Énergie textuelle de mémoires associatives

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 les Nancy FRANCE
[3] École Polytechnique de Montréal, CP 6079 Centre-ville, Montréal, Québec CANADA H3C3A7
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr

Résumé. Dans cet article[1], nous présentons une approche de réseaux de neurones inspirée de la physique statistique de systèmes magnétiques pour étudier des problèmes fondamentaux du Traitement Automatique de la Langue Naturelle. L'algorithme modélise un document comme un système de neurones où l'on déduit l'énergie textuelle. Nous avons appliqué cette approche aux problèmes de résumé automatique et de détection de frontières thématiques. Les résultats sont très encourageants.

Abstract. In this paper we present a neural networks approach, inspired by statistical physics of magnetic systems, to study fundamental problems in Natural Language Processing. The algorithm models documents as neural network whose textual energy is studied. We obtained good results on the application of this method to automatic summarization and thematic borders detection.

Mots-clés : réseaux de neurones, réseaux de Hopfield, résumé, frontière thématiques.

Keywords: neural networks, Hopfield network, summarization, thematic boundary.

1 Introduction

Hopfield (Hopfield, 1982; Hertz *et al*., 1991) s'est inspiré des systèmes physiques comme le modèle magnétique d'Ising (formalisme issu de la physique statistique decrivant un système avec des unités à deux états nommées spins) pour construire un réseau neuronal avec des capacités d'apprentissage et de récupération de patrons. Les capacités et limitations de ce réseau, appelé mémoire associative, ont été bien établies de façon théorique dans plusieurs études (Hopfield, 1982; Hertz *et al*., 1991) : les patrons doivent être non corrélés afin que leur récupération soit sans erreur, le système sature rapidement et seulement une fraction des patrons peut être stockée correctement. Dès que leur nombre dépasse $\approx 0,14N$, aucun des patrons n'est plus reconnu. Cette situation restreint fortement leurs applications pratiques. Cependant, dans le cas du traitement automatique de la langue naturelle (TALN), nous pensons que l'on peut exploiter ce comportement. Le modèle vectoriel de textes (Salton & McGill, 1983), transforme les

[1]Ce travail a été realisé en partie grâce au financement du CONACYT (Mexico), bourse 175225.

phrases d'un document en vecteurs. Ces vecteurs peuvent être traités comme un réseaux de neurones type Hopfield. Si l'on définit un vocabulaire de taille N, où N est le nombre de termes uniques d'un document, on peut représenter une phrase comme une chaîne de N neurones actifs, $i = 1, \cdots, N$ (le mot i étant présent) ou inactifs (le mot i étant absent). Un document de P phrases, est composé de P chaînes dans l'espace vectoriel Ξ de dimension N. Ces vecteurs sont plus ou moins corrélés, selon les mots qu'ils partagent. Si les thématiques sont proches, il est raisonable de supposer que le degré de corrélation sera très élevé. Cela pose des problèmes si on essaie de stocker et de récupérer ces représentations dans un réseau type Hopfield. Cependant notre intérêt porte non pas sur la récupération, mais sur les interactions entre les mots et entre les phrases. Cette interaction nous allons la définir comme l'énergie textuelle d'un document. Elle peut servir, entre autres, à pondérer les phrases ou à détecter des changements entre des chaînes de neurones. Nous dévéloppons une métaphore qui permet d'utiliser le concept d'enérgie textuelle pour son application dans le résume générique ou la segmentation thématique. Nous présentons en Section 2 une brève introduction au modèle de Hopfield. En Section 3, nous faisons une extension de cette approche dans le traitement automatique de la langue naturelle. Nous utilisons ainsi des notions élémentaires de la théorie des graphes pour donner une interprétation de l'énergie textuelle comme une nouvelle mesure de similarité. En Section 4 nous appliquons nos algorithmes à la génération de résumés automatiques et à la détection de frontières thématiques, avant de conclure et presenter quelques perspectives.

2 L'approche énergétique de Hopfield

La contribution la plus importante de Hopfield à la théorie de réseaux de neurones a été l'introduction de la notion d'énergie issue de l'analogie avec les systèmes magnétiques. Un système magnétique est constitué d'un ensemble de N petits aimants appelés spins. Ces spins peuvent s'orienter selon plusieurs directions. Le cas le plus simple est représenté par le modèle d'Ising qui considére seulement deux directions possibles : vers le haut ($\uparrow$, +1 ou 1) ou vers le bas ($\downarrow$, -1 ou 0). Le modèle d'Ising 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 d'Hebb[2] :

$$J^{i,j} = \sum_{\mu=1}^{P} s^i_\mu s^j_\mu \tag{1}$$

s^i et s^j sont les états des neurones i et j. Les autocorrelations ne sont pas calculées ($i \neq j$). La sommation porte sur les P patrons a stocker. Cette règle d'interaction est locale, car $J^{i,j}$ dépend seulement des états des unités connectées. Ce modèle est connu aussi comme mémoire associative. Elle 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) :

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s^i \; J^{i,j} \; s^j \tag{2}$$

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

L'énergie est fonction de la configuration du système, c'est-à-dire, de l'état (d'activation ou non activation) de toutes ces unités. Si on présente un patron ν, chaque spin subira un champ local $h^i = \sum_{j=1}^{N} J^{i,j} s^j$ induit par les autres N spins (voir figure 1). Les spins s'aligneront selon h^i

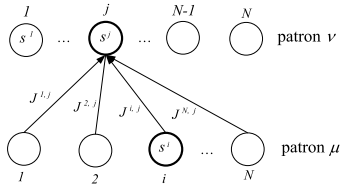

FIG. 1 – Champ h_i subi par le spin $s_j, \in$ la chaîne (patron) ν produit par les autres N spins $\in \mu$.

pour restituer le patron stocké qui est le plus proche au patron présenté ν. Nous n'allons pas détailler la méthode de recupération de patrons[3], car notre intérêt va porter sur la distribution et les propriétés de l'énergie du système (2). Cette fonction monotone et décroissante avait été utilisée uniquement pour montrer que l'apprentissage est borné. D'un autre côté, le modèle vectoriel (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 calculer l'énergie.

3 Applications au TALN

Les documents sont pré-traités avec des algorithmes classiques de filtrage de mots fonctionnels[4], de normalisation et de lemmatisation (Porter, 1980; Manning & Schutze, 2000) afin de réduire la dimensionnalité. Une représentation en sac de mots produit une matrice $S_{[P\times N]}$ de fréquences/absences composée de $\mu = 1, \cdots, P$ phrases (lignes) ; $\vec{\sigma}_\mu = \{s_\mu^1, \cdots, s_\mu^i, \cdots, s_\mu^N\}$ et un vocabulaire de $i = 1, \cdots, N$ termes (colonnes).

$$S = \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \\ \vdots & \vdots & \ddots & \vdots \\ s_P^1 & s_P^2 & \cdots & s_P^N \end{pmatrix} \; ; \; s_\mu^i = \begin{cases} TF^i & \text{si le terme } i \text{ existe} \\ 0 & \text{autrement} \end{cases} \tag{3}$$

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 $\vec{\sigma}_\mu$ est donc une chaîne de N spins. Nous allons nous différencier de (Hopfield, 1982) sur deux points : S est une matrice entière (ses éléments prennent des valeurs fréquentielles absolues) et nous utilisons les éléments $J^{i,i}$ car cette auto-corrélation permet d'établir l'interaction du mot i parmi les P phrases, ce qui est important en TALN. Pour calculer les interations entre les N termes du vocabulaire, on applique la règle de correlation de Hebb, qui en forme matricielle est égale à :

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

[3]Cependant le lecteur intéressé peut consulter, par exemple (Hopfield, 1982; Kosko, 1988; Hertz *et al*., 1991).
[4]Nous avons effectué le filtrage de chiffres et l'utilisation d'anti-dictionnaires.

Chaque élément $J^{i,j} \in J_{[N\times N]}$ est équivalent au calcul de (1). 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 d'interaction entre les patrons μ et ν (figure 1).

3.1 L'énergie textuelle : une nouvelle mesure de similarité

Nous allons expliquer théoriquement la nature des liens entre phrases que la mesure d'énergie textuelle induit. Pour cela nous utilisons quelques notions élémentaires de la théorie des graphes. L'interprétation que nous allons faire repose sur le fait que la matrice (5) peut s'écrire :

$$E = -\frac{1}{2} S \times (S^T \times S) \times S^T = -\frac{1}{2}(S \times S^T)^2 \tag{6}$$

Considérons les phrases comme des ensembles σ de mots. Ces ensembles constituent les sommets du graphe. On trace une arête entre deux de ces sommets σ_μ, σ_ν chaque fois qu'ils partagent au moins un mot en commun $\sigma_\mu \cap \sigma_\nu \neq \emptyset$. On obtient ainsi le graphe $I(S)$ d'intersection des phrases (voir un exemple à quatre phrases en figure 2). On value ces paires $\{\sigma_1, \sigma_2\}$ que l'on appelle arêtes par le nombre exact $|\sigma_1 \cap \sigma_2|$ de mots que partagent les deux sommets reliés. Enfin, on ajoute à chaque sommet σ une arête de réflexivité $\{\sigma\}$ valuée par le cardinal $|\sigma|$ de σ. Ce graphe d'intersection valué est isomorphe au graphe $G(S \times S^T)$ d'adjacence de la matrice carrée $S \times S^T$. En effet, $G(S \times S^T)$ contient P sommets. Il existe une arête entre deux sommets μ, ν si et seulement si $[S \times S^T]_{\mu,\nu} > 0$. Si c'est le cas, cette arête est valuée par $[S \times S^T]_{\mu,\nu}$, valeur qui correspond au nombre de mots en commun entre les phrases μ et ν. Chaque sommet μ est pondéré par $[S \times S^T]_{\mu,\mu}$ ce qui correspond à l'ajout d'une arête de reflexivité. Il en résulte

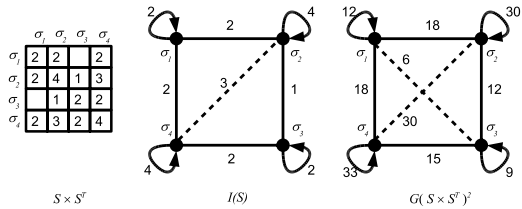

FIG. 2 – Graphes d'adjacence issus de la matrice d'énergie.

que la matrice d'énergie textuelle E est la matrice d'adjacence du graphe $G(S \times S^T)^2$ dont :
- les sommets sont les mêmes que ceux du graphe d'intersection $I(S)$;
- il existe une arête entre deux sommets chaque fois qu'il existe un chemin de longueur au plus 2 dans le graphe d'intersection ;
- la valeur d'une arête : a) boucle sur un sommet σ est la somme des carrés des valeurs des arêtes adjacentes au sommet et b) entre deux sommets distincts σ_μ et σ_ν adjacents est la somme des produits des valeurs des arêtes sur tout chemin de longueur 2 entre les deux sommets. Ces chemins pouvant comprendre des boucles.

De cette représentation on en déduit que la matrice d'énergie textuelle relie à la fois des phrases ayant des mots communs puisque elle englobe le graphe d'intersection, ainsi que des phrases qui partagent un même voisinage sans pour autant partager nécessairement un même vocabulaire. C'est à dire que deux phrases σ_1, σ_3 ne partageant aucun mot en commun mais pour lesquelles il existe au moins une troisième phrase σ_2 telle que $\sigma_1 \cap \sigma_2 \neq \emptyset$ et $\sigma_3 \cap \sigma_2 \neq \emptyset$ seront tout de même reliées. La force de ce lien dépend premièrement du nombre de phrases σ_2 dans leur voisinage commun, et donc du vocabulaire apparaissant dans un contexte commun.

4 Expériences et résultats

L'énergie textuelle peut être utilisée comme mesure de similarité dans les applications du TALN. De façon intuitive, cette similarité peut servir à scorer les phrases d'un document et séparer ainsi celles qui sont pertinentes de celles qui ne le sont pas. Ceci conduit immediatement à une stratégie de résumé automatique par extraction de phrases. Une autre approche, moins évidente, consiste à utiliser l'information de cette énergie (vue comme un spectre ou signal numérique de la phrase) et de la comparer au spectre de toutes les autres. Un test statistique peut alors indiquer si ce signal est semblable à celui d'autres phrases regroupés en segments ou pas. Ceci peut être vu comme une détection de frontières thématiques dans un document.

4.1 Résumé automatique

Sous l'hypothèse que l'énergie d'une phrase μ reflète son poids dans le document, nous avons appliqué (6) au résumé par extraction de phrases (Mani & Maybury, 1999; Radev *et al.*, 2002). Cette méthode est orientée, pour le moment, à la génération de résumés génériques monodocument. Cependant, nous pensons qu'une modification de l'approche (voir Section 5) pourrait nous permettre d'obtenir des résumés guidés par une requête ou un sujet défini par l'utilisateur (ce qui correspond au protocole des conféréences DUC[5]). L'algorithme de résumé comprend trois modules. Le premier réalise la transformation vectorielle du texte avec des processus de filtrage, de lemmatisation/*stemming* et de normalisation. Le second module applique le modèle de spins et réalise le calcul de la matrice d'énergie textuelle (6). Nous obtenons la pondération de la phrase ν en utilisant ses valeurs absolues d'énergie, c'est-à-dire, en triant selon $\sum_\mu |\boldsymbol{E}_{\mu,\nu}|$. Ainsi, les phrases pertinentes seront selectionnées comme ayant la plus grande énergie absolue. Finalement, le troisième module génère les résumés par affichage et concaténation des phrases pertinentes. Les deux premiers modules reposent sur le système Cortex[6]. Pour les tests en français[7] nous avons choisi les textes : « 3-mélanges » composé de trois thématiques, « puces » de deux thématiques et « J'accuse » (lettre d'Émile Zola). Deux textes de la wikipedia en anglais ont été analysés, « Lewinksky » et « Québec »[8]. Nous avons evalué les résumés produits par notre système avec ROUGE (Lin, 2004), qui mesure la similarité, suivant plusieurs stratégies, entre un résumé candidat (produit automatiquement) et des résumés de référence (créés par des humains). Nous comparons dans les tables 1 à 5 les performances de la méthode d'énergie, de

[5]Document Understandig Conferences http ://www-nlpir.nist.gov/projects/duc/index.html

[6]Le système Cortex (Torres-Moreno *et al.*, 2002) effectue une extraction non supervisée de phrases pertinentes en utilisant plusieurs métriques pilotées par un algorithme de décision.

[7]Recupérables à l'adresse http ://www.lia.univ-avignon.fr.

[8]http ://en.wikipedia.org/wiki/Monica_Lewinsky, http ://en.wikipedia.org/wiki/Quebec_sovereignty_movement

Cortex et d'une *baseline* où les phrases ont été choisies au hasard. Nous constatons que notre méthode est comparable au système Cortex en termes de précision, de rappel et de F-score.

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,49577	0,50635	**0,49676**	**0,50643**	0,29125	0,3117
Précision	**0,43229**	**0,44114**	0.42288	0,43068	0,32801	0,35191
F**-score**	**0,46186**	**0,47150**	0.45685	0,46549	0,30744	0,32936

TAB. 1 – Texte « 3-mélanges » (27 phrases, 826 mots ; résumé au 25% ; 8 résumés référence).

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,52040	0,53353	**0,53595**	**0,55878**	0,25938	0,27721
Précision	0,52469	0,53796	**0,53120**	**0,55380**	0,37589	0,40474
F**-score**	0,52254	0,53574	**0,53356**	**0,55628**	0,30530	0,32723

TAB. 2 – Texte « puces » (29 phrases, 653 mots ; résumé au 25% ; 8 résumés référence).

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,61457	0,64192	**0,63160**	**0,65987**	0,18690	0,20185
Précision	0,51425	0,53700	**0,52725**	**0,55071**	0,30920	0,37195
F**-score**	0,55995	0,58479	**0,57473**	**0,60037**	0,21766	0,26152

TAB. 3 – Texte « J'accuse » (206 phrases, 4936 mots ; résumé au 12% ; 6 résumés référence).

4.2 Détection de frontières thématiques

Plusieurs stratégies ont été développées pour segmenter thématiquement un texte. Elles peuvent être supervisées ou non. 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). De façon originale, nous avons utilisé la matrice d'énergie E (6). Ce choix permet de s'adapter à de nouvelles thématiques et de rester indépendant vis à vis de la langue des documents. Pour pouvoir comparer les énergies entre elles nous introduisons le coefficient de concordance W de Kendall (Siegel & Castellan, 1988) et le calcul de sa $p-$valeur. Ils permettent de définir un test statistique de concordance entre k juges qui classent un ensemble de P objets. Nous avons utilisé ce test pour trouver les frontières thématiques entre segments. Nous montrons en figure (3) l'énergie d'interaction entre quelques phrases d'un texte composé de deux thématiques. Étant donné que (6) est capable de détecter et de pondérer le voisinage d'une phrase, on peut constater une similarité entre les courbes de l'une (gras) et de l'autre thématique (pointillé). Voici le protocole de test que nous avons adopté.

1. Selon la nature du texte (homogène ou hétéroclite) on émet a priori l'une des deux hypothèses initiales H_0 qui suivent : *i)* la phrase $\mu + 1$ appartient à la même thématique que la phrase précédente μ ou au contraire *ii)* la phrase $\mu + 1$ marque une rupture avec μ.

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,56107	0,57859	**0,61832**	**0,62705**	0,24227	0,25584
Précision	0,39516	0,40658	**0,42587**	**0,43085**	0,32490	0,34393
F**-score**	0,46372	0,47757	**0,50436**	**0,51076**	0,27671	0,29248

TAB. 4 – Texte « Lewinsky » (30 phrases, 816 mots ; résumé au 20% ; 7 résumés référence).

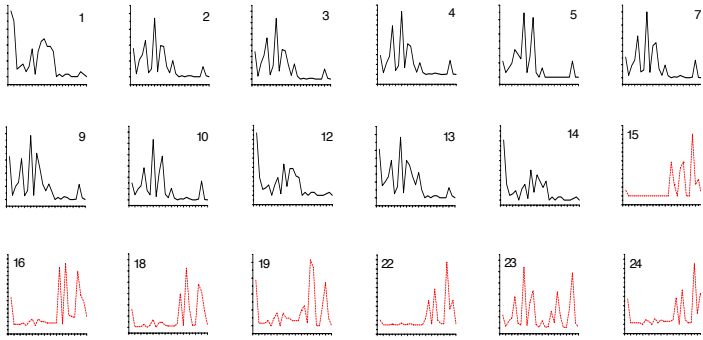

FIG. 3 – Énergie textuelle de « 2-mélanges ». En trait continu l'énergie des phrases de la 1ère thématique, en pointillé celle de la 2ème. Le changement d'allure des courbes entre les phrases 14-15 correspond à un changement thématique. L'axe horizontal indique le numéro de phrase dans l'ordre du document. L'axe vertical, l'énergie textuelle de la phrase affichée vs. les autres.

2. On estime alors la probabilité p que l'hypothèse H_0 choisie soit vérifiée en calculant le coefficient de concordance W de Kendall entre les deux classements par proximité induits par les phrases μ et $\mu + 1$ sur les autres phrases. Le coefficient W de Kendall vaut 1 en cas d'accord total entre les classements et 0 dans la cas de désaccord total. La probabilité p est alors estimée en utilisant l'approximation de la loi du W par une loi du χ^2.
3. Finalement, si $p < 0,1$ on rejette H_0 et l'on adopte l'hypothèse alternative (son complémentaire) avec un risque p de se tromper. Il est important de préciser que la valeur seuil $0,1$ est fixée a priori conformémént à la méthodologie statistique inférentielle.

Il s'agit donc de tests non-paramétriques qui ne requièrent aucune supposition sur une éventuelle distribution gaussienne des données. Pour chaque document, nous avons éliminé les phrases dont l'énergie est inférieure à un seuil. Ces phrases sont celles qui contiennent un nombre de mots < 2 (patrons à spins 0) ou trop longues (si l'on a suffisamment de phrases par segment), et qui induisent un fort bruit dans E. Les figures (4) et (5) montrent la détection

	Énergie		**Cortex**		**Baseline**	
	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4	Rouge-2	Rouge-SU4
Rappel	0,50945	0,53773	**0,56364**	**0,58716**	0,27344	0,32127
Précision	0,46276	0,48824	**0,50803**	**0,52899**	0,33254	0,39092
F**-score**	0,48498	0,51179	**0,53439**	**0,55656**	0,29991	0,35244

TAB. 5 – Texte « Québec » (44 phrases, 1190 mots ; résumé au 25% ; 8 résumés référence).

des frontières pour les textes à 2 et 3 thématiques. Les véritables frontières sont indiquées en pointillé. Ce protocole de test, en adoptant l'hypothèse *ii)* comme H_0, a détecté une frontière entre les phrases 14-15 pour le texte « 2-mélanges ». Pour le texte « 3-mélanges », le test a trouvé deux frontières entre les segments 8-9 et 16-18. Dans les deux cas, cela correspond effectivement aux frontières thématiques. Une troisième (fausse) frontière a été signalée entre les phrases 23-24 du texte « 2-mélanges ». Cela mérite d'être commenté : si on regarde sur la figure (3) l'énergie de la phrase 23, elle est bien différente de celle des phrases 22 ou 24. La phrase 23 présente une courbe chevauchant les deux thématiques. C'est pourquoi le test ne peut pas l'identifier comme appartenant à la même classe. Le même raisonnement tient pour toutes les fausses frontières. Pour le texte « physique-climat-chanel » le test du W de Kendall a détecté

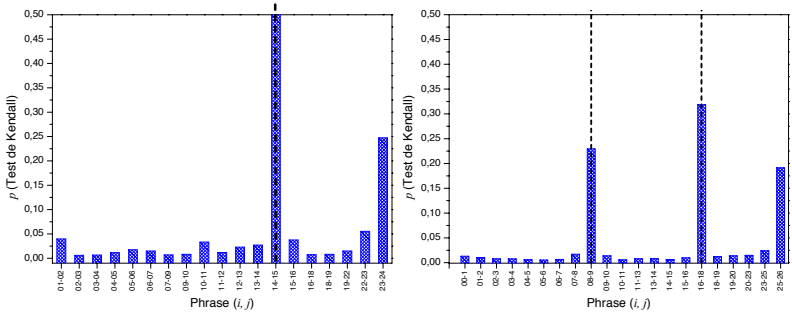

FIG. 4 – Détection des frontières pour le texte « 2-mélanges » (2 thématiques, à gauche) et « 3-mélanges » (3 thématiques, à droite).

trois frontières entre les phrases 5-6 et 12-15, qui correspondent aux frontières effectives. Pour le texte en anglais à deux thématiques le test a trouvé une frontière entre les segments 44-45 qui correspond à la vraie frontière. Nous avons calculé le F-score de façon similaire à DEFT 2005

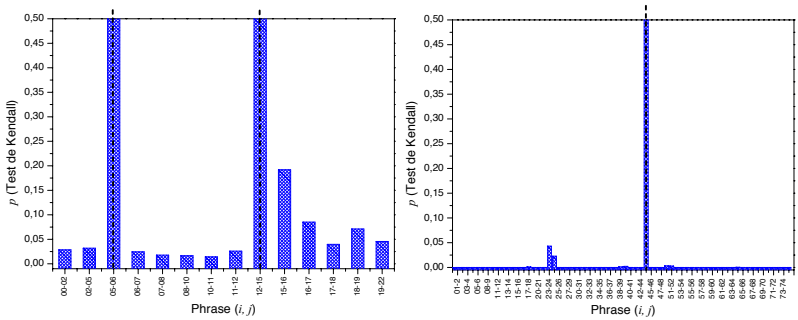

FIG. 5 – Détection des frontières pour le texte en français à 3 thématiques « physique-climat-chanel » à gauche et en anglais « québec-lewinsky » à droite.

(Alphonse *et al.*, 2005) [9]. Ainsi pour « 2-mélanges » F-score = 0,66 ; « 3-mélanges » F-score =

[9] En considérant $\beta = 1$, $F\text{-score}(\beta) = \frac{2 \times \text{Nb\_frontières\_correctes\_extraites}}{\text{Nb\_total\_frontières\_extraites} + \text{Nb\_total\_véritables}}$

0,66 ; « physique-climat-chanel » F-score = 0,80 et « québec-lewinsky » F-score = 1. Dans une autre expérience, nous avons comparé notre système à deux autres : LCseg (Galley *et al.*,) et LIA_seg (Sitbon & Bellot, 2005), qui utilisent tous les deux des chaînes lexicales. Le corpus de reférence a été construit à partir d'articles du journal Le Monde. Il est composé d'ensembles de 100 documents où chacun correspond à la taille moyenne des segments pré-définie. Un document est composé de 10 segments extraits d'articles thématiquement différentes tirés au hasard. Compte tenu des particularités de ce corpus nous avons adopté *i* comme hypothèse initiale H_0. Les scores sont calculés avec Windiff (Pevzner & Hearst, 2002), utilisée dans la segmentation thématique. Cette fonction mesure la différence entre les frontières véritables et celles trouvées automatiquement dans une fenêtre glissante : plus la valeur est petite, plus le système est performant. LIA_seg depend d'un paramètre qui donne lieu à différentes performances (d'où la plage de valeurs affichée). Notre méthode obtient des performances comparables aux systèmes dans l'état de l'art mais en utilisant bien moins de paramètres, en particulier nous ne faisons aucune supposition sur le nombre de thématiques à détecter.

Taille du segment (en phrases)	**LCseg**	**LIA_seg**	**Énergie**
9-11	0,3272	(**0,3187**-0,4635)	0,4419
3-11	0,3837	(**0,3685**-0,5105)	0,4403
3-5	0,4344	(0,4204-0,5856)	**0,4167**

TAB. 6 – Mesure Windiff pour LCseg, LIA_seg et Énergie (segments de différentes tailles).

5 Conclusion et perspectives

Nous avons introduit le concept d'énergie textuelle basé sur des approches des réseaux de neurones. Cela nous a permis de développer un nouvel algorithme de résumé automatique. Des tests effectués ont montré que notre algorithme est adapté à la recherche de segments pertinents. On obtient des résumés équilibrés où la plupart des thèmes sont abordés dans le condensé final. Les avantages supplémentaires consistent à ce que les résumés sont obtenus de façon indépendante de la taille du texte, des sujets abordés, d'une certaine quantité de bruit et de la langue (sauf pour la partie pré-traitement). Nous pensons que l'algorithme d'énergie pourrait être incorporé au système Cortex, où il jouerait le rôle d'une des métriques pilotée par un algorithme de décision. Ceci permettrait d'obtenir des résumés à l'aide d'une requête de l'utilisateur ou des résumés multi-documents. Une autre voie intéressante est le calcul des propriétés comme la « magnétisation » d'un document. (Shukla, 1997) a etudié des phénomènes magnétiques dans les réseaux de neurones type Hopfield dont on pense se servir. On étudiera la réponse du système face à l'application d'un champ externe. Ce champ, représenté par le vecteur des termes d'un texte decrivant une thématique (topique) sera mis en relation avec un document. Ainsi, les phrases du document pourraient, ou non, s'aligner selon leur degré de pertinence par rapport à la thématique. Ceci permettrait de générer des résumés personnalisés, telles que définis dans les tâches DUC. Nous avons également abordé le problème de la détection des frontières thématiques des documents. La méthode, basée sur la matrice d'énergie du système de spins, est couplée à un test statistique non-paramètrique robuste. Les résultats sont très encourageants. Une critique de la méthode d'énergie pourrait être qu'elle nécessite la manipulation (produits, transposée) d'une matrice de taille $[P \times P]$. Cependant la représentation sous forme de graphe nous permet d'exécuter ces calcus en temps $P \log(P)$ et en espace P^2.

Références

ALPHONSE E., AMRANI A., AZÉ J., HEITZ T., MEZAOUR A.-D. & ROCHE M. (2005). Préparation des données et analyse des résultats de DEFT'05. In *TALN 2005 - Atelier DEFT'05*, volume 2, p. 95–97.

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 *Thirteenth ACM conference on Information and knowledge management*, p. 127–136, Washington, D.C., USA.

GALLEY M., MCKEOWN K. R., FOSLER-LUSSIER E. & JING H. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, Sapporo, Japan.

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.

KOSKO B. (1988). Bidirectional associative memories. *IEEE Transactions Systems Man, Cybernetics*, **18**, 49–60.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)*.

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

MANI I. & MAYBURY M. T. (1999). *Advances in Automatic Text Summarization*. MIT Press.

MANNING C. D. & SCHUTZE H. (2000). *Foundations of Statistical Natural Language Processing*. The MIT Press.

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.

RADEV D., WINKEL A. & TOPPER M. (2002). Multi Document Centroid-based Text Summarization. In *ACL 2002*.

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

SHUKLA P. (1997). Response of the Hopfield-Little model in an applied field. *Physical Review E*, **56**(2), 2265–2268.

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

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

TORRES-MORENO J.-M., VELÁZQUEZ-MORALES P. & MEUNIER J. (2002). Condensés de textes par des méthodes numériques. In *JADT*, volume 2, p. 723–734.