Enertex : un système basé sur l'énergie textuelle

Silvia FERNÁNDEZ[1,2], Eric SANJUAN[1], Juan Manuel TORRES-MORENO[1,3]
[1] Laboratoire Informatique d'Avignon, BP 1228 84911 Avignon France
[2] LPM UHP-Nancy, BP 239 54506 Vandœuvre lès Nancy France
[3] École Polytechnique de Montréal, CP 6079 Montréal, Canada H3C3A7
{silvia.fernandez, eric.sanjuan, juan-manuel.torres}@univ-avignon.fr

Résumé. Dans cet article, nous présentons des applications du système Enertex au Traitement Automatique de la Langue Naturelle. Enertex est basé sur l'énergie textuelle, une approche par réseaux de neurones inspirée de la physique statistique des systèmes magnétiques. Nous avons appliqué cette approche aux problèmes du résumé automatique multi-documents et de la détection de frontières thématiques. Les résultats, en trois langues : anglais, espagnol et français, sont très encourageants.

Abstract. In this paper we present Enertex applications to study fundamental problems in Natural Language Processing. Enertex is based on textual energy, a neural networks approach, inspired by statistical physics of magnetic systems. We obtained good results on the application of this method to automatic multi-document summarization and thematic border detection in three languages : English, Spanish and French.

Mots-clés : Énergie textuelle, Réseaux de neurones, Modèle de Hopfield, Résumé automatique, Frontières thématiques.

Keywords: Textual Energy, Neural Networks, Hopfield Model, Automatic Summarization, Thematic Boundaries.

1 Introduction

Des idées empruntées à la physique ont déjà été utilisées dans l'analyse de textes. Les exemples plus notables sont l'approche entropique de (Shannon, 1948), les travaux de (Zipf, 1935; Zipf, 1949) et de (Mandelbrot, 1953) où les auteurs font des considérations thermodynamiques d'énergie et de température dans leurs études sur la Statistique Textuelle. Dernièrement (Takamura *et al.*, 2005) se sont servi des notions de polarisation des systèmes de *spins* pour trouver les orientations sémantiques des mots (désirable ou indésirable) à partir de mots amorce. La sortie de ce système est une liste de mots indiquant leurs orientations estimés selon l'approximation du champ moyen. Dans notre travail, nous avons utilisé différemment la notion de *spin*. Elle nous a permis de représenter les présences (↑) où absences (↓) des mots dans les documents. À partir de cet image, on aperçoit le document comme un matériaux composé d'un ensemble de unités en interaction dont l'énergie peut être calculée. Nous avons étudié les problèmes du Traitement Automatique de la Langue Naturelle (TALN) en utilisant la notion d'énergie textuelle. Récemment introduite (Fernández *et al.*, 2007a; Fernández *et al.*, 2007b), l'énergie textuelle a été appliquée au résumé automatique et à la détection de frontières sur des corpus en français

et en anglais. Elle est aussi un des algorithmes utilisés dans (da Cunha *et al.*, 2007) où des méthodes statistiques et linguistiques sont combinées pour résumer des articles médicaux en espagnol. Dans cet article nous étudions l'influence de deux facteurs, inspirés aussi de la physique, sur l'énergie textuelle. Il s'agit d'un champ externe et de la température. Cette démarche a permis d'améliorer les performances du modèle. Les résultats sur des corpus multi-documents et trilingues (français, anglais et espagnol) sont très encourageants. Nous présentons dans la Section 2 une brève introduction au modèle neuronal de Hopfield ainsi que son extension au TALN. Nous appliquons l'énergie textuelle à deux tâches bien distinctes : la génération de résumés multi-documents guidés par une thématique dans la Section 3 et l'amélioration de l'algorithme de détection de frontières thématiques dans la Section 4. Finalement nous présentons les conclusions et quelques perspectives.

2 L'énergie textuelle des documents

La contribution la plus importante de Hopfield à la théorie des réseaux neuronaux a été l'introduction du concept d'énergie issu de l'analogie avec les systèmes magnétiques : systèmes constitués d'un ensemble de N petits aimants appelés *spins* qui peuvent s'orienter selon plusieurs directions. Le cas le plus simple est représenté par le modèle d'Ising, avec deux directions possibles : vers le haut (↑, +1 ou 1) ou vers le bas (↓, -1 ou 0). Ce modèle a été utilisé dans une grande variété de systèmes qui peuvent être décrits par des variables binaires (Ma, 1985). Un système de N unités binaires possède $\nu = 1, ..., 2^N$ configurations (patrons) possibles. Dans le modèle de Hopfield, les *spins* correspondent aux neurones qui interagissent selon la règle d'apprentissage de Hebb[1] :

$$J^{i,j} = \sum_{\mu=1}^{P} s_\mu^i s_\mu^j \tag{1}$$

s^i et s^j sont les états des neurones i et j. La sommation porte sur les P patrons à stocker. Ce modèle est aussi connu sous le nom de mémoire associative. Il possède la capacité de stocker et de récupérer un certain nombre de configurations du système, car la règle de Hebb transforme ces configurations en attracteurs (minimaux locaux) de la fonction d'énergie (Hopfield, 1982) :

$$E_{\mu,\nu} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_\mu^i \; J^{i,j} \; s_\nu^j \tag{2}$$

Si on présente un patron proche à ν, chaque spin subira un champ local $h_\mu^i = \sum_{j=1}^{N} J^{i,j} s_\mu^j$ induit par les N *spins* des autres patrons μ. Les *spins* s'aligneront selon h_μ^i pour restituer le patron stocké ν. Hopfield a démontré que l'énergie du système diminue toujours pendant le processus de récupération. Nous ne développerons pas la méthode de récupération de patrons[2], car l'intérêt porte sur la distribution et les propriétés de l'énergie du système. Cette fonction monotone et décroissante a été utilisée uniquement pour montrer que l'apprentissage est borné.

D'un autre côté, le modèle vectoriel de textes (Salton & McGill, 1983) transforme un document dans un espace adéquat où une matrice S contient l'information du texte sous forme de sacs de mots. On peut considérer S comme l'ensemble des configurations d'un système dont on peut

[1] Hebb (Hertz *et al.*, 1991) a suggéré que les connexions synaptiques changent proportionnellement à la corrélation entre les états des neurones.

[2] Cependant le lecteur intéressé peut consulter, par exemple (Hopfield, 1982; Hertz *et al.*, 1991).

calculer l'énergie. Les documents sont pré-traités avec des algorithmes classiques de filtrage de mots fonctionnels[3], de normalisation et de lemmatisation (Porter, 1980; Manning & Schütze, 1999) afin de réduire la dimensionnalité. Une représentation en sac de mots produit une matrice $S_{[P\times N]}$ de fréquences/absences :

$$S = [s_\mu^i] = \begin{cases} TF^i & \text{si le terme } i \text{ existe} \\ 0 & \text{autrement} \end{cases} \tag{3}$$

où $\mu = 1, \cdots, P$ phrases et $i = 1, \cdots, N$ termes. La présence du mot i représente un spin $s^i \uparrow$ avec une magnitude donnée par sa fréquence TF^i (son absence par $\downarrow$ respectivement), et une phrase est donc une chaîne de N *spins*. Pour calculer les interactions entre les N termes du vocabulaire, on applique la règle de Hebb, qui sous forme matricielle se traduit par :

$$J = S^T \times S \tag{4}$$

Chaque élément $J^{i,j} \in J_{[N\times N]}$ est équivalent au calcul de (1). Enfin, l'énergie textuelle d'interaction (2) peut alors s'exprimer comme :

$$E = -\frac{1}{2} S \times J \times S^T \tag{5}$$

Un élément $E_{\mu,\nu} \in E_{[P\times P]}$ représente l'énergie textuelle entre les phrases μ et ν. La représentation sous forme de graphe (Fernández *et al.*, 2007b) nous a permis d'expliquer la nature des liens que la mesure d'énergie textuelle induit. On a déduit qu'elle relie à la fois des phrases ayant des mots communs, ainsi que des phrases qui partagent un même voisinage sans pour autant partager nécessairement un même vocabulaire. C'est pour cette raison que l'énergie textuelle peut être utilisée comme mesure de similarité dans les applications du TALN. Nous avons développé l'algorithme Enertex basé sur cette mesure de similarité. Les premières applications ont porté sur le résumé mono-document générique et sur la détection de frontières thématiques. Dans la section suivante nous montrons une modification qui consiste en mettre un champ externe en rapport avec un corpus multi-document. Cette stratégie permet de générer des résumés guidés par les besoins de l'utilisateur. Une autre approche, montrée dans la section 4, utilise l'énergie textuelle représentée comme un spectre de la phrase. Ceci permet la détection de frontières thématiques d'un document au moyen d'un test de concordance de Kendall. Nous montrons ici comment l'introduction d'une température modifie les spectres des phrases afin que le test de Kendall puisse mieux les identifier.

3 Résumé multi-document guidé par une requête

Les premiers systèmes de résumé automatique multi-documents ont été développés dans les années 90 (McKeown & Radev, 1995). Les conférences DUC portant sur la tâche de résumé automatique sont organisées depuis 2001 par le NIST[4]. La tâche principale de DUC consiste à traiter des questions complexes et réelles. Le type de réponse attendue ne peut pas être une entité simple (un nom, une date ou une quantité telle que classiquement défini dans les conférences TREC Question-Answering[5]). Le problème peut se poser comme ceci : étant donnée une thématique et un ensemble $\mathcal{L}$ avec D documents pertinents, générer un court résumé de 250 mots,

[3]Nous avons effectué le filtrage de chiffres et l'utilisation d'anti-dictionnaires.

[4]`http://www-nlpir.nist.gov/projects/duc`

[5]`http://trec.nist.gov/data/qamain.html`

cohérent et bien organisé qui répondra aux questions de la thématique. Les $D = 25$ documents proviennent du corpus AQUAINT : articles d'*Associated Press, New York Times* (1998-2000) et *Xinhua News Agency* (1996-2000). L'évaluation de la qualité des résumés mono-document reste une tâche difficile. En multi-documents le problème n'est pas plus simple. Des approches manuelles et semi-automatiques ont été utilisées à ce propos. Ainsi *Pyramid* (Passonneau *et al.*, 2005), *Basic Elements* (Hovy *et al.*, 2005) et ROUGE (Lin, 2004) ont été employées. Plusieurs mesures manuelles ont été évaluées : cohérence, grammaticalité, non-redondance, pertinence au sujet, qualité linguistique. ROUGE est utilisée par la communauté comme mesure d'évaluation semi-automatique. Elle mesure l'intersection d'ensembles de n-grammes entre les résumés candidats et ceux de référence. Les métriques les plus populaires sont ROUGE-2 (bigrammes) et SU4 (bigrammes séparés par un intervalle ≤ 4 mots). Nous avons utilisé l'énergie textuelle pour la tâche de résumé guidé par une thématique ou sujet. L'idée est d'observer la réponse du système face à un champ externe. Ce champ, représenté par le vecteur des termes d'un texte décrivant un sujet a été mis en relation avec le corpus multi-document. La figure 1 illustre le processus d'obtention du résumé guidé par un sujet. Les documents sont concaténés dans un seul document et un prétraitement standard (filtrage et stemming (Porter, 1980)) lui est appliqué. L'énergie textuelle entre le sujet et les phrases du document concaténé est calculée selon :

$$E(sujet, phrase) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s^{i}_{sujet}\ J^{i,j}\ s^{j}_{phrase} \tag{6}$$

Finalement, le résumé est formé avec les phrases présentant la plus haute énergie textuelle avec le sujet. Un post-traitement de diminution de la redondance lui a été appliqué.

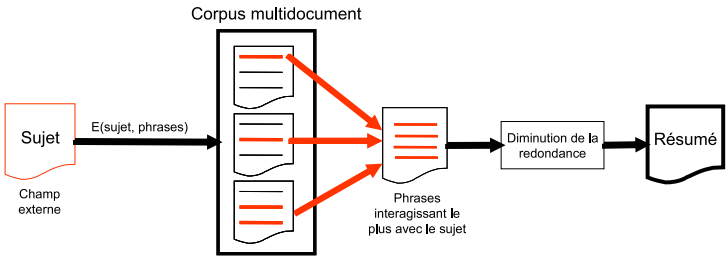

FIG. 1 – Résumé guidé par une thématique et un ensemble de documents.

Diminution de la redondance

Dans un résumé multi-document il y a une probabilité significative de re-inclure l'information déjà présente. Pour diminuer ce problème il faut une stratégie de diminution de la redondance. Notre système ne possède pas un traitement linguistique et la stratégie d'anti-redondance consiste à comparer les valeurs d'énergie des phrases candidates et leur longueur. Nous supposons que (dans de grands corpus) la probabilité que 2 phrases aient les mêmes valeurs d'énergie est très faible. Ainsi, nous avons éliminé la présence de doublons (phrases avec exactement la même valeur d'énergie). Peut-on aller encore plus loin et détecter avec ce même critère des phrases égales à quelques mots près ? Pour le tester, on considère que si 2 phrases partagent la plus grande partie de leurs mots, elles apportent la même information. On construit donc le résumé avec la phrase la plus énergétique (en valeur absolue), puis la suivante dans le score (la candidate) fera partie du résume si $|E_2 - E_1| \geq \epsilon$. E_1 est l'énergie de la phrase déjà présente.

La 3ème phrase candidate fera partie du résumé si $|E_3 - E_1| \geq \epsilon$ et si $|E_3 - E_2| \geq \epsilon$. Les énergies E_1 et E_2 sont considérées comme celles des phrases de référence. En général, une phrase candidate i sera ajoutée au résumé, si pour chaque phrase de référence $(i-1)$:

$$|E_i - E_{i-1}| = \Delta E \geq \epsilon; i = 2, 3, ... \tag{7}$$

Le cas contraire signifie que les énergies sont très proches avec une haute probabilité de redondance. On présente sur la figure 2 à gauche les valeurs du rappel du produit ROUGE-2 $\times$ SU4 pour différentes valeurs de ΔE. Le meilleur résultat sur les corpus DUC'05-07 est obtenu avec $\Delta E \approx 0,003$. Cela correspond aux phrases à 2 mots près. Une autre stratégie permettant de

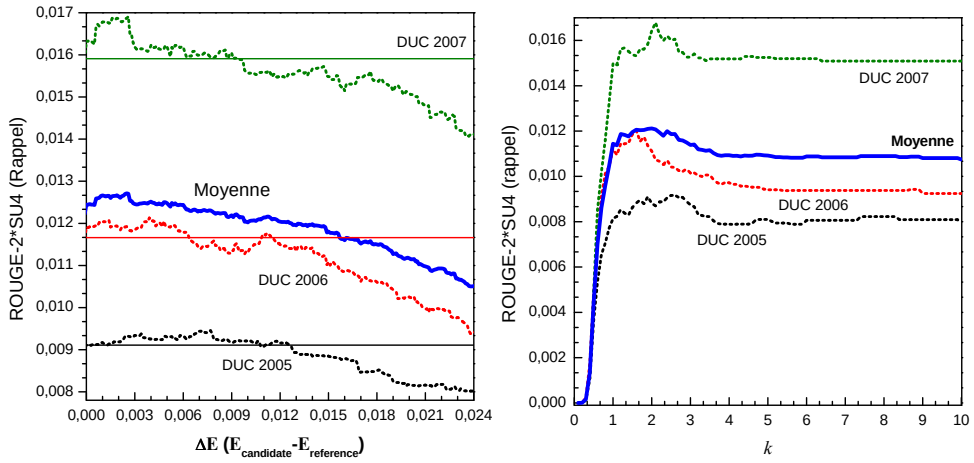

FIG. 2 – Diminution de la redondance : ΔE d'énergie des phrases et moyenne des longueurs de phrases.

diversifier le contenu, consiste à écarter du résumé les phrases longues (dans les documents il y a des phrases de taille $\approx$ à celle du résumé demandé). On a défini la taille maximale de phrase comme $k \times M$ où M = nombre moyen de mots par phrase dans les documents originaux. Nous avons fait varier k par petits pas en mesurant à chaque moment le produit de ROUGE-2 $\times$ SU4. Le comportement est montré sur la figure 2 à droite. Le meilleur résultat est avec $k \approx 1,6$. Nous avons fixé k, puis le seuil d'énergie $\Delta E = 0,003$ en maximisant le produit ROUGE-2$\times$SU4. En DUC'07 il y avait 2 *baselines* : la 1ère est tirée au hasard. La 2ème est un système de résumé générique. La figure 3 montre la position d'Enertex comparé aux participants après les campagnes DUC'05-07. Le cosinus obtient des performances ROUGE étonnament hautes, mais qui peuvent donner lieu à des résumés avec beaucoup de redondance, car toutes les phrases selectionnées sont proches de la thématique. Par contre l'énergie textuelle capture l'information entre 2 phrases calculéé parmi toutes les autres. De ce fait, la similarité tient en compte pas uniquement du nombre de mots partagés (le recouvrement et le cosinus sont des mesures locales) mais des interactions indirectes (chemins de longueur 2).

4 Frontières thématiques

Plusieurs stratégies ont été développées pour segmenter thématiquement un texte. On trouve PLSA (Brants *et al.*, 2002) qui estime les probabilités d'appartenance des termes à des classes sémantiques, des méthodes s'appuyant sur des modèles de Markov (Amini *et al.*, 2000), sur une classification des termes (Caillet *et al.*, 2004; Chuang & Chien, 2004) ou sur des chaînes lexicales (Sitbon & Bellot, 2005). Plus récemment, (Ferret, 2007) a proposé l'identification

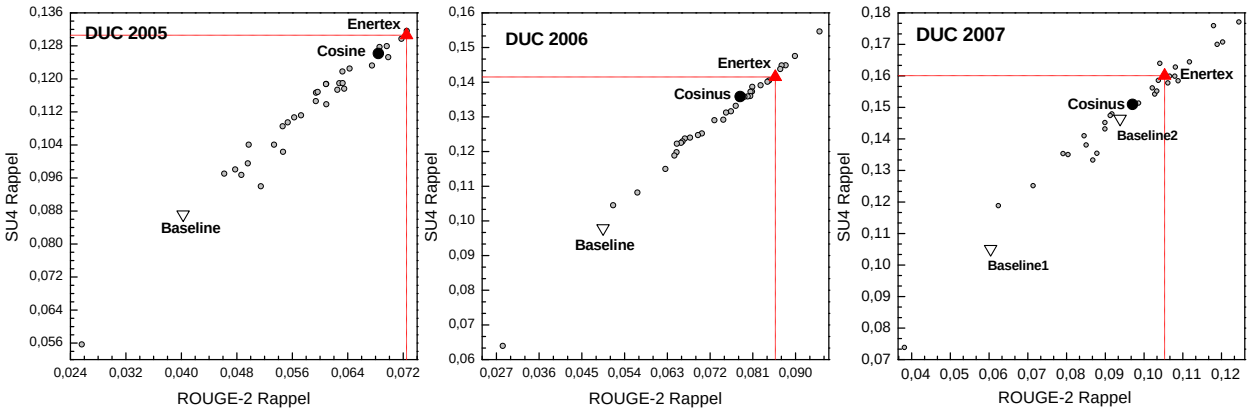

FIG. 3 – Aperçu du rappel SU4 vs ROUGE-2 des participants au-dessus des deux *baselines.*

préalable des sujets présentes dans le document comme stratégie pour améliorer la détection de ruptures thématiques. L'identification de sujets est faite à partir d'une analyse contextuelle basée sur la co-occurrence de mots. L'idée est que si deux segments n'ont pas une forte cohésion lexicale entre eux, mais ils apparaissent dans le même contexte, alors ils appartiennent au même sujet et la rupture thématique n'existe pas. Dans ce travil, nous avons utilisé la matrice d'énergie E (5). Chaque ligne de cette matrice produit un spectre qui répresente l'interaction de la phrase i avec les autres. La figure 4 montre les spectres de quelques phrases d'un texte composé de deux thématiques. Étant donné que l'énergie textuelle détecte et pondère le voisinage d'une phrase, on constate une similarité entre les courbes de l'une (en gras) et de l'autre thématique (en pointillées). Pour comparer les spectres nous avons utilisé (Fernández *et al.*, 2007a) le coefficient de concordance τ de Kendall (Siegel & Castellan, 1988) et le calcul de sa $p-$valeur qui permettent de définir un test statistique de concordance entre 2 juges qui classent un ensemble de P objets. Nous avons utilisé ce test pour trouver les frontières thématiques entre segments. Ces ruptures entre segments sont bien détectées si le voisinage commun entre les phrases est bien repéré. Mais il se trouve que des phrases chevauchant les thématiques présentent des courbes d'énergie que le test du τ de Kendall s'avère incapable de distinguer. C'est le cas du spectre de la phrase 23 de la figure 4. Pour diminuer cet effet nous avons proposé (Fernández *et al.*, 2007b) une variation du test de Kendall avec l'utilisation d'une fenêtre glissante : la phrase centrale est comparée aux autres dans la fenêtre Cette stratégie a permis une meilleure détection des ruptures. Mais nous pensons qu'on peut faire mieux. Dans ce travail nous introduisons une stratégie portant directement sur la modification des spectres des courbes : le lissage par un paramètre de bruit β qui peut être assimilé, en termes physiques, à l'inverse d'une température T.

Décroissance exponentielle : distance et température

La figure 4 montre que les spectres qui expriment correctement leur appartenance à une thématique ont une forme décroissante par rapport à un maximum. Ce maximum correspond à l'expression d'une forte interaction entre un couple de phrases. À partir de ce point maximal, les autres interactions diminuent rapidement jusqu'à la fin de la thématique. Cette décroissance de l'énergie textuelle peut être contrôlée avec un facteur $\exp^{-r/T}$ où r est la distance entre la phrase μ et la phrase voisine qui présente la plus haute interaction avec elle et T un paramètre de bruit présent dans les spectres[6]. La figure 5 montre le lissage induit dans les spectres pour deux

[6]Dans la littérature de réseaux de neurones on trouve souvent $\beta = 1/T$

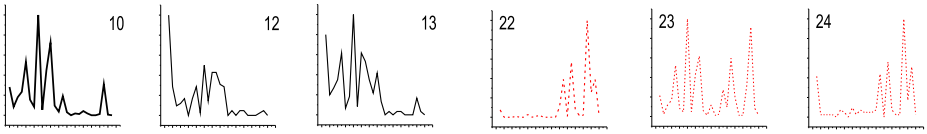

FIG. 4 – Énergie textuelle : en trait continu l'énergie des phrases de la 1ère thématique, en pointillé celle de la 2ème. Le changement d'allure correspond à un changement thématique. L'axe horizontal indique le numéro de phrase et l'axe vertical, l'énergie textuelle de la phrase par rapport aux autres.

phrases de la figure 4 par le facteur $\exp^{-r/T}$ (la phrase 23 est difficile à classer en fonction de ses pics). Nous avons diminué T progressivement afin d'analyser l'évolution du chevauchement des courbes. Cette diminution lisse les courbes de façon efficace : à $T \approx 8$ le bruit de la courbe 23 est réduit et un classement correct a été obtenu. Le spectre de la phrase 10 a aussi été lissé sans perte d'information. Nous faisons l'hypothèse que avec ce lissage, le test de concordance

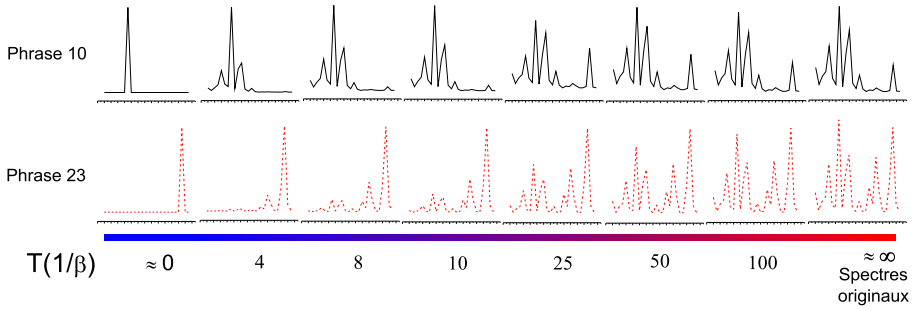

FIG. 5 – Lissage par $\exp^{-r/T}$ des spectres. En trait continu le spectre d'une phrase thématiquement bien définie et en pointillé celui d'une phrase inclassable. r est la distance au maximum et T la température.

de Kendall identifiera mieux les phrases selon leur thématique. Mais quelle est la valeur correcte de T ? Pour l'estimer nous avons réalisé des expériences sur des corpus multi-thématique en anglais, espagnol et français. Les corpus ont été construits à partir d'articles journalistiques du BROWN CORPUS, LA JORNADA et LE MONDE[7]. Les corpus comportent 4 ensembles de 100 documents qui correspondent à une taille de segments fixée. Un document est constitué de 10 segments extraits d'articles thématiquement différents tirés au hasard. Pour chaque document on a calculé l'énergie textuelle à différents températures : $T = 1, \cdots, 180$. Les spectres ont été comparés par le test de Kendall et les frontières détectées ont été mesurés par Windiff (WD) (Pevzner & Hearst, 2002)[8]. Plus la valeur WD est basse, mieux la segmentation a été réalisée. La figure 6 montre les résultats de 100 documents en français et une taille de segments de 6-8 phrases. En trait continu l'évolution de la valeur moyenne de WD et en pointillé le nombre de frontières trouvées. On observe qu'à températures très basses les courbes d'énergie perdent leurs pics (sauf le maximum). Le test de Kendall ne détecte plus de frontières et la valeur WD est élevée. En augmentant la température les courbes voisines se ressemblente de plus en plus et le nombre de frontières augmente. Nous avons retenu la valeur $T = 80$ qui maximise le nombre de frontières trouvées en minimisant la valeur WD.

[7]http ://khnt.aksis.uib.no/icame/manuals/brown, http ://www.jornada.unam.mx et http ://www.lemonde.fr
[8]Windiff mesure la différence entre les frontières véritables et celles trouvées dans une fenêtre glissante.

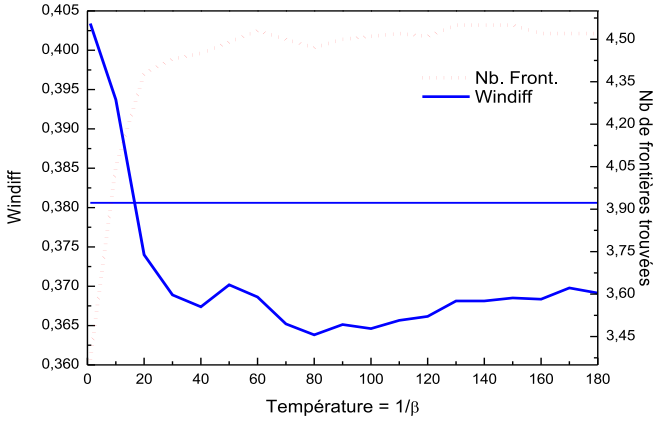

FIG. 6 – Évolution de WD et du nombre de frontières en fonction de T. La ligne horizontale représente la valeur de WD à température infinie. Taille des segment entre 6 et 8 phrases pour le corpus en français.

La mesure δ-Front

(Pevzner & Hearst, 2002) ont montré que WD est peu sensible aux variations de la taille de segments et plus équilibré que d'autres mesures dans la pénalisation des erreurs. Cependant elle a ses faiblesses. WD ne peut pas être assimilée à un taux d'erreur (sa valeur peut être > 1) et elle n'est qu'un élément de comparaison de la fiabilité des méthodes et non un paramètre absolu de sa qualité (Sitbon & Bellot, 2004). De plus, nous avons trouvé qu'une même valeur de WD pouvait correspondre aux segmentations différentes du document. Nous introduisons ici la mesure δ-Front qui calcule la distance euclidienne $d(\bullet)$ entre les vecteurs **A** et **B** de dimension P (nombre des phrases du document) : **A** correspond aux frontières véritables et **B** à celles détectées. La valeur de la composante i est le nombre de phrases séparant la phrase i de la frontière la plus proche (figure 7). Le facteur de normalisation est calculé avec le vecteur nul : ne contenant aucune frontière sauf les extrêmes. Plus la valeur δ-Front est basse, mieux la segmentation a été réalisée.

$$\delta\text{-Front}(\mathbf{A},\mathbf{B}) = \frac{d(\mathbf{A},\mathbf{B})}{d(\mathbf{A},\mathbf{C})} \tag{8}$$

On observe au tableau 1 que la valeur de T pour la meilleur segmentation dépend de la longueur

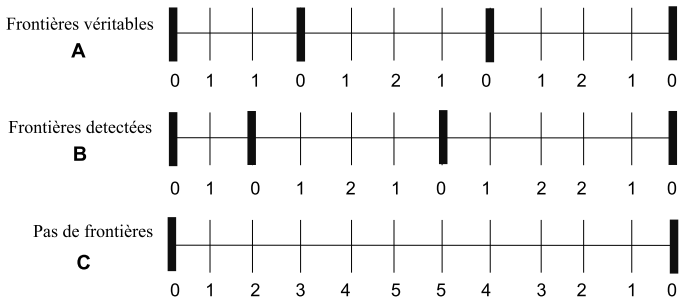

FIG. 7 – La mesure δ-Front.

du document. Plus la taille du segment est grande (plus le document est long) plus la valeur T est elevée. Les deux mesures ne sont pas toujours en accord. En français WD obtient la valeur la plus haute pour des segments de taille 9-11 et δ-Front pour 3-5. Cette différence peut être due

au nombre de véritables frontières trouvées : δ-Front considère plus finement ce facteur. Les méthodes cités rapportent des meilleures performances en anglais qu'en français, peut être dû aux différences structurales et de répétition de mots entre ces langues. Cependant nos résultats sont comparables dans les 3 langues. Cette stabilité découle du calcul d'interactions des mots combiné au processus de comparaison de segmentes. (Ferret, 2007) constate en partie cet effet.

Taille du segment	T	**Français**		**Espagnol**		**Anglais**		**Nb. frontières trouvées**
		WD	**δ-Front**	**WD**	**δ-Front**	**WD**	**δ-Front**	
9-11	120	0,4109	0,1817	0,3897	0,2069	0,3925	0,1524	$\approx$6/9
6-8	80	0,3638	0,1957	0,3601	0,2031	0,3804	0,1640	$\approx$5/9
3-11	40	0,3885	0,1974	0,3646	0,2043	0,3709	0,1634	$\approx$5/9
3-5	20	0,3851	0,4540	0,3598	0,3257	0,3786	0,3864	$\approx$3/9

TAB. 1 – Mesures WD et δ-Front pour des corpus en 3 langues et segments de tailles variables.

5 Conclusions

Nous avons présenté le système Enertex basé sur le concept d'énergie textuelle. L'énergie textuelle est bien adaptée à la recherche de segments porteurs d'information d'un texte et à sa pondération. L'extraction et l'assemblage de ces segments donne le condensé d'un document. Nous avons élargi la portée de cet idée pour développer un algorithme de résumé multi-document guidé par une thématique : un champ externe, représenté par le vecteur des termes décrivant une thématique a été mis en rélation avec les phrases d'un corpus multi-document. Ceci a permis de générer des résumés personnalisés que nous avons evalué dans le cadre des tâches DUC. La position d'Enertex est excellente par rapport à la trentaine de participants compte tenu que l'énergie textuelle est exprimée comme un simple produit matriciel. Aucune autre mesure de pondération de phrases a été incluse. En segmentation thématique, nous avons amélioré la détection de fausses frontières au travers d'une fonction pilotée par une témperature. Cela à permis de surpasser nos résultats précédents. Compte tenu des faiblesses de WD, nous avons introduit δ-Front, une nouvelle mesure d'évaluation de segmentation thématique. Nous envisageons de tester le modèle de Potts, autre modèle d'interaction entre *spins* qui favorise l'interaction entre mots de même fréquence, ainsi qu'une étude des chemins de longueur > 2 dans le graphe (intéractions d'ordre ≥ 3). Des applications en classification de textes sont aussi envisagées.

Références

AMINI M.-R., ZARAGOZA H. & GALLINARI P. (2000). Learning for sequence extraction tasks. In *RIAO 2000*, p. 476–489.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic anaysis. In *CIKM'02*, p. 211–218, McLean, Virginia, USA.

CAILLET M., PESSIOT J.-F., AMINI M. & GALLINARI P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *RIAO'04*, p. 648–657, France.

CHUANG S.-L. & CHIEN L.-F. (2004). A practical web-based approach to generating Topic hierarchy for Text segments. In *30th ACM IKM*, p. 127–136, Washington DC, USA.

DA CUNHA I., FERNÁNDEZ S., VELÁZQUEZ MORALES P., VIVALDI J., SANJUAN E. & TORRES MORENO J. M. (2007). A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. In *LNAI 4287, MICAI'07, Mexico*, p. 872–882.

FERNÁNDEZ S., SANJUAN E. & TORRES-MORENO J. M. (2007a). Energie textuelle des mémories associatives. In *TALN 2007*, p. 25–34.

FERNÁNDEZ S., SANJUAN E. & TORRES-MORENO J. M. (2007b). Textual Energy of Associative Memories : performants applications of ENERTEX algorithm in text summarization and topic segmentation. In *LNAI 4287, MICAI'07, Mexico*, p. 861–871.

FERRET O. (2007). Finding document topics for improving topic segmentation. In *ACL'07*, p. 480–487.

HERTZ J., KROGH A. & PALMER G. (1991). *Introduction to the theorie of Neural Computation*. Redwood City, CA : Addison Wesley.

HOPFIELD J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, **9**, 2554–2558.

HOVY E., LIN C. & ZHOU L. (2005). Evaluating DUC 2005 using Basic Elements. In *DUC 2005*.

LIN C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out : ACL-04 Workshop*, p. 74–81, Spain.

MA S. (1985). *Statistical Mechanics*. Philadelphia, CA : World Scientific.

MANDELBROT B. (1953). An informational theory of the statistical structure of languages. In *Communication Theory, ed. By Willis Jackson*, p. 486–502, New York : Academic Press.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : The MIT Press.

MCKEOWN K. & RADEV D. (1995). Generating summaries of multiple news articles. In *18^{th} ACM SIGIR*, p. 74–82.

PASSONNEAU R., NENKOVA A., MCKEOWN K. & SIGLEMAN S. (2005). Applying the Pyramid Method in DUC 2005.

PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. In *Computational Linguistic*, volume 1, p. 19–36.

PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

SALTON G. & MCGILL M. (1983). *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company.

SHANNON C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 79–423, 623–656.

SIEGEL S. & CASTELLAN N. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw Hill.

SITBON L. & BELLOT P. (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. In *TALN 2004*, p. 10–19.

SITBON L. & BELLOT P. (2005). Segmentation thématique par chaînes lexicales pondérées. In *TALN 2005*, volume 1, p. 505–510.

TAKAMURA H., INUI T. & MANABU O. (2005). Extracting semantic orientations of words using spin model. In *ACL'05*, p. 133–140.

ZIPF G. (1935). Psycho-biology of languages. *Houghton-Mifflin, Boston, MA.*

ZIPF G. (1949). Human behavior and the principle of least effort. *Addison-Wesley, MA.*