ramène beaucoup moins de phrases que Transeg, avec au moins une phrase de juste. Le calcul de la précision lui est donc favorable (les textes où c99 a une précision de 100 correspondent à des textes où l'algorithme n'a ramené que la première phrase).

Transeg est une méthode développée pour détecter les faibles variations intra-textuelles. Logiquement elle obtient, en général, un fort rappel et sa précision en pâtit quelque peu. Parfois trop sensible, Transeg a tendance à sur-segmenter. A l'opposé, c99 est un algorithme qui favorise la précision en ramenant peu de phrases et donc sous-segmente. On peut imputer cette tendance au fait que bon nombre de méthodes de segmentation thématique sont développées et testées sur des corpus très artificiels où le but est de retrouver des débuts de paragraphes, voir des débuts de textes courts dans un ensemble de textes concaténés. C'est encore plus évident si l'on regarde en détail les 6 textes où c99 a de meilleurs résultats. Ce sont soit des textes très courts, avec très peu de frontières thématiques identifiées, soit des énumérations sans réelle structure. Par exemple le texte 6 est un discours du porte-parole du gouvernement à l'issue d'un conseil des ministres. Ce discours n'est que l'énumération des différents sujets traités durant le conseil. Les mots sont très différents, et les sujets sont courts, ce qui est dans le domaine de compétence de c99.

3 Conclusion

D'après l'expérience décrite dans cet article, et que nous avons voulu la plus objective possible, Transeg semble plus performant que c99 lorsqu'il s'agit de retrouver les frontières thématiques intra-textuelles. Nous pouvons donc en déduire que l'usage d'informations sémantiques et syntaxiques a un effet appréciable sur la qualité de segmentation thématique d'un texte, lorsque celui-ci a une certaine taille, et qu'il est construit (avec des qualités stylistiques et rhétoriques). En revanche, ces informations peuvent parfois être trop sensibles et nous amener à une sursegmentation par rapport aux frontières données en référence. Ce qui nous amène légitimement à nous poser les questions suivantes : Les frontières thématiques ont été identifiées par des personnes pouvant raisonnablement être considérées comme des experts du domaine, mais toutes les frontières sont elles bonnes ? En faudrait il plus ? Ou moins ? Ou même, une question peut-être plus pertinente, leur solution est-elle l'unique solution ?

Le fait même de n'avoir aucune certitude sur le sujet pourrait remettre en cause la validité de l'évaluation absolue, celle qui se fait automatiquement, ou quasi-automatiquement, en rapport avec «une valeur de référence». La segmentation thématique est, par essence, subjective, et l'on peut dire la même chose d'autres domaines du traitement automatique de la langue. S'il faut évaluer, il serait plus approprié de proposer une procédure plus relative. Nous envisageons, à l'heure actuelle, d'autres manières d'évaluer nos résultats, en se basant, par exemple sur des avis à posteriori (faire «noter» nos résultats par des experts ou même des utilisateurs). Plutôt que d'affirmer la supériorité intrinsèque de tel ou tel outil, il serait plus adéquat d'en constater la plus grande adaptation, la plus grande souplesse, la plus grande satisfaction d'usage, etc.

Pour terminer, on pourrait remarquer que, toutes choses étant égales par ailleurs, c99 et Tran-

Comparaison de méthodes lexicale et syntaxico-sémantique dans la segmentation thématique non supervisée

	Nb. mots	Nb. phrases	Transeg			c99		
			Précision	Rappel	FScore	Précision	Rappel	FScore
Text 1	617	22	50	33.33	20	33.33	33.33	16.67
Text 2	3042	100	33.33	37.5	17.65	50	12.5	10
Text 3	2767	92	42.86	85.71	28.57	20	14.29	8.33
Text 4	1028	40	33.33	33.33	16.67	20	33.33	12.5
Text 5	4532	157	12.5	18.18	7.41	16.67	9.09	5.88
Text 6	5348	212	8.7	18.18	5.88	20	18.18	9.52
Text 7	1841	47	100	42.86	30	100	14.29	12.5
Text 8	1927	74	60	33.33	21.43	100	11.11	10
Text 9	1789	53	75	100	42.86	25	16.67	10
Text 10	1389	31	33.33	20	12.5	100	20	16.67
Text 11	2309	81	30	50	18.75	33.33	16.67	11.11
Text 12	7193	211	15.38	6.25	4.44	33.33	3.13	2.86
Text 13	6097	305	20.59	33.33	12.73	17.65	14.29	7.89
Text 14	1417	57	40	33.33	18.18	100	16.67	14.29
Text 15	3195	79	40	8	6.67	66.67	8	7.14
Text 16	1995	60	66.67	28.57	20	57.14	57.14	28.57
Text 17	558	16	33.33	33.33	16.67	50	66.67	28.57
Text 18	696	25	100	37.5	27.27	40	25	15.38
Text 19	678	26	33.33	33.33	16.67	50	66.67	28.57
Text 20	1388	57	50	66.67	28.57	100	16.67	14.29
Text 21	3127	110	62.5	25	17.86	40	10	8
Text 22	1618	40	60	75	33.33	100	25	20

TAB. 1 – Comparaison transeg / c99

seg font montre de propriétés complémentaires. A eux deux, ils couvrent (avec des améliorations futures à envisager) les possibilités de segmentation. Une fusion immédiate entre les deux méthodes étant paradigmatiquement et techniquement peu envisageable, une piste intéressante pourrait être celle d'un logiciel qui permettrait de détecter automatiquement lequel des deux algorithmes lancer en fonction des propriétés des textes à segmenter. C'est une étude qui reste à faire, et qui permettrait aussi d'examiner les améliorations en termes de performances à apporter aux deux algorithmes.

Références

- AZÉ J., HEITZ T., MELA A., MEZAOUR A., PEINL P. & ROCHE M. (2006). Présentation de deft'06 (defi fouille de textes). *Proceedings of DEFT'06*, **1**, 3–12.
- BESTGEN Y. & PIÉRARD S. (2006). Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matériel de référence. *Proceedings of TALN'06*.
- CHAUCHÉ J. (1984). Un outil multidimensionnel de l'analyse du discours. *Proceedings of Coling'84*, **1**, 11–15.
- CHAUCHÉ J. & PRINCE V. (2007). Classifying texts through natural language parsing and semantic filtering. *In Proceedings of LTC'03*.

- CHAUCHÉ J., PRINCE V., JAILLET S. & TEISSEIRE M. (2003). Classification automatique de textes à partir de leur analyse syntaxico-sémantique. *Proceedings of TALN'03*, p. 55–65.
- CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. *Proceedings of NAACL-00*, p. 26–33.
- CHOI F. Y. Y., WIEMER-HASTINGS P. & MOORE J. (2001). Latent semantic analysis for text segmentation. *Proceedings of EMNLP*, p. 109–117.
- HEARST M. A. & PLAUNT C. (1993). Subtopic structuring for full-length document access. *Proceedings of the ACM SIGIR-93 International Conference On Research and Development in Information Retrieval*, p. 59–68.
- KARATZAS D. (2003). *Text Segmentation in Web Images Using Color Perception and Topological Features*. UK : ECS Publications.
- LABADIÉ A. & CHAUCHÉ (2006). Segmentation thématique par calcul de distance sémantique. *Proceedings of DEFT'06*, **1**, 45–59.
- LAROUSSE (1992). *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Paris : Larousse.
- LELU A., M. C. & AUBAIN S. (2006). Coopération multiniveau d'approches non-supervisées et supervisées pour la détection des ruptures thématiques dans les discours présidentiels français. *In Proceedings of DEFT'06*.
- MORRIS J. & HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, **17**, 20–48.
- PONTE J. M. & CROFT W. B. (1997). Text segmentation by topic. *European Conference on Digital Libraries*, p. 113–125.
- PRINCE V. & LABADIÉ A. (2007). Text segmentation based on document understanding for information retrieval. *In Proceedings of NLDB'07*, p. 295–304.
- ROGET P. (1852). *Thesaurus of English Words and Phrases*. London : Longman.
- WU Z. & TSENG G. (1993). Chinese text segmentation for text retrieval : Achievements and problems. *Journal of the American Society for Information Science*, **44**, 532–542.
- YANG C. C. & LI K. W. (2005). A heuristic method based on a statistical approach for chinese text segmentation. *Journal of the American Society for Information Science and Technology*, **56**, 1438–1447.