Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic segmentation

Silvia Fernández[α,∇], Eric SanJuan[∇] and Juan Manuel Torres-Moreno[∂,∇]**

[∇] Laboratoire Informatique d'Avignon, BP 1228 F-84911 Avignon Cedex 9 FRANCE
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr
http://www.lia.univ-avignon.fr
[∂] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7, Montréal (Québec), Canada.
[α] Laboratoire de Physique des Matériaux, CNRS UMR 7556, Nancy, France.

Abstract. In this paper we present a Neural Network approach, inspired by statistical physics of magnetic systems, to study fundamental problems of Natural Language Processing (NLP). 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 Topic Segmentation.

Key words: Automatic Summarization, Topic Segmentation, Statistical Methods, Statistical Physics

1 Introduction

Hopfield [1, 2] took as a starting point physical systems like the magnetic Ising model (formalism resulting from statistical physics describing a system composed of units with two possible states named spins) to build a Neural Network (NN) with abilities of learning and recovery of patterns. The capacities and limitations of this Network, called associative memory, were well established in a theoretical frame in several studies [1, 2]: the patterns must be not correlated to obtain free error recovery, the system saturates quickly and only a little fraction of the patterns can be stored correctly. As soon as their number exceeds $\approx 0,14N$, any pattern is recognized. This situation strongly restricts the practical applications of Hopfield Network. However, in NLP, we think that it is possible to exploit this behavior. Vector Space Model (VSM) [3] represents the sentences of a document into vectors. These vectors can be studied as Hopfield NN. With a vocabulary of N terms of a document, it is possible to represent a sentence as a chain of N neurons actives (words are presents) or inactives (words are absents). A document with P sentences is formed of P chains in the vector space Ξ of

** Corresponding author.

dimension N. These vectors are correlated according to the shared words. If thematics are close, it is raisonable to suppose that the degree of correlation will be very high. That is a problem if we want to store and recover these representations from a Hopfield NN. However, our interest does not relate with recovery, but to study the interactions between the terms and the sentences. From these interactions we have defined the Textual Energy of a document. It can be useful, for example, to score or to detect changes between sentences. We have developed a metaphor which makes possible to use the concept of Textual Energy in automatic summarization or topic segmentation tasks. We present in Section 2 a short introduction to the model of Hopfield. In Section 3, we show an extension of this approach in Natural Language Processing. We use elementary notions of the graph theory to give an interpretation of Textual Energy like a new measurement of similarity. In Section 4 we apply our algorithms to the generation of automatic summaries and the detection of topic boundaries, before concluding and presenting some prospects.

2 The Model of Hopfield

Certainly the most important contribution of Hopfield to the theory of NN was the introduction of the notion of energy that comes from the analogy with the magnetic systems. A magnetic system is constituted of a set of N small magnets called spins. These spins can turn according to several directions. The simplest case is represented by the Ising model which considers only two possible directions: up (↑, +1 or 1) or down (↓, -1 or 0). The Ising model is used in several systems which can be described by binary variables [4]. A system of N binary units has $\nu = 1, ..., 2^N$ possible configurations (patterns). In the Hopfield model the spins correspond to the neurons, interacting with the Hebb learning rule[1]:

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

s^i et s^j are the states of neurons i and j. Autocorrelations are not calculated $(i \neq j)$. The summation concerns the P patterns to store. This rule of interaction is local, because $J^{i,j}$ depends only on states of the connected units. This model is also known as associative memory. It has the capacity to store and to recover certain number of configurations of the system, because the Hebb rule transforms these configurations into attractors (minimal local) of the energy function [1]:

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

Clearly the energy is a function of the system configuration, that is, of the state (of activation or non-activation) of all these units. If we present a pattern ν, every spin will undergo a local field $h^i = \sum_{j=1}^{N} J^{i,j} s^j$ induced by the others N spins (figure 1). Spins will align themselves according to h^i in order to restore

[1] The connections are proportionals to the correlation between neurons' states [2].

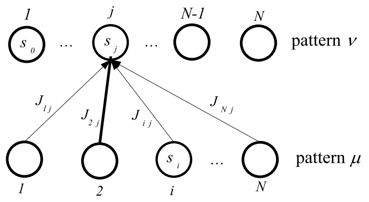


Fig. 1. Field h_i created by the units of the pattern μ affects the pattern ν.

the stored pattern that is the nearest one to the presented pattern ν. We will not detail the pattern recovery method[2], because our interest will concern the distribution and the properties of the energy of the system (2). This monotonic and decreasing function had only been used to show that the recovery is convergent. VSM [3] transforms documents in an adequate space where a matrix S contains the information of the text in the form of bags of words. We can consider S as the configuration set of a system which we can calculate its energy.

3 Applications in NLP

Documents are pre-treated with classical algorithms of functional words filtering[3], normalization and lemmatisation [6, 7] to reduce the dimensionality. A representation in bag of words produces a matrix $S_{[P\times N]}$ of frequencies/absences consisting of $\mu = 1, \cdots, P$ sentences (lines); $\boldsymbol{\sigma}_\mu = \{s_\mu^1, \cdots, s_\mu^i, \cdots, s_\mu^N\}$ and a vocabulary of $i = 1, \cdots, N$ terms (columns).

$$S = \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \\ \vdots & \vdots & \ddots & \vdots \\ s_P^1 & s_P^2 & \cdots & s_P^N \end{pmatrix} \; ; \; s_\mu^i = \begin{cases} TF^i & \text{if word } i \text{ exists} \\ 0 & \text{elsewhere} \end{cases} \tag{3}$$

Because the presence of the word i represents a spin $s^i \uparrow$ with a magnitude given by its frequency TF^i (its absence by $\downarrow$ respectively), a sentence $\boldsymbol{\sigma}_\mu$ is therefore a chain of N spins. We differ from [1] on two points: S is a whole matrix (its elements take absolute frequential values) and we use the elements $J^{i,i}$ because this autocorrelation makes possible to establish the interaction of the word i among the P sentences, which is important in NLP. We apply Hebb's rule (in matricial form) to calculate the interactions between N terms of the vocabulary:

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

[2] However the interested reader can consult, for example, [1, 5, 2].

[3] Filtering of numbers and stop-words.

Each element $J^{i,j} \in J_{[N \times N]}$ is equivalent to the calculation of (1). The Textual Energy of interaction between patterns (figure 1) (2) can be expressed:

$$E = -\frac{1}{2} S \times J \times S^T \; ; \; E_{\mu,\nu} \in E_{[P \times P]} \tag{5}$$

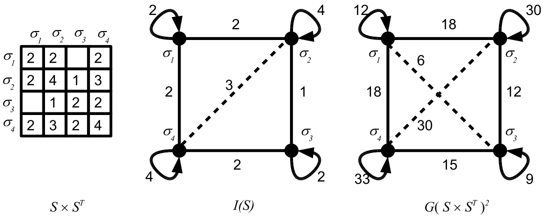
$E_{\mu,\nu}$ represents the energy of interaction between patterns μ and ν.

3.1 Textual Energy: a new similarity measure

At this level we are going to explain theoretically the nature of the links between sentences that Textual Energy infers. To do that, we use some elementary notions of the graph theory. The interpretation that we are going to do, is based on the fact that the matrix (5) can be written:

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

Let us consider the sentences as sets σ of words. These sets constitute the vertices of the graph. We draw an edge between two of these vertices σ_μ, σ_ν every time they share at least a word in common $\sigma_\mu \cap \sigma_\nu \neq \emptyset$. We obtain the graph $I(S)$ from intersection of the sentences (see an example of four sentences in figure 2). We evaluate these pairs $\{\sigma_1, \sigma_2\}$, which we call edges, by the exact number $|\sigma_1 \cap \sigma_2|$ of words that share the two connected vertices. Finally, we add to each vertex σ an edge of reflexivity $\{\sigma\}$ valued by the cardinal $|\sigma|$ de σ. This valued intersection graph is isomorphic with the adjacency graph $G(S \times S^T)$ of the square matrix $S \times S^T$. In fact, $G(S \times S^T)$ contains P vertices. There is an edge between two vertices μ, ν if and only if $[S \times S^T]_{\mu,\nu} > 0$. If it is the case, this edge is valued by $[S \times S^T]_{\mu,\nu}$ and this value corresponds to the number of words in common between the sentences μ and ν. Each vertex μ is balanced by $[S \times S^T]_{\mu,\mu}$, which corresponds to the addition of an edge of reflexivity. It results that the matrix of Textual Energy E is the adjacency matrix of the graph $G(S \times S^T)^2$ in which:



	σ_1	σ_2	σ_3	σ_4
σ_1	2	2		2
σ_2	2	4	1	3
σ_3		1	2	2
σ_4	2	3	2	4

Fig. 2. Adjacency graphs from the matrix of energy.

- the vertices are the same ones that those of the graph of intersection $I(S)$;
- there is an edge between two vertices each time that there is a way of length 2 in the graph of intersection;
- the value of an edge: a) loop on a vertex σ is the sum of the squares of the values of adjacent edges at the vertex, and b) between two distinct adjacent vertices σ_μ and σ_ν is the sum of the products of the values of the edges on any way with length 2 between both vertices. These ways can include loops.

From this representation we deduce that the matrix of Textual Energy connects at the same time sentences having common words because it includes the graph of intersection, as well as sentences which share the same neighbourhood without to necessarily share the same vocabulary. So, two sentences σ_1, σ_3 not sharing any word in common but for which there is at least one third phrase σ_2 as it is $\sigma_1 \cap \sigma_2 \neq \emptyset$ and $\sigma_3 \cap \sigma_2 \neq \emptyset$, will be connected all the same. The strength of this link depends in the first place on the number of sentences σ_2 in its common neighbourhood, and as well on the vocabulary appearing in a common context.

4 Experiments and results

Textual Energy can be used as a similarity measure in NLP applications. In an intuitive way, this similarity can be used in order to score the sentences of a document and thus separate those which are relevant from those which are not. This leads immediately to a strategy for automatic summarization by extraction of sentences. Another approach, less evident, consists in using the information of this energy (seen as a spectrum or numerical signal of the sentence) and to compare with the spectre of all the others. A statistical test can then indicate if this signal is similar to the signal of other sentences grouped together in segments or not. This can be seen as a detection of thematic boundaries in a document.

4.1 Mono-Document Generic Summarization

Under the hypothesis that the energy of a sentence μ reflects its weight in the document, we applied (6) to summarization by extraction of sentences [8, 9]. The summarization algorithm includes three modules. The first one makes the vectorial transformation of the text with filtering, lemmatisation/stemming and standardization processes. The second module applies the spins model and makes the calculation of the matrix of textual energy (6). We obtain the weighting of a sentence ν by using its absolute energy values, by sorting according to $\sum_\mu |\boldsymbol{E_{\mu,\nu}}|$. So, the relevant sentences will be selected as having the biggest absolute energy. Finally, the third module generates summaries by displaying and concatenating of the relevant sentences. The two first modules are based on the Cortex system[4]. French texts[5] choosed are: *3-melanges* made up of three topics, *Puces*

[4] The Cortex system [10] is an unsupervised summarizer of relevant sentences using several metrics controlled by an algorithm of decision.

[5] http://www.lia.univ-avignon.fr/chercheurs/torres/recherche/cortex

of two topics and *J'accuse* (Emile Zola's letter). Three texts of the Wikipedia in English were analysed, *Lewinksky*, *Quebec* and *Nazca Lines*[6]. We evaluated the summaries produced by our system with Rouge 1.5.5 [11], which measures the similarity, according to several strategies, between a candidate summary (produced automatically) and summaries of reference (created by humans). We compare in table 1 the performances of the energy method, Mead system[7] that produces only English summaries (symbols ⊘ in table), Copernic Summarizer[8], Cortex and a Baseline where the sentences were randomly selected. The compression rate was variable (following the size of the texts) and computed as a percent of number of sentences of text. The best performances are in fat-line and in italic those in 2d position (all scores). Enertex is a powerful summarizing system (it obtains 3 firsts places and 7 second), near of Cortex system.

Corpus	**Mead**		**Copernic**		**Enertex**		**Cortex**		***Baseline***	
	R2	SU4	R2	SU4	R2	SU4	R2	SU4	R2	SU4
3-melanges	⊘	⊘	0.4231	0.4348	*0.4958*	**0.5064**	**0.4968**	**0.5064**	0.3074	0.3294
Puces	⊘	⊘	**0.5775**	**0.5896**	0.5204	0.5336	*0.5360*	*0.5588*	0.3053	0.3272
J'accuse	⊘	⊘	0.2235	0.2707	*0.6146*	*0.6419*	**0.6316**	**0.6599**	0.2177	0.2615
Lewinsky	0.4756	0.4744	0.5580	0.5610	*0.5611*	*0.5786*	**0.6183**	**0.6271**	0.2767	0.2925
Quebec	0.4820	0.3891	0.4492	0.4859	*0.5095*	*0.5377*	**0.5636**	**0.5872**	0.2999	0.3524
Nazca	0.4446	0.4671	0.4270	0.4495	**0.6158**	**0.6257**	*0.5894*	*0.5966*	0.3041	0.3288

Table 1. Rouge-2 (R2) and SU4 score recall. 25%: *3-melanges* (8 ref), *Puces* (8 ref), *Québec* (8 ref) and *Nazca* (6 ref) ; 12%: *J'accuse* (6 ref); 20%: *Lewinsky* (7 ref).

4.2 Query Oriented Multi-Document Summarization

The main task of the NIST-Document Understanding Conference DUC'07[9] is given 45 topics and their 25 document clusters, to generate 250-word fluent summaries that answer the question(s) in the topics statements. In order to calculate the similarity between every topic and the phrases contained in the corresponding cluster we have used Textual Energy (2). Consequently the summary is formed with the sentences that present the maximum interaction energy with the topic. We describe now the process of summary constructionnd using the matricial forms of J and E (4,5). At first, the 25 documents of the cluster are concatenated into a single document in chronological order. Placing the topic like one more sentence of this long document (for exemple as the last sentence) the Textual Energy between the topic and each of the other sentences in the

[6] `http://en.wikipedia.org/wiki/Quebec_sovereignty_movement;Monica_Lewinsky;Nazca_lines`

[7] `http://tangra.si.umich.edu/clair/md/demo.cgi`

[8] `http://www.copernic.com`

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

document is computed using 5. Finally, we are only interested in recovering the row ρ of matrix E which corresponds to interaction energy of the topic vs. the document. We construct the summary by sorting the most relevant values of ρ.

Redundancy removal In general, in multi-document summarization there is a significant probability of including duplicated information. To avoid this problem, it must to implement a redundancy elimination strategy. Our system does not include any linguistic processing, then our non-redundancy strategy consists in comparing the energy values of sentences in the generated summary. We suppose that (in a long corpora) the probability that two sentences have the same values of energy is very small. Then we detected the duplicated phrases (with exactly the same energy value) and we replace them by the following ones in the score table. In effect we found that two phrases with the same textual value of energy are identical. Another strategy, enabling to diversify the content, is to omit the long sentences (there are sentences of comparable size to the one of the summary). The threshold with which we obtained the best results was two times the average of the number of words per sentence. Figure 3 shows the position of our system in the ROUGE automatic evaluation comparing to the 30 participants and two baselines –ID's 1 (random) and 2 (generic summarization system)–.

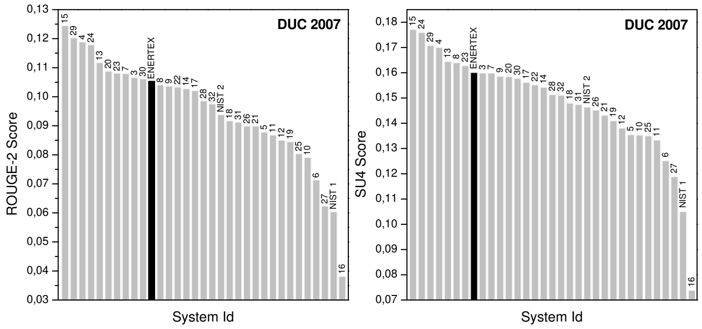


Fig. 3. Recall ROUGE-2 and SU4 of the 30 participants in DUC'07 and two baselines.

4.3 Topic Segmentation

Several strategies have been developed to segment a text thematically. Most of them are based on Markov models [12], classification of the terms [13, 14], lexical chains [15] or on PLSA model [16], that estimates the probabilities of the terms to belong to latent semantic classes. In an original way, we have used the matrix of energy E (6). This choice makes possible to adapt to new topics and to remain independent from document language. We show in figure 4 the energy of interaction between some sentences of a text made up of two topics. Given that (6) is capable of detecting and of balancing the neighbourhood of a sentence, we can notice a similarity between the curves of the one (fatty) and the other topics (dotted line). In order to compare energies between themselves we have used Kendall's τ coefficient of correlation. Given two sentences μ and ν, we estimate the probability $P[\mu \neq \nu]$ of being in distinct topics by the probability of $[\tau(x, y) > \tau(E_{\mu,.}, E_{\nu,.})]$. This is done using the normal approximation of Kendall's τ law valid if vectors $E_{\mu,.}, E_{\nu,.}$ have more than 30 terms. τ coefficient does not depend on exact energy values, only on their rank in the vectors $E_{\mu,.}, E_{\nu,.}$. Basically, it evaluates the degree of concordance between two rankings and makes possible robust non parametric statistical test of agreement between two judges classifying a set of P objects using that fact that $P[\tau(x, y) > \tau(E_{\mu,.}, E_{\nu,.})] = 1$ if the ranking vectors associated with $E_{\mu,.}$ and $E_{\nu,.}$ are two statistically independent variables. Here the judges are two sentences that classify all other sentences based on the interaction energy. We shall say that it is almost sure that two sentences μ and ν are in the same topic if $P[\mu \neq \nu] > 0.05$ We have used this test to find the thematic borders between segments. As illustrated in figure 5, a sentence is considered to be at the border of one segment if it is almost sure that: 1/ it is in the same topic as at least two over the three previous sentences; and 2/ it is not in the same topic as at least two over the three following sentences. We have implemented this approach of topic segmentation as a slippery window of seven sentences. As the window is moving on, the sencence on its center is compared to all other sentences in the window based on Kendall's τ coefficient. If a border is found then the window jumps over the next three sentences. Our programs have been optimized for standard PERL 5 libraries. Figures 6 and 7 show the detection of the boundaries for

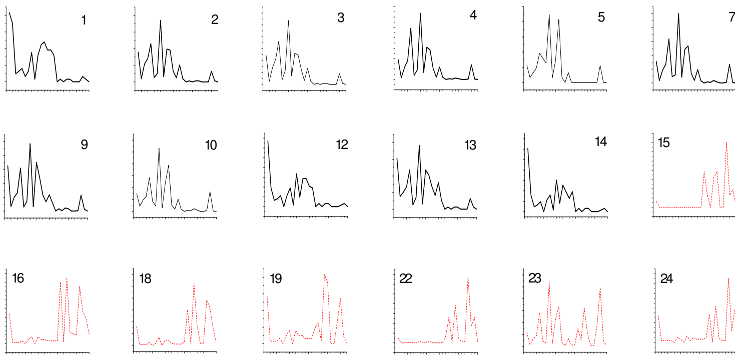


Fig. 4. Textual Energy of *2-melanges*. In continuous line, the energy of the sentences of 1^{th} topic, in dotted line that of 2^{th}. The change of shape of the curves between sentences 14-15 corresponds to a topic boundary. The horizontal axis indicates the number of sentence in the order of the document. The vertical axis, the Textual Energy of the showed sentence vs. others.

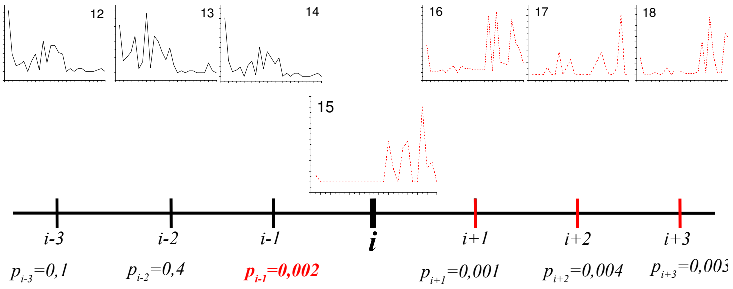


Fig. 5. Kendall's τ in window. $p_{i \pm k}$ = probability of concordance between $i \pm k$ and i.

the texts with 2 and 3 topics. The true boundaries are indicated in dotted line. For the text *3-mélanges*, the test found two borders between the segments 8-9 and 16-18. In both cases, that corresponds indeed to the thematic boundaries. The third (false) boundary was indicated between sentences 14-15 of the text *2-mélanges*. It deserves to be commented on. If we look at figure 4 we can notice that energy of the sentence 23 is very different from that of the sentences 22 or 24. Sentence 23 presents a curve overlapping the two topics. It is the reason why the test cannot identify it like pertaining to the same class. This reasoning can be extended to all other false borders. We show in figure 7 the boundary detection for texts with 3 and 4 thematics. For the text *physique-climat-chanel* we have detected three boundaries between the sentences 5-6 and 12-15, which corresponds to the real boundaries. For the text in English with two topics the test found one boundary between the segments 44-45 which also corresponds to the real one. In another experiment, we have compared our system to two oth-

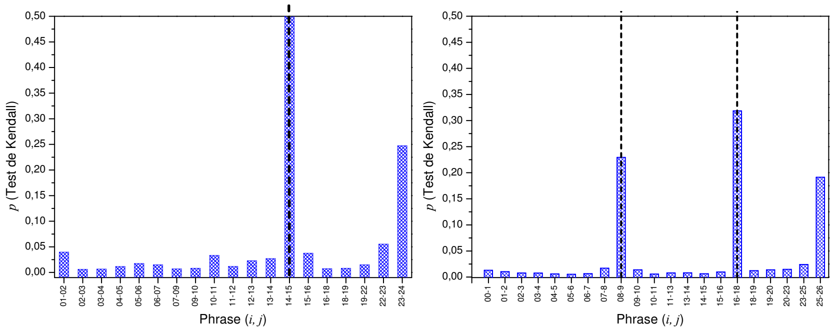


Fig. 6. Topic segmentation for the text *2-mélanges* (2 topics, on the left) and *3-mélanges* (3 topics, on the right).

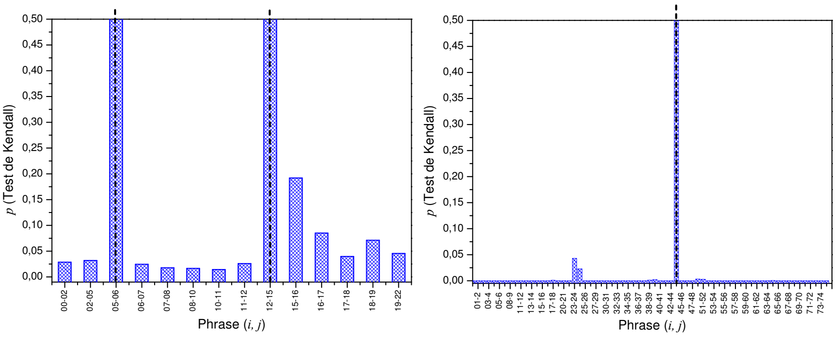


Fig. 7. Topic segmentation for the text in French with 3 topics *physique-climat-chanel* on the left and in English *Quebec-Lewinsky* on the right.

ers: LCseg [17] and LIA_seg [15] that are based on lexical chains. The corpus of reference was built by [15] from articles of the newspaper *Le Monde*. It is composed of sets of 100 documents where each one corresponds to the average size of the predefined segments. A document is composed of 10 segments (9 borders) extracted from articles of different topics selected at random. The scores are calculated with [18], used in the topic segmentation. This function measures the difference between the real boundaries and those found automatically in a slippery window: the smaller the value is, the more the system is performant. LIA_seg depends on a parameter which gives place to various performances (that is why the evaluation of this system gives rise to a range of values). Our method, that uses much less parameters as we do not make any assumption on the number of topics to detect, obtains performances comparable to the systems in the state of the art. In table 2 we show these results as well as the average number of found borders using Enertex.

Segment size (sentences)	**LCseg**	**LIA_seg**	**Enertex**	(Found boundaries)
9-11	0.3272	(**0.3187**-0.4635)	0.4134	7.10/9
3-11	0.3837	(**0.3685**-0.5105)	0.4264	7.15/9
3-5	0.4344	(0.4204-0.5856)	**0.4140**	5.08/9

Table 2. Windiff for LCseg, LIA_seg and Enertex (variable size segments).

5 Conclusion and perspectives

We have introduced the concept of Textual Energy based on approaches of NN that have enabled us to develop a new algorithm of automatic summarization.

Several experiments have show that our algorithm is adapted to extract relevant sentences. The majority of the topics are approached in the final digest. The summaries are obtained independently of the size of the text, subjects and language (except for the preprocessing part), and a few quantity of noise is tolerated. Query-guided summaries will be obtained by introducing the topic as the last sentence. Some performant tests on the DUC'07 corpora were realized. We also have studied the problem of topic segmentation of the documents. The method, based on the energy matrix of the system of spins, is coupled with a robust statistical non-parametric test Kendall τ. The results are very encouraging. A criticism of this algorithm could be that it requires the handling (produced, transposed) of a matrix of size $[P \times P]$. However the graph representation performs these calculations in time $P \log(P)$ and in space P^2.

References

1. Hopfield, J.: Neural networks and physical systems with emergent collective computational abilities. National Academy of Sciences **9** (1982) 2554–2558
2. Hertz, J., Krogh, A., Palmer, G.: Introduction to the theorie of Neural Computation. Addison Wesley, Redwood City, CA (1991)
3. Salton, G., McGill, M.: Introduction to modern information retrieval. Computer Science Series McGraw Hill Publishing Company (1983)
4. Ma, S.: Statistical Mechanics. World Scientific, Philadelphia, CA (1985)
5. Kosko, B.: Bidirectional associative memories. IEEE Transactions Systems Man, Cybernetics **18** (1988) 49–60
6. Porter, M.: An algorithm for suffix stripping. Program **14** (1980) 130–137
7. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (2000)
8. Mani, I., Maybury, M.T.: Adv. in Automatic Text Summarization. MIT (1999)
9. Radev, D., Winkel, A., Topper, M.: Multi Document Centroid-based Text Summarization. In: ACL 2002. (2002)
10. Torres-Moreno, J.M., Velázquez-Morales, P., Meunier, J.: Condensés de textes par des méthodes numériques. In: JADT. Volume 2. (2002) 723–734
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out (WAS 2004). (2004)
12. Amini, M.R., Zaragoza, H., Gallinari, P.: Learning for sequence extraction tasks. In: RIAO 2000. (2000) 476–489
13. Caillet, M., Pessiot, J.F., Amini, M., Gallinari, P.: Unsupervised learning with term clustering for thematic segmentation of texts. In: RIAO'04. (2004) 648–657
14. Chuang, S.L., Chien, L.F.: A practical web-based approach to generating Topic hierarchy for Text segments. In: ACM IKM, Washington (2004) 127–136
15. Sitbon, L., Bellot, P.: Segmentation thématique par chaînes lexicales pondérées. In: TALN 2005. Volume 1. (2005) 505–510
16. Brants, T., Chen, F., Tsochantaridis, I.: Topic-based document segmentation with probabilistic latent semantic anaysis. In: CIKM'02, Virginia, USA (2002) 211–218
17. Galley, M., McKeown, K.R., Fosler-Lussier, E., Jing, H.: Discourse segmentation of multi-party conversation. In: ACL-03, Sapporo, Japan (2003) 562–569
18. Pevzner, L., Hearst, M.: A critique and improvement of an evaluation metric for text segmentation. In: Computational Linguistic. Volume 1. (2002) 19–36