Condensés automatiques de textes

Juan-Manuel Torres-Moreno

École Polytechnique / Département de génie informatique
Équipe de recherche en micro-électronique et traitement informatique des signaux (ERMETIS), Université du Québec à Chicoutimi
Laboratoire d'ANalyse Cognitive de l'Information (LANCI), Université du Québec à Chicoutimi
juan-manuel.torres@polymtl.ca

Patricia Velázquez-Morales

Laboratoire d'ANalyse Cognitive de l'Information (LANCI), Université du Québec à Chicoutimi

Jean-Guy Meunier

Laboratoire d'ANalyse Cognitive de l'Information (LANCI), Université du Québec à Chicoutimi
meunier.jean-guy@uqam

ABSTRACT. Summarizing is a critical phase in the automatic analysis of texts. Abstract generation is a complex cognitive process. The state of art only allows the production of document condensations. This paper describes our method, Cortex, which uses an algorithmic-numerical approach to obtain a text abstract. The final condensation is independent of the subject and the size of the corpus. Cortex offers the added capability of producing condensations in French or Spanish in a short period of time.

KEYWORDS : Text condensation, automatic summarizing, text analysis, statistical methods.

RÉSUMÉ. L'obtention de résumés de textes constitue une phase critique dans l'analyse automatique de textes. La génération de résumes étant un processus cognitif difficile, l'état de l'art ne permet d'obtenir que des condensés des documents. Cet article décrit notre méthode Cortex, basé sur une approche numérique algorithmique, pour l'obtention d'un condensé d'un texte. Le condensé ainsi obtenu est indépendant du thème et de l'ampleur du corpus. Le système trouve en plus, des condensés de textes en français ou espagnol très rapidement.

MOTS-CLÉ : Condensés de textes, résumés automatiques, analyse de textes, méthodes statistiques.

---

1. Introduction

L'obtention des résumes s'avère très important, car le volume d'informations généré est de plus en plus important. L'élaboration des méthodes d'obtention de résumés automatiques de textes constitue une phase cruciale dans l'analyse automatique de textes. Le condensé est le premier pas vers l'obtention d'un vrai résumé, qui est la forme concrète la plus connue et la plus visible de la condensation de textes. L'utilisation des méthodes linguistiques est certes pertinente, mais leur utilisation concrète demeure

encore difficile ou limitée à des domaines restreints [SAGGION 00]. L'approche vectorielle des textes [SALTON 83] peut être utilisée pour obtenir des condensés de documents. Les méthodes statistiques et neuronales sont de plus en plus utilisées dans plusieurs domaines du traitement de l'information textuelle [SALTON 71 ; SALTON 83 ; DEERWESTER 90 ; LELOUP 97 ; VERONIS 91 ; BALPE 96 ; TORRES-MORENO 00 ; MEUNIER 98 ; MEUNIER 00]. Nos recherches portent sur l'obtention de résumés de type informatif, le seul type de condensés que l'état actuel de l'art permet d'obtenir [MORRIS 99]. Nous présentons un algorithme développé récemment, qui combine plusieurs traitements statistiques et informationnels avec un algorithme optimal de décision, pour choisir des phrases pertinentes du texte à traiter. L'ensemble de ces phrases constitue ce qu'on appelle le condensé du document.

2. Pré-traitement

Dans l'approche vectorielle on traite de textes en passant par une représentation numérique très différente d'une analyse structurale linguistique, mais qui permet des traitements performants [MEMMI 00]. Les textes sont représentés dans un espace vectoriel, et on leur applique des traitements numériques. Nous nous sommes inspirés de Conterm [SEFFAH 96, MEMMI 00] pour créer un module de pré-traitement qui comporte des processus de filtrage, de segmentation et de lemmatisation. Le texte original possède N_M termes incluant des mots fonctionnels, des noms, des verbes fléchis et des mots composés. On emploie la notion de terme pour designer un mot plus abstrait. Pour réduire la complexité, des processus de réduction de filtrage du lexique sont amorcés. Nous avons supprimé des mots et des verbes fonctionnels (être, avoir, pouvoir, falloir, …), des mots à haute et à très basse fréquence d'apparition, ainsi que le texte entre parenthèses, des chiffres et des symboles tels que @, %, #, etc. La lemmatisation de verbes est un processus très important pour la réduction du lexique, car elle permet de diminuer le phénomène de la *malédiction dimensionnelle* qui pose de sérieux problèmes dans des textes de grandes dimensions. La segmentation est faite en utilisant des séparateurs (<!>,<?>,<.>,<:>). Le critère de segments à taille fixée a été écarté, car on cherchait l'extraction des phrases complètes. Nos expériences ont été réalisées sur des textes bruts ; les titres, sous-titres et sections ne sont donc pas marqués explicitement. Mais nous considérons le titre comme un segment additionnel. Après le pré-traitement, le nouveau texte comporte P segments avec N_f termes totaux.

3. Condensation du texte

La segmentation transforme un document dans un ensemble de vecteurs :

$$\square = (\square_1, \square_2, ..., \square_{NM}) = \{0,1\}^{NM}$$

Chaque segment de texte est représenté par un vecteur à composantes binaires. La dimension N_M est le nombre total de mots différents dans le texte. L'ensemble des segments dont on dispose consiste en P vecteurs, et la matrice $X_i = X_i^{\square}$; $\square$ =1,…,P représente le texte. Seul les termes à fréquence supérieure à 1 ont été utilisés (TOR01), et donc un lexique de N_L termes est obtenu. On garde la relation $N_L < N_f < N_M$. Nous avons

donc défini $\square_L = N_L / N_M$, comme le ratio de réduction du lexique filtré/lemmatisé. La matrice Terme-Segment $\square = \square^{\square}$; $\square$=1,...,P, dérivée de $\square$ représente le lexique réduit du texte :

$$\xi = \begin{bmatrix} \xi_1^1 & \xi_2^1 & \xi_3^1 & \cdots & \xi_{N_L}^1 \\ \xi_1^2 & \xi_2^2 & \xi_3^2 & \cdots & \xi_{N_L}^2 \\ \vdots & & \ddots & & \vdots \\ \xi_1^\mu & \xi_2^\mu & \xi_3^\mu & \cdots & \xi_{N_L}^\mu \\ \vdots & & \ddots & & \vdots \\ \xi_1^P & \xi_2^P & \xi_3^P & \cdots & \xi_{N_L}^P \end{bmatrix}$$

Dans cette matrice chaque composante montre la présence (ξi^μ=1) ou l'absence (ξi^μ=0) du mot i dans un segment μ. De façon analogue, la matrice fréquentielle :

$$\gamma = \begin{bmatrix} \gamma_1^1 & \gamma_2^1 & \gamma_3^1 & \cdots & \gamma_{N_L}^1 \\ \gamma_1^2 & \gamma_2^2 & \gamma_3^2 & \cdots & \gamma_{N_L}^2 \\ \vdots & & \ddots & & \vdots \\ \gamma_1^\mu & \gamma_2^\mu & \gamma_3^\mu & \cdots & \gamma_{N_L}^\mu \\ \vdots & & \ddots & & \vdots \\ \gamma_1^P & \gamma_2^P & \gamma_3^P & \cdots & \gamma_{N_L}^P \end{bmatrix}$$

où chaque composante $\gamma=(\gamma_1,\gamma_2,...,\gamma_{NL})$ contient la fréquence γi^μ du terme i dans un segment μ. Cette matrice contient l'information fréquentielle essentielle du texte. La condensation du texte s'effectue à partir de ces deux matrices : l'espace des entrées. Nous avons défini la taille réduite des matrices γ et ξ comme $\alpha=P/N_L$, qui représente la proportion P de segments par rapport à la dimension N_L du lexique réduit à l'entrée. Les segments possèdent une quantité hétérogène du lexique : il y a des segments plus importants que d'autres, qui devront être extraits par l'algorithme pour obtenir un condensé. Notre méthode Cortex comprend donc une méthode de construction des métriques informationnelles indépendantes et un algorithme pour la récupération de l'information codée. Ce dernier prendra une décision sur les segments à choisir en fonction d'une stratégie des votes.

3.1 Métriques

Des informations mathématiques et statistiques importantes sont calculées à partir des matrices ξ et γ sous forme de neuf métriques [TORRES-MORENO 02]. Elles mesurent la quantité d'information contenue dans chaque segment : plus une métrique est importante, plus elle comporte des valeurs élevées. Des indicateurs de repérage explicites, comme les titres, sont aussi utilisés. Les mots du titre par exemple, sont considérés comme ayant une grande valeur informationnelle et pondérés en conséquence. Les métriques étant décrites en (TOR02, TOR01), nous allons tout simplement les énumérer ici :

- Les distances d'Hamming entre paires des mots
- Les poids d'Hamming lourd des segments
- Le poids d'Hamming des segments
- La somme des probabilités par fréquence
- Les interactions entre mots
- La somme des poids d'Hamming
- La fréquence relative des mots
- L'entropie des segments
- La somme fréquentielle des poids d'Hamming.

3.2 Algorithme de décision

Nous avons développé un algorithme pour récupérer l'information codée par les métriques. Le problème se pose comme suit : étant donné les votes pour un événement particulier qui provient d'un ensemble de k votants indépendants, chacun avec une certaine probabilité d'avoir raison, trouver la décision optimale. Notre méthode appelée Algorithme de décision (AD) [TORRES-MORENO 01], utilise deux probabilités mutuellement exclusives : p_0 et p_1. On présente les k votants en modifiant p_0 et p_1 en fonction des sorties normalisées π_j; j=1,…,k pour chacun des segments. L'algorithme de décision possède deux propriétés intéressantes : convergence et amplification. Les probabilités p_0 et p_1 sont modifiées en tout temps de façon mutuellement exclusive. Il est amplificateur parce que la probabilité de choisir un segment pertinent est au moins égale à la probabilité π_j du meilleur votant branché à ce moment. Le processus complet de génération des résumés selon notre approche, est illustré à la figure 1.

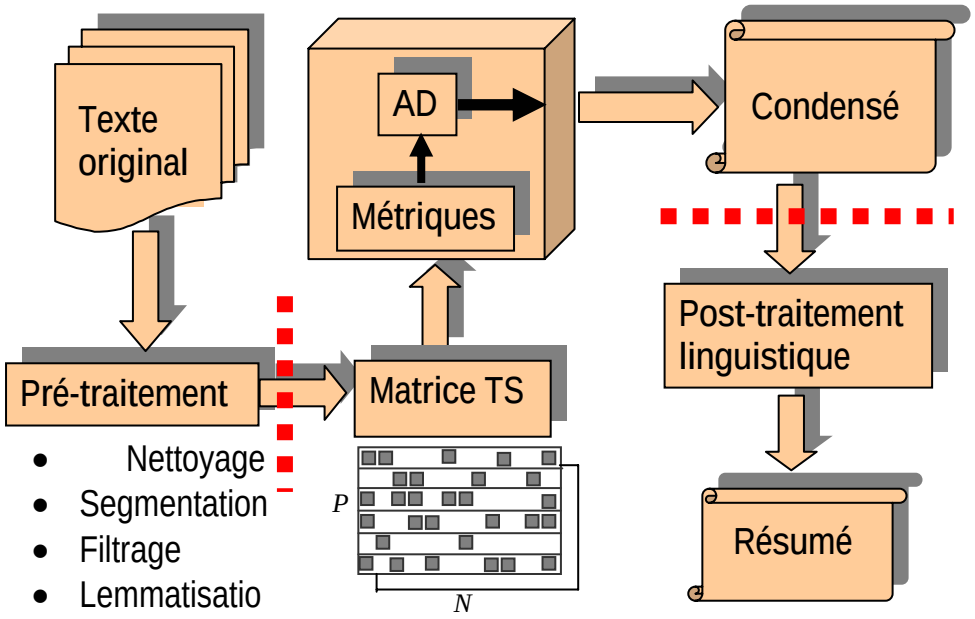

Figure 1. Génération automatique des résumés.

4. Résultats

Nous avons effectué des tests sur des articles de vulgarisation scientifique de petite taille, extraits de la presse sur Internet. L'objectif a été d'obtenir un condensé de 25% du nombre de segments totaux. Nous avons comparé nos résultats avec les condensés obtenus par d'autres logiciels : Minds (http://messene.nmsu.edu/minds/ SummarizerDemoMain.html), Summarizer (http://www.copernic.com), Word et Pertinence (www.pertinence.com). Dans les cas de Minds et Word, le paramètre utilisé a été d'obtenir une synthèse de 25% de la taille du texte. Pour Pertinence il nous a fallu choisir l'option de condensés d'entre 30% et 40% afin d'obtenir un nombre de segments similaires aux autres systèmes. Nous avons aussi demandé à 14 personnes de faire un condensé à la main en choisissant les phrases du texte qui leur semblaient les plus pertinentes. Tous nos sujets ont un niveau d'étude universitaire et sont habitués à faire des résumés.

4.1 Textes en français

Nous avons étudié le texte « Puces », téléchargeable à l'adresse électronique www.professeurs.polymtl.ca/juan-manuel.torres/homepage/investigacion/cortex, texte qui est artificiellement ambigu et composé d'un mélange non homogène de deux textes : sujets « puces biologiques » et « puces informatiques » dans le cadre de la classification de segments par leur contenu [TORRES-MORENO 00]. Les segments les plus importants sélectionnés par les humains ont été le 2, 5, 15 et le 17 (figure 2). Nous montrons à la figure 3 nos résultats, où on voit que ces segments importants ont été bien

repérés. Cortex montre un résumé équilibré, de même que celui obtenu par Minds, même si ce dernier ne trouve ni le segment 5 ni le 15 (voir la figure 4). Par contre, les résultats de Word sont biaisés et peu pertinents, comme on le voit à la figure 5.

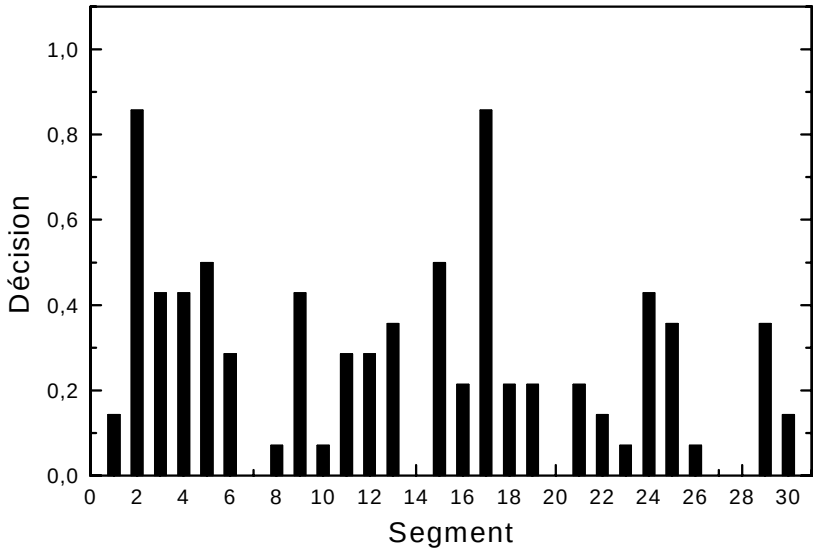

Figure 2. Choix de segments pertinents pour le texte « Puces » obtenu par les 14 sujets humains.

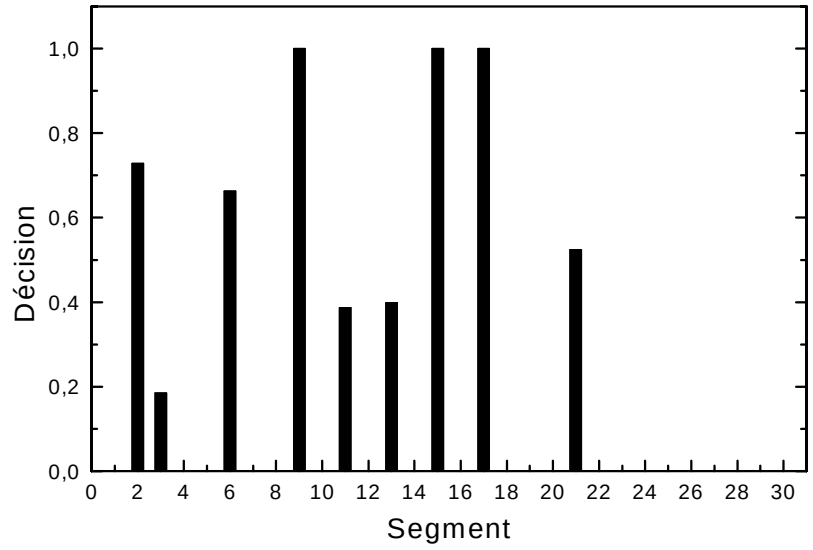

Figure 3. Choix de segments pertinents pour le texte « Puces » obtenu par le système Cortex : les segments 2 et 17 ont été bien choisis.

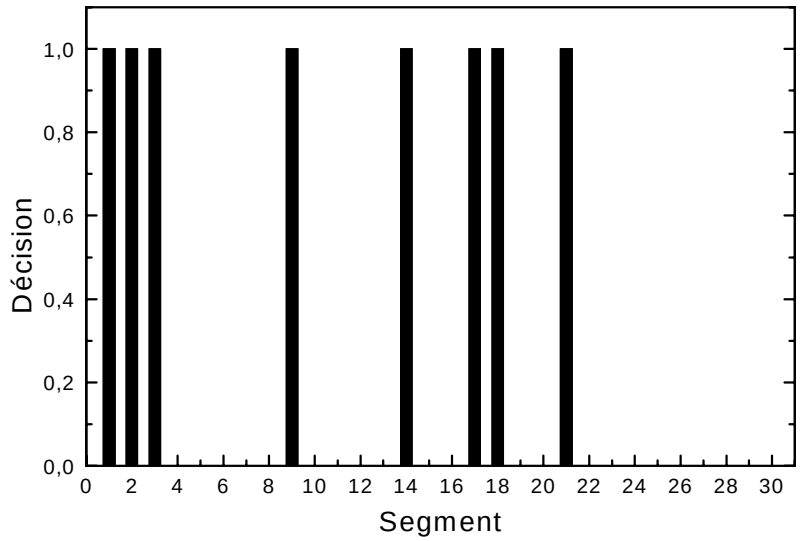

Figure 4. Choix de segments pour le texte « Puces » obtenu par le système Minds.

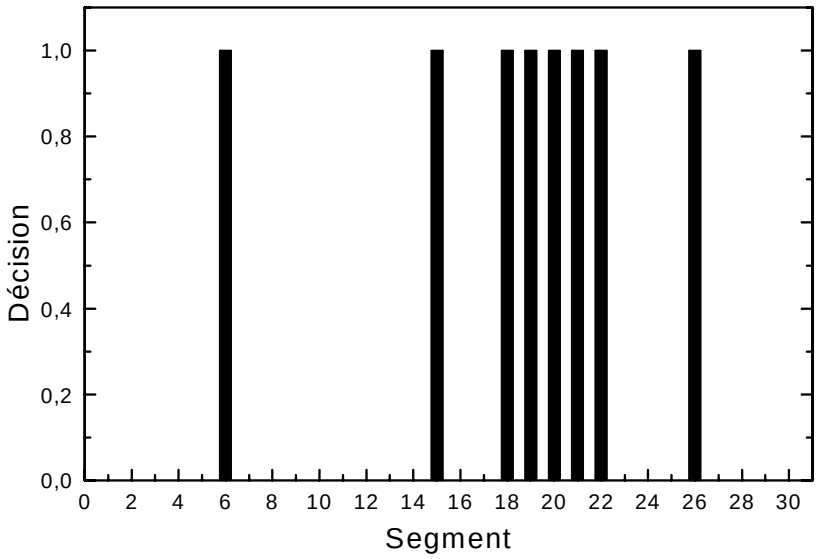

Figure 5. Choix de segments pour le texte « Puces » obtenu par le synthétiseur Word.

Pour le texte « Fêtes » (http://www.quebecmicro.com/6-12/6-12-28.html), les résultats préliminaires montrent que Cortex génère des condensés acceptables. Nous avons effectué des comparaisons avec Summarizer, et nos condensés sont comparables, voire de meilleure qualité [HUOT 01]. Dans d'autres tests les condensés trouvés par Cortex semblent être assez cohérents. Nous avons toujours constaté que les condensés obtenus par les sujets humains dépendent de l'expertise de la personne et de sa capacité d'abstraction, ce qui donne parfois des résultats assez différents. Malgré cela le choix fait

par des humains semble être une référence sur les segments importants et notre méthode donne des résultats comparables.

4.2 Textes en espagnol

Nous avons travaillé deux textes : « Nopal » (http://www.invdes.com.mx/suplemento/anteriores/Noviembre2000/htm/espina.html) et « Tabaco » http://www.invdes.com.mx/suplemento/anteriores/Diciembre2000/htm/tabaco.html). Les résultats sur « Nopal » sont illustrés aux figures 6 et 7. On constate la bonne qualité du condensé, même dans des textes contenant peu de mots. « Nopal » possède seulement N_L=5 mots, P=7 segments et Word est incapable de le traiter.

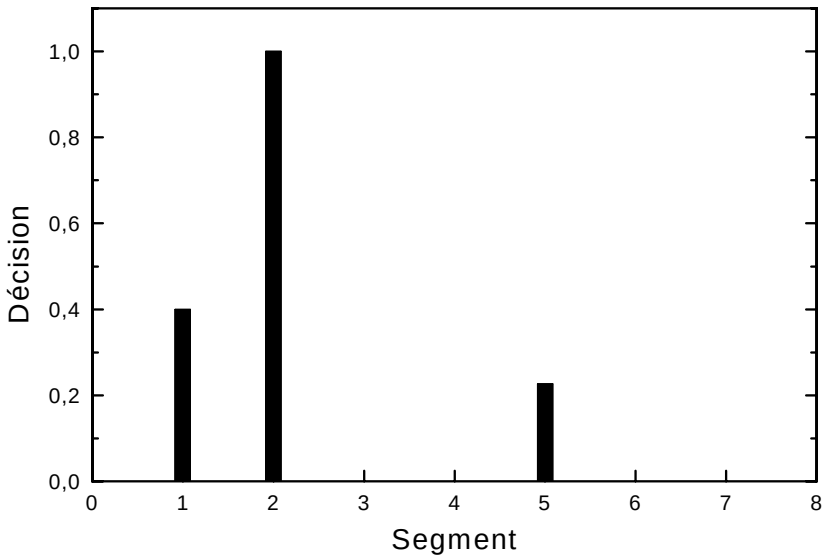

Figure 6. Choix de segments pour « Nopal » fait par Cortex. Les segments 1 et 2 qui ont une importance particulière ont été bien repérés.

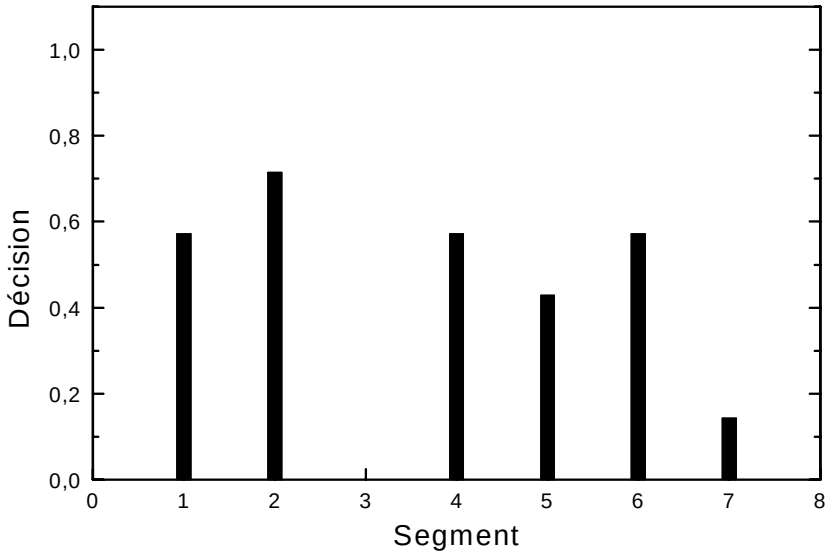

Figure 7. Choix de segments pour « Nopal » fait par les humains.

5. Discussion

5.1 Taille du lexique et autres considérations

Le pré-traitement réduit le lexique de plus en plus : $N_L \leq N_f \leq N_M$. Des études sur les ratios de réduction moyens du lexique ont été effectuées sur l'ensemble des textes. Ceci nous a permis d'établir expérimentalement des estimateurs ρ_f, ρ_L pour le lexique filtré/lemmatisé réduit et ρ_γ pour le lexique fréquentiel [TORRES-MORENO 02]. Nous avons constaté un comportement semblable dans les textes en espagnol et les textes en français. La réduction de la taille du lexique filtré/lemmatisé ρ_L suit un comportement linéaire par rapport au nombre de mots du texte original : pour obtenir un condensé d'un texte avec nos méthodes nous utilisons seulement un seizième du volume de termes totaux du document. Une étude a montré que l'ordre de présentation des métriques a un certain impact sur les performances de l'AD : leur pouvoir discriminatoire a été mesuré comme fonction des écart-types des métriques par rapport à chaque segment. Certaines métriques sont plus discriminantes que d'autres : la somme des distances d'Hamming entre mots est très discriminante, mais les métriques qui impliquent des calculs d'entropie semblent l'être moins. Un ordre optimal a donc été déterminé de façon expérimentale [TORRES-MORENO 02]. Nos expériences ont montré aussi que l'ordre de présentation des segments n'a aucune influence sur le choix final de l'Algorithme de décision. Nous avons segmenté les textes en les mélangeant au hasard pour obtenir un nouveau texte, qui a été présenté à nouveau à Cortex. Les mêmes résultats ont été retrouvés. Par contre, des tests sur Minds et Word montrent que ces méthodes sont plus dépendantes de l'ordre de présentation des segments. En effet, la segmentation de phrases par le séparateur <:> a tendance à perturber leur pertinence de choix.

5.2 Qualité des condensés

Nous avons établi une mesure objective de la qualité des condensés, défini par rapport aux probabilités de segments choisis par les personnes. Nous avons donc établi Q :

$$Q = \sum \text{prob(humains)} \times \theta \quad ; \quad \theta = \left\{ \frac{0 : \text{seg.non trouvé}}{1 : \text{seg. trouvé}} \right\}$$

Les tableaux 1, 2 et 3 montrent les détails des calculs de Q pour les textes « Puces » et « Hydro » (où seulement 25% des segments totaux ayant une valeur différente de 0 dans chaque méthode sont présentés). La figure 8 montre la qualité des condensés Q sur l'ensemble des textes, pour chacune des méthodes.

Tableau 1

PUCES : nombre de segments																					
Méthode	1	2	3	4	5	6	7	8	9	11	13	14	15	17	18	19	20	21	22	24	25	26
Humains		0,9	0,4	0,4	0,5				0,4				0,5	0,9						0,4		
Cortex		1				1			1	1	1		1	1				1				
Minds	1	1	1						1			1		1	1			1				
Word						1							1		1	1	1	1	1			1
Pertinence		1	1				1	1		1	1										1	

Tableau 2

HYDRO : nombre de segments									
Méthode	3	5	6	7	10	12	13	14	15
Humains	0,9		0,8	0,9	0,3				
Cortex	1		1	1			1		
Minds		1	1	1					
Word				1				1	1
Pertinence						1	1		

Tableau 3

Méthode	**Q**	
	PUCES (25% = 8 segments)	HYDRO (25% = 4 segments)
HUMAINS	4,43	2,77
CORTEX	2,64	2,46
MINDS	2,57	1,62
WORD	0,50	0,85
PERTINENCE	1,29	0

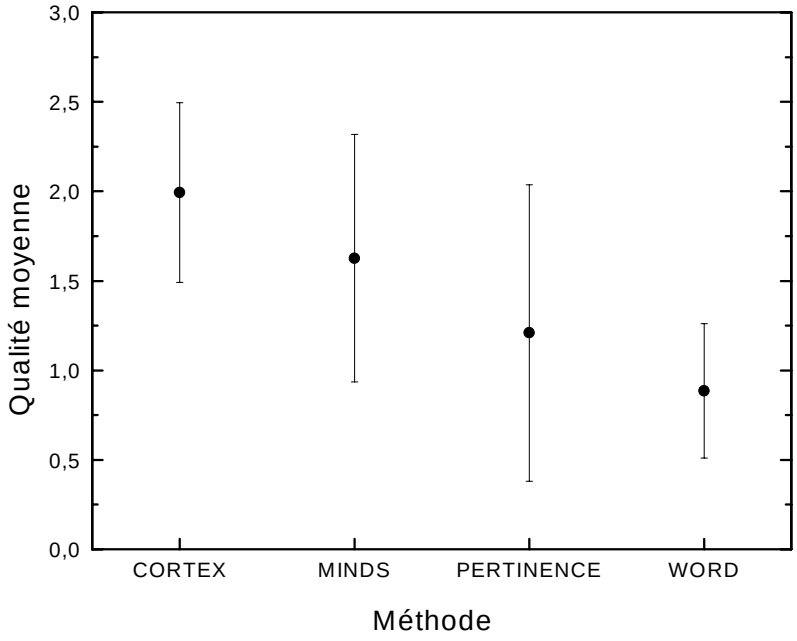

Figure 8. Qualité moyenne Q des condensés obtenus par différentes méthodes.

6. Conclusion

L'algorithme Cortex est un condensateur de textes très performant. Nous avons mesuré la qualité de nos condensations de façon objective et, en moyenne, nos condensés sont d'une qualité supérieure aux autres méthodes utilisées. Notre technologie permet de traiter des corpus importants, multilingues (français, espagnol), sans préparation, comportant une certaine quantité de bruit, et ce dans un délai de temps court. Plusieurs tests comparatifs menés avec des sujets humains ou d'autres méthodes de condensation, ont montré que Cortex retrouve des phrases pertinentes, indépendamment des sujets abordés. On obtient un condensé équilibré où la plupart des thèmes sont abordés. Notre algorithme de décision basé sur les votes de métriques est robuste, convergent, amplificateur et indépendant de l'ordre de présentation des segments. Nous pensons que l'ajout d'autres métriques comme l'entropie résiduelle, la détection des changements d'entropie, ou de maximum d'entropie, pourraient améliorer la qualité des condensations.

Remerciements

Les auteurs tiennent à remercier le Conseil de Recherches en Sciences Naturelles et en Génie (CRSNG Canada) et le Conseil de recherche en sciences humaines (CRSH Canada) pour leur soutien financier, de même que l'École Polytechnique de Montréal.

Références

American National Standards for Writing Abstracts. ANSI Inc., USA, 1979

[BALPE 96]. BALPE J., LELU A., PAPY F., and SALEH I. *Techniques avancées pour l'hypertexte*. Éditions Hermès, Paris, 1996

[DEERWESTER 90] DEERWESTER S., DUMAIS D., FURNAS T., LAUNDER G., and HARSHMAN T. *Indexing by latent semantic analysis*. Journal of the Amer. Soc for Infor. Science, 6(41):391—407, 1990

[HUOT 00] HUOT F. *Copernic summarizer ou la tentation de l'impossible*. Québec Micro, 6.12 (12):61—64, 2000

[LELOUP 97] LELOUP C., *Moteurs d'indexation et de Recherche*. Eyrolles, 1997

[MEMMI 00] MEMMI D., *Le modèle vectoriel pour le traitement de documents*. Cahiers Leibniz 2000-14, INPG, 2000

[MEMMI 98] MEMMI D., GABI K., and MEUNIER J.-G., « Dynamical knowledge extraction from texts by art networks ». In *Proc. of the NEURAP'98*, Marseille, 1998

[MEMMI 00] MEMMI D. and MEUNIER J.-G. « Proc. of nc'2000 ». In *Using competitive networks for text mining*, Berlin, May 2000

[MEUNIER 97] MEUNIER J.-G. and NAULT G. « Approche connexionniste au problème de l'extraction de connaissances terminologiques à partir de textes ». In *Les Techniques d'intelligence artificielle appliquées aux Technologies de l'Information*, pages 62—76, 1997. Les Cahiers scientifiques ACFAS 97.

[MORRIS 99] MORRIS A., KASPER G., and ADAMS D. « The effects and limitations of automated text condensing on reading comprehension performance ». In *Advances in automatic text summarization*, pages 305--323. The MIT Press, U.S.A., 1999

[SAGGION 00] SAGGION H. and LAPALME G. « Concept identification and presentation in the context of technical text summarization ». In *Automatic Summarization Workshop*, pages 1--10, Seattle. ANLP/NAACL, 2000

[SALTON 71] Salton G. *The SMART Retrieval System - Experiments un Automatic Document Processing.* Englewood Cliffs, 1971

[SALTON 83] SALTON G. and MCGILL M. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983

[SEFFAH 96] SEFFAH A. and MEUNIER J.-G. « Aladin: an integrated object-oriented environment for computer assited text analisys ». *Cahiers de recherche* 96.1, LANCI-UQAM, 1996

[TORRES-MORENO 02] TORRES-MORENO, J.M, VELÁZQUEZ-MORALES, P. et MEUNIER, J.G. (2002). « Condensés de textes par des méthodes numériques ». In *JADT 2002*, 2:723-734, A. Morin & P. Sébillot éditeurs, IRISA/INRIA, France.

[TORRES-MORENO 01]. TORRES-MORENO, J.M, VELÁZQUEZ-MORALES, P. et MEUNIER, J.G. (2001). « Cortex : un algorithme pour la condensation automatique des textes. » In *la cognition entre individu et société* ARCo 2001. Coord. Hélène PAUGAM-MOISSY, Vincent NYCKEES, Josiane CARON-PARGUE, Lyon, Hermès Science, pages 365 vol. 2, ISC-Lyon, France.

[TORRES-MORENO 00] TORRES-MORENO J.-M., VELAZQUEZ-MORALES P., and MEUNIER J. (9-11 Mars 2000). « Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes ». In *JADT 2000*, pages 365--372, Lausanne. EPFL M. Rajman & J.-C. Chappelier éditeurs.

[VERONIS 91] VERONIS J., Ide N., and HARIE S., « Very large neural networks as a model of semantic relations ». In *Proc. of the 3rd Cognitiva Symposium*, Madrid, 1991