Applications exploratoires des modèles de spins au Traitement Automatique de la Langue

Silvia Fernandez Sabido

UNIVERSITÉ HENRI POINCARÉ, NANCY I

THÈSE

présentée à l'Université Henri Poincaré, Nancy I
pour obtenir le grade de Docteur en Sciences Physiques

SPÉCIALITÉ : PHYSIQUE STATISTIQUE

École Doctorale EMMA (Energie Mécanique MAtériaux) «409 Nancy-Metz»
Département de Physique de la Matière et des Matériaux
Institut Jean Lamour (anciennement LPM)

Applications exploratoires des modèles de spins au Traitement Automatique de la Langue

par
Silvia Fidelina FERNÁNDEZ SABIDO

Soutenue publiquement le 22 mai 2009 devant un jury composé de :

M^{me}	Mirta B. GORDON	DdR CNRS, TIMC-IMAG, Grenoble	Rapportrice
M.	Phillipe LANGLAIS	Professeur, DIRO, Montréal	Rapporteur
M.	Horacio SAGGION	Research Fellow, NLPG, Sheffield	Examinateur
M^{me}	Eva BUCHI	DdR CNRS, ATILF, Nancy	Examinatrice
M.	Daniel MALTERRE	Professeur, IJL, Nancy	Examinateur
M.	Bertrand BERCHE	Professeur, IJL, Nancy	Co-directeur
M.	Eric SANJUAN	MdC, LIA, Avignon	Co-directeur
M.	Juan M. TORRES MORENO	MdC HDR, LIA, Avignon	Directeur

Remerciements

Je tiens tout d'abord à remercier les Profs. Mirta Gordon et Phillipe Langlais d'avoir accepté d'être les rapporteurs de cette thèse. Ils ont contribué par leurs nombreuses remarques et suggestions à améliorer la qualité de ce mémoire. Je remercie les Profs. Horacio Saggion et Eva Buchi pour participer au Jury de soutenance. Également le Prof. Daniel Malterre pour présider ce Jury. Je remercie le *Consejo Nacional de Ciencia y Tecnología* (CONACYT) du Mexique pour le financement de cette thèse. Aussi les laboratoires LPM de Nancy et LIA d'Avignon pour leurs supports.

Je voudrais souligner les rôles des Profs. Luis Martínez et Daniel Malterre dans le choix qui est devenu finalement mon chemin scientifique, même si, comme ils les savent bien, j'aurai préféré autrement. Je remercie spécialement le Prof. Marc El-Bèze de m'avoir si gentiment accueilli au LIA pendant deux ans et demi.

Merci à mes trois directeurs de m'avoir guidé pendant la élaboration de cette thèse : à Juan Manuel Torres, je le remercie de m'avoir confié un projet si intéressante et original, de m'avoir toujours donné la liberté d'action et les outils TAL nécessaires pour développer ce travail. À Eric SanJuan, toujours gentil et humble, je le remercie d'avoir mis la main à la pâte pour améliorer ma rudimentaire façon de programmer et surtout pour les nombreuses discussions dans lesquelles nous avons dû systématiquement diviser le tableau en deux pour confronter nos différents points de vue (le prix de la pluridisciplinarité !). Les meilleures idées ont venu, bien sûr, quand il avait de la bière et des cacahuètes ! Je tiens à remercier tout spécialement Bertrand Berche pour avoir été à mon côté pendant la révision, en temps record, de ce manuscrit, pour ses idées et précisions en quant les modèles de spins utilisés, et surtout pour son support sincère dans les moments difficiles et son humanité (de la vraie). Sans l'un d'entre vous, ce travail serait un arc-en-ciel sans couleurs.

Je suis très reconnaissante à Patricia Velázquez, Iria Dacuna, Sonia Mandin et Fidelia Ibekwe, avec qui j'ai eu la fortune de collaborer, pour partager avec moi la richesse de leurs recherches, leurs esprits et leurs cœurs. À mes principaux relecteurs, Rémi Lavalley et Raphael Rubino, je dois tout ce qu'il est bien écrit dans ce manuscrit (le reste c'est moi !). Merci aux 3-Florians du LIA : Boudin, Pinault et Verdet ; toujours prêts à aider une mexicaine en détresse, soit pour la relecture, pour apprivoiser le linux ou pour installer des outils TAL ; mais surtout pour des choses plus sérieuses comme les dégustations de chocolat suisse, les cafés faits machine IUT ou le très attendu atelier tarte ! À tous le trois, merci d'être de très bons camarades.

Aux personnes dont leur travail professionnel et gentillesse ont fait spécialmente agréable les séjours à Nancy et Avignon. Du LIA : Simonne Mouzac, Jocelyne Gourret, Afssana Nourmamode et Frank Benoit. Du LPM : Sylvie Roberts, Nicole Nussmann, Martine Barbier, Cristian Senet, Danielle Pierre, Aymeric Avisou, Christophe Chatelain, Christine Sartori et Martine Gaulier. Merci à vous tous pour votre aide et sympathie.

Aux nombreuses amis des tous les coins du monde avec qui j'ai partagé des bons moments. Fadawine, Essaid, Habib, Khalil et Abdellatif (Marroc) ; Tembine (Mali) et Piotr (Pologne) ; Sujit, Kavitha, Sreenath, Sunitha, Amar et Vijay (Inde) ; Nimann (Djibouti) et Peter (Allemagne) ; Gilles, Rémi, Ti'Fred, Nicolas, Thierry et beaucoup d'autres (France). À notre Profe. de français, Noëlle Matis, pour ses précieux conseils et son amitié. Merci à Remy Kessler pour nous avoir offert l'intéressante expérience d'assister à un mariage en France (le sien). Ce détail nous a beaucoup touché.

Se trouver dans un autre continent favorise l'occasion de faire la connaissance simultanée des gens de toute l'Amérique Latine. Je pense que cela est l'une des plus riches expériences qu'on peut vivre à l'étranger. Ainsi, un espace spécial ont dans mon cœur mes chers amis *latino-americanos* rencontrés en France. *De México* : Rebe, Hugo, Avenilde et le petit André (*La Barca, Jalisco*) ; Luis, Claudia, Ale et Fernando (*Sabinas, Coahuila*) ; Alma et Raúl (*Monterrey, Nuevo León*) ; Karen et Luis (*DF y Michoacán*) ; Yahir (*Ciudad Victoria, Tamaulipas*) ; Joel et Sinuhé(*San Luis Potosí*). *De Chile* : Rodrigo et Mariela (*Chillán y Concepción*) ; Fernando (*Pingüinolandia*). *De Venezuela* : Julio, Sulan, Tania et famille, Alfonso et Maira (*Mérida y el Vigía*) ; *De Perú* : Lucy, Guillermo et les jumeaux. *De Cuba* : Rafael (*La Isla*). Merci à vous tous pour la solidarité, le support ou tout simplement pour les fêtes ou réunions express pour nous relaxer de *las marmoteadas*.

Une mention particulière mérite l'association CALMECALC et notamment Manuel Adam, un des ses fondateurs, pour la labour d'accompagnement des nouveaux arrivés latin-américains à Nancy. Nous avons eu la fortune d'avoir été assistés par Manuel dans nos premiers moments en France. Ce sont les moments où on panique pour le logement, la santé, le titre de séjour...pour tout ! À partir de ce geste, qui nous a beaucoup aidé, nous avons essayé de faire pareil avec le gens qui sont arrivés après nous, en leur suggérant, à leur tour, de continuer la labour. Nous espérons que la chaîne soit déjà longue, très très longue.

Je voudrais dédier ce travail à ma famille lointaine et pourtant proche. À ma mère Rosalía dont la force et détermination a été l'exemple à suivre. À mes soeurs Laura et Yura et mon frère Carlos pour leurs différentes façons d'être là. À leurs compagnons, Roy, Carlos et Rosy ; et surtout à leurs enfants, Andrea, Andrés, Ale, Ixchel, Fer, Willy et Lea, que j'aime plus que quiconque dans le monde. Vous êtes mon inspiration la plus forte. Merci à mes amis de Mérida, Tere, Leo et Juan Antonio, et à la troupe de Conkal, représentée par Carlos et Carmen, pour se souvenir de nous de temps en temps.

Et à toi Pedro, avec le seul que j'aurai osé de vivre l'expérience du mariage. Toujours à mon côté, en attendant dans la nuit mon arrivé du labo. J'adore la quantité de films français qu'on a vu ensemble, les bouquins qu'on a découvert dans cette belle langue et le *tpbb* qu'on laisse ici, dans le chemin d'arbres qui joint ton bureau et le mien.

Table des matières

Chapitre 1

Introduction

1.1 La Physique statistique

La Physique statistique s'intéresse au comportement de systèmes contenant une grande quantité de particules. Vues de manière isolée, ces particules obéissent à des équations de mouvement simples. Elles sont cependant trop nombreuses pour que l'on puisse les résoudre simplement. Par exemple, pour décrire le comportement physique d'un litre d'air, il faut considérer le mouvement et les collisions d'environ 3×10^{22} molécules (Newman et Barkema, 1999) (de l'ordre du nombre d'Avogadro [1]). En revanche, il est possible d'approcher le comportement général ou moyen d'un tel système. La Physique statistique offre ainsi un raccourci vers le calcul des propriétés globales au travers d'un regard probabiliste.

Par l'étude des probabilités des états d'un système, la Physique statistique a montré que l'ordre de grandeur du nombre de comportements, envisageables pour un grand système, est moins important que ce que l'on pouvait le penser. Ce fait est fort intéressant car des quantités, telle que l'énergie, peuvent être calculées sur le petit ensemble d'états que le système parcours pendant une expérience (Newman et Barkema, 1999).

Les techniques de la Physique statistique ont été appliquées principalement aux systèmes physiques comme les solides, les liquides et les gaz (Nestler et al., 2005; Szolnoki, 1999; Moukarzel et al., 2007) ; mais on retrouve également des applications aux systèmes chimiques et biologiques (Binder et al., 2008). Au fil du temps, les études ont été élargies à de nouvelles applications qui concernent des problématiques issues d'autres domaines, par exemple aux systèmes économiques (Farmer, 1999; Bartolozzi et al., 2006) et sociaux (Castellano et al., 2000; Nadal et Gordon, 2005). Le groupe du LPT d'Orsay [2] utilisant des techniques de la Physique statistique pour des études sur le trafic routier affirme : *chaque technique de la physique statistique est tôt ou tard susceptible de servir à résoudre un problème pour lequel elle n'avait pas été conçue initialement*.

1. La constante d'Avogadro est le nombre d'entités élémentaires contenues par mole. Sa valeur approchée est $N_A \approx 6,022 \times 10^{23}$.
2. Laboratoire de Physique Théorique d'Orsay, http://www.th.u-psud.fr

Bien que les méthodes utilisées soient diverses, les domaines d'application ont en commun le fait qu'ils traitent de systèmes à plusieurs composants interagissant entre eux. Tel est le cas de ce travail de thèse où nous avons perçu le texte comme un système dont les constituants (par exemple les mots) opèrent ensemble pour fournir aux documents des vertus qui leur sont propres (par exemple leur signification). Nous avons supposé qu'à partir de cette analogie, il serait possible de profiter des idées et des outils de la Physique statistique pour accomplir l'analyse de grandes quantités de documents.

1.2 Le Traitement Automatique de la Langue

Définie comme l'ensemble de signes oraux (le discours) et écrits (le texte) qui permettent à un groupe de communiquer, la langue possède toujours un sens associé, un message à donner. En général, on analyse un discours ou un texte pour extraire et manipuler son contenu conceptuel. Bien que l'être humain soit capable de mener à bien ces tâches, la quantité de données disponibles dépasse de loin ses capacités d'assimilation. D'où l'intérêt des techniques capables d'automatiser l'analyse de quantités considérables d'information.

Le Traitement Automatique des Langues (TAL) ou Traitement Automatique de la Langue Naturelle [3] (TALN) est une discipline scientifique très récente. Né aux États-Unis vers 1949 [4], le TAL est dédié à la conception de méthodes et d'outils informatiques pour analyser la langue humaine.

Notre étude a été réalisée au sein de la thématique Traitement Automatique de la Langue Naturelle Écrite (TALNE) du Laboratoire Informatique d'Avignon (LIA) [5], et concerne seulement la langue écrite. L'analyse automatique de textes représente toujours un grand défi. Les textes constituent des données non structurées [6] qu'il est impossible de représenter efficacement par des objets informatiques classiques avec un nombre limité et prédéfini d'attributs. Par nature le texte a un nombre illimité de dimensions. Il en résulte que les outils classiques d'analyse de données ne s'appliquent pas automatiquement à l'exploration des textes. De plus les bases de données textuelles sont de beaucoup plus grande taille que les bases de données classiques : il y a une quantité beaucoup plus grande de textes que de données structurées. De ce fait, le texte libre est une mine d'information que les techniques du TAL commencent à peine à exploiter. Ces techniques, en effet, permettent d'automatiser l'exploration d'une grande quantité de textes, qu'il serait impossible d'analyser manuellement dans son ensemble (Ibekwe-SanJuan, 2007).

3. Le terme TALN provient de la traduction de NLP (*Natural Language Processing*).
4. Le TAL est né pendant la guerre froide et pendant longtemps s'est concentré sur la traduction automatique avec évidemment un grand intérêt pour le passage du russe à l'anglais.
5. Dirigé actuellement par Marc El-Bèze (http://www.lia.univ-avignon.fr).
6. L'adjectif « structuré » désigne une organisation sous forme de tableau avec des variables et leur attributs (Ibekwe-SanJuan, 2007).

Les approches du TAL

En général, on trouve en TAL deux types de méthodes : les approches symboliques et les approches numériques. Les premières exploitent l'aspect structurel des textes en utilisant des règles linguistiques (syntaxiques et grammaticales par exemple). L'objectif est de décoder, puis de reproduire le processus par lequel la concaténation de symboles ayant un sens propre (les mots) s'organisent pour produire un sens plus général (le texte). Ces méthodes cherchent ainsi à simuler le processus de compréhension. En revanche, les approches numériques privilégient surtout le caractère fréquentiel des textes et utilisent des méthodes statistiques pour « calculer » le sens. Cette approche ne cherche guère à comprendre mais à reproduire une sortie adéquate.

Il est courant de retrouver une véritable confrontation entre l'un et l'autre type de méthodes. Par exemple, (Poibeau, 2003) arrive à la conclusion qu'on ne s'attaque plus guère à l'enjeu de la compréhension, car elle implique la linguistique, le raisonnement et la cognition. De ce point de vue, la compréhension paraît inaccessible dans l'état actuel de nos connaissances. (Poibeau, 2003) argumente que les méthodes numériques offrent de nombreux champs d'application, bien que ses détracteurs affirment qu'elles ne sont pas tout à fait explicatives ou interprétables. Au milieu de ce champ de bataille, il existe des propositions plus ouvertes, comme celles de (Sébillot, 2005; Torres-Moreno, 2007), où les auteurs montrent que parfois combiner le numérique et le symbolique est avantageux. Les approches statistiques facilitent l'automaticité et celles symboliques augmentent la qualité des résultats en donnant plus d'interprétabilité.

Une exemple concret est le système proposé par (Wong et Mooney, 2007) qui montre que l'analyse sémantique peut être traité comme un problème de traduction entre la langue naturelle et une langue formelle[7]. L'algorithme combine le *λ-calculus*, un outil mathématique pour l'étude de la récursion, avec des techniques de traduction statistique pour apprendre une grammaire synchrone entre les deux langues.

Malgré la grande quantité de travaux dans le domaine, l'automatisation de l'analyse textuelle reste un problème ouvert, et les nouvelles idées et méthodes originales sont toujours les bienvenues dans la communauté.

Nous proposons d'explorer les théories de la Physique statistique comme une source possible d'outils applicables au traitement de grandes masses de textes. Les nouvelles approches que nous présentons sont majoritairement numériques avec un minimum de ressources linguistiques.

1.3 Les problématiques abordées

L'application historique du TAL, la traduction automatique, a été initialement abordée en sous estimant le degré de compréhension et de complexité nécessaire pour transposer une idée d'une langue à une autre. C'est pour cette raison que, dans ses débuts,

7. Une langue formelle est constituée de formes logiques qui représentent la sémantique des énoncés.

elle a connu une longue période d'échecs. Dans les années 1990, il y a eu un retour au pragmatisme : la production de ressources et d'outils pour le traitement de gros volumes de textes a été favorisée (Sabah, 2006). Le champ de compétences du TAL s'est élargi à une grande variété d'applications qui vont de la recherche de réponses à des questions spécifiques en passant par la génération automatique de texte. Notre travail concerne plusieurs de ces problématiques, à savoir : la production des résumés automatiques, la recherche d'information, la segmentation thématique, la classification de documents et la compression de phrases. Ces tâches seront définies en détail dans les chapitres à venir.

1.3.1 L'approche proposée

Pour aborder les problématiques évoquées, nous proposons d'utiliser les modèles magnétiques issus de la Physique statistique. Un solide magnétique est un système composé de spins assimilables à de petits aimants qui peuvent se trouver en différents états ou orientations. L'interaction entre spins définit en grande partie le comportement global du système.

Nous montrons que ce modèle s'applique aussi au texte. En effet, un texte peut être représenté comme un système composite fait de phrases qui interagissent entre elles au travers des mots. Ces derniers peuvent être assimilés à des spins à deux états selon qu'ils sont absents ou présents dans une phrase.

Il nous semble important d'insister ici sur le terme système. Le modèle qui prévaut dans les approches numériques du TAL est la représentation du texte comme un sac de mots. Ce dernier a l'avantage d'être facilement et très efficacement implémentable dans un ordinateur. Tenter d'aller vers des représentations plus structurelles du texte implique nécessairement des structures de données et des algorithmes plus complexes.

Nous avons trouvé une manière de correctement implémenter la modélisation du texte comme un modèle magnétique. Il s'agit d'une première mise en œuvre que nous avons optimisée au fur et à mesure de notre travail. Cependant, notre implémentation permet déjà d'estimer des quantités physiques comme les couplages d'échange ou l'énergie. Il s'agit de quantités que l'on peut mesurer directement et que nous utilisons. Nous nous distinguons en cela des approches composites qui utilisent des mélanges de mesures reposant sur des multiples modèles (vectoriel, probabiliste, réseaux de neurones, etc.) ou qui nécessitent des processus complexes de calculs itératifs.

Nous cherchons alors à montrer comment un concept unique de Physique statistique, l'énergie, permet d'atteindre facilement les performances des principales approches numériques du TAL sur de multiples applications. Comme les approches numériques du TAL mettent souvent en œuvre un grand nombre d'heuristiques, il est difficile de se comparer à elles autrement que par le biais d'expérimentations selon des protocoles établis par cette communauté. Nous avons adapté l'implémentation de notre modèle magnétique à ces tâches de référence et protocoles d'évaluation.

1.3.2 Prototype en langage Perl

Dans ce travail de thèse, nous avons ainsi développé un système complet de TAL pour extraire l'information essentielle contenue dans les collections de texte. Fondé sur les modèles de spins, le système proposé est capable de :

- représenter les textes comme des ensembles d'unités en interaction magnétique ;
- mesurer l'intensité de telles interactions ;
- en déduire des quantités qui soient des indices de l'importance de l'information véhiculée.

Tous les programmes de ce travail ont été écrits en langage Perl, créé par le linguiste Larry Wall (Schwartz et al., 2005). Le langage et son interpréteur multi-plateforme sont libres sous licence GNU. Il est devenu entre autres, très populaire dans la communauté TAL en raison de sa facilité pour manipuler des textes avec des structures de données adaptées (listes et tables de hâchage) et une implémentation complète des « expressions régulières »[8]. Les algorithmes en TAL prennent comme entrée principale de grands volumes de textes provenant de sources hétérogènes (journalistiques, encyclopédiques, littéraires, etc.). Ces textes nécessitent d'être nettoyés pour devenir exploitables (filtrage, normalisation). Perl offre d'énormes avantages pour réaliser facilement ces tâches.

1.3.3 Corpus d'expérimentation et protocole d'évaluation

Les recherches en TAL demandent toujours une étape d'évaluation pour mesurer les performances des logiciels construits et valider les hypothèses. En général, les mesures d'évaluation peuvent être classées en deux catégories : les méthodes extrinsèques et les méthodes intrinsèques. Dans les premières, les sorties du système à évaluer sont jugées en se basant sur leur aptitude à accélérer la complétion d'autres tâches (par exemple : l'utilisation des résumés automatiques à la place des documents sources, dans des systèmes question-réponse). À l'opposé, les mesures intrinsèques réalisent un jugement direct des résultats selon au moins l'une deux méthodes suivantes :

- manuellement en évaluant la qualité du texte produit comme la lisibilité, la complexité de la langue ou la présence des concepts majeurs du document source ;
- automatiquement en calculant des mesures de similarité vis à vis de références produite par des humains.

Les évaluations intrinsèques automatiques sont devenues un standard de la communauté. Tout nouveau système proposé doit nécessairement obtenir des résultats convenables sur ce type d'évaluation avant d'être publié. Nous utilisons donc un large spectre de mesures intrinsèques pour asseoir l'intérêt de notre modèle. Ces mesures seront détaillées dans les chapitres à venir.

Pour cela nous avons besoin de réunir des corpus d'expérimentation ayant des références humaines valides. En effet, en TAL, la validation des processus et des hy-

8. Ce sont des patrons utilisés pour effectuer la recherche et la manipulation de chaînes de caractères. L'extraction de sous-chaînes et les opérations de remplacement sont effectuées de façon efficace. L'utilisation des expressions régulières est un moyen puissant pour la recherche dans les corpus de textes.

pothèses se fait sur des collections de textes (les corpus) bien définies, adaptées et accessibles à l'ensemble de la communauté puisque l'expérimentation doit être facilement reproductible. Selon (Mellet, 2002), construire des corpus consiste à constituer des échantillons représentatifs d'une réalité plus large, c'est-à-dire un « observatoire » où appréhender et donner à voir cette réalité trop vaste pour être embrassée dans sa totalité .

À l'heure actuelle, la constitution de corpus suffisamment bien construits, en différentes langues, différents domaines et souvent visant des applications TAL spécifiques est un champ de recherche en soit. Il existe désormais des corpus d'expérimentation incontournables sur certaines problématiques qui permettent la comparaison directe des résultats entre groupes de recherche, ce que nous avons fait. Le tableau 1.1 donne les caractéristiques des principaux corpus utilisés dans ce travail.

Nous avons cependant estimé que ces ressources publiques restent insuffisantes. Nous en avons alors construit d'autres adaptés à nos besoins. Nous allons expliquer leur élaborations au fur et à mesure que nous nous en servirons.

Corpus	**Langue**	**Date de création**	**Taille approximative en phrases**	**Application visée**
DUC [8]	anglais	2005-2007	1200 par édition	résumé automatique
HANSARD [9]	bilingue	1970-1990	2,8 millions	traduction automatique
DEFT [10]	français	2008	500 000	classification de documents
CHOI [11]	anglais	2000	20 000	segmentation thématique

TABLE *1.1* – *Principaux corpus publics d'évaluation utilisés.*

1.4 Organisation de la thèse

Notre étude est développée en six chapitres organisés comme suit. Dans le chapitre 2, nous présentons un parcours historique des applications de la Physique dans l'analyse textuelle. Nous mentionnons les travaux pionniers, les plus importants et les

8. *Document Understanding Conferences* (DUC), sponsorisées par l'*Advanced Research and Development Activity* (ARDA) et organisées par le *National Institute of Standards and Technology* (NIST) ; http://duc.nist.gov

9. Les *proceedings* du *Canadian Hansard* en anglais et français, http://www.parl.gc.ca/common/Cham\discretionary{-}{}{}ber_Senate_Debates.asp

10. Depuis 2006, le DEFT (DÉfi Fouille de Texte), http://deft.limsi.fr, propose des campagnes d'évaluation dans le domaine du TAL. L'édition 2008 concerne la classification en thème et en genre de textes.

11. Construit par F. Choi à partir du *Brown Corpus*, corpus standard en anglais sur l'actualité des États-Unis, http://www.cs.man.ac.uk/~mary/choif/software.html

plus récents. Ensuite, nous décrivons les approches qui nous ont permis de représenter le texte comme des systèmes de spins.

Le chapitre 3 est consacré à la définition de la mesure d'énergie textuelle inspirée par le modèle magnétique d'Ising et les réseaux neuronaux du type Hopfield. Une interprétation en utilisant des graphes nous a permis de définir l'énergie textuelle comme une mesure de similarité pour les tâches du TAL et à illustrer ses avantages par rapport aux autres mesures utilisées.

Au chapitre 4, nous présentons le système ENERTEX, basé sur le calcul de l'énergie textuelle. Les premières applications portent sur le résumé générique monodocument en différentes langues et domaines. Une modification non triviale nous a permis de l'appliquer au résumé multidocument guidé par une requête et de définir une stratégie anti-redondance. Nous terminons ce chapitre par une étude des impuretés introduites dans le réseau textuel. Ces impuretés correspondent à des annotations insérées pour étiqueter le type d'information véhiculée par les phrases. En maîtrisant leurs effets sur le calcul d'énergie textuelle, nous avons obtenu un système de recherche d'information guidée.

Dans le chapitre 5 nous présentons une approche originale pour la détection de frontières thématiques, basée sur la comparaison des spectres énergétiques des phrases. Nous proposons également une méthode pour la classification de documents au travers de matrices d'échange.

Au chapitre 6, nous abordons la problématique de compression de phrases utilisant le modèle de verres de spin. Les phrases sont compressées en appliquant des couplages obtenus au préalable sur un corpus d'apprentissage. Nous utilisons des simulations Monte-Carlo et la dynamique de Metropolis pour produire des variantes intéressantes de compression.

Finalement, nous tirons les conclusions de ce travail et dressons une liste de perspectives.

Chapitre 2

Le texte vu comme un système de spins

Sommaire

2.1 Introduction

Nous commençons ce chapitre par un historique des applications des concepts issus de la Physique à l'analyse de textes. Ensuite, nous décrivons la démarche pour transformer les documents en une représentation numérique adaptée aux algorithmes du TAL. Nous exposons les possibilités d'élection des unités textuelles selon le niveau d'analyse désiré. Nous expliquons aussi l'importance des pré-traitements pour faire face au

problème dimensionnel et l'utilité des espaces vectoriels pour définir des mesures de comparaison d'objets textuels. L'analogie entre la représentation vectorielle des textes et certains systèmes de la Physique du magnétisme, permet l'utilisation de ces derniers pour construire des algorithmes pour les problématiques du TAL.

Avant de présenter la description général des méthodes proposées dans cette thèse, nous détaillons le travail de (Takamura et al., 2005) qui partage avec nous le fait d'utiliser un modèle de spins pour une application du traitement des langues. Cela offre un repère pour expliquer nos choix concernant la représentation des textes, le calcul d'interactions entre unités textuels et leur utilisation selon les tâches TAL à résoudre.

2.2 La Physique dans l'analyse textuelle : l'état de l'art

La loi de Zipf (Zipf, 1935) et la Théorie de l'information de Shannon (Shannon, 1948) sont certainement les résultats numériques les plus célèbres et les plus exploités dans l'étude des langues naturelles. Toutes deux ont des liens avec la physique. Zipf a découvert le rapport entre la fréquence d'un mot et son rang. Il a suggéré une interprétation physique qui néanmoins n'a jamais été formellement établie. D'autre part, en cherchant une mesure de la quantité d'information générée par un système, Shannon a proposé une fonction dont la forme mathématique est identique à celle de l'entropie en mécanique statistique. Cette mesure est désormais connue comme l'entropie de Shannon. Sur cette approche, (Jaynes, 1957) a établi le critère d'entropie maximale. Cet outil permet de faire des prédictions à partir de l'information partielle disponible sur le comportement d'un système. Depuis sa conception, le critère d'entropie maximale de Jaynes a été largement appliqué aux tâches du TAL et on le retrouve même dans les travaux les plus récents. Dans ce chapitre, nous faisons un parcours des recherches dont les résultats remarquables ont attiré l'attention des gens intéressés par l'étude et la modélisation de la langue.

2.2.1 La loi de Zipf et le principe du moindre effort

« *L'émergence d'un langage complexe est un des événements fondamentaux de l'évolution humaine et plusieurs propriétés remarquables suggèrent la présence de principes d'organisation. De ces principes qui semblent être communs à toutes les langues, le plus connu est la loi de Zipf* » (Ferrer i Cancho et Solé, 2003). Cette règle empirique établit que la fréquence d'un mot décroît en fonction de son rang (equation 2.1). Le mot le plus fréquent a un rang de 1, le deuxième le plus fréquent a un rang de 2, et ainsi de suite. $f(n)$ est la fréquence du n-ème mot dans le rang et K une constante.

$$f(n) = \frac{K}{n} \tag{2.1}$$

Intuitivement cela signifie que, dans un document, il existe un petit nombre de mots qui sont très utilisés et beaucoup d'autres qui le sont moins. Zipf a justifié ce comportement avec l'image d'une balance de vocabulaire, résultat de deux forces qui s'opposent : une

force d'unification exercée par le locuteur qui tend à réduire le vocabulaire (le principe du moindre effort) et en contrepartie, une force de diversification exercée pour l'auditeur qui demande un vocabulaire plus large pour mieux comprendre le message (Zipf, 1949). Zipf n'a jamais formalisé cette hypothèse (Harremöes et Topsoe, 2005) et de nos jours personne ne connaît vraiment la signification de cette loi (Nowak et al., 2000). Cependant, les efforts pour trouver une explication continuent (Ferrer i Cancho et Solé, 2003; Dahui et al., 2005).

D'autres auteurs considèrent la loi de Zipf comme une hypothèse nulle sans signification particulière (Nowak et al., 2000). (Mandelbrot, 1953) a démontré qu'un texte généré aléatoirement obéit aussi à la loi de Zipf. Dans le prolongement des résultats de Shannon sur la quantité d'information produite par une source discrète, Mandelbrot a aussi découvert une loi qui généralise la loi de Zipf :

$$f(n) = \frac{K}{(a+bn)^c} \tag{2.2}$$

où K, a, b et c sont des constantes dérivées du texte en question. La distribution de Zipf-Mandelbrot apparaît dans la plupart des travaux où une hiérarchisation du vocabulaire s'avère nécessaire. Par exemple, dans la génération statistique de texte (Biemann, 2007).

2.2.2 L'entropie de Shannon et les langues naturelles

En 1948, Claude Shannon a proposé une théorie mathématique générale de la communication (Shannon, 1948). L'objet d'étude est un système d'information classique (figure 2.1). **Une source** produit **un message** qui est transformé en **signal** par **le transmetteur**. Le signal voyage au travers d'**un canal** jusqu'à atteindre **le récepteur** qui le reconvertit en message avant de le délivrer au **destinataire**.

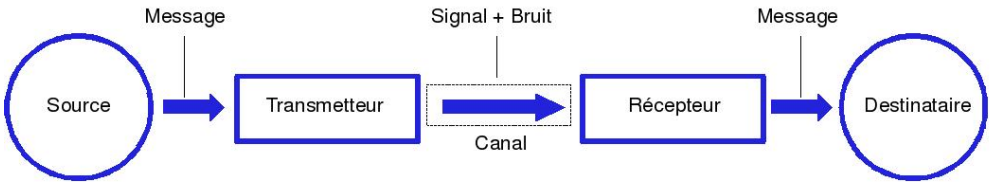

FIGURE 2.1 – *Système d'information classique étudié par Shannon.*

Shannon s'est intéressé à la diminution du bruit introduit dans le canal. L'hypothèse principale a été que la connaissance de la structure statistique de la source permet de minimiser ce bruit pour mieux profiter de la capacité du canal. Il a illustré ce principe par un exemple simple : en télégraphie il est très utile de connaître à l'avance la probabilité d'apparition des caractères dans une langue. Ainsi, il est possible de faire des économies sur le canal en assignant des symboles courts aux caractères fréquents, tels que « a » ou « e » et plus longs aux plus rares comme « q » ou « x ». En effet, Shannon a identifié les langues naturelles comme des sources discrètes dont la structure statistique peut être approchée. Pour quantifier l'information produite par ce type de source,

il a proposé la mesure d'**entropie H** dont la forme est identique à l'entropie de Gibbs d'un système thermodynamique en physique statistique :

$$H = -K \sum p_i \log p_i \tag{2.3}$$

où K est une constante correspondant au choix d'unités de mesure et les p_i sont l'ensemble des probabilités associées à la source d'information. Dans la formulation de Gibbs, $K = k_B$ est la constante de Boltzmann qui peut être vue comme un facteur de correction pour mesurer la température en unités arbitraires (Jaynes, 1957).

L'entropie est une notion introduite en théorie classique de la thermodynamique pour formaliser l'observation que les processus évoluent de manière naturelle dans une direction particulière. Un système isolé [1] qui n'est pas à l'équilibre évoluera vers des états d'entropie supérieure jusqu'à atteindre l'équilibre où l'entropie approche sa valeur maximale. L'entropie a été souvent associée à l'idée de désordre (ce qui semble être intuitivement le contraire de la stabilité). Cette image vient de l'idée qu'un processus thermodynamique entraîne un bouleversement dans l'arrangement des composantes d'un système (au niveau moléculaires), qui peut être quantifié par l'entropie (Landsberg, 1984).

Dans le cadre de la théorie de l'information, l'entropie de Shannon représente l'incertitude de l'information émanant de la source. Dans ce sens, l'information est le contraire de l'entropie. Une entropie élevée signifie un manque d'information, une indétermination associée à la nature probabiliste du processus de communication.

2.2.3 L'entropie maximale de Jaynes

(Jaynes, 1957) a constaté que la théorie de l'information de Shannon et la mécanique statistique sont des formes d'inférences statistiques basées sur l'information partielle disponible du système. L'entropie n'est qu'une manière de mesurer l'incertitude due à cette connaissance partielle.

La contribution principale de Jaynes a été le postulat du critère d'entropie maximale : « *le fait qu'une distribution de probabilité maximise l'entropie avec quelques contraintes qui représentent l'incomplétude de notre information, c'est la propriété qui justifie l'utilisation de cette distribution pour faire des inférences* » (Jaynes, 1957). Intuitivement, cela signifie que pour caractériser des évènements inconnus avec un modèle statistique, il faut toujours choisir celui d'entropie maximale.

Le critère de Jaynes ne doit pas être confondu avec la loi d'évolution des processus physiques vers des états d'entropie croissante (mentionnée en section 2.2.2). Le critère d'entropie maximale n'est qu'un outil pour choisir une distribution de probabilité entre un ensemble de distributions qui décrivent l'information disponible sur un processus (observations, hypothèses, contraintes, etc.). Selon Jaynes, dans le problème de prédiction, la maximisation de l'entropie n'est pas une application des lois de la physique mais seulement une méthode de raisonnement pour assurer que l'on n'ait pas fait

1. Un système isolé n'interagit pas avec son environnement.

des suppositions arbitraires. Cela veut dire qu'il faut modéliser tout ce qu'on connaît sur le processus sans rien supposer sur ce qu'on ne connaît pas.

Bien que Jaynes ait montré que l'expression mathématique de l'entropie a un sens indépendant de la thermodynamique, d'autres auteurs qui insistent sur une liaison profonde entre ces champs. (Bavaud et Xanthos, 2002) présentent un rappel historique des bases du formalisme thermodynamique dans un contexte de statistique textuelle selon un schéma à deux entrées : thermodynamique $\Longrightarrow$ théorie de l'information. Leur étude porte sur les applications textuelles utilisant les concepts de température, d'état cristallin et de mélange de langues. Ils concluent qu'il existe une équivalence claire entre ces deux formalismes qui assure que tout développement issu d'une approche thermodynamique trouvera son expression en théorie de l'information.

2.2.4 Applications au TAL

L'entropie de Shannon et le critère d'entropie maximale de Jaynes, sont des concepts qui ont dominé les applications du TAL depuis des années. Dans cette section, nous présentons quelques exemples de ces travaux. Nous décrivons aussi d'autres méthodes, inspirées aussi de la physique, mais qui s'éloignent de la voie entropique.

Shannon a combiné ses résultats avec la loi de Zipf pour déterminer l'entropie des mots de l'anglais (Shannon, 1951). La méthode proposée est connue sous le nom de *jeu de Shannon*. Elle consiste à demander à une personne de deviner la première lettre d'un texte, puis la seconde, puis la troisième, etc. Les bornes inférieures et supérieures de l'entropie du langage sont alors estimées à partir du nombre d'essais pour chaque lettre.

Des travaux postérieurs ont adapté le jeu de Shannon à la mesure des performances des modèles de langage. Un modèle de langage est l'ensemble de probabilités qui permettent la prédiction d'un événement linguistique (l'apparition d'une lettre, d'un mot ou d'une phrase). De telles probabilités sont déterminées en observant l'événement sur un corpus d'apprentissage de grande taille. La mesure introduite pour évaluer les qualités prédictives d'un modèle est la perplexité (Jelinek et al., 1977). Dans le cas des mots, la perplexité est l'inverse de la moyenne géométrique des probabilités de prédiction des mots du test. La perplexité indique le nombre moyen de mots candidats possibles après qu'un mot a été reconnu. Sa valeur maximale est V, la taille du vocabulaire et sa valeur minimale est 1, ce qui correspond à un modèle entièrement déterministe. Le modèle de langage est d'autant plus précis que la perplexité est faible. Le logarithme de la perplexité $PP(M)$ d'un modèle M est analogue à son entropie de Shannon $H(M)$ (Beaujard et Jardino, 1999) :

$$log PP(M) = H(M)\ log 2 \tag{2.4}$$

D'un autre coté, (Berger et al., 1996) décrivent une méthode pour la modélisation statistique basée sur le critère d'entropie maximale et ils l'appliquent avec succès aux problèmes du TAL tels que la désambiguïsation du sens et la segmentation de phrases.

Ils ont trouvé que le modèle qui maximise l'entropie appartient à une famille exponentielle avec un paramètre à ajuster à chaque contrainte existante et à estimer sur le corpus d'apprentissage.

(Stephens et Bialek, 2008) ont construit des modèles d'entropie maximale pour produire des séquences de caractères formant des mots. Leur objectif premier était de tester la capacité de cette approche pour caractériser l'ordre des caractères dans un mot. Le modèle prédit approximativement les probabilités d'apparition des mots de quatre caractères en anglais et capture bien la structure générale de la distribution selon Zipf.

Récemment, (Waszak et Torres-Moreno, 2008) ont utilisé un modèle de langage basé sur la probabilité des bigrammes des mots ainsi qu'un calcul d'entropie type Shannon pour retrouver la meilleure compression d'une phrase. Ils ont utilisé également un perceptron [2] pour définir si une phrase était suffisamment compressée.

(Taira et al., 2007) présentent un travail qui s'éloigne de la voie entropique. Ils ont fait appel à la théorie du champ moyen pour construire un parseur de textes médicaux. Leur système transforme les rapports médicaux rédigés en texte libre en une représentation structurée exploitable par un ordinateur. La sortie est un arbre de dépendances entre les mots d'une phrase. Cet arbre représente l'état d'énergie minimale du système.

(Takamura et al., 2005) se sont servis des systèmes de spins pour trouver les orientations sémantiques des mots : positive ou négative (désirable ou indésirable) à partir de mots amorces. La sortie est une liste de mots indiquant leurs orientations estimées selon l'approximation du champ moyen. Les auteurs signalent que leur approche est équivalente à celle de la maximisation d'entropie. Bien que l'application visé pour ce travail ne soit pas comprise dans l'ensemble de tâches que nous avons abordé, l'utilisation commun des systèmes de spins pour représenter et analyser les textes clame une description plus profonde laquelle sera présenté à la fin du chapitre, section 2.4.3, après d'avoir fait une révision des concepts basiques de la théorie des systèmes magnétiques.

2.3 Représentation numérique des textes

Pour appliquer des techniques numériques, comme celles décrites dans la section précédente, les textes doivent être transformés en une représentation permettant de faire des calculs. La représentation numérique que nous avons utilisée au cours de notre travail, est le modèle vectoriel de (Salton et al., 1975). Dans cette section nous expliquerons comment utiliser cet outil pour transformer les textes en vecteurs et comment faire face aux problèmes inhérents à une telle représentation.

2. Le perceptron est un modèle de neurone artificiel qui ajuste ses connexions en fonction d'un ensemble d'apprentissage (Hertz et al., 1991). En l'occurrence, un corpus fournit les données d'apprentissage pour le perceptron.

2.3.1 Le modèle vectoriel

Le premier pas vers l'application du modèle vectoriel sur un corpus est le choix des unités textuelles ou termes d'indexation (termes, n-grammes [3] de termes, expressions). Ces termes constitueront le vocabulaire. À chaque élément du vocabulaire est associé un index unique arbitraire. Ensuite, on accorde un vecteur v à chaque segment de texte (une phrase, un paragraphe, un document). La dimension de ce vecteur correspond à la taille du vocabulaire et chaque composante v_i associe un poids au terme d'indice i (par exemple la fréquence d'apparition du terme i dans le segment).

Du choix des termes et des segments dépend le niveau de l'analyse. Ainsi on peut souhaiter, par exemple, la comparaison entre paragraphes à partir des phrases qu'ils contiennent, ou des phrases à partir de leurs termes ou à comparer des termes à partir de leurs caractères. Pour nous, le choix a été généralement celui des phrases et leurs termes (néanmoins dans quelques expériences nous avons changé cette échelle d'analyse).

Pour illustrer la démarche, nous prenons le document de la figure 2.2. Nous choisissons comme vocabulaire les termes séparés par des espaces en blancs.

Dans la représentation vectorielle, chacune des cinq phrases sera un vecteur dont les composantes indiquent la présence (1) ou l'absence (0) d'un terme (voir tableau 2.1). L'arrangement consécutif des vecteurs forme la matrice terme-segment où l'on a perdu l'ordre des termes dans les phrases. C'est pourquoi cette représentation est aussi connue comme « sac de termes ». Perdre l'ordre des termes dans les textes peut devenir gênant

```
Les maisons bleues de ma tante. Ma tante
s'appelle Lulu. J'adore sa maison. Le
bleu c'est ma couleur préférée. J'ai des
chaussures bleues toutes neuves.
```

FIGURE 2.2 – *Petit texte en français.*

Phrase	Les	maisons	bleues	de	ma	tante.	Ma	tante	s'appelle	Lulu.	J'adore	sa	maison.	Le	bleu	c'est	couleur	préférée.	J'ai	des	chaussures	toutes	neuves.
1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

TABLE 2.1 – *Matrice de termes × phrases pour le texte de la figure 2.2. Chaque phrase devient un vecteur de présences/absences.*

quand on est intéressé par des tâches comme la traduction automatique. Cependant, pour faire des calculs de fréquence ou des distributions, il est très utile d'avoir une correspondance entre les termes et les composantes des vecteurs phrase.

3. Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée.

2.3.2 Les états et leur pondération

Dans la matrice terme-segment de l'exemple 2.1, les états des unités représentent leurs présences/absences (1/0). Nous pouvons aussi donner un poids aux états des unités selon leur importance dans les documents. L'option la plus simple est d'employer les fréquences d'occurrence *tf* indiquant le nombre d'occurrences des termes dans les segments textuels. Son utilisation produit une matrice d'occurrences dont les composantes sont les *tf* des termes (*tf*/0). Si'il s'agit de l'occurrence par phrase, ce choix peut n'est pas être très différent de celui de présences car les termes présentent très souvent des fréquences unitaires en particulière si l'on considère des phrases courtes.

D'autres formes de pondération sont aussi possibles. Un exemple est le TF.IDF (*Term Frequency-Inverse Document Frequency*), très utilisé dans l'analyse textuelle. Le TF représente la fréquence d'un terme dans un document pendant que l'IDF sert à pénaliser les termes apparaissant dans un grand nombre de documents. La partie IDF a été proposée par (Spärck Jones, 1972) et la fonction TF.IDF par (Salton et Yang, 1973).

Deux versions couramment utilisées sont présentées dans le tableau 2.2 où $tf_{i,j}$ est la fréquence d'ocurrence du terme i dans le document j ; N est le nombre de documents dans le corpus et n le nombre de documents dans lesquels le terme i est présent ; $max(tf_{ij})$ est la fréquence d'occurrence maximale d'un mot i dans le document j. Les facteurs de normalisation évitent de favoriser les documents longs.

	TF	**IDF**
version 1	$tf_{ij} / \sum_i tf_{ij}$	$log(N/n)$
version 2	$tf_{ij} / max(tf_{ij})$	$log((N-n)/n)$

TABLE 2.2 – *Deux versions de la mesure TF.IDF couramment utilisées.*

Finalement la pondération d'une terme est obtenue par le produit de ces deux facteurs :

$$(TF.IDF)_{i,j} = TF_{i,j} \cdot IDF_i \tag{2.5}$$

Cette mesure statistique évalue l'importance d'un mot par rapport à un document dans un corpus. L'importance augmente selon le nombre de fois que le mot est présent dans le document mais diminue selon sa fréquence totale dans le corpus. Elle est donc, un filtre des termes communs. Si l'on opte pour cette pondération, les composantes de la matrice terme-segment seront les TF.IDF des termes (TF.IDF/0).

Bien que les différentes implémentations de la matrice terme-segment que nous venons de présenter soient utilisées au long de notre travail, celle qui corresponde aux états *tf*/0 est à la base de la plupart de nos algorithmes. Par ailleurs, une autre option de pondération est proposée dans l'annexe C pour résoudre une problématique particulière. Le choix a dépendu de l'application et des résultats empiriques. Par souci de simplicité, nous utiliserons le terme « fréquence » à la place de « fréquence d'occurrence » .

2.3.3 Réduction dimensionnelle : pré-traitement des textes

Un des aspects de la représentation vectorielle qui pose souvent des problèmes est la dimension de l'espace de termes. Un corpus peut donner lieu à une matrice avec des centaines de millions de colonnes (le vocabulaire du corpus) et un million de lignes (les documents) (Ibekwe-SanJuan, 2007). Il est clair que chaque document contient un petit sous-ensemble du vocabulaire total du corpus, et les matrices terme-segment sont, en général, creuses. Les réduire aux seules valeurs présentes est une astuce technique qui ne répond pas vraiment au problème de la dimensionalité. Une diminution plus drastique du vocabulaire s'avère nécessaire. C'est pour cela qu'il est important d'appliquer aux textes une phase de prétraitement qui comprend généralement les actions suivantes.

1. **Uniformiser la casse :** transformation des majuscules en minuscules ;
2. **Filtrage :** élimination de termes fonctionnels dits « vides de sens » (prépositions, conjonctions, adverbes, chiffres, ponctuations, etc.) ;
3. **Normalisation :** ramener chaque terme à une forme invariable qui peut être sa racine (racinisation ou *stemming*) ou sa forme canonique (masculin singulier ou infinitif pour les verbes). Cette dernière opération est connue sous le nom de lemmatisation [4]. Par exemple, les termes *jette, jetions, jeté, jetât, jetassent, jetâmes*, seront tous réduits à la forme *jeter*.

Ces opérations permettent de réduire considérablement la dimension de l'espace tout en augmentant la fréquence des termes canoniques. La figure 2.3 montre les cinq phrases de la figure 2.2 après le pré-traitement. La matrice correspondante est montrée au tableau 2.3. La diminution dimensionnelle est évidente. La réduction de la taille du lexique

```
Phrase 1 : maison bleu tante
Phrase 2 : tante appeler lulu
Phrase 3 : adorer maison
Phrase 4 : bleu couleur préférer
Phrase 5 : chaussure bleu neuf
```

FIGURE 2.3 – *Texte réduit après l'uniformisation, le filtrage et la normalisation du vocabulaire.*

Phrase	maison	bleu	tante	appeller	lulu	adorer	couleur	préférer	chaussure	neuf
1	1	1	1	0	0	0	0	0	0	0
2	0	0	1	1	1	0	0	0	0	0
3	1	0	0	0	0	1	0	0	0	0
4	0	1	0	0	0	0	1	1	0	0
5	0	1	0	0	0	0	0	0	1	1

TABLE 2.3 – *Matrice réduite de termes × phrases après les opérations de prétraitement.*

filtré/lemmatisé suit un comportement linéaire par rapport au nombre de termes du texte original. Ainsi, pour réaliser une analyse textuelle, afin d'obtenir un résumé par

4. Définition prise de (Ibekwe-SanJuan, 2007).

exemple, on pourrait utiliser seulement un seizième du volume des termes du document (Torres-Moreno et al., 2002) .

Selon la langue et la nature des corpus analysés, dans ce travail nous avons utilisé différents outils de prétraitement :

- les modules du système Cortex du LIA(Torres-Moreno et al., 2002) qui réalisent la segmentation, le filtrage et la normalisation de textes en anglais, espagnol, français et somali ;
- le *stemmer* de Porter (Porter, 1980) pour la racinisation en anglais ;
- les outils de segmentation des campagnes DUC particulièrement performants dans le traitement des abréviations en anglais [5].

Dans d'autres cas, nous avons programmé des routines de prétraitement adaptées à nos besoins spécifiques.

2.3.4 La similarité vectorielle

À partir d'une représentation matricielle, une manière de calculer la proximité entre les unités textuelles est d'utiliser les mesures communes de similarité vectorielle comme le cosinus (Salton et al., 1975) :

$$Cos(D_\mu, D_\nu) = \frac{|D_\mu \cdot D_\nu|}{||D_\mu||\ ||D_\nu||} \tag{2.6}$$

où D_μ et D_ν sont les vecteurs obtenus pour deux documents μ et ν. Plus l'angle qui les sépare est petit, plus l'information qu'ils portent est proche. Un exemple est présenté dans la figure 2.4. Il correspond à la représentation vectorielle des trois documents :

> D_0 : L'intervention télévisée du président de France a réussi à rassurer aux ceux qui ont manifesté.
> D_1 : La France a déjà manifesté contre la reprise des massacres de baleines en Islande.
> D_2 : Chercheurs, enseignants, personnels des universités et étudiants ont protesté contre les mesures défendues par le président.

Le tableau 2.4 présente les résultats de la mesure cosinus prenant le document D_0 comme référence. On observe que, après filtrage et normalisation, le document plus proche de D_0 est D_1 et le plus éloigné est D_2.

L'utilisation des mesures de similarité vectorielle dans l'analyse textuelle est une pratique très courante. Par exemple, le système de résumé proposé par (Boudin et Torres-Moreno, 2008) se base sur des variantes de la mesure cosinus. Les auteurs ont combiné le cosinus avec la mesure de Jaro-Winkler (Winkler, 1999) (qui réalise une analyse au niveau des caractères dans les termes) pour proposer un système de traitement de textes de chimie où les formules des molécules apportent une information essentielle

5. http://www-nlpir.nist.gov/projects/duc/data.html

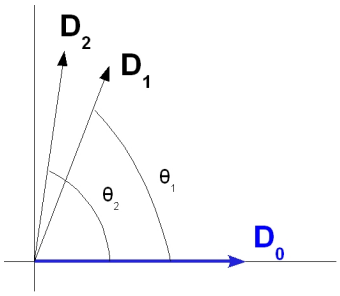

FIGURE 2.4 – *Représentation vectorielle des documents du tableau 2.4 en prenant le document D_0 comme référence.*

D_i	**Texte filtré et normalisé**	**Cos(D_i , D_0)**	θ_i
D_0	`intervention téléviser président france réussir rassurer manifester`	1,0000	0°
D_1	`france manifester reprise massacre baleine islande`	0,3086	$\approx 72°$
D_2	`chercheur enseignant personnel université étudiant protester mesure défendre président`	0,1140	$\approx 84°$

TABLE 2.4 – *Exemple d'application de la mesure cosinus. Le document plus proche à la référence D_0 est D_1 et le plus éloigné est D_2.*

sous une forme très complexe. Les outils vectoriels offrent donc de multiples manières de calculer le degré d'interaction entre segments textuels. Cependant le choix de telle ou telle métrique est souvent fait de manière ad-hoc par essais successifs. La Physique statistique du magnétisme offre un cadre méthodologique permettant de choisir les mieux adaptées à la prise en compte d'interactions plus complexes.

2.4 Représentation magnétique de textes

La mécanique statistique étudie les systèmes composés d'une grande quantité d'unités en interaction. Si l'on considère le texte comme un système de ce type, quel modèle physique est le mieux adapté à son analyse textuelle ?

Une idée directrice qui guide notre recherche est de trouver un modèle théorique le plus généraliste possible qui intègre en son cœur la notion d'interaction indirecte. En effet les mesures vectorielles les plus largement utilisées en TAL reposent sur des interactions directes entre mots, c'est à dire des co-occurrences. Les méthodes qui permettent de tenir compte d'interactions plus complexes reposent sur le coûteux calcul des principaux axes factoriels de la matrice terme-segment, ou encore sur des calculs de vraisemblance fondés sur des modèles probabilistes de mélanges de gaussiennes dont l'estimation des paramètres requiert un grand nombre d'itérations, ou enfin des modèles de langage surtout adaptés au traitement de très larges corpus de documents.

La Physiques statistique permet de proposer des modèles alternatifs plus simples qui révèlent les mécanismes structurants fondamentaux d'un texte.

2.4.1 Le texte codé comme un système de spins

Les modèles théoriques du magnétisme étudient des systèmes constitués d'un ensemble de N spins assimilables à de petits aimants. Les spins peuvent s'orienter selon plusieurs directions, mais Ising [6] a fait une simplification en considérant seulement deux directions possibles : vers le haut (↑, +1 ou 1) ou vers le bas (↓, -1 ou 0).

On peut constater qu'une matrice terme-segment S de présences/absences (figure 2.5 à gauche) correspond de manière naturelle à un système de spins binaires (à droite). La correspondance est claire : un terme est un spin, une phrase est une chaîne de spins et un document, un ensemble de ces chaînes. Ainsi, les objets physiques avec lesquels nous traitons dans cette thèse seront modélisés comme des chaînes de spins, même si dans la plupart des cas nous avons pondéré les états des spins selon la fréquence tf des termes .

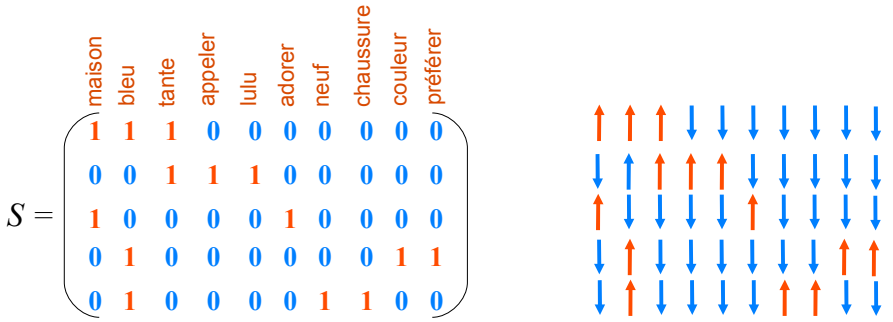

FIGURE 2.5 – *À gauche, une exemple de matrice terme-segment. Chaque ligne est une phrase et chaque colonne un terme du vocabulaire. À droite le système de spins binaires correspondant. Les spins vers le haut (↑) représentent la présence des termes dans les phrases (1) et les spins vers le bas (↓) leur absence (0).*

2.4.2 L'interaction d'échange

Dans un aimant, les spins interagissent entre eux au travers de l'interaction d'échange. Introduite par Heisenberg en 1929 dans le cadre de la mécanique quantique, l'interaction d'échange est très intense entre spins voisins et s'attenue très vite avec la distance. Ce comportement autorise l'indépendance mutuelle entre spins éloignés (Trémolet et al., 2000). L'énergie associée à cette interaction peut s'exprimer de la façon suivante :

$$E = -\sum_{<i,j>} J_{i,j} s_i s_j \tag{2.7}$$

6. Le physicien allemand Ernst Ising (Ising, 1925) a proposé dans sa thèse de doctorat la solution au modèle linéaire de moments magnétiques, connu comme le modèle d'Ising de ferromagnétisme.

Selon les signes des coefficients $J_{i,j}$, les spins s_i et s_j tendent à s'aligner parallèlement (ferromagnétisme) ou antiparallèlement (antiferromagnétisme) (figure 2.6), $< i, j >$ dénote qu'ils sont voisins. De ces interactions découle un comportement collectif qui se manifeste par l'arrangement microscopique des spins.

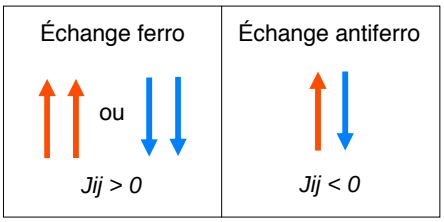

FIGURE 2.6 – *Interaction d'échange entre spins. Si le coefficient $J_{i,j}$ est positif l'interaction est ferromagnétique (orientation parallèle) et s'il est négatif l'interaction est antiferromagnétique (orientation antiparallèle).*

Pour appliquer le modèle de spins aux documents, afin d'étudier le comportement collectif des termes qui les composent, il faut au préalable trouver des réponses aux questions suivantes.

1. Comment obtenir les valeurs des coefficients d'échange entre les termes ?
2. Puisque la représentation vectorielle induit la perte d'ordre original des termes dans le texte, comment peut-on l'utiliser pour calculer l'interaction entre voisins proches ?
3. Dans le cas des textes, est-ce une bonne idée de limiter les interactions uniquement aux voisins proches ?

Dans les sections et chapitres à venir nous répondrons à ces questions en tirant les conclusions de notre étude sur la portée et les valeurs des interactions d'échange. Ultérieurement, nous décrirons nos méthodes pour exploiter le comportement collectif des textes pour résoudre les tâches du TAL.

2.4.3 Le système de spins de Takamura

En (Takamura et al., 2005) nous retrouvons une approche pour extraire les orientations sémantiques des mots en anglais basé sur le modèle de spins binaires d'Ising. Étant donné la base commun que ce travail détient avec nos méthodes, nous en faisons ici une description détaillé. Une comparaison parallèle sera présenté dans la section suivante.

La représentation des mots et l'interaction d'échange

Dans le modèle de Takamura, les états $\uparrow/\downarrow$ des spins représentent les orientations positive/négative, désirable/indésirable des mots. Les auteurs proposent trois méthodes pour calculer l'interaction d'échange $J_{i,j}$ entre couples de mots i et j utilisant différen-

tes bases de termes. Elles génèrent trois ensembles de couplages qui peuvent être utilisés dans les calculs de manière isolée ou combinée :

1. **Dictionnaire.** On établit une interaction entre deux mots si l'une est contenue dans la définition de dictionnaire de l'autre. La valeur de cette interaction est décidée comme suit : si un mot précède une particule de négation (*not* en anglais) dans la définition d'une autre mot, alors $J_{i,j} = -1$; $J_{i,j} = +1$ autrement. Si le mot i ne fait jamais partie de la définition d'un autre mot j, alors $J_{i,j} = 0$. Ces valeurs d'échange sont normalisées selon $\sqrt{d(i)d(j)}$ où $d(i)$ et le nombre de mots interagissant avec le mot i.
2. **Thésaurus.**[7] Un autre ensemble d'interactions d'échange est construit entre synonymes et hyperonymes pour lesquels $J_{i,j} = +1$. Entre antonymes $J_{i,j} = -1$.
3. **Corpus.** L'information de co-ocurrence des mots dans le corpus, combinée avec l'identification des particules conjonctives, est aussi utilisée. Par exemple, entre deux adjectifs connectés par *and*, $J_{i,j} = +1$ et s'ils sont connectés par *but* alors $J_{i,j} = -1$.

Cette étape utilise une quantité considérable de ressources linguistiques. L'apprentissage est faite sur 88 000 mots dont les définition, synonymes, antonymes et hyperonymes sont prises du WORDNET[8]. Les auteurs ont collecté aussi 804 expressions conjonctives du WALL STREET JOURNAL[9] et le BROWN CORPUS[10]. L'étiqueteur grammatical TREETAGER (Schmid, 1994) est utilisé pour identifier les adjectifs. Des mots et des phrases à sens négatif sont identifiés manuellement.

Estimation de l'orientation des mots

On commence avec un petit ensemble L de 14 mots amorce dont l'orientation a été assignée manuellement : {*good, nice, excellent, positive, fortunate, correct, superior*} sont tous (↑) ;{*bad, nasty, poor, negative, unfortunate, wrong, inferior*} sont (↓). Pour estimer l'orientation s_i d'un nouveau mot i, il faut calculer son orientation moyenne. Le processus à suivre est :

1. Initier les orientations moyennes $\bar{s_i}$ de tous les mots. Pour les mots amorces, $\bar{s_i} = s_i$ (l'orientation assignée manuellement) et pour les autres mots $\bar{s_i} = 0$.
2. Calculer la nouvelle orientation moyenne de chaque mot au travers de l'approximation de champ moyenne. Le modèle est réduit à la seule règle d'actualisation :

$$\bar{s_i}^{new} = \frac{\sum_{s_i} s_i exp\left[\beta s_i \sum_j J_{i,j}\bar{s_j}^{old}\right] - \alpha(s_i - a_i)^2}{\sum_{s_i} exp\left[\beta s_i \sum_j J_{i,j}\bar{s_j}^{old}\right] - \alpha(s_i - a_i)^2} \tag{2.8}$$

7. Un thésaurus est une sorte de dictionnaire sur la base de termes génériques. Il ne fournit qu'accessoirement des définitions. Un dictionnaire de synonymes est un exemple de thésaurus.
8. WORDNET est une base de données lexicales en langue anglaise développée par des linguistes du laboratoire des sciences cognitives de l'Université de Princeton (http://wordnet.princeton.edu).
9. http://www.wsj.com
10. http://khnt.aksis.uib.no/icame/manuals/brown

où β est une constante qui donne la notion de température inverse ; $\bar{s_j}^{old}$ et $\bar{s_j}^{new}$ sont les orientations moyennes du mot j avant et après de l'actualisation ; a_i est l'orientation du mot amorce i et α est une constante positive dont le rôle n'est pas spécifié. La valeur de β est ajusté en utilisant un critère de magnétisation.
Cette règle est récursive. Le critère de convergence est basé sur la différence d'énergie libre F avant et après l'actualisation. En Physique statistique, l'énergie libre est définie comme la différence entre l'énergie moyenne et l'entropie de la distribution de Gibbs-Boltzmann qui donne la probabilité P_μ de trouver le système dans un état μ :

$$P_\mu = \frac{exp(-\beta E_\mu)}{Z} \tag{2.9}$$

$$Z = \sum_\nu exp(-\beta E_\nu) \tag{2.10}$$

où E est l'énergie du système (equation 2.7) et Z un facteur de normalisation connu sous le nombre de fonction de partition. Le calcul de Z devient difficile car la somme parcourt sur les 2^N possibles configurations ν d'un système de N spins. L'approximation de champ moyenne propose d'utiliser une fonction Q qui approxime la distribution P. Pour établir l'équation 2, les auteurs se sont appuyés sur le travail de (Inoue et Carlucci, 2001) qui utilise un système de spins et l'approximation de champ moyen pour restaurer des images. Cependant, dans le cas de (Takamura et al., 2005), le calcul de F ainsi comme le seuil utilisé pour déterminer la convergence de l'algorithme ne sont pas clairement indiqués.

3. L'orientation finale des mots sera positive si $\bar{s_i}^{new}$ est haute et négative si $\bar{s_i}^{new}$ est basse. Nous pouvons observer que ces critères sont aussi ambigus.

Dans les expériences d'extraction d'orientation sur un corpus d'environ 4 000 mots, l'algorithme a montré des bonnes performances avec une précision entre 60% et 90% selon le nombre de mots amorces et l'ensemble des valeurs d'interaction utilisés.

2.4.4 Les approches que nous proposons

Étant donné que notre travail est aussi inspiré dans les modèles de spins, il nous semble convenable de mener une comparaison entre nos approches et le modèle de spins de Takamura. Cela va nous permettre d'exposer le cadre théorique dans lequel nous avons développé nos algorithmes.

Nous proposons dans cette thèse l'utilisation de deux modèles de spins pour résoudre les tâches du TAL. Le premier, basé sur le calcul d'énergie du système et la théorie des réseaux de neurones, est décrit au chapitre 3. Il a montré être utile dans les applications où une mesure de pertinence d'information est nécessaire. Ces applications sont l'objet des chapitres 4 et 5. Cette approche est désormais appelé ENERTEX.

Le deuxième modèle est développé au chapitre 6 et vise la compression automatique de phrases. Il repose sur l'idée que l'état fondamental (ou d'énergie minimale) d'une chaîne de spins (une phrase), sous contraintes d'interaction entre termes, est une bonne option de compression de la phrase originale. Cette approche a été nommée VERRE TEXTUEL par analogie avec les verres de spins qui seront décrits en détail dans le chapitre correspondant.

Voici donc notre comparaison selon :

La représentation. Bien que les trois approches utilisaient comme base le modèle de spins binaires d'Ising, les états représentent des notions différentes. Pour Takamura, il s'agit de l'orientation sémantique des mots positive/négative. Dans le cas d'ENERTEX les deux états correspondent à la fréquence/absence des termes dans le texte entant que le VERRE TEXTUEL fait référence à son présence/absence.

Les interactions d'échange. Dans les trois modèles les interactions entre couples de termes jouent une rôle essentiel pour accomplir les objectifs visés. Or, les algorithmes de calcul ont des différences importantes. Takamura propose trois alternatives de calcul (qui peuvent être combinées ou pas) pour résoudre une même problématique (l'estimation de l'orientation sémantique des mots) et qu'utilisent une quantité considérable de ressources linguistiques.

Dans le cas d'ENERTEX, le calcul des coefficients d'échange sont tout simplement basé sur l'information de co-ocurrence des termes dans les corpus de texte. Tous les couples de termes i et j sont connectés et les $J_{i,j}$ sont toujours positifs. Ils nous ont servi à résoudre une gamme d'applications telles que le résumé automatique et la segmentation thématique. Dans la plupart de cas, il n'existe pas une étape d'apprentissage préalable car les $J_{i,j}$ sont intégrés au calcul unique de l'énergie.

En revanche, pour le VERRE TEXTUEL les $J_{i,j}$ sont apprises sur un corpus de phrases complètes/compressées à deux niveaux : lexical et grammatical. Ce dernier niveau utilise l'étiqueteur grammatical TREETAGER. L'algorithme produit des valeurs de $J_{i,j}$ positives et négatives qui caractérisent un système verre de spins. Cette dernière caractéristique est présente aussi dans le modèle de Takamura.

Évolution du système. Takamura utilise une règle récursive provenant du méthode d'approximation de champ moyen pour calculer des nouvelles orientations moyennes des mots. Ces orientations peuvent varier de $\uparrow$ à $\downarrow$ et vice verse pendant le processus jusqu'à la convergence de l'algorithme.

En revanche, ENERTEX n'utilise aucun algorithme d'évolution car nous ne sommes pas intéressés à modifier les phrases en faisant disparaître ou même apparaître des mots (retournement des spins).

Par contre, le VERRE TEXTUEL, cherchant à compresser les phrases, fait en sorte de disparaître les termes non essentielles au contenu (spins changeant de $\uparrow$ à $\downarrow$). Dans ce cas nous utilisons des simulations Monte-Carlo avec la dynamique de Métropolis qui introduisent des fluctuations thermiques pour retourner les spins et permettent de

trouver des nouveaux états des chaînes de spins, c'est-à-dire des candidates à phrases compressées.

Les détails de ces implémentations seront présentés dans les chapitres à venir.

2.5 Conclusion

Il n'est pas évident qu'un physicien s'intéresse à un domaine aussi éloigné de ses préoccupations habituelles telles que l'astrophysique ou de la physique de la matière condensée. Cependant, les approches numériques ont mis en évidence que la langue véhicule une information qui peut être codée comme un ensemble d'unités en interaction dont les propriétés peuvent être calculées. Nous avons commencé ce chapitre avec un parcours historique sur l'utilisation de concepts issus de la physique dans le TAL numérique et nous avons remarqué la prévalence des critères entropiques. Nous pensons que cette tendance a peut-être occulté d'autres voies telles que celle de la physique statistique, qui peuvent être particulièrement pertinentes pour l'analyse de textes.

Plus précisément, la Physique statistique permet de dégager des comportements moyens de systèmes d'interaction, sans pour autant estomper complètement les phénomènes de nature discrète apparaissant à l'échelle des spins. Au contraire, l'approche entropique fond complètement le système physique dans un modèle régis par des lois de probabilité continues. Or, les corpus de textes considérés dans les tâches les plus caractéristiques du TAL sont certes de taille importante, ce qui justifie l'assistance de traitement automatiques, mais pas au point de gommer tout phénomène discret englobant des termes de très faible fréquence, mais dont l'interaction peut produire une information essentielle à la compréhension du texte. Par exemple, dans un corpus de textes journalistiques, deux termes synonymes n'apparaîtront pas ensemble dans une même phrase et auront une fréquence d'apparition inférieure à celle des autres termes. Par contre, ils apparaîtront dans des contextes proches. C'est ce type d'interaction qu'il nous faut capter.

Nous avons décrit comment le modèle vectoriel peut faciliter le codage des documents comme des systèmes de spins. Une telle représentation nous permettra d'élargir l'étude des liens entre segments de texte au-delà de l'algèbre vectorielle. En utilisant des modèles théoriques qui ont été proposés pour décrire le comportement des systèmes magnétiques, nous réaliserons le calcul des interactions entre termes et segments de texte afin d'exploiter les propriétés d'un système textuel.

Au travers de la comparaison avec une approche TAL existante, basé aussi sur un modèle de spins binaires, nous avons justifié le cheminement qui nous a amené à utiliser l'énergie et les verres de spins comme points de départ de nos recherches. Aux chapitres suivants nous ferons une description plus détaillée de l'utilisation de ces approches dans la construction des algorithmes d'analyse textuelle.

Chapitre 3

L'énergie textuelle

Sommaire

3.1 Introduction

Le modèle magnétique d'Ising a été utilisé dans une grande variété de systèmes qui peuvent être décrits par des variables binaires (Ma, 1985). Une approche très connue de la théorie de réseaux de neurones, qui utilise le modèle d'Ising pour garder et récupérer des patrons binaires, est la mémoire associative de Hopfield. Nous avons décidé d'adapter cette variante du modèle d'Ising à l'analyse de textes pour deux raisons. Premièrement, elle propose une interaction totale entre unités et dans ce cas là, il n'est pas nécessaire de repérer les voisins proches. Alors, la représentation de sac de termes, qui induit une perte d'ordre, ne pose pas de problème. Deuxièmement, elle offre une méthode pour calculer les couplages d'échange (déjà définis au section 2.4.2) entre les termes. Cela rendra possible une étude du comportement des documents à partir de l'interaction de leurs composantes. Cette démarche, qui sera détaillée dans le présent chapitre, nous a amené à la construction d'un système d'analyse textuelle basée sur

le calcul de son énergie. Nous ferons appel à la théorie de graphes pour expliquer la nature des liens capturés par la mesure de cette énergie textuelle.

3.2 Le modèle d'Ising et le réseau de Hopfield

Un réseau de neurones est un modèle de calcul composé d'éléments en parallèle et densément reliés, inspiré du fonctionnement de neurones biologiques. Il ressemble au cerveau sur deux aspects : i) la connaissance est obtenue de l'environnement par un processus d'apprentissage ; ii) les connexions entre neurones, connues sous le nom de poids synaptiques, sont utilisées pour stocker les connaissances acquises (Hertz et al., 1991).

Hopfield (Hopfield, 1982; Hertz et al., 1991) s'est inspiré des systèmes physiques comme le modèle magnétique d'Ising pour construire un réseau de neurones capable de fonctionner comme une mémoire associative. Une mémoire associative est un réseau de neurones qui stocke des liens entre couples d'information. Par exemple, on pourrait associer les noms de personnes aux images de leurs visages, ou les noms de scientifiques célèbres à l'information sur leurs contributions, etc. Idéalement, ce type de mémoire a deux propriétés essentielles. D'abord, elle est adressable par son contenu, c'est-à-dire, qu'on n'a pas besoin d'index pour ranger et récupérer l'information. Deuxièmement, elle est peu sensible aux petites erreurs. Le système produira une information exacte même si l'entrée peut être incomplète ou partiellement erronée. Par exemple, le patron *évolution* en entrée devrait être suffisant pour récupérer toute l'information en rapport à *Darwin*, et $E = mc^3$ doit nous répondre avec *Einstein* malgré la « petite erreur » sur la formule (Hertz et al., 1991).

Cependant, le modèle de Hopfield s'est révélé ne pas être aussi robuste qu'espéré. Après de nombreuses études qui ont exploré diverses modifications de son algorithme, sa faible performance à causé l'abandon des recherches à son sujet (Dreyfus et al., 1992). Nous faisons ici une description de son fonctionnement et de ses limitations.

3.2.1 Une approche énergétique

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 de spins d'Ising. Comme nous l'avons signalé dans la section 2.4.1, le système d'Ising est constitué d'un ensemble de N spins qui peuvent s'orienter suivant deux directions : vers le haut (↑, +1 ou 1) ou vers le bas (↓, -1 ou 0). Dans le modèle de Hopfield il existe une connectivité totale (figure 3.1), et les spins interagissent tous entre eux selon la règle d'apprentissage d'Hebb (équation 3.1). Cette règle suggère que les connexions changent proportionnellement à la corrélation entre les états des spins (Hertz et al., 1991), ce qui donne une manière directe de calculer l'interaction d'échange $J_{i,j}$ entre deux unités i et

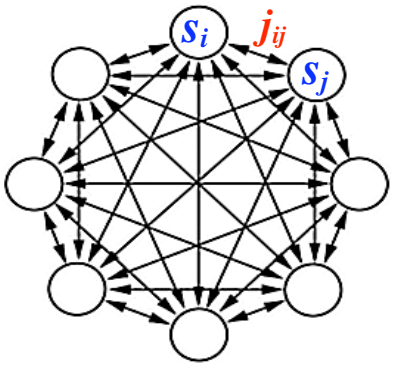

FIGURE 3.1 – *Réseau de Hopfield. Il existe une connectivité totale entre les unités et les connexions sont pondérées selon la règle d'Hebb.*

j :

$$J_{i,j} = \sum_{\mu=1}^{P} s_i^{\mu} s_j^{\mu}; \quad i \neq j \tag{3.1}$$

où s_i et s_j sont les états des neurones i et j. Les autocorrélations ne sont pas calculées ($i \neq j$) car on suppose qu'une unité n'a pas d'influence sur elle même. La sommation porte sur les P patrons à stocker. Cette règle d'interaction est locale, car $J_{i,j}$ dépend seulement des états des unités connectées. Ce modèle est une mémoire associative capable de stocker et de récupérer un certain nombre de configurations du système car la règle de Hebb transforme ces configurations en attracteurs (minima locaux) de la fonction d'énergie (Hopfield, 1982) :

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_i \, J_{i,j} \, s_j \tag{3.2}$$

L'énergie est une fonction de la configuration du système, c'est-à-dire, de l'état (activation ou de non activation) des unités. Si on présente un patron ν, chaque spin subira un champ local h_i induit par les autres N spins :

$$h_i = \sum_{j=1}^{N} J_{i,j} \, s_j \tag{3.3}$$

Les spins s'aligneront selon h_i de la façon suivante (nous avons choisi la notation $s_i \in \{1, 0\}$) :

$$s_i = \Theta(h_i) \text{ où } \quad \Theta(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{autrement} \end{cases} \tag{3.4}$$

pour restituer le patron stocké qui est le plus proche du patron présenté ν. Hopfield a démontré que l'énergie de ce système (équation 3.2), diminue toujours pendant le processus de récupération. Nous n'allons pas détailler la méthode de récupération de patrons [1], car notre intérêt va porter plutôt sur la distribution et les propriétés de l'énergie

1. Cependant le lecteur intéressé peut consulter, par exemple (Hopfield, 1982; Kosko, 1988; Hertz et al., 1991).

du système. Cette fonction monotone et décroissante a été utilisée uniquement pour montrer que l'apprentissage est borné.

3.2.2 Adaptation au Traitement Automatique de la Langue

Les capacités et limitations de la mémoire associative de Hopfield ont été bien établies de façon théorique à maintes reprises (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 faible 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 TAL, nous pensons que l'on peut exploiter autrement ce comportement.

En utilisant la vectorisation des textes (Salton et McGill, 1983), les documents ainsi transformés en vecteurs sont susceptibles d'être traités comme un réseau d'unités binaires. 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 spins $\uparrow$, $i = 1, \cdots, N$ (le terme i étant présent) ou $\downarrow$ (le terme 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 termes qu'ils partagent. Si les thématiques sont proches, il est raisonnable 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 énergétiques entre les termes et les phrases.

3.3 Le calcul de l'énergie des textes

Les documents sont donc prétraités avec les algorithmes classiques mentionnés dans la section (2.3.3) afin de réduire la dimensionalité. Puis le modèle vectoriel transforme le document en une matrice S qui contient l'information du texte sous forme de sacs de termes. On considère S comme l'ensemble des configurations d'un système dont on peut calculer l'énergie.

3.3.1 La version matricielle de l'énergie

La représentation vectorielle d'un document produit une matrice $S_{[P\times N]}$ de fréquences/absences :

$$S = \begin{pmatrix} s_1^1 & s_2^1 & \cdots & s_N^1 \\ s_1^2 & s_2^2 & \cdots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^P & s_2^P & \cdots & s_N^P \end{pmatrix} \; ; \; [s_i^\mu] = \begin{cases} \mathrm{tf}_i^\mu & \text{si le terme } i \text{ existe dans le segment } \mu \\ 0 & \text{autrement} \end{cases} \tag{3.5}$$

où $\mu = 1, \cdots, P$ phrases et $i = 1, \cdots, N$ termes ; tf_i^{μ} est la fréquence du terme i dans le segment textuel μ.

La présence du terme 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. En utilisant la fréquence, nous introduisons des amplitudes pour les couplages d'échange. Ce choix permet de mieux estimer l'intensité de l'interaction entre segments de texte.

Nous allons nous démarquer de (Hopfield, 1982) sur deux points : S est une matrice entière (ses éléments prennent des valeurs indiquant le nombre d'occurrences) et nous utilisons les éléments $J_{i,i}$ car cette auto-corrélation permet d'établir l'interaction du terme i parmi les P phrases, ce qui est important en TAL. Pour calculer les interactions d'échange entre les N termes du vocabulaire, on applique la règle de Hebb (3.1), que sous forme matricielle se traduit par :

$$J = S^T \times S \tag{3.6}$$

où S^T est la transposée de la matrice S. Chaque élément $J_{i,j} \in J_{[N \times N]}$ est équivalent au calcul de (3.1). Dans l'application à un texte, s_i^{μ} est la fréquence du terme i dans la phrase μ et $J_{i,j}$ mesure l'interaction entre les termes i et j par le produit des fréquences d'occurrences simultanées dans les phrases. Par exemple, pour la matrice suivante, $J_{mot_1, mot_2} = (2 \times 1) + (1 \times 2) + (3 \times 1) = 7$.

$$S = \begin{array}{c} \begin{matrix} mot_1 & mot_2 & mot_3 & mot_4 \end{matrix} \\ \begin{pmatrix} 2 & 1 & 0 & 2 \\ 1 & 2 & 2 & 0 \\ 3 & 1 & 2 & 1 \end{pmatrix} \end{array}$$

Ainsi, l'énergie textuelle d'interaction (3.2) peut alors s'exprimer comme :

$$E = -\frac{1}{2} S \times J \times S^T \tag{3.7}$$

L'élément $E^{\mu,\nu} \in E^{[P \times P]}$ représente donc l'énergie textuelle entre les phrases μ et ν. Comme nous observons dans la figure 3.2, les valeurs de la diagonale représentent l'énergie d'interaction entre les termes d'une même phrase et celles qui se trouvent en dehors, symbolisent les interactions entre les termes de phrases différentes.

D'un autre côté, il est intéressant de noter dans la figure 3.3, qu'une ligne (ou colonne, car E est symétrique) de cette matrice peut donner lieu à un spectre qui correspond à l'énergie d'interaction d'une phrase avec toutes les autres. L'énergie textuelle donne un aperçu du panorama global du document, vu depuis la perspective d'une phrase. La somme des valeurs absolues d'un spectre correspond à l'énergie d'interaction totale d'une phrase μ avec le document :

$$E^{\mu,doc} = \sum_{\nu=1}^{P} |e^{\mu,\nu}| \tag{3.8}$$

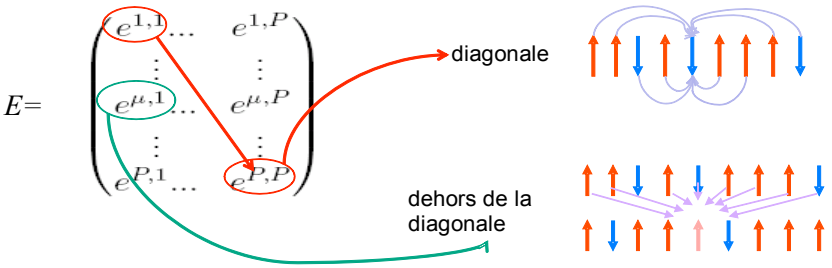

FIGURE 3.2 – *La matrice d'énergie textuelle E. Les valeurs de la diagonale sont les sommes des interactions des termes de la même phrase, et celles hors de la diagonale, les interactions entre les termes de deux phrases différentes.*

où $e^{\mu,\nu}$ est l'énergie d'interaction entre un couple de phrases, et la somme porte sur toutes les phrases ν du document. Les valeurs $E^{\mu,doc}$ peuvent servir à pondérer les phrases les unes par rapport aux autres. Ainsi, les phrases les plus énergétiques seront les plus représentatives du contenu du document.

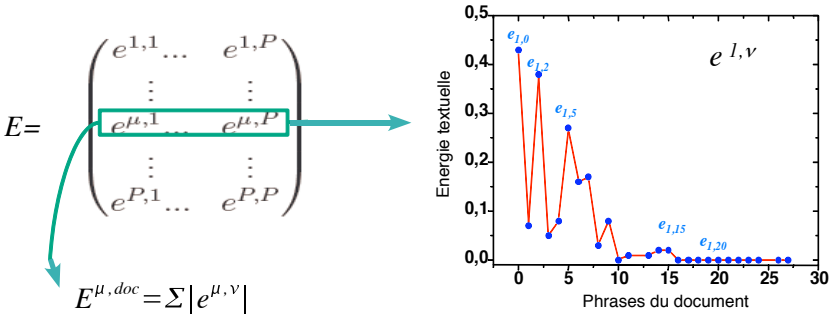

FIGURE 3.3 – *Exemple de spectre d'énergie d'une phrase. La courbe correspond à l'énergie d'interaction de la phrase 1 avec toutes les autres phrases ν, $e^{1,\nu}$.*

Dans la section suivante nous faisons appel à la théorie des graphes pour expliquer les propriétés de l'énergie textuelle comme mesure de similarité entre documents. Grâce à ses propriétés, l'énergie textuelle a été appliquée avec succès à plusieurs tâches du TAL telles que le résumé automatique, la recherche d'information, la segmentation thématique et la classification documentaire. Ces applications seront détaillées dans les chapitres à venir.

3.3.2 Interprétation sur les graphes

Nous allons expliquer théoriquement la nature des liens entre phrases induits par l'énergie textuelle. Pour cela nous allons utiliser quelques notions élémentaires de la théorie des graphes.

Chemins d'ordre deux

L'interprétation que nous allons faire repose sur le fait que la matrice d'énergie (3.7) peut s'écrire :

$$E = S \times (S^T \times S) \times S^T = (S \times S^T)^2 \tag{3.9}$$

En raison de l'utilisation des valeurs absolues de l'énergie (équation 3.8), nous avons négligé le coefficient $-1/2$ de la formule 3.7.

Soit le texte artificiel de $P = 4$ phrases σ_i $(i = 1, .., P)$ et un vocabulaire normalisé de $N = 5$ mots $\{A, B, C, D, E\}$:

σ_1= *(A B)*
σ_2= *(A B C E)*
σ_3= *(C D)*
σ_4= *(A B C D)*

qui produit la matrice terme-segment $S_{[P\times N]}$:

$$S = \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Si on multiplie la matrice S par sa transposée, on obtient la matrice de partage de mots entre phrases $(S \times S^T)_{[P\times P]}$:

$$\mathrm{S}\times S^T = \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \end{array} \\ \begin{pmatrix} 2 & 2 & 0 & 2 \\ 2 & 4 & 1 & 3 \\ 0 & 1 & 2 & 2 \\ 2 & 3 & 2 & 4 \end{pmatrix} \end{array}$$

Le carré de cette matrice est la matrice d'énergie textuelle $E_{[P\times P]} = (S \times S^T)^2$:

$$(\mathrm{S}\times S^T)^2 = \begin{array}{c} \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \end{array} \begin{array}{c} \begin{array}{cccc} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \end{array} \\ \begin{pmatrix} 12 & 18 & 6 & 18 \\ 18 & 30 & 12 & 30 \\ 6 & 12 & 9 & 15 \\ 18 & 30 & 15 & 33 \end{pmatrix} \end{array}$$

Considérons que les phrases σ constituent les sommets du graphe $I(S \times S^T)$ d'intersection (voir la figure 3.4). On trace une arête entre deux sommets σ_μ et σ_ν chaque fois qu'ils partagent au moins un terme. C'est-à-dire $\sigma_\mu \cap \sigma_\nu \neq \emptyset$.

En effet, $I(S \times S^T)$ 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

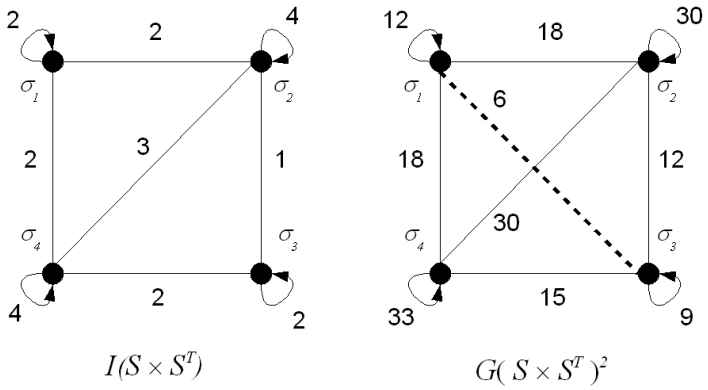

FIGURE 3.4 – *Graphes d'adjacence issus de la matrice d'énergie.*

qui correspond au nombre de termes 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 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 $I(S \times S^T)$;
- il existe une arête entre deux sommets s'il existe un chemin de longueur au plus deux 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 deux entre les deux sommets. Ces chemins pouvant comprendre des boucles.

De cette représentation on déduit que la matrice d'énergie textuelle relie à la fois des phrases ayant des termes communs puisqu'elle englobe le graphe d'intersection, ainsi que des phrases qui partagent un même voisinage sans pour autant partager nécessairement un même vocabulaire. Comme on peut observer dans la figure 3.4, deux phrases qui ne partagent aucun terme en commun, comme σ_1 et σ_3, 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 dans leur voisinage commun, et donc du vocabulaire apparaissant dans un contexte commun. Pour donner un exemple sur des textes réels, soient les trois phrases :

1. `La biologie étudie la complexité de la vie.`
2. `L'information génétique contenue dans l'ADN.`
3. `L'ADN est le porteur de l'information de la vie.`

Après la phase de prétraitement elles deviennent :

1. `biologie étudier complexité vie`
2. `information génétique contenir adn`
3. `adn porteur information vie`

On voit bien que, après le prétraitement (section 2.3.3), les phrases 1 et 2 n'ont pas de termes en commun et le cosinus entre elles, calculé selon l'équation (2.6), est nul. L'existence de la phrase 3 dans la collection ne change rien à cette valeur. Par contre l'énergie textuelle entre 1 et 2 n'est pas nulle grâce à la phrase 3 qui partage le terme *vie* avec la phrase 1, et les termes *adn* et *information* avec la phrase 2.

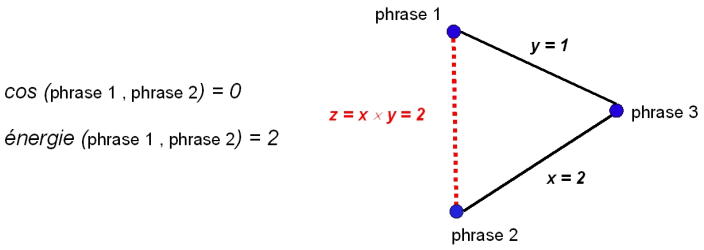

FIGURE 3.5 – *Chemin d'ordre deux entre les phrases 1 et 2. L'énergie textuelle entre les phrases n'ayant pas de termes en commun n'est pas nulle.*

Dans ce sens, cosinus est une mesure locale qui ne prend pas en compte l'interaction entre termes. En revanche, l'énergie textuelle, en incluant cette interaction, est une mesure globale qui fait intervenir le voisinage commun entre documents.

Les chemins d'ordre supérieur

L'énergie textuelle pondère les segments textuels selon leur pertinence grâce au calcul des chemins d'ordre deux. Cette mesure capture des relations indirectes que les mesures vectorielles locales, comme le cosinus, s'avèrent incapables de détecter. Par conséquent, il est logique de penser que les chemins d'ordre trois et supérieurs iront plus loin dans ce processus et ils pourraient, *a priori*, améliorer les résultats.

Pour tester cette hypothèse, nous avons réalisé des expériences en calculant différentes puissances n de la matrice $(S \times S^T)^n$ pour des petits textes en anglais et en français et on n'a pas observe de changements de rang significatifs. Le corpus utilisé est celui décrit dans la première expérience de la section 4.2.3.

Comme illustration, nous présentons les résultats du classement des 26 phrases du document généraliste « 3-mélanges ». Le texte complet se trouve dans les annexes (section A.1). Dans le tableau 3.1 à gauche, nous montrons les phrases rangées selon la valeur descendante de $(S \times S^T)^n$ (pertinence) (la phrase au rang 1 est la plus importante et celle de rang 26 la moins importante) en variant n de 2 à 5. On observe que les changements de rang ne sont pas significatifs surtout si on pense à extraire, par exemple, 25% du total de phrases comme représentatives du contenu. Dans ce cas, nous choisirons toujours les même sept phrases : 9, 12, 11, 5, 7, 6, 14 (sauf pour $n = 5$ où l'on change la phrase 7 pour la 13). Dans le même tableau à droite, le résumé produit par concaténation des sept phrases pertinentes est affiché. Ce même effet a été constaté sur les autres textes du corpus (de tailles, langue et thématiques différentes).

$(S \times S^T)^n$				
Rang	$n=2$	$n=3$	$n=4$	$n=5$
1	**9**	**9**	**9**	**9**
2	**12**	**12**	**12**	**12**
3	**11**	**11**	**11**	**11**
4	**5**	**5**	**14**	**14**
5	**7**	**14**	**5**	**5**
6	**6**	**6**	**6**	**13**
7	**14**	**7**	**7**	**6**
8	13	13	13	7
9	8	8	8	15
10	0	0	4	8
11	4	4	0	4
12	2	15	15	10
13	15	2	10	0
14	10	10	2	2
15	20	3	3	3
16	3	1	1	1
17	1	20	20	16
18	19	18	16	20
19	18	16	18	18
20	16	19	19	19
21	23	25	25	25
22	25	26	26	26
23	26	23	23	23
24	21	21	21	21
25	17	17	17	17
26	22	22	22	22

Pour n=2,3,4, le résumé au 25% du nombre de phrases est :
5 Le secrétaire général des Nations Unis a appelé la communauté internationale à apporter une contribution financière au gouvernement fédéral de transition somalien.
6 Le nouveau président du gouvernement fédéral de transition somalien, Abdillahi Youssouf Ahmed, a lancé un appel, lors de son investiture la semaine dernière, à la communauté internationale pour qu'elle participe à la reconstruction de la somalie.
7 Une délégation conduite par le secrétaire du bureau des affaires étrangères et économiques de la Grande Bretagne, Chris Mullin, a rencontré aujourd'hui le nouveau président somalien et des membres du parlement fédéral de transition de la somalie.
9 De manière générale, les objectifs scientifiques du LIA concernent le traitement automatique du langage naturel, l'optimisation vise le développement de méthodes numériques spécifiques pour le traitement des langues naturelles et les systèmes de télécommunication dans lesquels ces méthodologies peuvent être appliquées ou intégrées.
11 La représentation et le traitement de ces informations sont effectués au moyen d'outils réalisés au laboratoire ou choisis pour leur efficacité dans le contexte de nos travaux.
12 Les travaux réalisés ou envisagés, s'ils concernent fréquemment des difficultés pratiques posés par le traitement de très grandes quantités d'informations complexes, sont abordés de manière à mettre en lumière des problématiques qui généralisent ces questions spécifiques.
14 Nos activités scientifiques et leurs retombées pratiques sont systématiquement et régulièrement confrontées aux travaux concurrents dans le cadre des évaluations internationales des systèmes de traitement automatique des données.
Pour $n=5$, la phrase 13 remplace la 7 dans le classement :
13 Des interactions permanentes et approfondies avec d'importants groupes industriels nationaux ou internationaux permettent d'assurer une meilleure continuité entre les travaux académiques et leurs applications rapides dans les systèmes opérationnels réalisés par nos partenaires.

TABLE *3.1* – *Score des 26 phrases du texte « 3-mélanges » obtenu avec le calcul des chemins d'ordre* $2 \leq n \leq 5$.

Il semble que le calcul de chemins d'ordre deux (donc, de l'énergie textuelle) est une stratégie de pondération suffisamment efficace qui englobe les interactions les plus importantes contenues dans les documents. De plus, la complexité du calcul matriciel avec $n > 2$ est considérable. Il n'est pas donc intéressant d'alourdir le processus sans obtenir des bénéfices clairs.

3.4 Comparaison avec des méthodes basées sur les graphes

L'interprétation de l'énergie textuelle en théorie des graphes révèle des similarités avec les algorithmes de résumé automatique par extraction adaptés du célèbre PAGERANK (Page et al., 1998) utilisé pour classer les pages du Web par ordre de pertinence. Ces adaptations du PAGERANK ont obtenu de bons résultats lors des différentes campagnes d'évaluation DUC. Cependant pour expliquer leurs résultats leurs auteurs ne disposent comme modèle que celui d'un large réseau de co-référencement. Cela pose deux problèmes. D'une part on assimile les mots à des références de pages Web. D'autre part on applique cet algorithme à des textes qui n'ont certainement pas les dimensions du Web, l'outil semble donc disproportionné vis à vis de l'objectif. Nous allons montrer ici que l'énergie textuelle suffit à expliquer les résultats de ces approches. C'est une

illustration de comment un concept fondamental de la Physique Statistique peut aussi contribuer à comprendre et à simplifier des algorithmes complexes établis en TAL.

Dans cette section nous commençons par rappeler les adaptations de l'algorithme PAGERANK pour le résumé par extraction de phrases. Nous procédons ensuite à une comparaison sur des matrices aléatoires, puis sur des textes réels dont les corpus des campagnes DUC.

3.4.1 Les approches fondées sur l'algorithme de PAGERANK

L'algorithme PAGERANK (Page et al., 1998) a été proposé pour calculer l'importance des pages Web liées par hyperliens. De façon intuitive, une page aura un score PAGERANK haut s'il existe de multiples pages ayant elles mêmes un score élevé qui la référencent. Le score PAGERANK $R(\mu)$ d'une page Web μ qui a un ensemble B_μ de pages signalant vers elle est calculé récursivement selon la formule :

$$R(\mu) = (1-d) + d \sum_{\nu \in B_\mu} \frac{R(\nu)}{N_\nu} \tag{3.10}$$

où N_ν est le nombre de liens qui sortent de la page ν ; d est un paramètre d'amortissement qui prend des valeurs entre 0 et 1. Il est mis à 0,85 habituellement et représente la probabilité qu'un internaute ne suive pas un hyperlien de μ et choisisse une autre page au hasard. Ainsi, PAGERANK prend en compte le comportement d'un internaute aléatoire qui, à partir d'une page choisie au hasard, commence à suivre les liens contenus dans ce site. Éventuellement il peut sortir de ce chemin et recommencer aléatoirement dans une autre page.

La résolution de l'équation 3.10 se fait de manière itérative, par recherche de point fixe en choisissant, avec des valeurs arbitraires pour les scores initiaux. Les scores finaux sont calculés itérativement jusqu'à satisfaction du critère de convergence :

$$\delta = |\overrightarrow{r}_{i+1} - \overrightarrow{r}_i| < \epsilon \tag{3.11}$$

où ϵ est une valeur positive proche de 0. Un tel processus propage récursivement les poids des liens au travers de la structure du Web (Page et al., 1998).

Vu d'un autre angle, les scores donnés par PAGERANK correspondent aux composantes du premier vecteur propre $\overrightarrow{r}$ de la matrice carrée M des probabilités qu'un internaute passe d'une page à une autre (Page et al., 1998). $M_{\mu,\nu} = 1/N_\mu$ s'il existe un lien entre μ et ν et $M_{\mu,\nu} = 0$ autrement. Si on voit $\overrightarrow{r}$ comme un vecteur non nul, alors :

$$M \times \overrightarrow{r} = \overrightarrow{r} \tag{3.12}$$

En effet la matrice M est stochastique, il en résulte que l'application linéaire correspondant à M est contractante et donc que sa plus grande valeur propre est 1. $\overrightarrow{r}$ correspond donc au vecteur propre principal de M qui est aussi son point fixe.

Les méthodes itératives de l'algèbre linéaire (comme le *power method*) permettent de calculer le vecteur $\overrightarrow{r}$ dont les éléments donnent les pondérations PAGERANK des pages de manière à les ordonner par ordre décroissant.

L'algorithme PAGERANK a été transposé au traitement de textes par (Mihalcea, 2004). L'auteur a assimilé les phrases aux pages Web et les liens aux ensembles de termes partagés. Leur système TEXTRANK calcule les rangs des phrases dans les documents. Une première différence avec notre approche d'Énergie textuelle, est que nous proposons une méthode qui se limite au calcul du carré de la matrice $(S \times S^T)$, alors que TEXTRANK décrit un processus itératif (de 30 pas approximativement) basé sur le calcul du premier vecteur propre de la matrice de liens entre phrases. Pour vérifier qu'une seule itération suffit effectivement, nous avons réalisé deux types d'expériences :

1. Sur un ensemble de matrices aléatoires, nous utilisons le logiciel de Statistique **R** [2] pour calculer le vecteur propre principal de ces matrices.
2. Sur un ensemble de textes réels, en calculant les scores de TEXTRANK en utilisant l'algorithme tel que publié.

Dans les deux cas, les rangs obtenus sont comparés à ceux calculés par notre système d'énergie sur les mêmes matrices et documents.

3.4.2 Comparaison sur des matrices aléatoires (texte artificiel)

Pour comparer les classements issus du vecteur propre principal avec ceux obtenus par l'énergie textuelle, nous avons réalisé l'expérience suivante.

Nous avons défini un ensemble de matrices entières positives M de taille arbitraire P comme le produit matriciel $S \times S^T$ où S est une matrice binaire de P lignes (phrases) et N colonnes (mots). Ceci peut être assimilé à du texte artificiel. Nous supposons que pour $0 < i \leq P, 0 < j \leq N$, la probabilité d'avoir $S_{i,j} = 1$ est une constante p. Nous avons que :

1. P est le nombre de phrases à scorer,
2. N est le nombre de termes différents,
3. p est la probabilité qu'un terme t se trouve dans une phrase μ.

Pour chaque matrice nous avons calculé la matrice d'énergie $E = (S \times S^T)^2$ et le vecteur $\overrightarrow{e} = E \times \overrightarrow{1}$, où $\overrightarrow{1} = (1, ..., 1)$ et $\overrightarrow{e}$ est le score donné par l'énergie textuelle. Additionnellement nous avons obtenu le vecteur propre principal $\overrightarrow{r}$ de M. Enfin, nous avons comparé les scores induits par chacun des vecteurs en calculant la valeur τ de Kendall [3] telle que implémentée dans la fonction `cor.test` de **R**. Le coefficient τ de Kendall est proche de 0 dans le cas d'une indépendance totale entre les scores, proche de 1 pour une concordance parfaite et -1 pour des rangs opposés. La p-valeur donne la

2. http://www.r-project.org/

3. Les tests non paramétriques de corrélation utilisant le τ de Kendall s'imposent lorsque qu'on ne peut pas faire l'hypothèse d'une distribution normale bivariée. Sur des échantillons de plus de 10 paires de valeurs, la distribution de la statistique du τ de Kendall suit approximativement une loi normale ce qui permet d'estimer sa vraisemblance.

probabilité de l'hypothèse nulle d'indépendance statistique. Une description détaillée du test de Kendall se trouve dans l'annexe D.

Dans la figure 3.6 nous montrons le code que nous avons utilisé pour **R**.

```
# Génération des matrices aléatoires
S=rbinom(P*N,1,p) ; dim(S)=c(P,N) ; M=S%*%t(S)
# Calculer l'énergie textuelle et l'énergie total par phrase
E=M%*%M; X=matrix(c(1),P,1) ; e=E%*%X
# Calculer le premier vecteur propre
L=eigen(M,TRUE) ; r=L[["vectors"]][,1]
# Test de comparaison des rangs
cor.test(e,r,method = c("kendall"))
```

FIGURE 3.6 – *Comparaison entre les classements des phrases induits par l'Energie textuelle et par le premier vecteur propre, sur une matrice aléatoire, en utilisant le logiciel statistique R. Les sorties sont comparées en calculant le τ de Kendall.*

Nous avons expérimenté les triplets suivants de valeurs (P, N, p) : (100 ;100 ;0,01), (500 ;100 ;0,01), (500 ;100 ;0,001), (1000 ;100 ;0,01) et (1000 ;100 ;0,001) en répétant le processus 30 fois pour chaque triplet. Nous avons alors calculé la valeur minimale obtenue pour le τ de Kendall. Nous avons obtenu $|\tau| > 0,8$, ce qui induit une p valeur inférieure à 10^{-5}.

Il en résulte que les rangs induits par $\overrightarrow{e}$ et $\overrightarrow{r}$ sont fortement corrélés.

3.4.3 Comparaison sur des textes

Pour mener une comparaison sur des textes réels nous avons implémenté l'algorithme TEXTRANK tel que décrit dans (Mihalcea, 2004). Nous avons utilisé les deux systèmes, TEXTRANK et Énergie textuelle, pour classer les phrases d'une vingtaine de documents issus du corpus DUC 2002[4] choisis aléatoirement. Les classements obtenues sont très similaires surtout dans l'assignation des premières places. Un exemple est montré dans le tableau 3.2. Il correspond au document sur l'ouragan *Gilbert* utilisé par (Mihalcea, 2004) pour illustrer le fonctionnement du système. À gauche, les scores normalisés des 25 phrases du texte obtenus pour la méthode d'énergie et TEXTRANK. À droite, les quatre phrases les plus pertinentes qui seraient sélectionnées pour produire, par exemple, un résumé d'environ 100 mots. Le texte complet se trouve dans les annexes (section A.2).

Dans le chapitre suivant, nous élargirons cette comparaison à la totalité du corpus DUC 2002 en appliquant les algorithmes à la tâche de résumé automatique. Nous y réaliserons une évaluation plus formelle en utilisant les protocoles du TAL. Cependant, à partir des résultats présentés ici, il est possible de tirer quelques conclusions préliminaires.

4. La conférence DUC 2002 a proposé des tâches de résumé automatique. Pour le résumé mono-document, le corpus contient $\approx$ 600 documents non spécialisés d'une trentaine de phrases chacun disponibles à http://www-nlpir.nist.gov/projects/duc/data.html

Rang	Phrase	Score Énergie	Phrase	Score TextRank
1	**9**	1,00	**9**	1,00
2	**15**	0,89	**16**	0,90
3	**18**	0,83	**18**	0,86
4	**16**	0,73	**15**	0,74
5	20	0,32	5	0,66
6	14	0,31	14	0,60
7	10	0,31	21	0,56
8	17	0,28	10	0,54
9	5	0,23	12	0,51
10	13	0,21	20	0,46
11	21	0,20	23	0,44
12	11	0,19	13	0,42
13	22	0,18	4	0,39
14	4	0,17	8	0,38
15	23	0,13	17	0,38
16	24	0,12	22	0,38
17	12	0,11	11	0,31
18	8	0,10	24	0,27
19	7	0,03	6	0,08
20	19	0,02	7	0,08
21	3	0,01	19	0,08
22	6	0,00	3	0,00
23	2	0,00	2	0,00
24	1	0,00	1	0,00
25	0	0,00	0	0,00

Les deux systèmes classent en premières places les même quatre phrases :

9 Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.

15 The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

16 The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

18 Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico ?s south coast.

TABLE 3.2 – *Score des 25 phrases d'un des documents du corpus DUC'02 obtenus par le calcul d'énergie textuelle et le système* TEXTRANK. *Les résultats sont similaires, surtout pour les phrases classées en premières places.*

Bien que l'idée d'adapter l'algorithme PAGERANK pour scorer les phrases d'un document est tout-à-fait intéressante, nous pensons qu'un document et une page Web sont deux systèmes de natures différentes. La connectivité est clairement plus forte entre les phrases d'un même texte qu'entre les hyperliens entre pages Web (figure 3.7). Notamment, l'utilisation du parametre $d = 0,85$ est un choix arbitraire dans le contexte d'un texte. C'est probablement la raison pour laquelle les itérations de TEXTRANK, tout comme le calcul de chemins d'ordre supérieur du graphe (section 3.3.2), ne semblent pas apporter grand chose sur le classement des phrases d'un document. Il semble qu'un calcul direct, comme celui de l'énergie textuelle, soit suffisant pour cette tâche.

3.5 Conclusion

Un des apports principaux de ce travail de thèse est l'introduction du concept d'énergie textuelle. Il représente l'énergie d'Ising calculée sur l'ensemble des phrases et des termes des documents. Dans ce chapitre nous avons expliqué le cheminement qui nous a amené à la définition de ce concept en partant de la mémoire associative de Hopfield dont nous avons pu exploiter les faiblesses. Également, nous avons introduit le calcul de l'énergie textuelle en utilisant la représentation vectorielle d'un document. Nous avons montré qu'elle offre un moyen de pondérer des segments textuels entre eux.

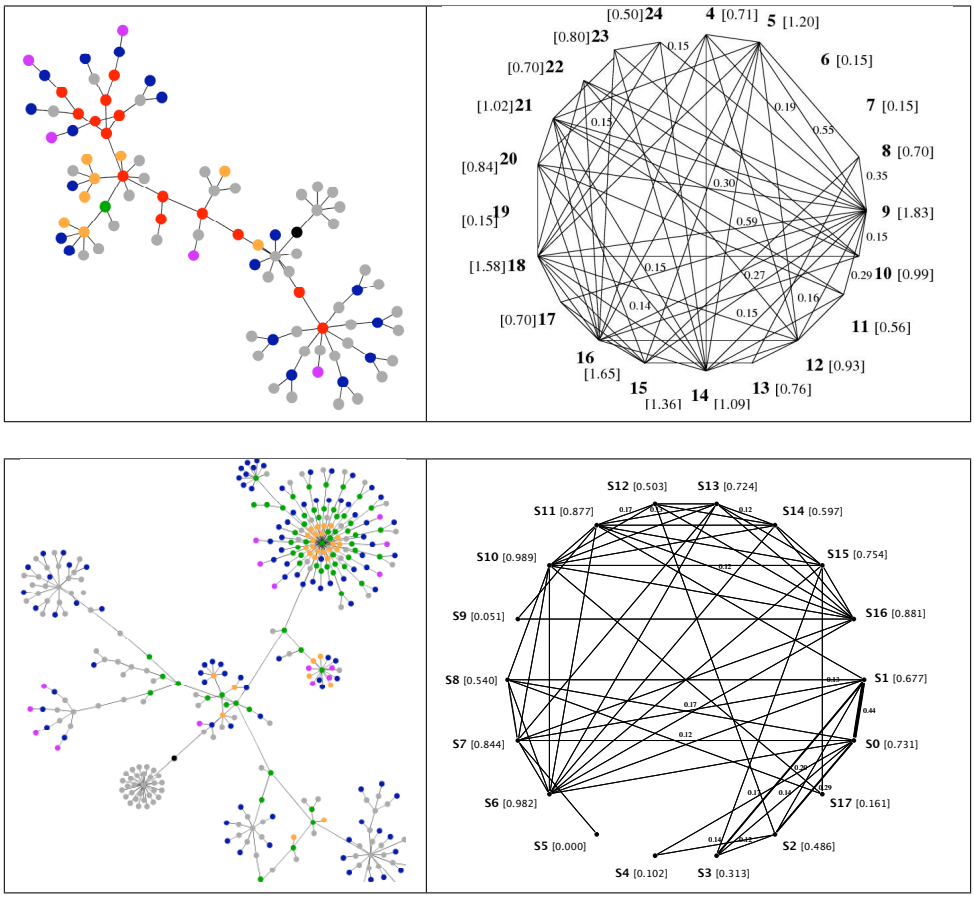

FIGURE 3.7 – *Comparaison entre graphes de sites Web et graphes des documents. En haut à gauche le graphe du site Web sur l'ouragan Gilbert publié par le National Hurricane Center (`www.nhc.noaa.gov/1988gilbert.html`). En haut à droite, le graphe d'un document traitant aussi sur l'ouragan Gilbert (voir les annexes, section A.2). En bas à gauche le graphe d'un site Web sur l'actualité au Tibet (`www.tibet-info.net/www/index.php`). En bas à droite le graphe d'un cluster de trois articles journalistiques sur le même sujet (voir le document dans les annexes A.3).*

La théorie des graphes nous a permis d'expliquer la nature des liens capturés par l'énergie textuelle. Grâce au fait qu'elle englobe le calcul de chemins d'ordre deux entre les segments d'un texte, cette nouvelle mesure de similarité capture des relations indirectes entre documents qui peuvent échapper aux autres mesures locales comme le cosinus. Finalement on a comparé notre approche d'énergie avec TEXTRANK, un système basé sur les graphes et qui est une adaptation de l'algorithme de classement de pages Web PAGERANK. Ce type de système, qui utilise la notion de vecteur propre principal comme outil de classement, propose un processus itératif pour obtenir les rangs des phrases d'un document. Nous avons montré des résultats suggérant que les processus itératifs de type PAGERANK et les calculs de chemins d'ordre supérieur, ne modifient significativement le score obtenu avec l'énergie textuelle. Il semble que dans un système avec un haut niveau de connectivité, comme celui des textes, le calcul des chemins

d'ordre deux est un moyen simple et efficace pour établir la pertinence de l'information portée. Cependant, une telle constat mérite un étude plus conséquente.

Chapitre 4

ENERTEX : un système basé sur l'énergie textuelle

Sommaire

4.1 Introduction

Nous avons montré au chapitre précédant que l'énergie textuelle peut être utilisée pour classer les phrases d'un document par ordre de pertinence. Nous avons développé le système ENERTEX basé sur cette nouvelle mesure de similarité entre les phrases d'un document. La première application abordée a été le résumé automatique dont nous décrivons les variantes dans la section 4.1.1. La démarche d'évaluation des systèmes de résumé sera aussi présentée. Nous avons mené une palette d'expériences en différentes

langues et domaines qui ont permis de constater les performances du système ENERTEX et de le positionner par rapport aux systèmes de l'état de l'art. Les résultats pour le résumé monodocument sont présentés au section 4.2. Nous montrerons en section 4.3 une modification qui consiste à mettre un champ externe en rapport avec un corpus multidocument. Cette stratégie permet de générer des résumés guidés par les besoins de l'utilisateur. Finalement, en section C nous ferons face à deux situations : un changement d'échelle et l'introduction d'éléments étrangers dans le réseau textuel. L'objectif est d'implémenter un système de recherche d'information guidé par annotations, qui fonctionne au niveau d'*abstracts*.

4.1.1 Le résumé automatique de documents

Les différentes modalités

Résumer, c'est le processus qui transforme un texte source en texte cible, de taille plus réduite et dans lequel l'information pertinente est conservée (Torres-Moreno, 2007). Les outils pour résumer automatiquement les textes peuvent être classés selon les critères suivants :

La méthode utilisée : On retrouve les approches par **compréhension** et les approches par **extraction**. La première postule que pour obtenir un résumé de qualité, il faut passer par une étape de compréhension. Elle produit des contractions du texte basées sur des re-formulations en utilisant un lexique nouveau (résumé ou *abstract*). En revanche, la méthode par extraction se limite à repérer les phrases importantes (par exemple, en analysant leur position dans le texte, la fréquence d'apparition des mots ou d'autres indicateurs statistiques), et de les extraire pour produire un résumé (extrait ou *extract*).

Le texte source : Il existe deux possibilités, le résumé **monodocument** où il s'agit de synthétiser un seul texte, et le résumé **multidocument** où l'idée est de condenser une grande quantité de documents souvent venus de sources très variées.

L'information à extraire : L'utilisateur peut souhaiter un résumé **générique** qui couvre l'ensemble thématique du texte source, ou un résumé **guidé** (ou personnalisé) permettant de capturer l'information concernant un sujet exprimé sous la forme d'une requête.

L'approche par extraction

L'avantage des méthodes de résumé par extraction, bien que moins proches du fonctionnement humain, est qu'elles sont plus faciles à mettre en œuvre et plus prolifiques.

La recherche en résumé automatique est devenue très dynamique ces derniers temps. Elle a été initiée par Luhn à la fin des années 50 (Luhn, 1958) avec un système de résumé par extraction adapté aux articles scientifiques. L'algorithme de Luhn utilisait la distribution des fréquences de mots dans le document pour pondérer les phrases.

Les approches par extraction proposent un résumé par sélection des phrases importantes. Il a été observé qu'environ 70% des phrases utilisées dans des résumés créés manuellement sont empruntées du texte source sans aucune modification (Lin et Hovy, 2003).

La plupart des travaux sur le résumé par extraction appliquent des techniques statistiques (analyse de fréquence, recouvrement de mots, etc.) aux unités telles que les termes, les phrases, etc. D'autres approches sont basées sur la structure du document (mots repères, indicateurs structuraux) (Edmundson, 1969; Paice, 1990), l'utilisation des SVM (*Support Vector Machine*) (Mani et Mayburi, 1999; Kupiec et al., 1995), les chaînes lexicales (Barzilay et Elhadad, 1997) ou encore la théorie de la structure rhétorique (Mann et Thompson, 1987).

Soit pour produire des condensés mono ou multidocuments, génériques ou guidés, la démarche par extraction se retrouve dans la majorité des travaux actuels, y compris ceux qui cherchent à produire des résumés par reformulation (Minel, 2004; Monod et Prince, 2006) ou ceux qui étudient les processus cognitifs pour résumer un texte (Lemaire et al., 2005; Mandin et al., 2005; Fayol, 1985). La communauté internationale reconnaît la pondération et l'extraction de phrases comme une étape importante dans la production de résumés. Notre travail a suivi cette voie en utilisant l'énergie d'une phrase pour indiquer son importance dans le document. Ceci nous a conduit immédiatement à une stratégie de résumé par extraction.

Nous décrivons ci-après les outils, issus des campagnes d'évaluation internationales, que nous utiliserons dans ce chapitre.

4.1.2 Les campagnes d'évaluation DUC

Le *National Institute of Standards and Technology* [1] (NIST) organise depuis 2001 les campagnes d'évaluation *Document Understanding Conference* [2] (DUC). Un des objectifs de ces campagnes est de permettre aux chercheurs de confronter leurs méthodes dans le cadre des expérimentations à grande échelle. Les campagnes DUC ont successivement introduit différentes tâches concernant le résumé. Nous nous intéressons aux campagnes des années 2002, pour la tâche de résumé générique monodocument, et 2005-2007 [3] pour des résumés orientés multidocuments. Elles seront décrites en détail dans les sections suivantes. En raison de l'évolution des tâches, nous n'avons pas participé directement à ces campagnes, cependant nous avons utilisé les corpus pour tester nos approches et comparer leurs performances à celles des participants.

1. http://www.nist.gov
2. http://duc.nist.gov
3. À partir de 2008, DUC devient TAC (*Text Analysis Conference*) et la tâche de résumé multidocument guidé à été remplacée par celle de mise à jour de résumés. Elle consiste à écrire un court résumé à partir d'un ensemble d'articles journalistiques avec l'hypothèse que l'utilisateur a déjà lu un premier ensemble d'articles plus anciens. Il s'agit de detécter la nouveauté.

4.1.3 Les mesures ROUGE

L'évaluation des systèmes dans les campagnes DUC, a été principalement réalisée à travers des mesures intrinsèques qui confrontent les sorties avec des modèles de référence. Ces modèles sont toujours des productions humaines. Cette section présente une de ces mesures, largement utilisée dans ce type d'évaluations.

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004), est un outil permettant d'évaluer la qualité d'un résumé à partir de sa ressemblance avec d'autres résumés considérés comme idéaux. Évidemment, ces modèles de référence sont écrits par des humains. L'unité de texte utilisée par ROUGE est le n-gramme ou séquence de n termes. ROUGE propose tout un ensemble de métriques basées sur la taille et la structure des n-grammes :

ROUGE-**(*n*)** : Mesure de rappel calculée sur les co-occurrences de n-grammes entre un résumé candidat r_{cand} et un ensemble R_{ref} de résumés de référence (équation 4.1). Co-occurrences(n-grammes) correspond au nombre maximum de co-occurrences de n-grammes entre r_{cand} et R_{ref} et Nombre(n-grammes) au nombre de n-grammes apparaissant dans un résumé idéal.

$$\text{ROUGE-}(n) = \frac{\sum_{s \in R_{ref}} \sum_{n\text{-grammes} \in s} \textit{Co-ocurrences}(n\text{-grammes})}{\sum_{s \in R_{ref}} \sum_{n\text{-grammes} \in s} \textit{Nombre}(n\text{-grammes})} \tag{4.1}$$

ROUGE-**SU(*m*)** : Adaptation de ROUGE-2 utilisant des bigrammes à trous (SU = *skip units*) de taille maximale m et comptabilisant les unigrammes.

Les métriques les plus courantes, utilisées lors des campagnes DUC, sont ROUGE-2 et SU4. Nous illustrons leurs fonctionnements avec un exemple adapté de (Lin, 2004).

Soit R un document de référence, et D_1, D_2 et D_3 les candidats à évaluer. Nous montrons dans le tableau 4.1 le découpage des documents en unités textuelles de tailles différentes. Les documents D_i ont quatre unigrammes, trois bigrammes, et six bigrammes à trous. R et D_1 partagent 1/3 des bigrammes (*the-gunman*). La valeur ROUGE-2 est de 0,33. Ils partagent aussi 3 des 6 bigrammes à trous (*police-the*, *police-gunman* et *the-gunman*) et 3 des 4 unigrammes (*police*, *the*, *gunman*) ; ce qui donne 6/10 des unités partagées. Alors, pour D_1, ROUGE-SU4 vaut 0,6. Les résultats pour les autres documents se trouvent dans le tableau 4.2. Le document le plus proche à la référence R est D_1.

(Lin, 2004) soutient que les valeurs ROUGE ont une forte corrélation avec les jugements humains [4]. Même si les humains ont tendance à utiliser des synonymes, ROUGE fait intervenir l'entourage des mots éventuellement substitués. Un autre résultat remarqué par l'auteur, est que la corrélation avec les jugements humains augmente avec le nombre de modèles. D'où l'importance de disposer de nombreuses références.

Cependant, ROUGE n'est pas une mesure de la qualité du résumé mais seulement de la pertinence du contenu. Il existe des problèmes liés à la cohérence du texte et à la

4. Il reporte des coefficients de Pearson pour ROUGE-2 et SU4 d'une valeur comprise entre 0,94 et 0,99.

	Document	**Bigrammes**	**Bigrammes à trous par 4 mots maximum**
R	*police killed the gunman*	*police-killed, killed-the, the-gunman*	*police-killed, police-the, police-gunman, killed-the, killed-gunman, the-gunman*
D_1	*police kill the gunman*	*police-kill, kill-the, the-gunman*	*police-kill, police-the, police-gunman, kill-the, kill-gunman, the-gunman*
D_2	*the gunman kill police*	*the-gunman, gunman-kill, kill-police*	*the-gunman, the-kill, the-police, gunman-kill, gunman-police, kill-police*
D_3	*gunman the killed police*	*gunman-the, the-killed, killed-police*	*gunman-the, gunman-killed, gunman-police, the-killed, the-police, killed-police*

TABLE 4.1 – *Exemple de découpage de documents en unités textuelles.*

Documents	**Rappel ROUGE-2**	**Rappel ROUGE-SU4**
D_1	0,33	0,60
D_2	0,33	0,40
D_3	0,00	0,40

TABLE 4.2 – *Rappel* ROUGE *pour les documents* D_1 *à* D_4 *du tableau 4.1.*

résolution des anaphores [5] que ROUGE est incapable de détecter. L'évaluation automatique reste donc un problème ouvert.

4.2 L'énergie textuelle comme critère de pertinence

4.2.1 Résumé monodocument générique

ENERTEX a été utilisée pour la génération de résumés génériques monodocument. L'algorithme (figure 4.1) prend en entrée la représentation vectorielle (matrice S) du texte après filtrage, lemmatisation/*stemming* et normalisation (voir section 2.3.3). Ensuite, il applique le modèle de spins et réalise le calcul de la matrice d'énergie textuelle (équation 3.9). Finalement, il fait la pondération et le tri des phrases en utilisant les valeurs absolues d'énergie (équation 3.8). Les phrases pertinentes seront sélectionnées comme celles ayant la plus grande énergie d'interaction avec le document. Enfin, le système génère les résumés par affichage et concaténation des phrases pertinentes.

5. Les anaphores sont des expressions qui assurent la reprise des personnages ou objets.

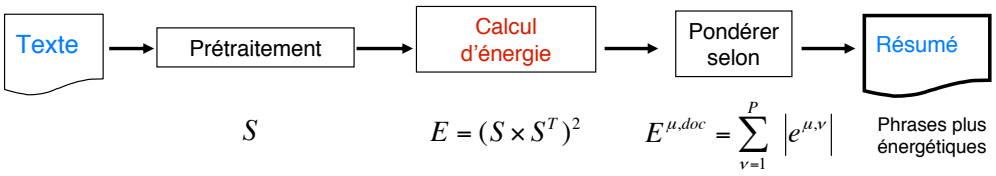

FIGURE 4.1 – ENERTEX *appliqué à la tâche de résumé générique. L'algorithme prend en entrée le texte source. Après la phase de prétraitement on produit la matrice terme-segmente S qui est utilisée pour le calcul d'énergie. La pondération et le tri des phrases sont faits en utilisant les valeurs absolues d'énergie. Le système génère les résumés avec la concaténation des phrases plus énergétiques.*

4.2.2 Évaluation sur le corpus DUC 2002

La tâche de résumé monodocument de l'édition DUC 2002 a consisté à la génération, pour chaque document du corpus, d'un court résumé d'environ 100 mots contenant l'information la plus pertinente. Le corpus est composé de 567 documents journalistiques en langue anglaise [6]. Quinze équipes ont participé à cette tâche.

Les documents ont été traités par ENERTEX selon le schéma 4.1 pour produire leurs résumés. Nous comparons dans le tableau 4.3 à gauche la performance d'ENERTEX avec :

1. les systèmes qui ont obtenu les trois premières places de la campagne (S27, S28 et S31 dans le tableau) ;
2. l'approche de graphes TEXTRANK décrite en section 3.4 ;
3. un référent de base ou *baseline*, proposé par les organisateurs, construit avec les premières 100 mots du document. Une telle *baseline* obtient des résultats intéressants dû à la nature journalistique du corpus. Dans ce type de documents les premières phrases sont, en général, porteuses d'informations importantes.

La mesure utilisée pendant la campagne DUC 2002 (et reportée par ces systèmes en (Mihalcea, 2004)) correspond à ROUGE-1. Nous observons que, selon cet indicateur, notre performance est similaire à celle des meilleurs participants du défi et supérieure à celle de TEXTRANK.

Or à l'heure actuelle est bien connu que l'évaluation faite uniquement avec ROUGE-1 (basée sur les uni-grammes de termes) est trop basique. C'est pour cette raison que dans les éditions suivantes le comité DUC a privilégié l'utilisation de ROUGE-2 et SU4. Nous avons élargi la comparaison entre TEXTRANK et ENERTEX en utilisant ces mesures. Étant donné qu'elles ne sont pas rapportées par (Mihalcea, 2004), nous avons obtenu les sorties de TEXTRANK en utilisant notre implémentation mise au point pour les expériences de la section 3.4. Les résultats sont présentés dans le tableau 4.3 à droite. Puisque les sorties des algorithmes participants ne sont pas disponibles pour cette édition, les performances des systèmes S27, S28 et S31 selon ROUGE-2 et SU4 sont omises. Pour la construction de la *baseline* nous avons suivi la même stratégie du comité DUC : prendre les premiers 100 mots des documents. Nous pouvons observer que la performance de notre système est à nouveau supérieure à celle de TEXTRANK.

6. http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

Système	ROUGE-1
S27	0,4405
S28	0,4346
ENERTEX	0,4342
TEXTRANK (Mihalcea, 2004)	0,4229
Baseline DUC	0,4162
S31	0,4160

Système	ROUGE-2	ROUGE-SU4
ENERTEX	0,1813	0,2045
TEXTRANK implémenté	0,1748	0,1988
Baseline implementée	0,1682	0,1855

TABLE 4.3 – *Résultats* ROUGE *obtenus pour : les trois premières places du défi DUC 2002 (S27, S28, S31), et les systèmes* TEXTRANK *et* ENERTEX *dans la tâche de résumé générique monodocument.*

4.2.3 Évaluation sur des corpus en plusieurs langues et domaines

Dans la section précedante nous avons produit des résumés mono-document en langue anglaise pour un corpus journalistique. Ici nous sommes intéressés en : *i)* appliquer notre algorithme sur des documents en autres langues et domaines ; *ii)* comparer les performances de la méthode ENERTEX à celles d'autres systèmes de résumé par extraction.

Comme une première expérience nous avons choisi les textes en français [7] : *3-mélanges* composé de trois thématiques (la politique, le Laboratoire d'Informatique d'Avignon, les trolls), *Puces* de deux thématiques (l'informatique et une invasion de puces) et *J'accuse* (lettre d'Émile Zola). Trois textes de l'encyclopédie Wikipédia en anglais ont été analysés, *Lewinksky* [8], *Québec* [9] et *Nazca Lines* [10].

En suivant le protocole décrit en section 4.2.1, chaque texte a été pre-traité selon les opérations décrites en section 2.3.3 pour produire la matrice terme-segment S sur laquelle le calcul d'énergie textuelle (équation 3.9) a été effectué. Ensuite, les phrases ont été ponderées selon l'équation 3.8. Le résumé est enfin construit en concaténant les r phrases les plus énergetiques du document. Le taux de compression est variable (selon la taille des textes) et calculé en rapport avec les nombre de phrases P du texte :

$$\textit{Taux de compression} = \frac{r : \textit{Nombre de phrases du résumé}}{P : \textit{Nombre de phrases du document}} \tag{4.2}$$

Les systèmes de résumé CORTEX (Torres-Moreno et al., 2002), MEAD [11], COPERNIC SUMMARIZER [12] ont été utilisés pour produire des résumés aux mêmes taux de compression appliqués par ENERTEX. Nous avons ajouté une *baseline* où les phrases ont été choisies au hasard.

7. Récupérables à l'addesse http://www.lia.univ-avignon.fr/fileadmin/documents/Users//Intranet/chercheurs/torres
8. http://en.wikipedia.org/wiki/Monica_Lewinsky
9. http://en.wikipedia.org/wiki/Quebec_sovereignty_movement
10. http://en.wikipedia.org/wiki/Nazca_lines
11. http://tangra.si.umich.edu/clair/md/demo.cgi
12. http://www.copernic.com

Nous avons évalué les résumés générés par ces systèmes avec les mesures ROUGE-2 et SU4. Pour obtenir les références humaines, nécesaires pour les tests ROUGE, l'ensemble de textes a été distribué à un petit groupe de doctorants et chercheurs du LIA et de l'INRA d'Avignon. Chacun d'entre eux a lu les documents et a signalé les r phrases considerées comme les plus pertinentes. Les résultats sont présentés dans le tableau 4.4. Les taux de compression et le nombre de références humaines utilisées pour l'évaluation sont affichés.

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 4.4 – *Rappel* ROUGE-2 *(R2) et SU4. Taux de compression de 25% : 3-melanges (8 ref), Puces (8 ref), Québec (8 ref) and Nazca (6 ref) ; 12% : J'accuse (6 ref) ; 20% : Lewinsky (7 ref).*

Les meilleures performances sont en gras et les deuxièmes positions en italique. ENERTEX, basé sur le seul calcul de l'énergie, obtient des bons résultats avec trois premières places et sept deuxièmes, proches de CORTEX qui se sert de 13 metriques (la position, l'entropie, la perplexité, le poids de Hamming, entre autres) et un algorithme de décision pour classer les phrases des documents (Torres-Moreno et al., 2002). Il est également intéressant de noter que certains systèmes sont limités à une seule langue. C'est le cas du MEAD qui traite seulement des textes en anglais (d'où les symboles ⊘ dans le tableau). La raison est que ce type de système repose sur une grande base de ressources linguistiques. Cependant, leurs performances ne sont pas forcément meilleures que celles de systèmes majoritairement numériques.

Additionnellement à cette expérience, trois collaborations nous ont permis d'étudier l'approche de résumé mono-document ENERTEX dans différents domaines. Nous présentons ici une brève description de ces travaux. Les détails se trouvent dans l'annexe B.

Compréhension vs. extraction. Issu d'une collaboration avec le Laboratoire des Sciences de l'Education de Grenoble [13] (LSE), ce travail a eu comme objectif de comparer notre approche par extraction avec une approche cognitive, proposée pour l'équipe grenobloise, basée sur l'analyse sémantique latente ou LSA (Landauer et Dumais, 1997). En plus d'ENERTEX et LSA, nous avons choisi d'inclure dans cette comparaison les systèmes : CORTEX, COPERNIC, PERTINENCE, et MICROSOFT WORD. Nous avons également ajouté deux *baselines* : *Baseline 1* (sélection aléatoire des phrases) et *Baseline 2* (sélection des phrases du début et de la fin du texte). Nous avons réalisé deux types d'analyse :

i) Au niveau de phrases. Nous avons calculé la précision des systèmes dans le repérage des phrases estimées importantes par les humains.

13. http://web.upmf-grenoble.fr/sciedu

ii) Au niveau de n-grammes de mots. En utilisant ROUGE-2 et SU4, nous avons mesuré la similarité entre les résumés automatiques et les humains.

La figure 4.2 présente la performance finale des systèmes comme la moyenne entre i) et ii). LSA_adultes (une des variantes du système LSA), CORTEX et ENERTEX obtiennent les meilleures performances. Encore une fois nous observons que le seul calcul de l'énergie permet d'approcher la performance de systèmes complexes et spécialisés comme CORTEX et LSA. Pour ce dernier, chaque mot est représenté dans un espace de grande dimension (milles ou millions de mots) et une réduction (à environ 300) est réalisée au travers de la décomposition en valeurs singulières. Pour plus de détails, voir l'annexe B.1.

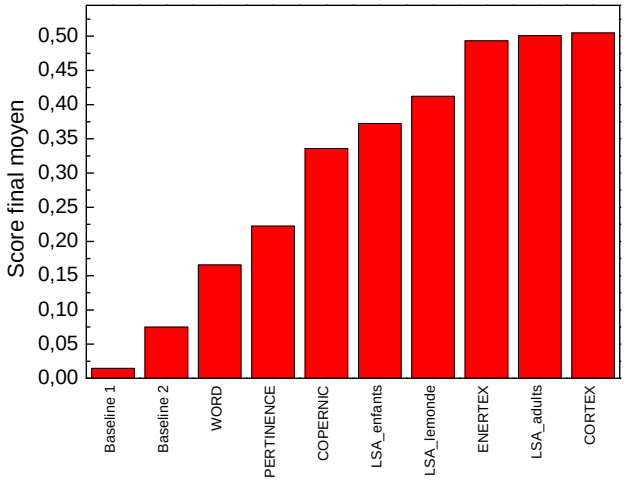

FIGURE 4.2 – *Compréhension vs. extraction : score final des systèmes.*

Cette expérience a été réalisée dans un cadre très particulier qui nous a permis d'avoir de nombreuses références humaines (plus de 600) classées par niveaux scolaires. Nous avons profité de cette situation pour classifier les systèmes automatiques par niveau scolaire selon la qualité des resumés qui produisent. Pour accomplir un tel objectif, nous avons mesuré avec ROUGE la proximité entre résumés générés automatiquement et les références humaines séparées par classes : collégiens, lycéens et étudiants universitaires. Pour chaque système et chaque classe, nous avons calculé le produit ROUGE-2 $\times$ SU4. Nous observons dans le tableau 4.5 que les résumés qui correspondent aux meilleurs systèmes sont plus similaires aux ceux des niveaux les plus avancés : 1ère et Master (M2). Par contre, les systèmes avec de piètres performances (MICROSOFT WORD) sont plus proches des collégiens. Les détails de ces expériences se trouvent dans l'annexe B.1.

Textes médicaux en espagnol. Une deuxième collaboration, avec l'IULA [14] (*Institut Universitari de Lingüística Aplicada*) de l'Université Pompeu Fabra de Barcelona España,

14. http://www.iula.upf.edu

Texte 1		
Système	**Niveau scolaire**	**ROUGE** $2 \times SU4$
LSA_adultes	M2	0,6719
CORTEX	M2	0,6719
LSA_lemonde	M2	0,6676
ENERTEX	M2	0,6362
LSA_enfants	M2	0,3538
COPERNIC	M2	0,2674
BASELINE 2	M2	0,3028
WORD	$3^{ème}$	0,1786
PERTINENCE	$4^{ème}$	0,1662
BASELINE 1	$4^{ème}$	0,0907

Texte 2		
Système	**Niveau scolaire**	**ROUGE** $2 \times SU4$
PERTINENCE	M2	0,3080
LSA_adultes	M2	0,2865
ENERTEX	$1ère$	0,2827
CORTEX	M2	0,2628
LSA_enfants	M2	0,2370
COPERNIC	M2	0,2056
LSA_lemonde	$4^{ème}$	0,1579
BASELINE 2	$3^{ème}$	0,0719
WORD	$4^{ème}$	0,0420
BASELINE 1	$3^{ème}$	0,0214

TABLE 4.5 – *Niveaux scolaires correspondant aux systèmes automatiques pour les deux textes étudiés.*

a eu pour objectif de combiner des méthodes statistiques (ENERTEX et CORTEX) et linguistiques (DISICOSUM) pour résumer des articles médicaux en espagnol. Le système DISICOSUM , proposé par l'équipe espagnole, se fonde sur l'hypothèse que dans les domaines spécialisés, les professionnels utilisent des techniques concrètes pour résumer leurs documents. Après une analyse manuelle d'un ensemble d'articles médicaux, des règles ont été déduites et intégrées à DISICOSUM (da Cunha et al., 2007).

La combination de ces trois systèmes a été faite de la façon suivante. D'abord, des règles linguistiques de DISICOSUM sont appliquées au document original. La sortie contient uniquement des phrases principales libres d'information supplémentaire. Ce document réduit a été résumé séparément par CORTEX , ENERTEX et DISICOSUM . La sortie est analysée par un algorithme de décision qui garde les phrases à extraire pour le résumé final.

Pour analyser la performance de cette approche hybride, nous l'avons appliquée à un corpus de 10 articles médicaux. Nous avons procédé à l'évaluation avec ROUGE en utilisant comme références les *abstracts* rédigés par les auteurs. Deux *baselines* aléatoires ont aussi été inclues. La première a été extraite du document original et la deuxième du document réduit par l'élimination des phrases accessoires détectées par DISICOSUM. Le tableau 4.6 présente la médiane du score ROUGE sur les dix articles analysés. Ils

Système	**ROUGE-2**	**ROUGE-SU4**
Hybride	**0,3638**	**0,3613**
DISICOSUM	0,3572	0,3359
ENERTEX	0,3598	0,3457
CORTEX	0,3218	0,3169
Baseline 1	0,2539	0,2489
Baseline 2	0,2813	0,2718

TABLE 4.6 – *Score ROUGE-2 et SU-4 pour les systèmes individuels et le système hybride.*

montrent que les performances individuelles des trois systèmes sont similaires mais inférieures à celle du système hybride. Les résultats corroborent que la combinaison des

techniques statistiques et linguistiques améliore les performances des systèmes isolés. Les détails de cette expériences se trouvent dans l'annexe B.2.

Des langues à structure éloignée

Les langues somalienne et maya, présentant des structures grammaticales éloignées du français et de l'espagnol, permettent d'évaluer une propriété souhaitable dans les systèmes du TAL : l'indépendance de la langue.

Le français et le somali : nous avons comparé statistiquement la performance d'ENERTEX et CORTEX pour la production de résumés de textes alignés français/somali. Ce travail a été réalisé en collaboration avec l'Institut des Sciences et des Nouvelles Technologies de Djibouti [15] (ISNT). Les détails de cette expérience se trouvent dans l'annexe B.3.1.

L'espagnol et le maya : des textes parallèles espagnol/maya ont été analysés au travers des algorithmes numériques qui ont permis de filtrer, normaliser et résumer les documents. Ce corpus a été proportionné par le Profr. Miguel Güémez de l'Université Autonome de Yucatán (UADY) [16]. Le système de résumé choisi pour cette expérience a été ENERTEX. Les détails sont dans l'annexe B.3.2.

Nous avons constaté la pertinence des phrases retenues par les systèmes ENERTEX et/ou CORTEX sur les corpus alignés français/somali et espagnol/maya. Ces résultats ont mis en évidence la capacité des systèmes numériques d'analyser des textes dont la structure est très differente de celles du français, l'anglais ou l'espagnol. Un tel analyse est possible grâce au traitement vectoriel qui est à la base de ces approches et qui permettent aux algorithmes une considérable indépendance de la langue.

4.3 Application d'un champ externe au système textuel

Un système magnétique peut être soumis à un champ externe B. Selon le signe et l'intensité de B, les spins du système tendront à s'orienter vers ce champ ou en direction opposée. L'énergie d'interaction entre les N spins s_i et le champ local B_j est :

$$E = -\sum_{j=1}^{N} B_j s_j \tag{4.3}$$

De façon analogue, nous avons utilisé l'énergie textuelle pour la tâche de résumé guidé par une thématique (ou un sujet). L'idée est d'observer la réponse du système (un corpus) face à un champ externe (une thématique ou sujet). Ce champ, représenté par le vecteur des termes d'un texte décrivant un sujet a été mis en relation avec le corpus multidocument formé avec les D documents concaténés (voir figure 4.3). L'énergie entre le

15. L'ISNT est un des cinq instituts du Centre d'Étude et de Recherche de Djibouti (CERD), http://www.cerd.dj.

16. Le Profr. Güemez est chercheur du Centre de Recherches Régionales (CIR) en Sciences Sociales de la UADY, http://www.cirsociales.uady.mx

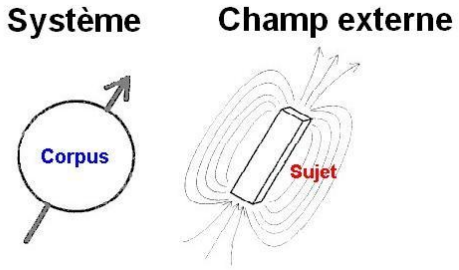

FIGURE 4.3 – *Le champ produit par un sujet sur un corpus de textes.*

sujet et chacune des phrases du corpus est calculée selon :

$$E(sujet, phrase \in corpus) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s_i^{sujet}\ J_{i,j}\ s_j^{phrase} \tag{4.4}$$

où le champ B produit par le sujet est :

$$B_j = \sum_{i=1}^{N} s_i^{sujet}\ J_{i,j} \tag{4.5}$$

4.3.1 Résumé multidocument guidé par une thématique

Les premiers systèmes de résumé automatique multidocuments ont été développés dans les années 90 (McKeown et Radev, 1995). Il s'agit de produire un résumé à partir d'une grande quantité de documents. Les difficultés introduites avec la dimension multidocument sont la redondance et la contradiction qui peuvent émaner des phrases extraites de documents différents. À cette complexité on ajoute un autre facteur : guider le résumé selon une requête de l'utilisateur. Cette requête doit permettre au système d'isoler les parties du document concernant une ou plusieurs thématiques pour ensuite produire un résumé n'incluant que ces dernières. C'est le résumé multidocument guidé, tâche principale des campagnes DUC 2005-2007.

Le problème peut se poser comme ceci : étant donnée une thématique et un ensemble d'environ 25 documents [17] pertinents, générer un court résumé de 250 mots, cohérent et bien organisé qui répondra aux questions de la thématique. Les thématiques sont composées de deux parties : le titre et une partie narrative contenant les questions. Un prétraitement standard (section 2.3.3) est appliqué à l'ensemble des documents. L'énergie textuelle entre le sujet et chaque phrase du corpus est calculée selon l'équation 4.4. Finalement, le résumé est composé des phrases présentant la plus haute énergie textuelle par rapport au sujet. Un post-traitement de diminution de la redondance est appliqué à la fin de la chaîne. La figure 4.4 illustre l'adaptation d'ENERTEX pour l'obtention d'un résumé guidé par un sujet.

17. Les documents proviennent du corpus AQUAINT : articles d'*Associated Press, New York Times* (1998-2000) et *Xinhua News Agency* (1996-2000).

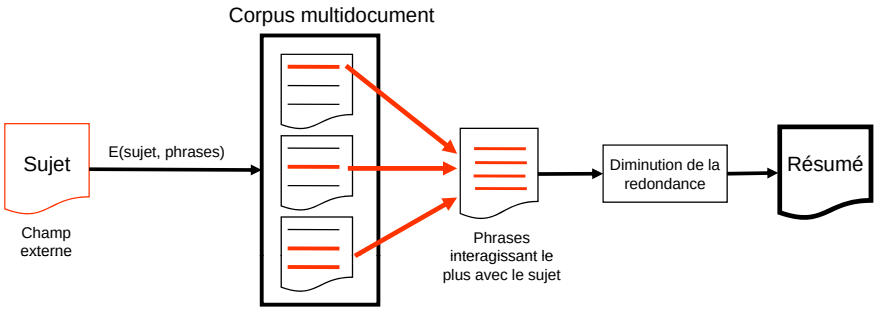

FIGURE 4.4 – *Algorithme d'*ENERTEX *pour le résumé guidé par un sujet sur un ensemble de documents.*

4.3.2 ΔE comme mesure de la redondance

Dans un résumé multidocument il y a une probabilité significative de re-inclure de l'information déjà présente. Pour diminuer ce problème il faut incorporer une stratégie de diminution de la redondance. ENERTEX n'utilise pas de traitement linguistique. La stratégie anti-redondance consiste uniquement à comparer les valeurs d'énergie des phrases candidates.

Nous supposons que (dans des grands corpus) la probabilité que deux 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). On a observé que dans un corpus suffisamment grand, deux phrases 1 et 2, avec la même valeur d'énergie textuelle E_1 et E_2 par rapport à la thématique sont égales. Peut-on aller encore plus loin et détecter avec ce même critère des phrases trop proches à quelques mots près ?

Pour le tester, on considère que si deux phrases partagent une grande partie du vocabulaire, elles apportent la même information. On construit donc le résumé avec la phrase la plus énergétique (en valeur absolue), puis la suivante en 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 $(i-1)$:

$$|E_i - E_{i-1}| = \Delta E \geq \epsilon; i = 2, 3, ... \tag{4.6}$$

Le cas contraire signifie que les énergies sont très proches avec une haute probabilité de redondance. Deux exemples extraits du corpus DUC 2006, sont montrés dans la figure 4.5. Les mots en désaccord dans les textes ont été soulignés, la différence d'énergie est de 0,0016 pour le premier couple et 0,0025 pour la deuxième. On présente en figure 4.6 à gauche les valeurs du produit ROUGE-2 $\times$ SU4 pour différentes valeurs de ϵ. Le meilleur résultat sur les corpus DUC'05-07 a été obtenu avec $\epsilon = 0,003$. Cela correspond aux phrases à deux mots près.

Phr 398 ''Star Wars : Episode I The Phantom Menace'' was screened Tuesday night in eight North American cities for movie theater executives, their families, and apparently a bundle of ''Star Wars'' fans who somehow finagled some of the prized tickets.
Phr 427 ''Star Wars : Episode I The Phantom Menace'' was screened Tuesday night in eight North American cities for movie theater executives, their families and apparently a bundle of ''Star Wars'' fans who somehow finagled some tickets.

Phr 409 While the early amateur critics loved the action sequences and most of the effects, including what they found to be an incredible underwater sequence, some reviewers thought the 2-hour-plus movie dragged a bit in places, that it was a little too kid-oriented and that one of the computer-generated characters, Jar Jar Binks, was annoying.
Phr 438 While the amateur critics loved the action sequences and most of the effects, including what they found to be an incredible underwater sequence, some reviewers thought the two-hour-plus movie dragged a bit in places, that it was a little too kid-oriented and that one of the computer-generated characters, Jar Jar Binks, was annoying.

FIGURE 4.5 – *Deux couples des phrases redondantes du corpus DUC 2006. La différence d'énergie textuelle entre elles est de* $\Delta E = 0,0016$ *et* $\Delta E = 0,0025$ *respectivement. Les petites différences entre les textes sont soulignées.*

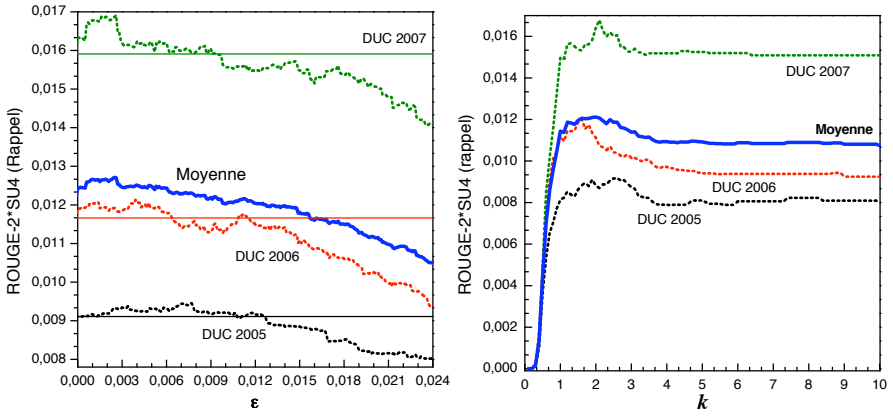

FIGURE 4.6 – *Diminution de la redondance :* ΔE *d'énergie des phrases et moyenne des longueurs de phrases.*

Le choix du paramètre ϵ est empirique. Il semble dépendant de la nature du corpus, mais reste encore difficile à déterminer [18].

Une autre stratégie permettant de diversifier le contenu, consiste à écarter du résumé les phrases longues (dans les documents il y a des phrases de taille comparable à celle du résumé demandé). Pour cela, on a défini la taille maximale des phrases comme $k \times M$, où M = nombre moyen de mots par phrase dans les documents originaux. Pour trouver la valeur optimale, nous avons fait varier k de 0 à 10 par petits pas de 0,1 en mesurant le produit de ROUGE-2×SU4. Le comportement est montré sur la figure 4.6 à droite. Le meilleur résultat est obtenu avec $k \approx 1,6$.

18. La difficulté réside dans le fait que les résumés produits avec une forte redondance peuvent avoir une valeur énergétique favorable.

4.3.3 Expériences

Nous avons testé ENERTEX guidé par une thématique sur les corpus DUC 2005-07. La figure 4.7 montre la position d'ENERTEX (triangle plein) dans l'évaluation ROUGE-2 vs. SU4, comparé aux participants. Pour raisons de clarté, on affiche uniquement les performances des systèmes au-dessus des deux évaluations *baselines* (triangles creux). En DUC'07 le comité a inclus deux *baselines*. La 1ère est une *baseline* tirée au hasard (DUC 05-07) et la 2ème est un système de résumé générique. Nous avons inclus une troisième mesure de comparaison, le cosinus (cercles noirs) (équation 2.6). Les résultats montrent que le cosinus obtient des performances ROUGE étonnamment hautes, mais les résumés peuvent contenir beaucoup de redondance, car toutes les phrases sélectionnées sont proches de la thématique. Par contre, la similarité de l'énergie textuelle tient compte non seulement du nombre de mots partagés, mais aussi des interactions indirectes. Le système ENERTEX a des résultats performants. Le cosinus (et le recouvrement [19]) sont des mesures locales dont les valeurs restent inchangées si l'on ajoute ou on enlève des phrases, tandis que l'énergie textuelle est sensible à ces variations.

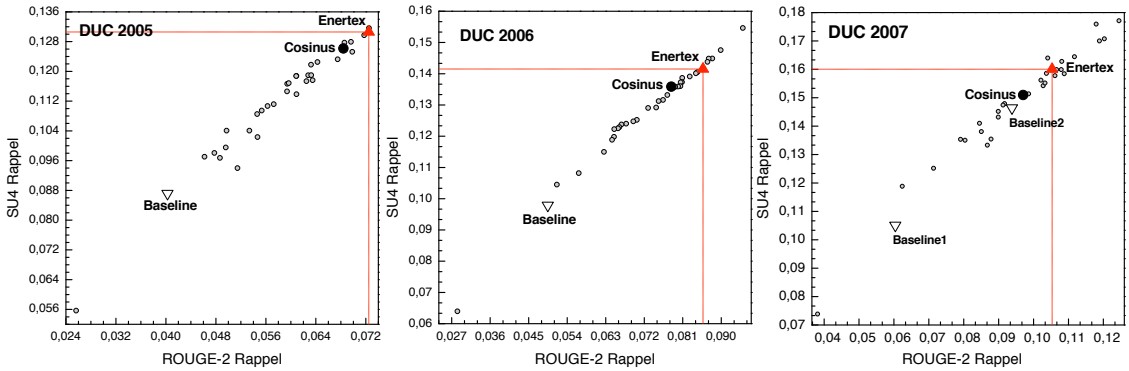

FIGURE 4.7 – *Aperçu du rappel SU4 vs* ROUGE-2 *des systèmes au-dessus des deux baselines.*

On fait noter dans la figure 4.7 que les performances d'ENERTEX le situent au niveaux des meilleurs candidats de 2005, mais que les équipes participant en 2006 et 2007 ont produit des résumés plus performants que ceux issus d'ENERTEX. Ce fait est lié à l'utilisation de plus en plus importante d'éléments linguistiques.

En effet, l'utilisation des ressources externes augmente quelques milliémes le score ROUGE mais le prix à payer est non négligeable : listes d'expressions régulières pour la réécriture des phrases en post-traitement ; la détection, expansion ou introduction des acronymes ; la résolution des anaphores, etc. Ces stratégies rendent les systèmes de plus en plus dépendants de la langue. ENERTEX n'utilise pas de post-traitement linguistique, ce qui permet de garder une certaine indépendance vis-à-vis de la langue. Les participants des campagnes DUC sont en général des systèmes de résumé automatique complexes. ENERTEX est une mesure de similarité qui pourrait éventuellement être couplée à des post-traitements appropriés, mais ceci sort du cadre de cette thèse.

19. Le recouvrement ou *overlap* est le niveau d'inclusion d'un segment textuel dans un autre.

Évaluation qualitative

Additionnellement à l'évaluation semi-automatique faite avec la mesure ROUGE, nous avons demandé à deux juges humains (chercheurs du LIA) de faire une évaluation manuelle sur cinq des résumés guidés produits par ENERTEX. Les critères d'évaluation ont été : structure, cohérence entre les phrases et pertinence du contenu (qui doit répondre aux questions du narratif). En général, les juges ont conclu que :

1. Les nuances de l'instruction donnée dans le narratif, comme : « discuter », « décrire » ne semblent pas avoir été prises en compte ;
2. il existe quelques problèmes de cohérence entre phrases consécutives ;
3. environ 80% de l'information qui répond aux questions est présente dans les résumés.

Dans la figure 4.8 on montre un exemple de ces résumés. Le sujet et le narratif sont aussi présentés. Nous pouvons observer la pertinence des phrases retenues bien que les problèmes de cohérence et de non résolution d'anaphores inhérentes aux approches par extraction sont aussi évidents.

Document : d301i
Titre : International Organized Crime.
Narratif : Identify and describe types of organized crime that crosses borders or involves more than one country. Name the countries involved. Also identify the perpetrators involved with each type of crime, including both individuals and organizations if possible.
Résumé : Colombian drug cartels and the Mafia are building a cocaine empire in Western European countries, which have replaced the United States as the biggest market for the drug, Bogota's El Tiempo newspaper reported, citing a joint study by the international police organization Interpol and the Colombian intelligence police. The drugs organisation used intricate methods - including bank accounts, couriers and ships as well as dummy and real companies in many countries - to smuggle cocaine from South America to Europe. Nigerian drug barons have also flourished in a country where it is easy to launder drug money through the free-wheeling banking sector and where the recently established National Drug Law Enforcement Agency is inadequately funded. Stories of shoot-outs in the country's second largest airport, and tunnels dug for drug traffickers between the and Mexico, dominated coverage of the country, and fuelled concerns over the North American Free Trade Agreement. Modelled on the Drug Enforcement Agency, the institute will lead investigations of the drug trade, arms trafficking and money laundering, create an intelligence system and run operations to intercept drug traffickers and shipments. He cited the recent example of Russian organised gangs working with the Italian Mafia to funnel a big drug consignment into the 'We have to go where the crime is. . and not do what we did with respect to Italian organised crime and that was wait for 50 years before we got involved,' Mr Freeh said .

FIGURE 4.8 – *Résumé guidé généré par* ENERTEX *pour un des sujets du DUC 2005.*

4.3.4 Effet du TF.IDF sur le calcul de l'énergie textuelle

Nous avons décrit dans la section 2.3.2 le TF.IDF comme une mesure de l'importance des mots dans les documents d'un corpus. Cette importance croit avec la fréquence des mots dans le document et diminue inversement à leur fréquence dans le corpus.

Le TF.IDF est devenu le moyen de pondération le plus populaire dans le domaine du TAL, en conséquence il était intéressant d'observer ses effets sur le calcul de l'énergie textuelle. Nous avons utilisé les deux versions apparaissant dans le tableau 2.2 afin de pondérer les valeurs fréquentielles de la matrice terme-segment S avant de faire le calcul de l'énergie. Les résultats montrent qu'en raison du double produit matriciel impliqué dans le calcul de l'énergie, il se produit un effet de bord. Toute pondération sur les valeurs de la matrice terme-segment S favorise les cas extrêmes (phrases trop longues ou trop courtes ; termes très fréquents ou rares). Ce biais induit fait diminuer la pertinence des phrases rangées en premières places.

Nous avons utilisé les versions de TF et IDF du tableau 2.2. La tâche a été le résumé multidocument guidé uniquement sur cinq des clusters du corpus DUC'07. Les scores ROUGE-2 et SU4 obtenus par ENERTEX avec et sans normalisation sont affichés dans la figure 4.9. Les scores sont plus élevés sans normalisation.

Cette expérience est indicative. Elle fait en sorte de comparer l'effet des différentes implémentations de la matrice terme-segment sur le calcul de l'énergie textuel. Puisqu'elle est faite sur un petit sous-ensemble du corpus DUC'07, les résultats ne peuvent pas être comparés directement avec ceux de la section 4.3.3.

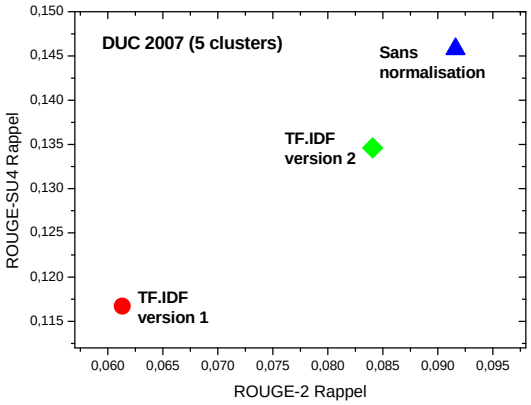

FIGURE 4.9 – *L'effet du TF.IDF sur la performance du système* ENERTEX *dans la tâche de résumé multidocument guidé. Les résultats correspondent à un petit sou-ensemble de cinq des clusters du corpus DUC'07.*

Les expériences semblent indiquer que la pondération TF.IDF ne s'adapte pas au calcul de l'énergie textuelle. C'est pourquoi, nous n'avons fait aucun type de normalisation des éléments de la matrice S. Nous avons laissé les termes manifester librement leurs interactions dans le « matériau textuel » par leurs fréquences, sauf dans les cas particulièrs (changement d'échelle ou l'introduction des impuretés, section C) qui nous ont conduit à faire autrement.

4.4 Changement d'échelle et dopage du réseau textuel

En physique des matériaux, il est bien connu qu'un changement d'échelle peut entraîner des modifications dans le comportement d'un système, même si les composants primaires restent les mêmes. Des matériaux à l'état massif peuvent exhiber un ensemble de propriétés (électriques, magnétiques, optiques, etc.) qui sont considérablement modifiées quand on les mesure sur un petit échantillon nanométrique.

D'un autre côté, il est courant de trouver des impuretés dans les matériaux. En fait, un solide n'est jamais complètement pur. Les impuretés sont des éléments étrangers à la structure d'un matériau qui peuvent aussi modifier profondément ses propriétés. C'est pourquoi, plusieurs techniques ont été développées pour réduire la teneur en impuretés. En revanche, il existe aussi le dopage, qui correspond à l'adjonction volontaire et dosée d'éléments déterminés avec le but d'obtenir des effets spécifiques. En général, quand la quantité des éléments étrangers dans un matériau qui sert de réseau hôte est non négligeable ($> 0,5\%$), les conséquences peuvent être importantes (Marucco, 2004).

Dans le domaine du TAL, comment peut-on assimiler ces concepts ? Qu'est-ce qui se passe avec l'analyse de textes quand l'on change d'échelle ? Ou quand on introduit des termes qui ne font pas partie du document d'origine ? Ce sont les problématiques que nous abordons dans l'annexe C où nous présentons l'ensemble des expériences réalisées en collaboration avec le *College of Information Sciences* [20] de l'Université de Drexel à Philadelphie. Ce travail fait en sorte d'utiliser des annotations sémantiques (impuretés) pour aider un chercheur ou spécialiste à accéder à une catégorie particulière d'information dans les textes scientifiques. L'analyse a été fait au niveau des *abstracts* et non des phrases.

L'algorithme prend en entrée un corpus d'*abstracts* scientifiques et permet de guider le classement et l'extraction d'information par des requêtes contenant de termes conceptuels concernant le sujet scientifique et les étiquettes introduites. Nous avons mis en place une nouvelle combinaison de fonctions de pondération dans le système ENERTEX pour résoudre ces tâches. Bien que, en raison du manque de modèles de référence, l'évaluation de cette partie aie été faite manuellement, les expériences ont produit des résultats intéressants.

4.5 Conclusions

Dans ce chapitre nous avons présenté le système ENERTEX basé sur la mesure d'énergie textuelle. La première problématique abordée a été le résumé monodocument générique. Nous avons confronté notre approche aux systèmes de la littérature. Le système ENERTEX a obtenu de bonnes performances sur de tests en plusieurs langues et indépendants de la thématique. Les travaux que nous avons présentés ont donné lieu à trois publications. La méthode de résumé par extraction basé sur l'énergie textuelle et les

20. http://www.ischool.drexel.edu

résultats sur le résumé monodocument en français et anglais (Fernández et al., 2007a) ; la confrontation d'ENERTEX contre une approche cognitive (Fernández et al., 2008b) et la construction d'un système hybride combinant des méthodes linguistiques et numériques dont ENERTEX (da Cunha et al., 2007).

Nous avons montré une modification d'ENERTEX qui consiste à mettre un champ externe en rapport avec un corpus multidocument. Ceci a permis de générer des résumés guidés par les besoins de l'utilisateur. Une stratégie de réduction de la redondance, basée sur la comparaison de valeurs d'énergie, a été mise en place. L'évaluation sur les corpus DUC 2005-07 indique que notre système est aussi performant que les participants ayant obtenu les meilleures places. Cette contribution a donne lieu à deux publications, (Fernández et al., 2007b) et (Fernández et al., 2008a).

Finalement, nous avons effectué des expériences exploratoires de recherche d'information orientée en travaillant au niveau des résumés. Des impuretés sous forme d'étiquettes sémantiques ont été introduites dans le réseau textuel. Il a fallu adapter ENERTEX à cette tâche avec l'action combinée de deux fonctions qui agissent sur les termes rares propres aux résumés et sur les termes trop communs, comme les étiquettes introduites. Ce travail a été publié dans (Ibekwe-SanJuan et al., 2008a).

Notre travail montre comment le concept d'énergie de la Physique statistique permet d'atteindre les performances des principales approches numériques du TAL sur de multiples applications.

Chapitre 5

Les spectres des phrases et l'échange discriminatoire

Sommaire

5.1 Introduction

Dans ce chapitre nous présentons une approche non évidente consistant à représenter l'énergie textuelle des phrases comme des spectres. Un test statistique, en l'occurrence celui de Kendall, peut indiquer si ces signaux sont semblables entre eux ou non. Ceci permet de réaliser une détection de frontières thématiques dans un document. L'introduction d'une longueur de corrélation modifie les spectres des phrases et ainsi le test de Kendall peut mieux les identifier.

D'un autre côté, nous présentons une stratégie de classification de documents basée sur la capacité des matrices d'échange entre mots pour caractériser les différentes catégories.

5.2 La segmentation thématique

Plusieurs stratégies ont été développées pour segmenter thématiquement un texte. Elles consistent à identifier dans un texte des fragments avec une certaine cohérence et de leur associer une étiquette thématique (Sabah, 2006). Parmi elles, on trouve PLSA (*Probabilistic Latent Semantic Analysis*) (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) ; d'autres utilisant une classification des termes (Caillet et al., 2004; Chuang et Chien, 2004) ou sur des chaînes lexicales (Sitbon et Bellot, 2005). Plus récemment, (Ferret, 2007) a proposé l'identification préalable des sujets présents dans le document comme stratégie pour améliorer la détection de ruptures thématiques. L'identification de sujets est réalisée par une analyse contextuelle basée sur la co-occurrence de mots. L'idée est que si deux segments apparaissent dans le même contexte, même s'ils n'ont pas une forte cohésion lexicale entre eux, ils appartiennent au même sujet sans rupture thématique.

De façon originale, nous avons utilisé la matrice d'énergie textuelle E (équation 3.7) pour idéntifier les frontières thématiques. Dans la section 3.3.1 nous avons montré que chaque ligne de cette matrice produit un spectre qui représente l'interaction d'une phrase avec toutes les autres (figure 3.3). Dans cette section nous montrons comment la comparaison entre spectres permet de détecter les ruptures thématiques dans les documents. Ce choix s'adapte au fait d'avoir de nouvelles thématiques et de rester indépendant vis-à-vis de la langue des documents.

5.3 Le spectre énergétique : une signature thématique

Nous montrons en figure 5.1 l'énergie d'interaction entre quelques phrases du texte *2-mélanges* composé de deux thématiques (le texte au complet se trouve dans l'annexe A.4). On peut constater une similarité entre les courbes de l'une (gras) et de l'autre thématique (pointillé). En effet, le changement d'allure des courbes entre les phrases 14-15 correspond au changement de thématique. Nous proposons d'utiliser le spectre d'une phrase comme sa signature thématique et la comparaison des spectres comme technique pour détecter la transition entre deux thématiques.

5.3.1 Comparaison de spectres par le test de Kendall

Afin de comparer les courbes objectivement, nous avons utilisé le coefficient de concordance τ de Kendall et le calcul de sa $p-$valeur. Ils permettent de définir un

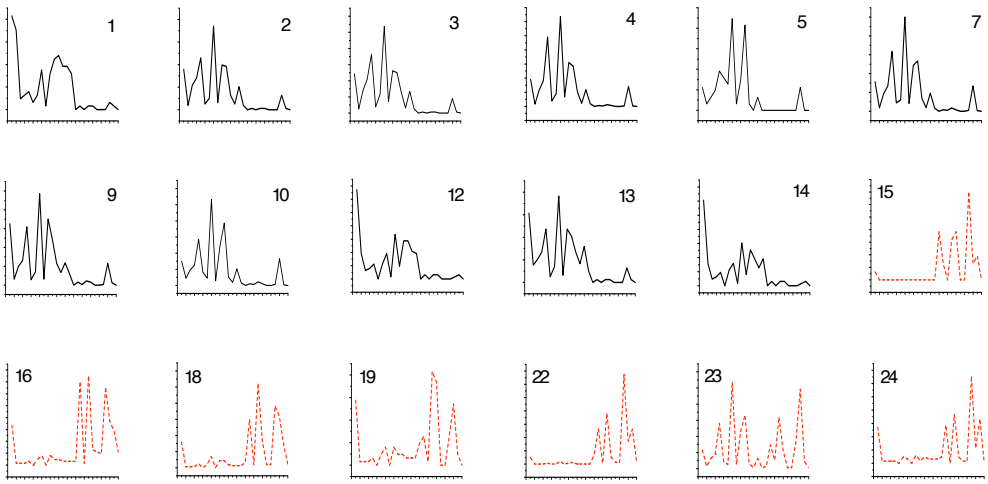

FIGURE 5.1 – *Énergie textuelle du texte à deux thématiques « 2-mélanges ». En trait continu l'énergie des phrases de la 1ère thématique, en pointillé celle de la 2ème. Le changement d'allure des courbes entre les phrases 14-15 correspond à un changement de thématique. L'axe horizontal indique le numéro de phrase dans l'ordre du document. L'axe vertical, l'énergie textuelle de la phrase affichée par rapport aux autres.*

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. Une description détaillée du test de Kendall se trouve dans les annexes, section D. Nous présentons ici notre adaptation. Le protocole est le suivant :

1. On émet les hypothèses qui suivent :
   H_0 : la phrase $\mu + 1$ marque une rupture avec μ ;
   H_1 : la phrase $\mu + 1$ appartient à la même thématique que la phrase précédente μ.
2. On calcule le coefficient de concordance τ de Kendall entre les deux classements induits par les phrases μ et $\mu + 1$ sur les autres phrases. τ vaut 1 en cas d'accord total entre les deux classements et -1 dans le cas où un classement est l'inverse de l'autre. Pour notre test, nous ne utilisons pas directement les valeurs de τ mais seulement les p-valeurs associées. Comme dans tous les tests statistiques, il est nécessaire de vérifier si le score obtenu par une méthode est significativement différent d'un score au hasard sur les mêmes données. La p-valeur donne la probabilité que deux classements soient indépendants, c'est-à-dire, qui ils aient été faits complètement au hasard. Lorsque le texte a plus de 10 phrases, le τ induit par des classements aléatoires suit une loi normale, ce qui permet l'estimation de sa probabilité sans faire des hypothèses sur le vocabulaire utilisé dans les phrases.
3. Plus la p-valeur est petite, moins on peut supposer l'indépendance et une conclusion de corrélation est produite. Pour savoir si la p-valeur est suffisement petite, on la compare avec un seuil (niveau de signification) α fixé *à priori*. Si $p < \alpha$ on rejette H_0 et l'on adopte l'hypothèse alternative H_1 avec un risque p de se tromper. Il est important de préciser que le choix du niveau de signification auquel on rejet H_0 est arbitraire. Conventionnellement les valeurs de α utilisés sont $0,05$, $0,01$ et $0,001$ (Saporta, 1990). Nous avons utilisé un seuil beaucoup plus grand ($\alpha = 0,1$)

qui nous a permis de détecter les bonnes frontières. Si on imagine un texte totalement aléatoire, alors notre méthode trouve presque autant thématiques que de phrases différentes.

Nous avons calculé le coefficient τ de Kendall et sa p-valeur avec le logiciel *R* pour calculs statistiques (R Development Core Team, 2006) et le module Concord [1].

Les figures 5.2 et 5.3 montrent la détection des frontières pour les textes en français à deux et trois thématiques [2]. Les véritables frontières sont indiquées en pointillée. Ce protocole de test 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 bien d'être commenté : si on regarde sur la figure 5.1 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.

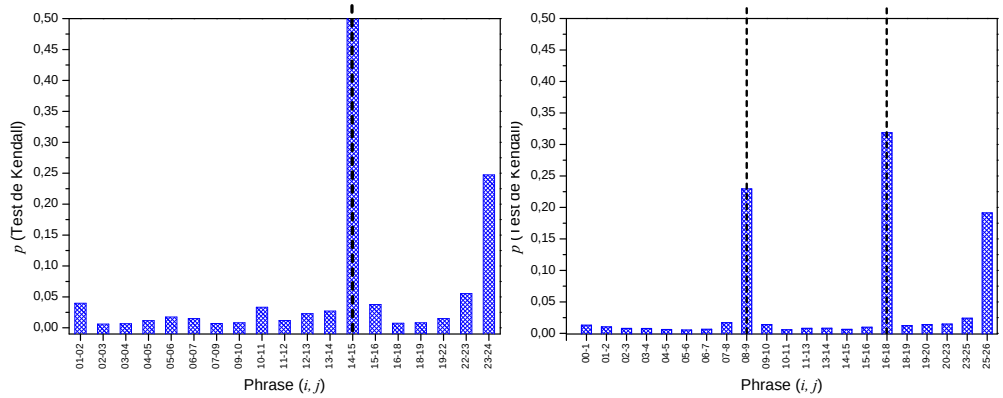

FIGURE 5.2 – *Détection des frontières pour le texte 2-mélanges (deux thématiques, à gauche) et 3-mélanges (trois thématiques, à droite). Une p-value* $> 0,1$ *indique la présence d'une rupture thématique.*

Nous montrons en figure 5.3 à gauche la détection des frontières pour le texte en français *physique-climat-chanel* à trois thématiques. Le test Kendall a détecté deux frontières entre les phrases 5-6 et 12-15, qui correspondent aux frontières effectives. Pour le texte en anglais à deux thématiques *québec-lewinsky*, figure 5.3 à droite, le test a trouvé une frontière entre les segments 44-45 qui correspond à la vraie frontière.

1. http ://www.R-project.org
2. Les textes utilisés dans cette section sont des documents composites extraits de Wikipédia.

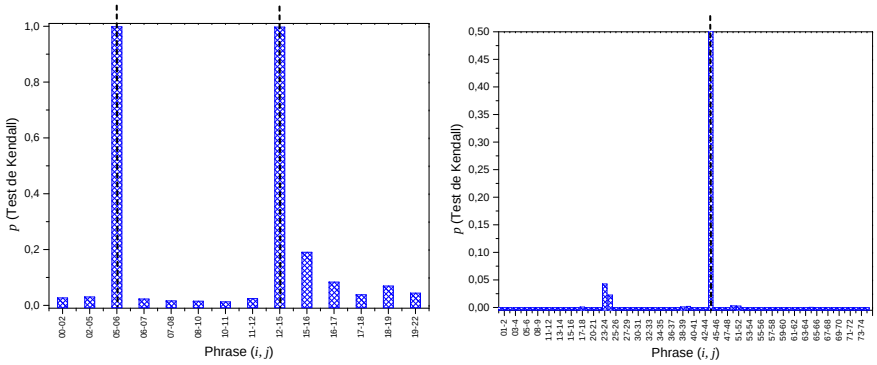

FIGURE 5.3 – *Détection des frontières pour le texte en français à 3 thématiques physique-climat-chanel à gauche et en anglais québec-lewinsky à droite.*

5.3.2 Les premières évaluations

La précision et le rappel

Les mesures classiques qui donnent un aperçu global de l'erreur commise par un système sont la précision (*precision*) et le rappel (*recall*). Ces indicateurs permettent de mesurer la quantité d'information correcte retrouvée parmi l'information retournée. Voici les définitions de ces outils dans le cadre de la détection de ruptures thématiques :
- la précision est le rapport entre le nombre de frontières correctes trouvées et le nombre total de frontières trouvées ;
- le rappel est le rapport entre le nombre de frontières correctes trouvées et le nombre total de frontières existantes.

Ainsi, si l'on note S l'ensemble des phrases qu'un système automatique considère comme des frontières, V l'ensemble des phrases qui sont effectivement des frontières thématiques, P et R respectivement la précision et le rappel du système, on aura :

$$P = \frac{|S \cap V|}{|S|} \tag{5.1}$$

$$R = \frac{|S \cap V|}{|V|} \tag{5.2}$$

Une précision de 100% signifie que toutes les frontières trouvées sont véritables. Un rappel de 100% que tous les frontières existantes ont été trouvées. La F-mesure fait une synthèse entre rappel et précision, en favorisant les systèmes dont les mesures de précision et rappel sont voisines (voir équation 5.3). Le coefficient β permet de privilégier soit les systèmes qui ont une meilleure précision, soit ceux qui ont un meilleur rappel (Poibeau, 2003). En général, β est mis à 1, afin de ne pas privilégier ni la précision ni le rappel.

$$F\text{-mesure} = \frac{(\beta^2 + 1) * P * R}{(\beta^2 * P) + R} \tag{5.3}$$

Ainsi, une première idée de la performance de notre système de détection de frontières thématiques est donnée par la F-mesure (en considérant $\beta = 1$) :

$$F\text{-mesure} = \frac{2 \times \text{Nb\_frontières\_correctes\_trouvées}}{\text{Nb\_total\_frontières\_trouvées} + \text{Nb\_frontières\_existantes}} \tag{5.4}$$

Les résultats sont :

Texte	Nb. de thématiques	*F*-mesure
2-mélanges	2	0,66
3-mélanges	3	0,66
physique-climat-chanel	3	0,80
québec-lewinsky	2	1,00

TABLE 5.1 – *F-mesure pour le détection de frontières thématiques. Textes en anglais et en français.*

La mesure WINDIFF et les fausses frontières

Dans cette section, nous présentons une comparaison entre notre système de segmentation thématique et autres approches existantes. Nous choisissons LCSEG (Galley et al., 2003) et LIA_SEG (Sitbon et Bellot, 2005) qui utilisent tous les deux des chaînes lexicales. Une chaîne lexicale relie les termes suffisamment proches dans le texte, éloignés d'une distance inférieure à une valeur fixe appelée hiatus. Classiquement, une chaîne est rompue quand elle dépasse la valeur du hiatus.

LCSEG et LIA_SEG ont été testés sur un corpus en français construit à partir d'articles du journal LE MONDE [3] (Sitbon et Bellot, 2005). Ce corpus est composé de trois ensembles de 100 documents où chacun correspond à la taille moyenne des segments pré-définie (de 3 à 5, de 3 à 11 et de 9 à 11 phrases par segment). Un document est composé de 10 segments extraits d'articles thématiquement différents tirés au hasard. Les performances des algorithmes ont été évaluée en (Sitbon et Bellot, 2005) avec la mesure WINDIFF (Pevzner et Hearst, 2002), utilisée en segmentation thématique. Cette fonction calcule la différence entre les frontières véritables et celles trouvées automatiquement dans une fenêtre glissante : plus la valeur est petite, plus le système est performant.

Nous avons appliqué notre strategie de segmentation, basée sur la comparaison de spectres par le test de Kendall, sur ce même corpus et nous avons évalué la pertinence des frontières trouvées au moyen de WINDIFF. Nous comparons dans le tableau 5.2 les scores obtenus par notre approche avec ceux rapportés par (Sitbon et Bellot, 2005). On observe que notre méthode obtient des performances comparables aux autres systèmes mais en utilisant beaucoup moins de paramètres, en particulier nous ne faisons aucune supposition sur le nombre de thématiques à détecter. LIA_SEG depend d'un paramètre qui donne lieu à différentes performances, d'où la plage de valeurs affichées.

3. http://www.lemonde.fr

Les ruptures entre segments thématiquement différents 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 de Kendall s'avère incapable de distinguer. C'est le cas du spectre de la phrase 23, figure 5.1. Pour diminuer cet effet nous avons proposé deux types de solutions : 1) étendre le test de Kendall à une fenêtre glissante ; 2) contrôler le bruit des spectres au moyen d'une longueur de correlation.

5.3.3 Kendall en fenêtre

Nous présentons une variation du test de Kendall afin de réduire la détection de fausses frontières. Il s'agit de l'utilisation d'une fenêtre glissante. À mesure qu'elle se déplace, la phrase centrale est comparée aux autres par le test τ de Kendall. Si un bord est détecté, on abandonne la fenêtre et on recommence sur un nouvel ensemble de phrases. Comme dans le cas précédent, le test τ-Kendall suppose que deux phrases μ et ν appartient aux thématiques différentes et il calcule la p-valeur associée. Si $p < \alpha$ les phrases μ et ν appartienent à la même thématique, aux thématiques différentes autrement. Le seuil α de la p-valeur est maintenant fixé à $0,01$. La diminution du niveau α permet de satisfaire simultanément le triple test de manière significative.

Nous illustrons cet effet dans la figure 5.4 pour une fenêtre de taille 7. Une phrase représente une rupture si :

1. elle n'appartient pas à la même thématique qu'au moins deux des trois phrases précédentes ; et
2. son spectre est similaire à ceux d'au moins deux des trois phrases suivantes.

Nous montrons dans le tableau 5.3 que cette stratégie a permis une meilleure détection des ruptures par rapport aux résultats précédentes (tableau 5.2). Mais nous pensons qu'on peut faire mieux en réduisant le bruit présent dans les spectres d'énergie.

5.3.4 Filtrage des spectres : distance et longueur de corrélation

Dans la nature, les objets sont soumis à toutes sortes de forces qui s'exercent à distance et en général, de telles interactions ont été décrites par des lois de puissance ou exponentielles. De façon similaire, nous avons observé empiriquement que l'énergie textuelle des phrases diminue exponentiellement avec la distance à un maximum.

Taille du segment en phrases	**LCSEG**	**LIA_SEG**	**ENERTEX** Kendall
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**

TABLE 5.2 – *Mesure* WINDIFF *pour les systèmes* LCSEG, LIA_SEG *et* ENERTEX *(segments de différentes tailles).*

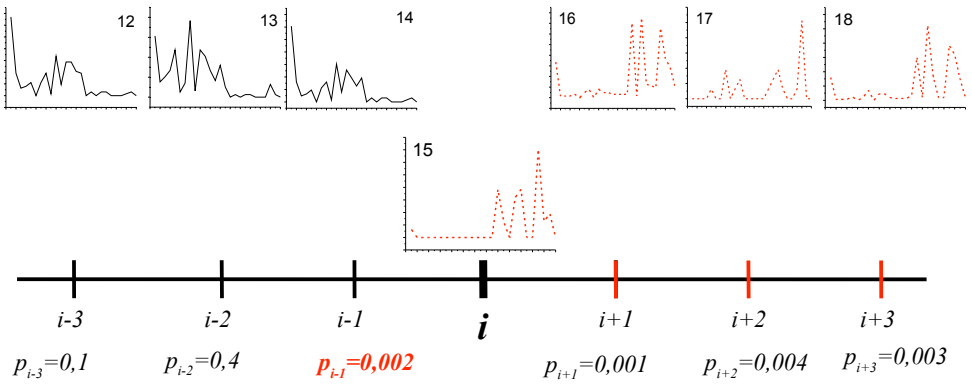

FIGURE 5.4 – *Test τ-Kendall en fenêtre ; $p_{i\pm k}$ = probabilité de se tromper avec l'hypothèse que $i \pm k$ et i appartient au même segment thématique (c'est à dire d'accepter H_1 quand H_0 est vraie) ; pred = nombre de phrases parmi les 3 qui précèdent à i qui ne sont pas concordants avec elle ($p > 0,01$), et succ = nombre de phrases parmi les 3 suivants à i qui sont concordantes avec elle ($p < 0,01$). Si $pred(i) > 2/3$ et $succ(i) > 2/3$, alors i est une frontière.*

Taille du segment en phrases	**ENERTEX** Kendall en fenêtre
9-11	0,4134
3-11	0,4264
3-5	**0,4140**

TABLE 5.3 – *Mesure* WINDIFF *pour les systèmes* ENERTEX *avec le test de Kendall en fenêtre (segments de différentes tailles) ; plus le coefficient est petit meilleure est la segmentation.*

Effectivement, les spectres qui expriment correctement leur appartenance à une thématique ont en général 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 maximum, les autres interactions diminuent rapidement jusqu'à la fin de la thématique. Comme nous le montrons avec la figure 5.5, cette décroissance peut être modélisée par la fonction exponentielle $exp(-r/\xi)$ où r est la distance entre une phrase μ et la phrase voisine qui présente la plus haute interaction avec elle. ξ est un paramètre que nous utilisons pour caractériser le bruit des spectres et qui peut être assimilé, en termes physiques, à une longueur de corrélation. La longueur de corrélation est une mesure de la plage sur laquelle les fluctuations dans une région de l'espace sont en corrélation avec ceux dans une autre région (Newman et Barkema, 1999). Cela signifie qu'on peut assimiler à du bruit les interactions entre phrases au delà de la distance ξ.

Notre stratégie pour la détection de fausses frontières porte directement sur la modification de l'allure de courbes des spectres : le filtrage par le paramètre de bruit ξ. La figure 5.6 montre le filtrage induit dans les spectres pour les phrases 10 (thématiquement bien définie) et 23 (difficile à classer en fonction de ses pics) de la figure 5.1. Leurs valeurs ont été multipliées par le facteur $exp(-r/\xi)$ pour différentes valeurs du paramètre ξ. Nous avons diminué ξ progressivement afin d'analyser l'évolution du chevauche-

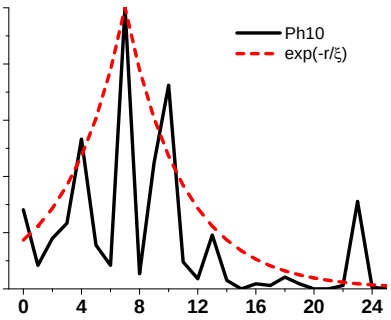

FIGURE 5.5 – *Forme générale d'un spectre d'énergie textuelle, bien défini par rapport à sa thématique et la fonction $exp(-r/\xi)$ pour $\xi = 4$.*

ment des courbes. Cette diminution lisse les courbes de façon efficace : à $\xi \approx 8$ le bruit de la courbe 23 est réduit et un classement correct est obtenu. Le spectre de la phrase 10 a aussi été lissé sans perte d'information. La figure 5.7 montre l'application du filtrage

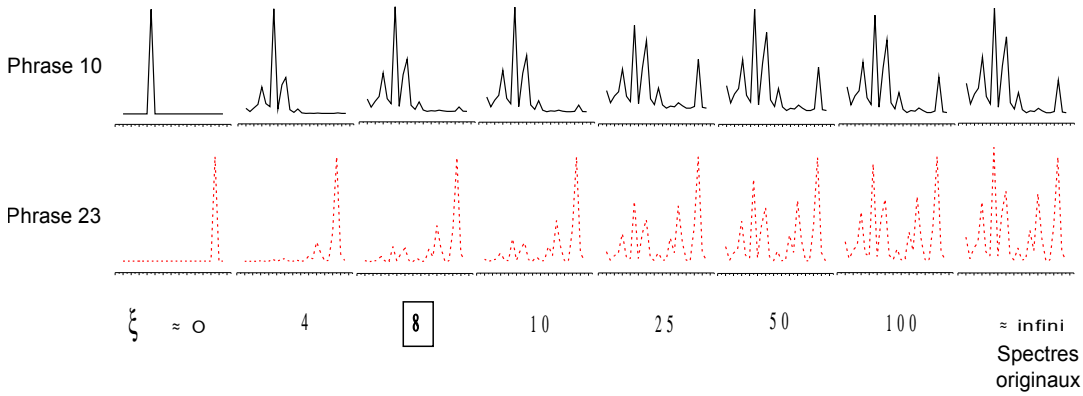

FIGURE 5.6 – *Filtrage des spectres par $exp(-r/\xi)$. En trait continu le spectre d'une phrase thématiquement bien définie et en pointillé celui d'une phrase inclassable. r est la distance au maximum et ξ la longueur de corrélation.*

sur l'ensemble de phrases de la figure 5.1.

Nous avons fait l'hypothèse qu'avec ce filtrage, le test de concordance de Kendall identifiera mieux les phrases selon leur thématique. Mais, comment estimer la valeur de ξ ? Dépend-t-elle de la taille ou du type de document ? Nous avons réalisé quelques expériences sur des corpus multi-thématiques en anglais, espagnol et français.

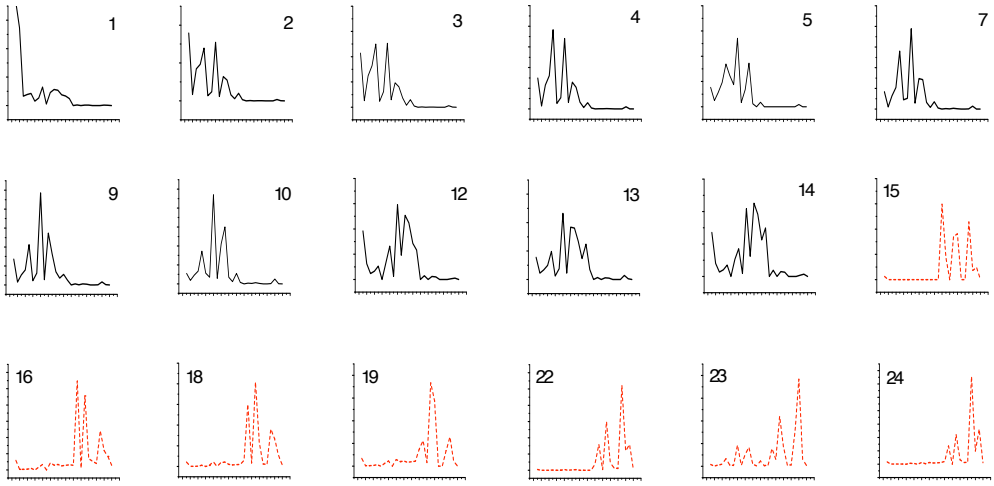

FIGURE 5.7 – *Spectres du texte à deux thématiques 2-mélanges lissés à $\xi = 8$.*

5.3.5 Expériences et résultats

Corpus multithématique et en plusieurs langues

Pour comparer les résultats issus du modèle avec filtrage aux résultats précédents (tableau 5.2 et 5.3), nous avons utilisé le même corpus en français issu du journal LE MONDE en ajoutant une taille de segment. Nous incluons aussi des corpus en anglais et espagnol construits à partir d'articles journalistiques du BROWN CORPUS [4] et du journal mexicain LA JORNADA [5] suivant le même protocole. Chaque corpus comporte quatre ensembles de 100 documents qui correspondent à une taille de segments fixée (de 3 à 5, de 3 à 11, de 6 à 8 et de 9 à 11 phrases par segment). Un document est constitué de 10 segments extraits d'articles thématiquement différents tirés au hasard.

Détermination des longueurs de corrélation optimales

Pour chaque document on a calculé l'énergie textuelle à différentes longueurs de corrélation : $\xi = 1, \cdots, 180$. Les spectres ont été comparés par le test de Kendall et les frontières détectées ont été mesurées par WINDIFF (WD). Nous rappelons que plus la valeur WD est basse, meilleure est la segmentation. La figure 5.8 montre les résultats sur un ensemble de 100 documents en français et une taille de segments allant de 6 à 8 phrases. En trait continu on observe l'évolution de la valeur moyenne de WD et en pointillé le nombre de frontières trouvées. On observe qu'à longueur de corrélations 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 longueur de corrélation les courbes voisines se ressemblent de plus en plus et le nombre de frontières augmente. Nous avons retenu la valeur $\xi = 80$ qui maximise à la fois le

4. http://khnt.aksis.uib.no/icame/manuals/brown
5. http://www.jornada.unam.mx

nombre de frontières trouvées tout en minimisant la valeur de WD. Notons que la valeur de ξ est ici largement augmentée par la taille du corpus beaucoup plus importante que dans l'exemple analysé précédement.

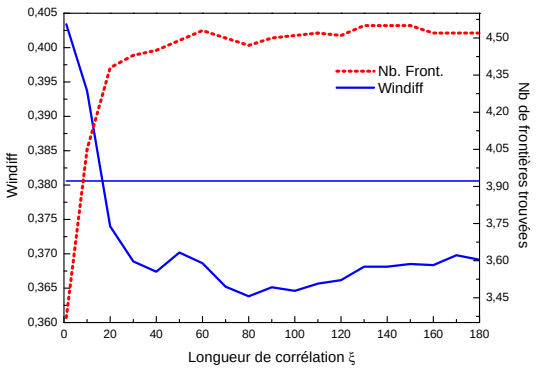

FIGURE 5.8 – *Évolution de WD et du nombre de frontières en fonction de ξ. La ligne horizontale représente la valeur de WD à longueur de correlation infinie. Taille des segments entre 6 et 8 phrases pour le corpus en français.*

On observe au tableau 5.4 que la valeur de ξ pour la meilleure segmentation dépend de la longueur du document. Plus la taille du segment est grande (plus le document est long) plus la valeur ξ est élevée. Sous l'hypothèse que les trois langues occidentales étudiées possèdent une certaine proximité, nous avons décidé de garder la même longueur de corrélation calculée en français dans les textes en espagnol et en anglais. Les résultats du tableau au tableau 5.4 confirment cette hypothèse.

D'après les valeurs pour la mesure WINDIFF (colonnes WD), nous constatons une amélioration par rapport à ceux du tableau 5.2. Ainsi, nos stratégies combinées (Kendall en fenêtre et filtrage par longueur de corrélation) ont permis d'améliorer nos résultats précédents.

La mesure δ-Front

(Pevzner et 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 (car 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 et Bellot, 2004). De plus, nous avons trouvé qu'une même valeur de WD peut correspondre à des segmentations différentes du document. Compte tenu de ces faiblesses, nous avons proposé δ-Front, une nouvelle mesure d'évaluation pour la segmentation thématique.

δ-Front calcule la distance euclidienne $d(\bullet)$ (équation 5.6) 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 5.9). La normalisation est faite avec le vecteur nul **C** : ne contenant aucune frontière sauf les extrêmes. Plus la valeur δ-Front est basse, meilleure est la segmentation.

$$\delta\text{-Front}(\mathbf{A},\mathbf{B}) = \frac{d(\mathbf{A},\mathbf{B})}{d(\mathbf{A},\mathbf{C})} \tag{5.5}$$

$$d(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{P}(a_i - b_i)^2} \tag{5.6}$$

a_i et b_i sont les composantes des vecteurs A et B respectivement.

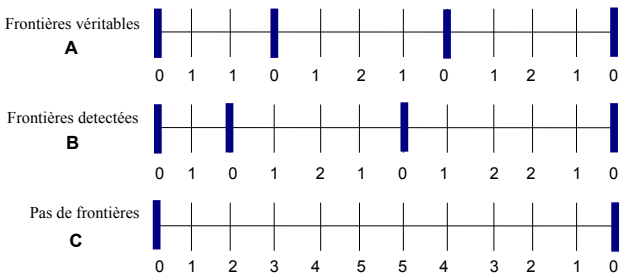

FIGURE 5.9 – *La mesure δ-Front entre A et B, calculée selon l'équation (5.5), vaut 0,3307.*

Nous observons dans la colonne δ-Front du tableau 5.4 que les deux mesures ne sont pas toujours en accord. En effet, 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 citées 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 trois langues. Cette stabilité découle du calcul d'interactions des mots combiné au processus de comparaison de segments. (Ferret, 2007) constate en partie cet effet. Nous avons utilisé les comparaisons des spectres énergéti-

Taille du segment	ξ	**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	≈6/9
6-8	80	0,3638	0,1957	0,3601	0,2031	0,3804	0,1640	≈5/9
3-11	40	0,3885	0,1974	0,3646	0,2043	0,3709	0,1634	≈5/9
3-5	20	0,3851	0,4540	0,3598	0,3257	0,3786	0,3864	≈3/9

TABLE 5.4 – *Mesures WD et δ-Front pour des corpus en 3 langues et segments de tailles variables.*

ques des phrases pour détecter les changements thématiques dans les documents. L'introduction d'une longueur de correlation artificielle a permis d'effectuer un filtrage des courbes. Cette stratégie a facilité leur comparaison par le test de Kendall.

5.4 La matrice d'échange et la classification documentaire

Dans cette section nous abordons la classification automatique des documents. Nous avons construit un algorithme basé sur les propriétés des coefficients d'échange entre les unités d'un système magnétique.

Entre chaque paire d'atomes i et j contenus dans un matériau magnétique, existe une interaction $J_{i,j}$, dite d'échange, qui a été décrite en section 2.4.2. Expérimentalement, les valeurs de telles interactions sont déterminées à l'aide des techniques comme la diffraction de neutrons [6] (*neutron scattering*), et des modèles théoriques [7] (Chaboussant et al., 2004; Onishi et al., 2003). Par exemple, dans l'alliage intermétallique $DyFe_2$, l'échange entre deux atomes de Fe (fer) est de $J_{Fe,Fe} \approx +120$; entre deux atomes de Dy (dysprosium) $J_{Dy,Dy} \approx +2$; et entre un atome de Fe et un autre de Dy le couplage d'échange vaut $J_{Fe,Dy} \approx -15$ (Dumesnil et al., 2000) [8]. Mais la valeur d'échange entre un couple spécifique d'atomes n'est pas constante. Elle peut varier d'un alliage à l'autre du fait que l'environnement n'est pas le même. Par exemple, si dans l'alliage précédent, on substitue les atomes de Dy pour les atomes de Y (Ittrium) pour produire l'alliage YFe_2, le couplage fer-fer change à $J_{fer,fer} \approx +98$ (Dumesnil et al., 2000). Ainsi, d'une certaine façon, la valeur de l'échange pourrait donner des pistes sur le type du matériau auquel le couple d'atomes appartient (figure 5.10). Cette observation est à la base de notre méthode de classification de documents.

Nous représentons le texte comme un alliage de mots. La règle de Hebb, utilisée en section 3.3 pour le calcul de l'énergie textuelle, produit la matrice d'échange J où chaque élément $J_{i,j}$ représente le couplage d'échange entre deux mots i et j. Nous proposons d'utiliser ces matrices pour déterminer la classe des documents. Comme dans le cas des matériaux, un même couple de mots i et j aura une valeur d'échange dans un texte traitant le sujet A ($J^A_{i,j}$) et une autre valeur dans un autre texte traitant un sujet différent B ($J^B_{i,j}$). Nous comptons profiter de cette différence.

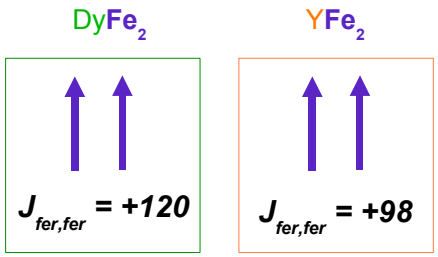

FIGURE 5.10 – *Interaction d'échange entre atomes de fer dans deux matériaux différents.*

6. Les neutrons, étant sensibles au magnétisme de la matière, sont utilisés pour étudier les propriétés et la structure magnétique des matériaux.
7. Un exemple est le modèle de double échange proposé par (Zener, 1951) pour expliquer les propriétés électriques et magnétiques de certains alliages de manganèse.
8. Les unités sont $1 \times 10^{-8} ergs/cm$.

5.4.1 La classification automatique de documents

Les méthodes de classification automatique de documents ont pour but de représenter les proximités entre les textes par des regroupements ou classes (Lebart et Salem, 1994). Les approches peuvent se différencier selon la méthode ou par les unités textuelles utilisées (mots, lemmes, *n*-grammes, étiquettes, longueurs de phrase, etc.). Il existe trois méthodes classiques de classification : les arbres de décisions sémantiques (Kuhn et De Mori, 1995), les algorithmes de boosting (Freund et Schapire, 1996) et les machines à support vectoriel (SVM) (Cortes et Vapnik, 1995). Nous décrivons quelques exemples de leurs applications.

LIA_SCT (Bechet et al., 2000) suit le principe d'un arbre de décision où chaque nœud contient une question sur la structure du texte. Les réponses, basées sur des règles statistiques apprises, produisent une subdivision dans les nœuds fils jusqu'aux feuilles de l'arbre. BoosTexter (Schapire et Singer, 2000) utilise l'algorithme de *boosting*. L'idée est de combiner des hypothèses faibles qui servent à pondérer et re-pondérer itérativement un ensemble de textes d'entraînement. La sortie est un ensemble de règles de classification plus précises afin de trier les documents à classe inconnue. SVMTorch (Collobert et al., 2001) applique des SVM à la résolution des problèmes de classification textuelle à grande échelle. Ce système est entraîné en résolvant un problème d'optimisation quadratique.

5.4.2 Le DÉfi de Fouilles de Texte (DEFT)

DEFT [9] propose depuis 2005 des tâches qui concernent l'analyse automatique de textes en langue française. Il s'agit d'une conférence d'évaluation qui permet de confronter, sur un même corpus, des méthodes d'équipes de recherche différentes. En 2008 ce défi a concerné la classification de documents en genre et en thème. Nous avons participé en proposant une méthode basée sur le calcul des couplages d'échange entre mots.

Description des tâches

DEFT'08 [10] a proposé deux tâches. **Tâche 1 :** classification en deux genres possibles (Le Monde ou Wikipédia) et en quatre thèmes (économie, art, télévision, sport). **Tâche 2 :** classification en cinq thèmes (société, actualité française, international, sciences, littérature). Le protocole de participation a été le suivant. Dans un premier temps, les organisateurs ont distribué un ensemble de documents avec étiquettes de genre et de thème. Cette première collection de documents constitue le corpus d'apprentissage et permet aux systèmes de s'entraîner. Puis, un deuxième ensemble de documents à genre et thème inconnus a été envoyé aux participants. Ce dernier constitue l'ensemble de test. Le corpus de la tâche 1 est composé d'un total de 25 819 documents (15 223 pour

9. http://deft.limsi.fr
10. http://deft08.limsi.fr

l'apprentissage et 10 596 pour l'évaluation) et celui de la tâche 2 de 39 243 (23 550 pour l'apprentissage et 15 693 pour l'évaluation).

5.4.3 L'échange discriminatoire

Dans l'expression de l'énergie d'Ising (section 3.3) la matrice J contient les $N \times N$ valeurs d'échange du vocabulaire d'un texte. On rappelle que $J = S^T \times S$. Notre stratégie est la suivante :

Phase d'apprentissage : Nous avons k ensembles de documents à thématiques différentes que nous considérons comme k matériaux différents. Pour chaque matériau, nous calculons la matrice d'échange J_k à partir d'un vocabulaire réduit qui a plus de probabilité de caractériser chaque thématique de manière inéquivoque.

Phase de test : Un nouveau document x à catégorie inconnue est présenté au système. En utilisant les matrices d'échange J_k :

- nous calculons k valeurs d'énergie textuelle :

$$E_k = x \times J_k \times x^T; \tag{5.7}$$

- le document x sera affecté à la thématique pour laquelle l'énergie textuelle obtient la valeur la plus grande.

Nous allons illustrer ce processus par un exemple. Les corpus ont été prétraités avec les algorithmes classiques décrits en section 2.3.3. Puis, les documents d'apprentissage ont été séparés par catégorie thématique (économie, art, télévision, sport). Chaque ensemble partage une partie du vocabulaire avec les autres (figure 5.11). Ce vocabulaire commun peut être une source d'erreur pour séparer les catégories. C'est pourquoi nous avons décidé d'utiliser le vocabulaire permettant de les différencier. Un choix raisonnable consiste à considérer les vocabulaires exclusifs (les mots n'apparaissant que dans une seule catégorie), mais compte tenu de leur petite taille il s'avérait insuffisant pour bien représenter chaque ensemble. Cela nous a amené à considérer aussi les mots partagés dont la fréquence est suffisamment différente entre catégories.

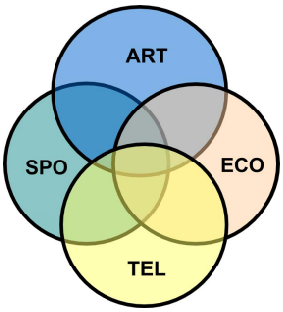

FIGURE 5.11 – *Vocabulaire du corpus d'apprentissage par catégories.*

Nous avons utilisé ces vocabulaires réduits pour construire les quatre matrices termesegment S_{ART}, S_{SPO}, S_{ECO}, S_{TEL}. Puis nous avons calculé les matrices d'échange d'Hebb

qui capturent les relations entre les mots de chaque catégorie : J_{ART}, J_{SPO}, J_{ECO}, J_{TEL} ; où chaque $J_i = S_i^T \times S_i$. Pour classer chaque document x de l'ensemble de test, nous avons calculé son énergie en utilisant les k échanges précédents (équation 5.7) :

$$\begin{aligned} E_{ART} &= x \times J_{ART} \times x^T \\ E_{SPO} &= x \times J_{SPO} \times x^T \\ E_{ECO} &= x \times J_{ECO} \times x^T \\ E_{TEL} &= x \times J_{TEL} \times x^T \end{aligned}$$

Le document x sera considéré comme un échantillon du matériau dans lequel se présente la plus grande énergie E_i.

5.4.4 Évaluation et résultats

L'évaluation proposée en DEFT [11] a été la F-mesure. Chaque fichier a été évalué en calculant la F-mesure pour chacun des corpus, pour la catégorie ou le genre, avec $\beta = 1$:

$$F\text{-}mesure(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel}, \tag{5.8}$$

La précision et le rappel correspondent aux macro-moyennes sur l'ensemble des classes. Elles sont calculées sur la précision et le rappel de chaque classe i en faisant la moyenne sur les n classes. Ainsi, chaque classe, qu'elle soit de grande ou de petite taille, compte alors à égalité dans le calcul de la précision et du rappel :

$$Précision = \sum_{i=1}^{n} \frac{Précision_i}{n} \tag{5.9}$$

$$Rappel = \sum_{i=1}^{n} \frac{Rappel_i}{n} \tag{5.10}$$

$$Précision_i = \frac{nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ attribués\ à\ la\ classe\ i} \tag{5.11}$$

$$Rappel_i = \frac{nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ à\ la\ classe\ i} \tag{5.12}$$

Les F-mesures obtenues par notre système selon (5.8) sont de 0,8307 et 0,8320 pour la tâche 1 (genre et catégorie respectivement) et 0,7561 pour la tâche 2. Ces résultats ont été un peu décevants à nos yeux. Ceux qui ont remporté le défi ont atteint des F-mesures supérieures à 0,9 pour les deux tâches. Tout de même, nous sommes au dessus de la moyenne des participants.

Nous avons observé que notre approche est très sensible au prétraitement. Cela veut dire que le fait d'appliquer des antidictionnaires pour réduire la dimension des matrices

11. http://deft08.limsi.fr/resultat.php#evaluation

(élimination des adverbes ou chiffres par exemple), a un impact négatif sur le score final. Cet effet est plus visible sur les thèmes proches comme « actualité française » et « international ».

Du fait que les opérations sur les matrices J_k sont lourdes, l'apprentissage à été limité à un ensemble d'environ 2 600 documents par thème. Cette réduction du vocabulaire améliore la vitesse du processus mais à un certain coût sur les performances. Malgré l'utilisation d'une base d'apprentissage réduite, les résultats obtenus par notre méthode sont honorables.

5.5 Conclusions

Dans ce chapitre nous avons proposé des méthodes inspirées des concepts de la Physique pour traiter deux problématiques TAL de nature très différente. Dans un premier temps, nous avons abordé le problème de détection de ruptures thématiques dans les documents. Pour le résoudre, nous avons utilisé la comparaison, par le test de Kendall, des spectres énergétiques des phrases. Ces spectres sont générés à partir des lignes de la matrice d'énergie textuelle sans avoir besoin des calculs supplémentaires. L'introduction d'une longueur de corrélation a permis de modifier de manière efficace l'allure des courbes afin que le test de Kendall puisse mieux identifier les frontières. Les résultats obtenus sont comparables à l'état de l'art. Les travaux présentés ici ont été publiés dans (Fernández et al., 2007a) et (Fernández et al., 2008a). Une analyse fine à partir des spectres pourrait consister à étudier les oscillations intra-thématique. Cette analyse devrait peut être faire appel aux méthodes linguistiques afin de pouvoir expliquer ce comportement.

Chapitre 6

Compression thermodynamique de phrases en français

Sommaire

6.1 Introduction

La compression d'une phrase consiste en la suppression de certains de ses constituants non essentiels avec le but d'obtenir une phrase plus courte tout en conservant le sens et la grammaticalité. Quel est l'intérêt d'une telle tâche ? Dans le résumé automatique par extraction, où les phrases les plus importantes sont concaténées pour produire le condensé, aucun traitement n'est effectué au niveau intra-phrase. Ainsi, une phrase longue est soit conservée dans son intégralité, soit totalement supprimée. La compression peut alors combler ce manque afin de supprimer les constituants les moins pertinents (Monod et Prince, 2006). Il existe deux grandes approches pour la compression de phrases : l'approche linguistique qui consiste à définir des règles et celle statistique qui détecte des régularités afin de produire automatiquement les règles.

Pour cette dernière approche, il est nécessaire de disposer d'un corpus d'apprentissage contenant des phrases et une version acceptable de leur compression.

Dans ce chapitre nous présentons une approche statistique-thermodynamique pour la compression automatique de phrases. L'idée est d'établir une concordance entre la compression d'une phrase à N termes et le processus par lequel, une chaîne de N spins magnétiques, tous orientés initialement vers le haut (tous les termes sont présents), subissent des fluctuations thermiques qui inversent quelques spins (suppression de quelques termes). Le problème est qu'un tel système possède 2^N configurations possibles parmi lesquelles seulement un petit sous-ensemble correspond aux compressions acceptables de la phrase initiale. Par exemple, en partant d'une phrase de 25 termes, il existe $2^{25} = 33\,554\,432$ sous-phrases possibles. Réduire un espace si énorme, tout en favorisant les configurations correctes, est le défi commun aux méthodes de compression.

Nous proposons d'utiliser les interactions entre termes (spins) voisins pour contrôler leurs retournements et réduire ainsi l'espace des configurations. Ces couplages seront mesurés préalablement sur un corpus aligné de phrases complètes/compressées. Nous consacrons notre étude exclusivement à la langue française. Ce choix est motivé par le fait que la plupart des travaux sur la compression de textes concernent la langue anglaise. Quelques systèmes qui compressent les phrases françaises sont le modèle linguistique de (Monod et Prince, 2006), l'approche statistique de (Waszak et Torres-Moreno, 2008) et la méthode de (Gagnon et Sylva, 2006) basée sur l'analyse syntaxique des phrases.

Cette dernière partie de la thèse a un caractère exploratoire. En modélisant la phrase comme un système thermodynamique sujet aux contraintes d'interaction entre unités, notre objectif ne vise pas à résoudre le problème de la compression de texte qui s'avère un des plus complexes du TAL, mais plutôt de l'étudier dans un nouveau cadre qui puisse donner des pistes pour des recherches futures.

Nous faisons un parcours des principales méthodes statistiques qui ont été proposées pour compresser les phrases. Ensuite, nous décrivons le modèle magnétique des verres de spins (*spin-glasses*) qui s'adapte bien à notre conception de la problématique. Puis, nous présentons une stratégie pour mesurer le couplage entre les termes des textes et entre leurs étiquettes grammaticales.

6.2 Les approches classiques pour la compression statistique de phrases

Parmi les premières approches statistiques de compression de phrases on trouve le modèle du canal bruité et le modèle des arbres de décision, introduits par (Knight et Marcu, 2000). La première approche considère que la compression c est la phrase originale, qui a été bruitée pour former la phrase longue l. Le modèle est constitué d'une source $P(c)$ où les phrases bien formées ont la plus grande probabilité ; du canal $P(l/c)$, qui privilégie les phrases en préservant l'information essentielle ; et de $P(c/l)$ le

décodeur. Celui-ci cherche la meilleure compression : la phrase c qui maximise $P(c/l)$. Ces probabilités sont appliquées aux arbres syntaxiques représentant les phrases [1]. Du fait que les probabilités sont pondérées selon la longueur de la phrase compressée, le taux de compression reste modeste.

La méthode des arbres de décision part d'un arbre représentant la structure d'une phrase et produit un autre arbre plus petit correspondant à la compression. L'ordre des termes est conservé, mais leurs catégories syntaxiques peuvent changer. Ces travaux ont servi de référence à beaucoup d'autres, comme celui de (Gagnon et Sylva, 2006) présentant une méthode de compression de texte basée sur la taille des arbres syntaxiques des phrases. Les auteurs proposent l'utilisation de filtres pour tailler des relations ciblées (des éléments subordonnés) tout en appliquant des anti-filtres qui empêchent l'élimination des éléments importantes (par exemple, le verbe principal de la phrase). Par ailleurs, (Clarke et Lapata, 2007) proposent une méthode qui utilise des arbres de décision et une autre qui évalue l'importance de chaque terme (selon des critères sémantiques et fonctionnels) pour décider s'il doit être effacé. D'autre part, (Jing, 2000) utilise plusieurs sources de connaissance pour la compression ; à savoir la syntaxe, le contexte et l'analyse statistique d'un corpus. L'idée est de supprimer les éléments ne se rapportant pas au sujet.

L'analyse syntaxique a été la stratégie privilégiée pour déterminer les éléments dont la disparition affectera le moins le sens et la grammaticalité des phrases. Or, les arbres syntaxiques peuvent ne pas être suffisamment robustes et le niveau supérieur (sémantique, fonctionnel) est encore plus difficile à déterminer (Waszak et Torres-Moreno, 2008). De plus, les analyseurs syntaxiques ne sont pas toujours disponibles pour toutes les langues. Il y a des études qui ne font pas appel aux arbres syntaxiques et qui obtiennent des résultats comparables. Le travail de (Nguyen et al., 2004), basé sur des *templates* de traduction, considère que les phrases non compressées sont écrites dans une langue source et les phrases compressées dans une autre langue cible. Un corpus aligné de phrases complètes/compressées est utilisé pour générer des règles qui considèrent les similarités entre phrases comme constantes et les différences comme variables. L'algorithme cherche les meilleures variables pour une phrase donnée mais, en raison de la grande quantité de règles possibles, le temps de calcul peut être exponentiel. Récemment, le système ENTROPIE proposé par (Waszak et Torres-Moreno, 2008) utilise aussi un corpus aligné de phrases complètes/compressées pour apprendre un modèle de langage bigramme et trigramme qui sert à déterminer quels termes ont une forte probabilité d'être supprimés. Le choix de la meilleure compression est réalisé en utilisant des critères entropiques. Un perceptron est utilisé pour déterminer si la phrase est suffisamment compressée. Ce système fonctionne pour des phrases en anglais et en français.

1. Un arbre syntaxique est la représentation hiérarchique entre les constituants d'un texte.

6.3 Les verres de spin

Les verres de spins sont des matériaux constitués d'unités magnétiques entre lesquelles les interactions, dites d'échange, sont aléatoirement positives ou négatives. Si le couplage entre deux spins est positif, ils ont tendance à s'orienter vers la même direction (interaction ferromagnétique). Par contre, si le couplage entre eux est négatif ils auront tendance à s'orienter en sens opposés (interaction antiferromagnétique). Ainsi, il existe une compétition locale entre ces forces et les spins ne peuvent pas toujours satisfaire simultanément les interactions contradictoires auxquelles ils sont soumis par leurs voisins. Ce comportement peut donner lieu à ce qu'on appelle la frustration, schématisée dans la figure 6.1. Dans un triangle constitué par trois spins, si les trois interactions sont négatives, elles ne peuvent jamais être satisfaites en même temps (Trémolet et al., 2000). Comment le système va-t-il finalement s'ordonner ? Les expériences ont montré qu'en dessous d'une température spécifique, les spins vont geler dans des directions variées. En effet, cette frustration est une contrainte pour la minimisation

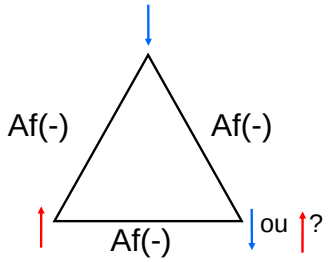

FIGURE 6.1 – *Frustration des interaction antiferromagnétiques (Af) entre trois spins (Trémolet et al., 2000).*

de l'énergie d'un système verre de spins et conduit à l'existence d'une multitude d'états métastables (Dupuis et al., 2005). Les états métastables correspondent aux minima locaux de la fonction d'énergie (figure 6.2) dans lesquels un système peut rester bloqué. L'appellation « verre » vient du fait qu'ils présentent un comportement similaire aux verres structuraux (comme ceux en silice). Le verre est un matériau qui n'est jamais en équilibre thermodynamique, il se fixe dans un état désordonné à partir duquel il évolue sans arrêt vers différentes configurations [2]. On entend parfois que le verre vieillit car ses propriétés instantanées dépendent de son âge. Les verres de spins magnétiques, désordonnés et frustrés présentent aussi du vieillissement (Dupuis et al., 2005).

6.3.1 Le texte vu comme un verre textuel

Un terme peut être vu comme un spin à deux états : ↑ (+1) indiquant sa présence dans une phrase ou ↓ (-1) son absence. Une phrase de N termes sera donc codée comme

2. On dit que le temps d'atteindre l'équilibre est tellement long que l'âge de l'Univers ne sera pas suffisant pour remarquer un changement dans sa structure.

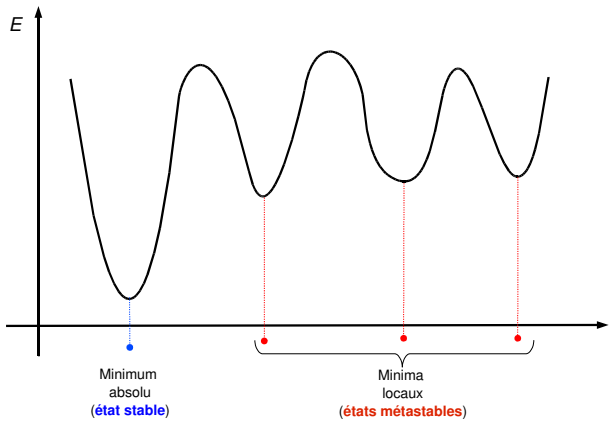

FIGURE 6.2 – *États métastables d'un système thermodynamique. Ils correspondent aux minima locaux de la fonction d'énergie.*

une chaîne de N spins tous orientés ver le haut et sa compression correspond à la même chaîne où quelques spins ont changé d'orientation.

Dans le calcul de l'énergie textuelle, nous avons établi une connectivité totale entre les termes. Or, dans notre modèle actuel nous limitons les interactions aux couples de voisins proches. C'est pourquoi, conserver l'information sur l'ordre des termes dans les phrases s'avère important. Nous abandonnons la représentation de sac de termes où la dimension vectorielle correspond à la taille du vocabulaire total du document. Dans ce chapitre, la dimension de chaque vecteur est le nombre de termes de la phrase représentée. Nous n'appliquons aucun prétraitement : les mots et les signes de ponctuation sont considérés comme des termes Nous voulons que notre algorithme détecte les termes les moins pertinents et les supprime sans perte de grammaticalité.

Le système de compression de phrases que nous proposons utilise le corpus aligné de phrases complètes/compressées en français MYRIAM [3] pour mesurer les couplages entre termes adjacents (voisins proches). Ces règles sont assimilées aux couplages entre les spins magnétiques qui interagissent dans un matériau. Il s'agit du même corpus utilisé par (Waszak et Torres-Moreno, 2008), ce qui nous a permis de faire des comparaisons.

Ainsi, nous voulons établir des règles d'interactions qui, à partir de la phrase originale, amènent à une compression correcte. Il est clair que pour supprimer les termes accessoires tout en gardant ceux pertinents, il faut en même temps des interactions positives et négatives. Par exemple, si l'on veut que la chaîne *la maison rouge* soit compressée

3. Le corpus MYRIAM a été construit par Michel Gagnon de l'École Polytechnique de Montréal, http://www.professeurs.polymtl.ca/michel.gagnon. Il est disponible à l'adresse http://www.lia.univ-avignon.fr/fileadmin/documents/Users//Intranet/chercheurs/torres. Malgré le fait qu'il contienne quelques erreurs grammaticales, il est à la base de l'approche avec lequel nous nous comparons. D'ailleurs, nous n'avons pas connaissance de l'existence d'autre corpus en français disponible et adapté à la tâche de compression.

en *la maison*, les interactions $J_{i,j}$ entre termes voisins doivent être : $J_{la,maison} = +x$ et $J_{maison,rouge} = -y$. Cette variété en valeur et en signe des interactions entre termes peut produire des compétitions internes dans la phrase.

En l'absence d'autre facteur affectant le système (température, champ externe), obéir aux règles d'échange conduit à une configuration appelée état fondamental du système où l'énergie est minimale. Par contre, si on essaie d'appliquer sur un terme des règles qui se contredisent, on sera dans un cas de « frustration de termes » ou de « phrase frustrée » ayant pour conséquence la production d'états métastables. Cette situation fait de notre système une sorte de verre de spins ou encore, un verre textuel.

6.4 Calcul des règles d'échange

Le corpus MYRIAM est composé de 219 phrases (contenant de 7 à 135 termes) issues de sources journalistiques variées. Pour chaque phrase, une version compressée a été produite manuellement. Un échantillon est montré dans le tableau 6.1. Nous avons éclaté le corpus en deux ensembles : 80% pour l'apprentissage des couplages et 20% pour faire les tests de compression.

Phrases complètes	**Phrases compressées**
1. Enfin, je souhaite à notre terre la paix.	1. Je souhaite la paix.
2. Le logement est par nature porteur d'une contradiction.	2. Le logement est porteur d'une contradiction.
3. Un livre enfin qui se dévore comme un roman.	3. Un livre qui se dévore comme un roman.
4. Les automobiles coréennes sont désormais vendues en France.	4. Les automobiles coréennes sont vendues en France.
5. Le président sortant, Jerry Rawlings, se succédant à lui-même.	5. Le président sortant, Jerry Rawlings, se succédant.
6. Le déficit actuel pourrait doubler d'ici l'an 2000.	6. Le déficit pourrait doubler d'ici 2000.

TABLE *6.1* – *Exemples de phrases parallèles complètes/compressées du corpus* MYRIAM.

6.4.1 Le couplage entre termes

En utilisant le corpus d'apprentissage, nous avons déduit des relations $J_{\text{terme}_i,\text{terme}_j}$ entre les termes voisins i et j, selon leurs états dans les versions compressées des phrases. Par exemple, soit la première phrase du tableau 6.1 où nous barrons les termes qui ont disparu dans le processus de compression, ce qui change son état de ↑ à ↓ :

↓	↓	↑	↑	↓	↓	↓	↑	↑	↑
~~Enfin,~~	~~,~~	je	souhaite	~~à~~	~~notre~~	~~terre~~	la	paix	.

Nous pouvons observer que les signes de ponctuation sont aussi considérés comme des termes. De ceci, nous avons déduit les règles suivantes entre termes adjacents :

$J_{\text{terme}_i,\text{terme}_j} = +1$ (ferromagnétique) si les deux termes sont présents ou absents ;

$J_{\text{terme}_i,\text{terme}_j} = -1$ (antiferromagnétique) si l'un des termes est présent et l'autre absent.

De cette façon, nous établissons les huit couplages entre voisins proches présentés dans le tableau 6.2. Ces règles indiquent à chaque terme de la phrase de suivre ou non l'orien-

Couplage ferromagnétique $\uparrow\uparrow$ **ou** $\downarrow\downarrow$	**Couplage antiferromagnétique** $\uparrow\downarrow$ **ou** $\downarrow\uparrow$
$J_{\text{~~Enfin~~ , ,}} = +1$	$J_{\text{, je}} = -1$
$J_{\text{je , souhaite}} = +1$	$J_{\text{souhaite , ~~à~~}} = -1$
$J_{\text{à , ~~notre~~}} = +1$	$J_{\text{~~terre~~ , la}} = -1$
$J_{\text{~~notre~~ , ~~terre~~}} = +1$	
$J_{\text{la , paix}} = +1$	
$J_{\text{paix , .}} = +1$	

TABLE 6.2 – *Les couplages entre termes voisins.*

tation de ses voisins. À partir de la phrase complète comme configuration initiale (tous les spins $\uparrow$), la satisfaction de toutes les règles amène à l'état fondamental d'énergie minimale qui, dans ce cas, correspond à une compression correcte. Cependant, les couplages du tableau 6.2 sont valables uniquement pour cette phrase particulière. Mais nous recherchons des règles générales qui puissent servir pour compresser des phrases non vues auparavant. Nous proposons de suivre la même démarche avec toutes les phrases du corpus d'apprentissage. Deux termes peuvent être voisins proches dans plusieurs phrases et donc, pour avoir une valeur unique pour chaque couple de termes, nous faisons la somme de leurs occurrences. Cela permet de jouer sur l'amplitude des couplages d'échange. Ce processus a produit environ 6 000 règles. Quelques exemples sont présentés dans le tableau 6.3 :

Terme i	**Terme** j	$J_{i,j}$
cependant	accueillies	-1
cependant	,	2
durée	et	-1
durée	de	1
que	les	+ 7
que	(	-1
occupent	,	-1
occupent	une	+ 1
plus	démunis	-2
beaucoup	plus	+1
sait	.	-3
sont	désormais	-2
on	se	+4

TABLE 6.3 – *Quelques exemples des règles d'échange entre termes apprises sur le corpus d'apprentissage.*

Application sur le corpus de test

Nous avons appliqué les règles apprises pour compresser les phrases du corpus de test. Cependant, les résultats obtenus sont mitigés pour les raisons suivantes :

1. une grande partie du vocabulaire des phrases à compresser n'existe pas dans le corpus d'apprentissage et, par conséquent, aucune règle ne les concerne ;
2. même si deux termes sont présents dans les corpus, il n'est pas sûr qu'ils soient voisins adjacents, donc leur règle d'échange est inexistante.

Il résulte de cette situation que, pour un grand ensemble des phrases de test, il n'existe aucune règle à appliquer et pour d'autres phrases, ces règles sont peu nombreuses. Un exemple est montré dans le tableau 6.4, où nous avons seulement deux règles pour une phrase de huit termes.

$J_{\text{la,pénurie}} = +1$ $J_{\text{fait,sentir}} = +1$ les autres $\not\exists$								
Conf. initiale	↑ Mais	↑ partout	↑ la	↑ **pénurie**	↑ se	↑ **fait**	↑ sentir	↑ .
Etat fond.	?	?	↑ la	↑ pénurie	?	↑ fait	↑ sentir	?

TABLE 6.4 – *Exemple d'application du couplage entre termes pour une phrase du corpus de test. Seulement deux des sept couplages possibles entre termes voisins ont été déterminés pendant le processus d'apprentissage. Ce manque d'information peut être une conséquence de la taille réduite du corpus.*

À notre connaissance, il n'existe pas, pour la langue française, d'autres corpus alignés disponibles plus représentatifs. Pour surmonter les problèmes liés au manque de termes, nous avons décidé de grouper les termes selon leur catégorie grammaticale. Nous pensons que cette stratégie pourrait nous permettre d'élargir la validité des règles obtenues.

6.4.2 Le couplage grammatical

L'idée est de calculer des règles d'échange plus générales que celles établies entre termes. Pour cela, nous avons utilisé le logiciel TREETAGGER (Schmid, 1994). Cet outil nous a permis d'étiqueter automatiquement les termes selon leurs catégories grammaticales pour produire des relations du type :

$$J_{\text{mais,partout}} \rightarrow J_{\text{PREP , ADV}}$$
$$J_{\text{fait,sentir}} \rightarrow J_{\text{VERB :PRE , VERB :INF}}$$
$$J_{\text{une,fois}} \rightarrow J_{\text{ART , NOM}}$$
$$J_{\text{la,pénurie}} \rightarrow J_{\text{ART , NOM}}$$

Ainsi, on regroupe plusieurs règles en une seule qui représente le couplage entre deux types de termes. Nous avons choisi d'utiliser la valeur moyenne. Par exemple, $J_{\text{une,fois}} = +1$ et $J_{\text{la,pénurie}} = +2$, alors $J_{\text{ART,NOM}} = +1.5$. Ces opérations ont réduit les presque 6 000 règles entre termes à environ 400 relations entre étiquettes grammaticales.

Les distributions des valeurs des couplages dans les deux cas, termes et étiquettes, sont montrées dans la figure 6.3. Par souci de clarté, nous ne montrons que la partie de

la courbe où les valeurs de fréquence sont significatives. On observe que dans le cas des termes (trait pointillé), environ 80% correspond aux valeurs +1. Elles sont produites en grande partie par des ocurrences uniques des termes voisins de la même orientation. En revanche, on observe pour les étiquettes (trait continu) que cet effet a été adouci. Le pic concerne environ 48% des couples.

Pour les termes, les valeurs des couplages sont comprises entre -9 et +58. Le couplage antiferromagnétique (négatif) le plus fort correspond au couple { **)** , **.** } ce qui se comprend car une information placée entre parenthèse n'est *a priori* pas essentielle. Le couplage ferromagnétique (positif) le plus important à { ***de*** , ***la*** }. Pour les étiquettes grammaticales, les valeurs des couplages s'étendent de -3 à +7. L'interaction antiferromagnétique la plus forte est produite pour { PUN , SENT } et la ferromagnétique (positive) a été obtenue pour { PUN :cit,SENT } (voir le tableau 6.5 pour la description de ces étiquettes). Nous observons dans les deux cas, une prédominance des valeurs

Étiquette	**Description**
ART	article
NOM	nom
PREP	préposition
ADV	adverbe
VERB :PRE	verbe présent
VERB :INF	verbe infinitif
PUN	signe de ponctuation (, ; :)
SENT	final de phrase (. ! ?)
PUN :cit	guillemets (")
KON	conjonction (et ou or mais ni)

TABLE 6.5 – *Quelques exemples d'étiquettes grammaticales.*

positives sur les négatives. Il semble que les termes voisins qui restent ou disparaissent ensemble pendant le processus de compression sont plus nombreux que ceux qui restent tandis que le voisin disparaît.

6.5 Application des règles à la compression de phrases

6.5.1 Les états fondamentaux de la chaîne de spins

Nous avons appliqué l'ensemble réduit des règles sur les phrases du corpus de test. Le tableau 6.6 montre un exemple de cette application. On observe que, pour la même phrase du tableau 6.4, on a maintenant les sept valeurs de couplages entre voisins proches. Appliqués sur la phrase originale, ces couplages produisent une compression acceptable. Malheureusement ce n'est pas le cas pour toutes les autres. Même en ayant toutes les valeurs d'échange permettant d'obtenir les états fondamentaux, nous serions confrontés à deux problèmes :

1. Les sous-phrases obtenues avec les états fondamentaux ne sont pas systématiquement de bonnes compressions. Cet effet peut être lié à la petite taille du corpus

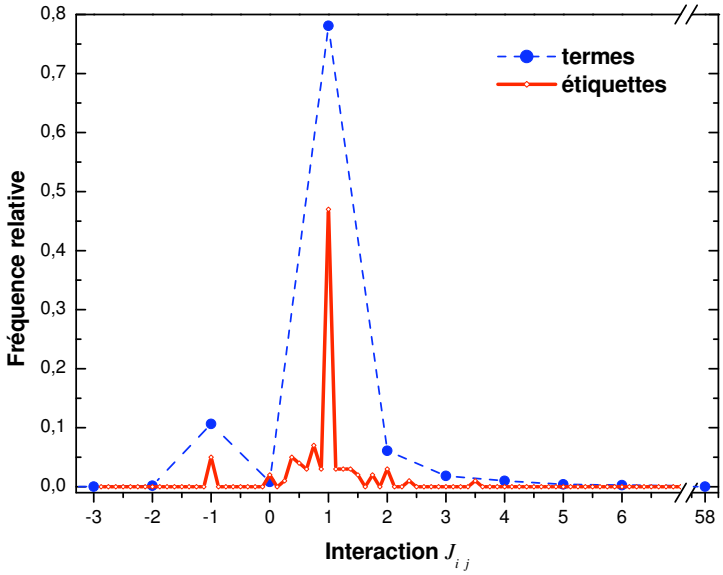

FIGURE 6.3 – *Fréquences relatives des couplages entre termes et entre étiquettes calculées à partir du corpus* MYRIAM.

$J_{KON,ADV} = +0,6154$
$J_{ADV,DET} = -0,2381$
$J_{DET,NOM} = +1,1725$
$J_{NOM,PRO} = +0,4500$
$J_{PRO,VER} = +1,0000$
$J_{VER,VER} = +1,0000$
$J_{VER,SENT} = +1,0000$

Conf. initiale	↑ Mais KON	↑ partout ADV	↑ la DET	↑ **pénurie NOM**	↑ se PRO	↑ **fait VER**	↑ sentir VER	↑ . SENT
Etat fond.	↓	↓	↑ la	↑ pénurie	↑ se	↑ fait	↑ sentir	↑ .

TABLE 6.6 – *Exemple d'application du couplage entre étiquettes grammaticales pour la même phrase du tableau 6.4. Maintenant nous avons les sept valeurs des couplages entre voisins proches. Appliqués sur la phrase originale, ces couplages produisent une compression bien acceptable.*

d'apprentissage qui engendre des règles rigides qui ne s'ajustent pas à tous les cas.

2. La frustration, c'est à dire, l'impossibilité de satisfaire en même temps toutes les règles d'échange, est présente dans environ 13% des phrases de test. Dans ce cas, il y a plus d'une solution pour une même phrase.

Il faut une stratégie qui accorde un certain degré de souplesse dans l'application des règles qui puisse en même temps traiter les phrases frustrées. C'est pour ces raisons que nous avons décidé de réaliser des simulations du type Métropolis Monte-Carlo. Cela nous permettra, dans un premier temps, d'introduire des fluctuations thermiques qui apporteront de la flexibilité dans l'application des règles et d'utiliser le recuit simulé pour faire face à la frustration des verres de termes.

6.5.2 Simulations Métropolis Monte-Carlo

L'idée principale d'une simulation Monte-Carlo est d'imiter les fluctuations thermiques aléatoires d'un système qui parcourt plusieurs états pendant une expérience. La probabilité p_μ de trouver le système dans un état μ est donnée par la distribution de Gibbs-Boltzmann :

$$p_\mu \propto exp(-E_\mu/kT). \tag{6.1}$$

où E_μ est l'énergie du système dans l'état μ, k est la constante de Boltzmann et T la température.

Pour faire la transition entre états nous avons utilisé la dynamique de Métropolis :

1. Soit une chaîne de N spins dans un état initial μ d'énergie E_μ ;
2. à chaque pas de la simulation (on fait N pas afin de donner à tous les spins la possibilité de se retourner), choisir un spin au hasard dont le retournement amène à un nouvel état ν d'énergie E_ν ;
3. calculer $\Delta E = E_\nu - E_\mu$ pour savoir si un tel retournement (*flip*) de spin fait diminuer ou augmenter l'énergie du système ;
   - si l'énergie diminue ($\Delta E < 0$), on accepte de manière définitive le *flip* car l'état produit est plus stable que le précédant ;
   - si l'énergie augmente ($\Delta E > 0$), on génère un numéro aléatoire r, tel que $0 \leq r \leq 1$;
   - si $r < exp(-\Delta E/kT)$ on accepte le *flip*, autrement, on reste dans le même état μ.
4. répéter la simulation un nombre suffisant de fois, pour permettre au système d'atteindre l'équilibre à une température établie.

Être en équilibre signifie que le système ne fera plus de transitions importantes et la valeur de l'énergie devient quasi constante. Dès que l'équilibre est atteint, l'objectif est de mesurer des quantités comme l'énergie ou la magnétisation. Dans notre cas, nous sommes plutôt intéressés à récupérer les états dans lesquels le système se stabilise à chaque température, car ils représentent des variantes potentielles de la compression.

Les paramètres de la simulation

La température est une perturbation qui fait varier l'énergie du système. De façon simple nous pouvons dire que dans notre système il y a deux facteurs en concurrence : l'interaction entre les spins (couplages d'échange) et la température. À basse température, l'échange domine et les spins tendront à rester dans la configuration fondamentale. À haute température, les fluctuations thermiques favorisent les états aléatoires qui ne répondent pas forcément aux demandes du facteur d'échange. Notre but est d'utiliser ce comportement pour faire sortir les phrases des états rigides dictés par les couplages d'échange et de produire des variantes pour en choisir la meilleure. Les simulations ont été faites sous les conditions suivantes :

État initial : Nous avons fait une simulation pour chaque phrase en commençant par l'état ferromagnétique. Ainsi, tous les termes de la phrase sont présents.

Spins fixés ↑ : Pour ne pas confondre une configuration avec la configuration symétrique de même énergie, où tous les spins ont l'état opposé, nous avons fixé quelques spins de la phrase. Nous avons fixé dans l'état ↑ le symbole de ponctuation final qui ne disparaît jamais.

Spins fixés ↓ : Vu la prédominance des échanges positifs sur les négatifs, nous avons fixé un spin dans l'état ↓. Pour choisir l'élément avec la possibilité la plus haute de disparaître, nous avons introduit un indice de suppression (IS) :

$$IS(terme_{j,i}) = \sum_{i=1}^{P} \frac{ns(terme_{j,i})}{|phr_i|} \tag{6.2}$$

où $ns(terme_{j,i})$ est le nombre de fois que le terme j a été supprimé de la phrase i, et $|phr_i|$ est le nombre de termes de la phrase. La somme parcourt les P phrases du corpus d'apprentissage. Par exemple, pour le texte suivant à trois phrases, où on a barré les termes qui ont été supprimés lors d'une compression manuelle, nous calculons l'IS du mot *bleu* :

1	Le livre ~~bleu~~ de ma tante .	IS(bleu,1)=1/7=0,14
2	Le bleu c ' est ma couleur préférée .	IS(bleu,2)=0/9=0,00
3	J ' ai un ordinateur ~~bleu~~ et un sac - à - dos ~~, aussi bleu , tout neufs~~ .	IS(bleu,3)=2/20=0,10
		IS(bleu,corpus)=0,24

Ces indices, calculés sur tous les termes du corpus d'apprentissage, sont utilisés dans nos simulations pour choisir le spin qui sera fixé ↓ dans la configuration initiale. Celui correspond au terme d'IS plus élevé. Si dans la phrase à compresser, il a y des termes qui n'existent pas dans le corpus d'apprentissage, leur IS sera égal à zéro.

Température : Nous faisons varier la température T de 1 à 0 par pas de 0,01 en adaptant les retournements de spin selon la dynamique de Métropolis. Chaque valeur de T accorde différents degrés de flexibilité à l'application des règles d'échange et pour cette raison les orientations des spins s'arrangent différemment.

Frustration : Pour faire face à la frustration et éviter ainsi que le système reste bloqué dans des états métastables, nous avons utilisé la technique du recuit simulé. Elle consiste à faire monter et descendre la température plusieurs fois dans un intervalle de températures suffisamment basses.

Nombre d'itérations : Pour estimer le nombre d'itérations nécessaires pour atteindre l'équilibre à une température déterminée, nous avons réalisé des simulations pendant un temps suffisement long. Le but étant de déterminer le moment où l'énergie du système se stabilise. Dans la figure 6.4 nous présentons l'évolution de l'énergie selon le temps (en nombre de itérations) pour deux phrases du corpus de test. La première à gauche avec N=17 termes atteint l'équilibre à une température de 0,15, en environ 70 itérations. La deuxième à droite, la phrase plus longue du corpus, avec N=135 termes à une température de 0,2 se stabilise en environ 500 itérations. À partir de ces résultats, et étant donné la petite taille du système, nous avons considéré suffisant d'utiliser 1 000 itérations par température.

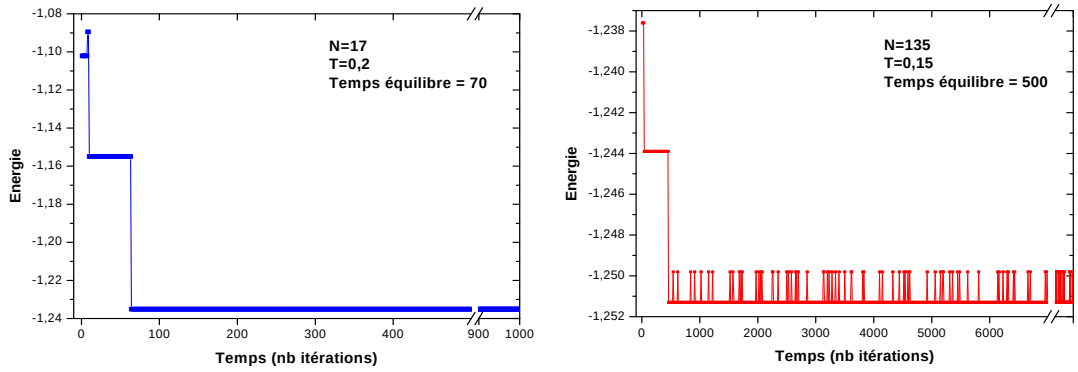

FIGURE 6.4 – *Temps pour atteindre l'équilibre pour deux phrases. N est le nombre de spins ou termes.*

Les configurations retenues : À la fin du processus, nous récupérons les états finaux à chaque température. Cela produit un ensemble de variantes de compression de la phrase initiale. Nous avons utilisé deux critères différents pour choisir les meilleures compressions des phrases, à savoir : l'état d'énergie minimale et la magnétisation maximale (qui correspond au taux de compression minimum).

L'énergie d'une chaîne de spins est :

$$E = \sum_{i,j} s_i s_j J_{i,j} \tag{6.3}$$

où s_i et s_j sont les états des spins i et j, $J_{i,j}$ l'intéraction d'échange entre eux. La magnétisation est définié :

$$M = \sum_{i} s_i. \tag{6.4}$$

6.6 Évaluation de la compression : mesures BLEU

Le système *Bilingual Language Evaluation Understudy* (BLEU) (Papineni et al., 2001), conçu à l'origine pour juger la précision des résultats de la traduction automatique, est aussi utilisé dans le cadre de la compression de textes (Dorr et al., 2003; Egawa et al., 2008). Il mesure la concordance entre une phrase candidate (la compression faite par un système automatique) et une référence (celle faite par un humain). Comme dans le cas de ROUGE (section 4.1.3), BLEU se base sur la comparaison des n-grammes contenus dans les deux textes. Les différences principales entre ces deux systèmes d'évaluation sont les suivantes :

1. ROUGE est une mesure du rappel et BLEU de la précision. Le rappel mesure le pourcentage de n-grammes des références qui sont aussi présents dans le candi-

dat. En revanche, la précision est le pourcentage de n-grammes du candidat qui apparaît dans la référence.

2. BLEU fait une comparaison alignée, cela veut dire qu'elle est faite phrase à phrase et non sur les documents en entier comme ROUGE.
3. Les longueurs des phrases, candidates et références, sont prises en compte dans le calcul du score BLEU ce qui n'est pas le cas pour ROUGE.

Les auteurs ont montré une forte corrélation entre les indices BLEU et les jugements humains sur la qualité de la compression.

Les tableaux 6.7 et 6.8 montrent les scores BLEU obtenus par notre système sans et avec recuit simulé. En ce qui concerne l'unité de comparaison, nous utilisons les 3-grammes et 4-grammes tel que suggéré par (Papineni et al., 2001). Le critère de sélection est celui d'énergie minimale et de magnétisation maximale (donc compression minimale). Nous avons réalisé trois simulations dans chaque cas (s1, s2 et s3 dans les tableaux). Nous comparons nos résultats avec ceux produits par le système ENTROPIE de (Waszak et Torres-Moreno, 2008). Nous ajoutons aussi une *baseline* construite à partir d'une simulation où les couplages sont des valeurs aléatoires entre -1 et +1. Plus les valeurs BLEU sont élevées, plus les compressions candidates sont proches du modèle de référence. On observe que, pour la plupart des simulations, le critère de magnétisation maximale obtient des scores plus élevés que celui d'énergie minimale et légèrement supérieurs à ceux obtenus par le système ENTROPIE. Le recuit simulé ne semble pas avoir un effet significatif sur le résultat.

Deux spins fixés : symbol de ponctuation final (↑) et terme d'IS$_{max}$ (↓) Critère : énergie minimale								
Unité BLEU	***Baseline***	**ENTROPIE**	**VERRE TEXTUEL**			**VERRE TEXTUEL avec recuit**		
			s1	s2	s3	s1	s2	s3
3-gramme	0,3767	0,7479	0,7470	0,6676	0,7200	0,7083	0,7446	0,7337
4-gramme	0,2990	0,7018	0,7158	0,6319	0,6936	0,6821	0,7150	0,7057

TABLE 6.7 – *Scores* BLEU *pour notre système* VERRE TEXTUEL *utilisant le critère d'énergie minimale. Pour chaque simulation si, nous montrons aussi les résultats pour le système* ENTROPIE *proposé par (Waszak et Torres-Moreno, 2008) et une baseline où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.*

Afin d'augmenter la pertinence de l'information conservée dans la phrase compressée, nous avons augmenté à quatre le nombre de spins fixés. Nous avons considéré que le premier substantif et le premier verbe (dans l'ordre original du texte) d'une phrase, restent en état ↑ dans la phrase compressée.

À nouveau, nous comparons nos résultats avec ceux du système ENTROPIE et le *baseline* de couplages aléatoires. On observe dans les tableaux 6.9 et 6.10 que les deux critères, d'énergie et de magnétisation, sont comparables selon la mesure BLEU. Le recuit simulé ne semble pas avoir un effet significatif dans le résultat. Dans le deux cas nos scores sont légèrement supérieurs à ceux obtenus par le système ENTROPIE.

Deux spins fixés : symbol de ponctuation final (↑) et terme d'IS$_{max}$ (↓) Critère : magnétisation maximale								
Unité BLEU	***Baseline***	**ENTROPIE**	**VERRE TEXTUEL**			**VERRE TEXTUEL avec recuit**		
			s1	s2	s3	s1	s2	s3
3-gramme	0,3767	0,7479	0,7560	0,7597	0,7527	0,7331	0,7567	0,7381
4-gramme	0,2990	0,7018	0,6715	0,7097	0,7058	0,6825	0,7064	0,6836

TABLE *6.8* – *Scores* BLEU *pour notre système* VERRE TEXTUEL *utilisant le critère de magnétisation maximale. Pour chaque simulation si, nous montrons aussi les résultats pour le système* ENTROPIE *proposé par (Waszak et Torres-Moreno, 2008) et une baseline où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.*

Quatre spins fixés : ponctuation final, nom et verbe premiers (↑) et terme d'IS$_{max}$ (↓) Critère : énergie minimale								
Unité BLEU	***Baseline***	**ENTROPIE**	**VERRE TEXTUEL**			**VERRE TEXTUEL avec recuit**		
			s1	s2	s3	s1	s2	s3
3-gramme	0,3767	0,7479	0,7854	0,7827	0,7642	0,7647	0,7759	0,7752
4-gramme	0,2990	0,7018	0,7528	0,7501	0,7298	0,7344	0,7418	0,7423

TABLE *6.9* – *Scores* BLEU *pour notre système* VERRE TEXTUEL *utilisant le critère d'énergie minimale. Pour chaque simulation si, nous montrons aussi les résultats pour le système* ENTROPIE *proposé par (Waszak et Torres-Moreno, 2008) et une baseline où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.*

La fixation des deux spins suplémentaires a favorisé le score BLEU. Cependant, comme dans le cas de ROUGE, BLEU est d'avantage une mesure de la pertinence de l'information que de la qualité grammaticale des textes. Nous pouvons dire que notre système produit des compressions dans lesquelles l'information essentielle est conservée, et ceci semble plus évident quand on fixe quatre spins. Or pour vérifier la qualité des phrases une évaluation manuelle s'avère nécessaire.

Pour avoir un aperçu de la performance globale des systèmes, nous avons calculé le nombre de phrases compressées correctes, le nombre de phrases compressées incor-

Quatre spins fixés : ponctuation final, nom et verbe premiers (↑) et terme d'IS$_{max}$ (↓) Critère : magnétisation maximale								
Unité BLEU	***Baseline***	**ENTROPIE**	**VERRE TEXTUEL**			**VERRE TEXTUEL avec recuit**		
			s1	s2	s3	s1	s2	s3
3-gramme	0,3767	0,7479	0,7614	0,7587	0,7827	0,7457	0,7397	0,7514
4-gramme	0,2990	0,7018	0,7400	0,7167	0,7382	0,7011	0,6937	0,7070

TABLE *6.10* – *Scores* BLEU *pour notre système* VERRE TEXTUEL *utilisant le critère de magnétisation maximale. Pour chaque simulation si, nous montrons aussi les résultats pour le système* ENTROPIE *proposé par (Waszak et Torres-Moreno, 2008) et une baseline où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.*

rectes et le nombre de phrases non compressées pour chaque système. Les résultats sont présentés dans le tableau 6.11. On observe que les fixations du symbole de ponctuation final et du terme d'indice de supression maximal (2 spins) produisent des meilleurs compressions que quand on fixe aussi le premier substantif et le premier verbe de la phrase (4 spins). Dans le cas où la phrase commence par un complément non essentiel, il est évident que le dernier choix est inadapté et qu'il a un effet négatif sur la compression.

Système	% des phrases non compressées	% des phrases compressées correctes	% des phrases compressées incorrectes
ENTROPIE	≈ 30%	≈ 30%	≈ 40%
VERRE TEXT. (2 spins fixés)	≈ 40%	≈ **40**%	≈ **20**%
VERRE TEXT. (4 spins fixés)	≈ 40%	≈ 20%	≈ 40%
Baseline	≈ 5%	≈ 5%	≈ 90%

TABLE *6.11* – *Pourcentages de phrases du corpus qui ont été compressées par le système* ENTROPIE *et* VERRE TEXTUEL *pendant une simulation.*

Des exemples de compressions sont montrées dans le tableau 6.12. Il faut noter que les termes comme *d'ici* ont été separés en *de* et *ici* pendant le processus de compression. Pendant ces expériences, on a observé que le système ENTROPIE semble être plus

Originale	De ci de là, certains fabricants adoptent des mesures .
Humain	certains fabricants adoptent des mesures .
ENTROPIE	de là, certains fabricants des mesures .
VERRE TEXTUEL 2 SPINS	**certains fabricants adoptent des mesures .**
VERRE TEXTUEL 4 SPINS	De ci de là certains fabricants adoptent des mesures .
Originale	Comme on le voit , on revient de loin .
Humain	on revient de loin .
ENTROPIE	**on le voit , on revient de loin .**
VERRE TEXTUEL 2 SPINS	**on le voit , on revient de loin .**
VERRE TEXTUEL 4 SPINS	voit .
Originale	CETTE fois nous y sommes , ce est la crise du logement .
Humain	ce est la crise du logement .
ENTROPIE	CETTE fois ce est la crise du logement .
VERRE TEXTUEL 2 SPINS	**ce est la crise du logement .**
VERRE TEXTUEL 4 SPINS	**ce est la crise du logement .**
Originale	Moyennant quoi , la culture " intégrée " utilise beaucoup moins de intrants .
Humain	la culture " intégrée " utilise moins de intrants .
ENTROPIE	**la culture " intégrée " utilise moins de intrants .**
VERRE TEXTUEL 2 SPINS	Moyennant quoi , la culture " intégrée " utilise moins de intrants .
VERRE TEXTUEL 4 SPINS	, la culture " intégrée " utilise .
Originale	Et , mieux encore , je vous souhaite une meilleure santé économique .
Humain	je vous souhaite une meilleure santé économique .
ENTROPIE	Et , , je vous souhaite une santé économique .
VERRE TEXTUEL 2 SPINS	**Et , je vous souhaite une meilleure santé économique .**
VERRE TEXTUEL 4 SPINS	souhaite une meilleure santé économique .

TABLE *6.12* – *Exemples de compressions générées par notre système* VERRE TEXTUEL. *Nous montrons la phrase originale, la compression faite par un humain et celle produite par le système* ENTROPIE. *En gras, la meilleure compression. Les termes comme « d'ici » ont été separés en « de » et « ici » pendant le processus de compression.*

robuste pour garder la grammaticalité que le nôtre. Il est possible que cela soit dû à l'utilisation de bigrammes et trigrammes de termes comme unité de base. Dans notre cas, nous nous intéressons à l'exploration des interactions des termes isolés (unigrammes).

6.7 Conclusions

Nous avons proposé un système thermodynamique de compression de phrases en français. Les phrases ont été codées comme des chaînes de verres de spins. Les couplages positifs et négatifs entre termes et entre leurs étiquettes grammaticales, ont été calculés sur un corpus d'apprentissage composé des phrases complètes/compressées.

Les phrases provenant d'un deuxième corpus de test ont été compressées en appliquant les couplages appris et en les soumettant à une dynamique thermique de Metropolis Monte-Carlo. Pour chaque phrase donnée, cette approche apporte un ensemble de choix de compression, une par température. Cet ensemble est différent par simulation. Ainsi le système n'est pas entièrement déterministe. Ce comportement est plus en accord avec la tâche de compression de texte, qui n'a pas de solution unique. Deux personnes différentes ne compressent pas une phrase de la même façon, et plus encore, la même personne peut donner deux compressions différentes à des moments différents. Les phrases compressées, choisies selon des critères d'énergie et de magnétisation ont été évaluées par rapport aux compressions faites par des humains. Les scores BLEU obtenus sont comparables à ceux du système ENTROPIE.

La stratégie de grouper les termes selon leur catégorie grammaticale a généré des règles d'échange plus générales qui génèrent des compressions où l'information essentielle est conservée. Or, la précision perdue avec cette démarche impacte dans la qualité grammaticale des phrases.

Nous pensons qu'il faut impérativement un corpus assez large (voir par exemple les tailles des corpus du tableau 1.1) pour que les règles puissent être établies de façon plus précise. Une fois réglé l'aspect de la représentativité du corpus, il sera possible de faire intervenir d'autres facteurs dans la simulation. Par exemple, on pourrait ajuster la simulation aux différents types de documents, et en différentes langues, en changeant les valeurs du champ externe appliqué. La frustration des phrases pourrait ainsi être traitée avec des algorithmes plus complexes comme le *simulated tempering* (Newman et Barkema, 1999) qui est plus adapté aux systèmes type verres de spins.

Le *simulated tempering* consiste à réaliser deux ou plusieurs simulations simultanées sur le même système mais à différentes températures. Le plus souvent possible, on change les états entre deux des simulations avec une certaine probabilité. Cette probabilité est choisie de manière à ce que les systèmes continuent à suivre la distribution de Boltzmann à la température adéquate. Étant donné qu'à hautes températures il est plus facile de vaincre les barrières d'énergie, cet échange d'états aide la simulation de températures plus basses à sortir des puits d'énergie où le système peut se trouver.

Ce travail a été accepté pour publication dans (Fernández et Torres-Moreno, 2009).

Chapitre 7

Conclusions et perspectives

Dans ce travail de thèse, nous avons utilisé les modèles de spins de la physique du magnétisme pour étudier le « comportement » des systèmes textuels. Nous avons modélisé les textes de façon à trouver des analogies entre les chaînes de spins et les segments de texte. Ceci a permis d'introduire des notions intéressantes. Ainsi, les matrices d'échange entre mots, l'énergie textuelle entre documents, le dopage d'un corpus de textes, le spectre énergétique d'un phrase et l'état fondamental d'une chaîne de termes ont été abordés.

Les apports expérimentaux de cette thèse sont très divers. Nous avons utilisé ces nouveaux concepts pour la construction d'algorithmes d'analyse textuelle pour résoudre une palette de problématiques du Traitement Automatique de la Langue (TAL) : le résumé automatique, la recherche d'information orientée, la détection de ruptures thématiques et la classification documentaire. La compression de phrases a été aussi abordée mais de façon encore exploratoire. Ces tâches nous ont obligé à respecter des différentes exigences et contraintes inhérentes au domaine. Pour les surmonter, nous avons fait appel aux stratégies qui combinent des intuitions et des outils issus des deux domaines : la Physique statistique et le TAL. Nos algorithmes ont toujours été confrontés aux approches classiques en obtenant de bonnes performances.

Dans le chapitre 3 nous avons introduit le concept d'énergie textuelle. Il représente l'énergie d'Ising calculée sur l'ensemble des phrases et des mots des documents. Nous avons montré que l'énergie textuelle offre un moyen efficace de pondération des segments textuels. Nous avons utilisé la théorie des graphes pour expliquer la nature des liens parmi les phrases. Grâce au fait qu'elle effectue le calcul de chemins d'ordre deux, cette nouvelle mesure de similarité capture des relations indirectes entre documents qui peuvent échapper aux mesures locales comme le cosinus. On a comparé notre approche avec les algorithmes de classement basés sur les graphes tels que PAGERANK et TEXTRANK. Nous avons montré que ni les processus itératifs, ni le calcul de chemins d'ordre supérieur, ne modifient significativement le score obtenu avec l'énergie textuelle.

Nous avons construit le système ENERTEX basé sur l'énergie textuelle. Nous avons ainsi abordé la génération des résumés automatiques. Il s'agit d'un problème qui reste ouvert, malgré la quantité d'efforts déployés, auquel la communauté scientifique n'a su répondre que partiellement.

Le chapitre 4 présente nos résultats obtenus en résumé automatique. Pour le résumé générique, nous avons mené une palette d'expériences en différentes langues et domaines qui ont permis de constater les performances du système ENERTEX et de le positionner par rapport aux systèmes présents dans la littérature. Sur les corpus des campagnes DUC 2002, ENERTEX est bien placé par rapport aux participants. Nous avons apporté une modification qui consiste à mettre un champ externe en rapport avec un corpus multidocument. Cette stratégie génère des résumés guidés par les besoins de l'utilisateur. L'évaluation a été réalisée sur les corpus des campagnes DUC 2005-07. ENERTEX a obtenu des bonnes performances par rapport aux participants, sans avoir besoin aux ressources externes ou à des lourds post-traitements linguistiques.

Nous avons exploré aussi le champ de la recherche d'information guidée par des annotations. Nous avons adapté nos systèmes à un changement macroscopique d'échelle et à l'introduction des impuretés dans le réseau textuel. Il a fallu ajouter à ENERTEX l'action combinée de deux fonctions (agissant sur les termes rares et/ou trop communs comme les annotations). Cela a permis d'obtenir des résultats préliminaires intéressants.

Le chapitre 5 est consacré aux tâches de segmentation thématique et de classification documentaire. Notre stratégie de détection de ruptures thématiques dans les documents a consisté à la comparaison, par le test de Kendall, des spectres énergétiques des phrases. L'introduction d'une longueur de corrélation a permis de modifier l'allure des courbes afin que le test de Kendall puisse mieux les identifier. Les résultats obtenus sont intéressants et comparables à ceux d'autres systèmes du domaine. Nous avons proposé aussi une méthode de classification de documents basée sur la capacité discriminatoire des matrices d'échange entre mots. Testée sur le corpus de la campagne DEFT'08, notre approche à obtenu une performance supérieur à la moyenne des participants, même en utilisant une faible fraction du corpus d'apprentissage.

Enfin, dans le chapitre 6 nous avons utilisé le modèle magnétique de verres de spin pour étudier une approche exploratoire de compression de phrases. Les couplages entre termes, et entre leurs étiquettes grammaticales ont été calculés sur un corpus d'apprentissage. Les phrases de test ont été compressées en appliquant les couplages appris à l'aide d'une dynamique thermique de Métropolis. Les compressions, évaluées par rapport à celles faites par des humains, ont des scores BLEU comparables à d'autres systèmes. Nous pensons qu'un corpus plus large peut donner lieu à des règles plus précises. Un aspect intéressant est de faire intervenir dans les simulations le type et la langue des documents. La frustration pourrait être traitée avec des algorithmes comme *simulated tempering* (Newman et Barkema, 1999), bien adaptés aux systèmes type verre de spins.

Dans une perspective optimiste, nous estimons que le système ENERTEX, pour sa capacité de fouiller le contenu informatif du textes et de récupérer des liens peu évidents entre segments de texte, pourrait être appliqué à la découverte à travers la littérature

(ou *Literature Based Discovery* (LBD)). Il s'agit d'extraire de façon automatique de découvertes potentielles, à partir d'une littérature existante, en cherchant à lier les concepts d'articles disjoints (Swanson et al., 2006).

Il serait possible de faire appel au calcul de chemins d'ordre supérieur à niveau de termes pour mettre en valeur les relations faibles entre eux. Ceci sera avantageux en la séparation de graphes intriqués issus des outils de cartographie de réseaux de termes comme TermWatch (Ibekwe-SanJuan et al., 2008c).

Les apports théoriques et pratiques les plus importants de cette thèse, consistent à avoir utilisé comme point de départ le modèle vectoriel pour introduire des modèles simples de la Physique statistique. Cela nous a permis de modeler et de manipuler efficacement l'information contenue dans les textes. À partir de nos expériences et nos résultats, on peut envisager l'utilisation de modèles de la Physique encore plus complexes et plus riches, capables de capturer des propriétés de la langue qui échappent à nos modèles actuels.

La recherche pluridisciplinaire n'est pas toujours facile à réaliser. Cependant, elle donne lieu à des nouvelles idées qui, en plus d'être intéressantes, s'avèrent efficaces. Se priver d'une telle opportunité aurait été vraiment regrettable.

Annexe A

Exemples de textes complets

Textes disponibles à l'adresse : http://www.lia.univ-avignon.fr/fileadmin/documents/Users//Intranet/chercheurs/torres

A.1 3-mélanges

Titre : La politique, le LIA et les trolls.
0 : Le secrétaire général des Nations Unis, KOFI ANNAN, a déclaré que les Nations Unis devaient augmenter leurs efforts vis à vis de la Somalie.
1 : Il a rappelé qu'auparavant les responsables politiques somaliens devraient faire des efforts pour améliorer la sécurité.
2 : Dans un communiqué qu'il a adressé au conseil de sécurité des Nations Unis, il a déclaré, qu'il est nécessaire que les Nations Unis entreprennent des mesures progressives pour s'acquitter de leur devoir vis à vis de la Somalie.
3 : Il a rajouté que ces mesures-actions découlent des récents progrès issus de la conférence pour la paix qui a abouti à la création d'un gouvernement intérimaire et l'élection d'un Président.
4 : Il a appelé les chefs somaliens de profiter de cette occasion pour compléter la constitution d'un état fédéral somalien de transition.
5 : Le secrétaire général des Nations Unis a appelé la communauté internationale à apporter une contribution financière au gouvernement fédéral de transition somalien.
6 : Le nouveau président du gouvernement fédéral de transition somalien, Abdillahi Youssouf Ahmed, a lancé un appel, lors de son investiture la semaine dernière, à la communauté internationale pour qu'elle participe à la reconstruction de la somalie.
7 : Une délégation conduite par le secrétaire du bureau des affaires étrangères et économiques de la Grande Bretagne, Chris Mullin, a rencontré aujourd'hui le nouveau président somalien et des membres du parlement fédéral de transition de la somalie.
8 : La délégation anglaise, a félicité le parlement de la manière démocratique qu'ils ont élu le président, le président du parlement et le vice-président, ceci constitue un pas vers la démocratie.
9 : De manière générale, les objectifs scientifiques du LIA concernent le traitement automatique du langage naturel (écrit et oral, et tout particulièrement l'interprétation des contenus linguistiques de messages,

de dialogues ou de textes), l'optimisation (programmation en nombres entiers) vise le développement de méthodes numériques spécifiques pour le traitement des langues naturelles et les systèmes de télécommunication dans lesquels ces méthodologies peuvent être appliquées ou intégrées.
10 : Il s'agit notamment d'élaborer sans a priori méthodologique, des connaissances, des modèles et des techniques permettant de représenter l'ensemble des informations disponibles dans de grandes quantités de données de parole, de textes ou de séquences audiovisuelles.
11 : La représentation et le traitement de ces informations sont effectués au moyen d'outils réalisés au laboratoire ou choisis pour leur efficacité dans le contexte de nos travaux (probabilités et statistiques, recherche opérationnelle, réseaux de neurones formels, heuristiques, réseaux et systèmes distribués).
12 : Les travaux réalisés ou envisagés, s'ils concernent fréquemment des difficultés pratiques posés par le traitement de très grandes quantités d'informations complexes (disponibles sur les réseaux et dans des bases de données), sont abordés de manière à mettre en lumière des problématiques qui généralisent ces questions spécifiques.
13 : Des interactions permanentes et approfondies (sous la forme de contrats de recherche ou de bourses) avec d'importants groupes industriels nationaux ou internationaux permettent d'assurer une meilleure continuité entre les travaux académiques et leurs applications rapides dans les systèmes opérationnels réalisés par nos partenaires.
14 : Nos activités scientifiques et leurs retombées pratiques sont systématiquement et régulièrement confrontées aux travaux concurrents dans le cadre des évaluations internationales des systèmes de traitement automatique des données.
15 : Le découpage en équipes du laboratoire, s'il facilite la classification de nos thèmes de recherche, reste très souple de manière à ce que les solutions élaborées dans un cadre particulier puissent être immédiatement appliquées avec profit dans un autre contexte.
16 : Pour chacune des équipes on peut effectuer un inventaire non exhaustif des grandes catégories de problèmes qui sont abordés.
17 : Dans la mythologie nordique, les trolls sont des êtres vivants dans les montagnes ou les buttes.
18 : Ce sont des géants incarnant les forces naturelles, au même titre que les Titans.
19 : Odin avait dû tuer Ymir, le géant dont il était né, pour assurer le règne des dieux et des hommes, selon un scénario rappelant la castration d'Ouranos par son fils Cronos et la victoire des dieux olympiens sur les Titans.
20 : Les trolls étaient des géants qui avaient surgi du corps d'Ymir ; ils symbolisent les forces naturelles dans leur énergie élémentaire.
21 : La christianisation de la Scandinavie a profondément diminué la taille des trolls et altéré la réputation de ces êtres qui étaient jadis plutôt considérés comme bêtes et naïfs que comme malfaisants.
22 : Il est maintenant claire que n'importe quelle créature peux être un Troll.
23 : Dans le jeu de rôle Donjons et Dragons, les trolls sont des géants solitaires à l'appétit insatiable.
24 : Ces créatures presque deux fois plus grandes qu'uH.
25 : Les trolls n'ont peur que du feu, leurs capacités de régénération hors du commun leur permettent de se remettre en quelques instants de n'importe quelle blessure à l'exception de celles dues aux flammes et aux acides.
26 : Les membres tranchés d'un troll repoussent en quelques heures et seule la décapitation ou l'incinération peuvent en venir à bout.

A.2 Hurricaine Gilbert

Les phrases 0-2 ne contiennent pas d'information significative.
3 : BC-HurricaineGilbert, 09-11 339.
4 : BC-Hurricaine Gilbert, 0348.
5 : Hurricaine Gilbert heads toward Dominican Coast.
6 : By Ruddy Gonzalez.
7 : Associated Press Writer.
8 : Santo Domingo, Dominican Republic (AP).
9 : Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10 : The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11 : "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12 : Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13 : An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14 : Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15 : The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16 : The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17 : The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18 : Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19 : There were no reports on casualties.
20 : San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21 : On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22 : Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23 : Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24 : The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

A.3 Tibet

0. La police a procédé aux arrestations lors d'une manifestation à Lhassa, coïncidant avec le 49e anniversaire du départ forcé du dalaï lama.
1. Une soixantaine de moines bouddhistes ont été arrêtés lundi à Lhassa, la capitale du Tibet, à l'occasion d'une manifestation coincidant avec le 49e anniversaire du départ forcé du dalaï lama, a affirmé mardi Radio Free Asia (RFA).
2. Entre 50 et 60 manifestants ont été arrêtés par la police par les forces de l'ordre, qui ont également bloqué les routes et encerclé les monastères pour empêcher les manifestations de se propager.
3. Cependant, onze personnes ont réussi à protester dans le centre de Lhassa avant d'être arrêtées, selon les mêmes sources citées par RFA.
4. Des responsables de la police et des affaires religieuses à Lhassa ont refusé de s'exprimer.
5. Le dignitaire religieux de 72 ans, qui a fui le Tibet en 1959 après l'échec d'un soulèvement anti-chinois, a abandonné ses revendications d'indépendance, se bornant à réclamer "une large autonomie" pour sauvegarder la langue, la culture et l'environnement de ce territoire himalayen.
6. La Chine, qui en a pris le contrôle à partir de 1950 -avant d'y mener une sanglante répression- n'a cessé de rejeter ces demandes qualifiées par le dalaï lama de diplomatie de la "voie moyenne".
7. Depuis lundi, des moines bouddhistes manifestent au Tibet et dans les régions avoisinantes, à l'occasion du 49e anniversaire du soulèvement de Lhassa, qui a conduit à l'exil du dalaï-lama.
8. Depuis Dharamsala, dans le nord de l'Inde, le dalaï-lama a demandé à Pékin de "renoncer à l'usage de la force" contre les manifestants.
9. Son porte-parole a jugé sans fondement les accusations chinoises selon lesquelles il aurait fomenté les manifestations violentes.
10. Ce nouvel embrasement, dans une région sensible, sous contrôle chinois depuis 1951, devrait accentuer la pression que subit déjà le gouvernement chinois pour améliorer les droits de l'homme, comme il s'est engagé à le faire en obtenant l'organisation des JO de Pékin, dont l'ouverture aura lieu dans cinq mois.
11. Le gouvernement chinois a proposé d'indemniser les familles des civils qui ont, selon lui, été tués lors des violences dans la capitale tibétaine ce mois-ci, a rapporté vendredi soir l'agence de presse officielle chinoise Chine nouvelle.
12. Selon le décompte du gouvernement chinois, 18 civils ont été tués le 14 mars lors de manifestations contre la tutelle chinoise à Lhassa, au cours desquelles des manifestants ont lancé des pierres en direction des forces de l'ordre, brûlé et pillé des magasins.
13. Les familles des victimes recevront 200.000 yuans (18.000 euros), a indiqué Chine nouvelle sur la foi d'une circulaire du gouvernement régional du Tibet.
14. "Des mesures sont prises pour aider les gens à réparer leurs maisons et leurs magasins détruits pendant les troubles ou à en construire d'autres", précise la circulaire selon Chine nouvelle.
15. Toute personne blessée lors des émeutes pourra être soignée gratuitement, a ajouté Chine nouvelle.
16. Le bilan officiel de deux semaines de violences au Tibet et dans l'ouest de la Chine est de 19 morts, mais le gouvernement tibétain en exil fait état de 140 morts.
17. La répression des émeutes par les autorités chinoises a provoqué des protestations internationales à l'approche des Jeux olympiques de Pékin.

A.4 2-mélanges (informatique et puces)

0. Et si l'ordinateur pouvait fonctionner un jour, sans électricité ou presque ? La démarche de chercheurs américains de l'université de Notre Dame, dans l'Indiana, montre que l'on peut manipuler des électrons pour construire des circuits élémentaires avec des quantités d'énergie infimes.
1. Leurs expériences, relatées dans l'édition du 9 avril du magazine Science, ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées que celles des puces actuelles qui sont bridées par des problèmes de dissipation de chaleur.
2. Les travaux de l'équipe dirigée par Greg Snider portent sur le puits quantique, un piège infinitésimal dans lequel un électron peut être enfermé.
3. Les scientifiques ont créé des cellules carrées formées de quatre puits quantiques, dans laquelle ils ont introduit une paire d'électrons.
4. Les forces de répulsion provoquent le déplacement des électrons qui trouvent leur équilibre lorsqu'ils se trouvent placés aux deux extrémités de l'une ou l'autre des diagonales de la cellule.
5. La première représente l'état 0, tandis que l'autre indique le 1.
6. Chaque cellule représente donc un bit, la plus petite quantité d'information que l'on peut manipuler dans les ordinateurs.
7. Tout déplacement d'un électron sous l'effet d'une force extérieure provoque automatiquement le déplacement du second électron de manière à retrouver l'équilibre, et donc le basculement de la cellule entre les états 0 et 1.
8. L'utilisation d'une cellule unique ne prouve rien.
9. Les chercheurs américains ont réussi à en assembler plusieurs, provoquant, suivant leurs besoins, le déplacement des électrons sans devoir fournir d'énergie, ou presque.
10. Dans les transistors actuels, le passage de l'état 0 à l'état 1 n'est possible qu'au prix du déplacement de plusieurs milliers d'électrons, ce qui génère un important flux de chaleur.
11. En regroupant cinq cellules élémentaires, les chercheurs ont mis au point un circuit baptisé "majoritaire" capable de réaliser les deux fonctions logiques de base, ET et OU, à la demande.
12. Ils ont ensuite vérifié son bon fonctionnement et espèrent assembler plusieurs de ces circuits pour effectuer des additions et des multiplications sur des nombres.
13. En cas de succès, la technique des cellules logiques quantiques pourrait permettre d'entasser des centaines de milliards de circuits dans une seule puce électronique.
14. Pour l'instant, le dispositif fonctionne seulement à une température voisine du zéro absolu, mais les chercheurs ne désespèrent pas de parvenir à le réchauffer tout en maîtrisant son comportement.
15. Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis parles puces et les poux.
16. Des piqûres de puces ont été relevées sur plus d'untiers des militaires.
17. On a aussi retrouvé des cadavres de poux sur 3 militaires.
18. Des mesures d'urgence ont été prises en conséquence.
19. Des piqûres de puces ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213.
20. Des cadavres de poux, mais aucun oeuf, ont également été décelés sur 3 militaires.
21. Ces insectes sont transmis par contact personnel.
22. La cause de cette invasion n'est pas claire ; ces insectes semblent toutefois avoir essaimé à partir du local de garde.
23. Le médecin de troupe a donné immédiatement les soins nécessaires aux militaires concernés et il a

ordonné les mesures d'hygiène qui s'imposaient.
24. Des produits spéciaux ont été remis pour les soins corporels.
25. Tout le matériel personnel de la compagnie a été emballé hermétiquement et apporté à l'arsenal cantonal de Fribourg.
26. La troupe sera déplacée dans un complexe industriel.
27. Une section d'hygiène de l'école de recrues d'hôpital 268, stationnée à Moudon, va désinfecter tous ces cantonnements.
28. On estime qu'avec ces mesures sanitaires appropriées la troupe pourra réintégrer ses cantonnements vendredi au plus tard.

A.5 *Experiencias de las parteras de Kaua Yucatán* (extrait)

Des expériences des sage femmes au Kaua Yucatán, texte proportionné par le Profr. Miguel Güémez du Centre de Recherches Régionales (CIR) en Sciences Sociales de l'Université Autonome de Yucatán (UADY). La traduction de l'espagnol au maya a été réalisé par Feliciano Sánchez Chan.

Langue : maya
Titre : Péek yéetel jets'eknak : bey u tukulta'al wíinklal.
0 : Ku ya'alik u kajnáalilo'ob K'auae' u wíinklal máake' junp'éel ba'ax tu'ux ku táakpajal ya'abach ba'alo'ob jach k'a'ana'an u yantal tuláakal tu kúuchil, tumen chen ja'alil bey ku yantal toj óolal ti' máako'.
1 : Ku ts'o'okole', ichil xan u jobnel máake' yaan ba'alo'ob ku tukulta'al ku péeko'obi'.
2 : Ko'ox a'alike', junp'éel wa ba'ax ti' máake' ku péek ka'alikil u ch'úuya'al wa ba'ax aal, wa ku lúubul máak tu k'aan, ka'alikil u meyaj máak ich kool wa beeyxan ikil u yáalintikubáa.
3 : Wa ku péek wa ba'ax ti' u wíinklal máake', ma' chen u káat ya'al u jóobol bix tsola'anil u wíinklali', ku taasik muk'yajo'ob wa u jejelasil k'oja'anilo'ob chen ku yutstalo'ob wa ku ka'a su'utul tu kúuchil
4 : Lela' chen ku béeytal yéetel yoot'.

Langue : espagnol
Titre : Cuerpo : movilidad y equilibrio.
0 : Según los habitantes de Kaua, el cuerpo se presenta como un sistema integrado en el que cada órgano tiene una posición propia que debe ser mantenida para preservar el estado de salud.
1 : Asimismo, los órganos internos son percibidos como partes móviles.
2 : Un órgano puede moverse si se alzan cosas muy pesadas, por una caída de la hamaca, mientras se trabaja en la milpa o se hacen sobreesfuerzos durante el parto.
3 : El desplazamiento de un órgano no incide sólo en la percepción de la conformación del cuerpo, que se desordena o descompone, sino que provoca malestares y enfermedades de distinta naturaleza que desaparecen solamente cuando el mismo órgano es regresado a la posición considerada correcta.
4 : El reposicionamiento de los órganos se produce exclusivamente a través de sobadas.

Annexe B

Différentes collaborations en plusieurs langues

B.1 Compréhension vs. extraction

Il existe des systèmes dont le but est d'apprendre les liens sémantiques entre unités textuelles afin de simuler les processus cognitifs humains. Nous présentons une analyse comparative entre une approche cognitive et ENERTEX. L'expérience a été faite sur un cadre très spécial qui nous a permis d'avoir de nombreuses références humaines classées par niveaux scolaires.

Dans une étude précédente (Lemaire et al., 2005), des sélections de phrases importantes (dont la concaténation produit des extraits) et des résumés rédigés ont été produits par des élèves allant de la 4ème à la 1ère année de différents établissements français. Nous avons soumis le même protocole aux étudiants en Master 2 (M2) d'informatique. Un feuillet d'exercices comprenant les deux tâches a été ainsi distribué : a) souligner de 3 à 5 phrases estimées les plus importantes ; b) résumer les textes. Les documents proposés ont été MIGUEL DE LA FAIM [1] (Miguel) et LA PHARMACIE DES ÉLÉPHANTS [2] (Eléphants). Nous obtenons au total 296 extraits et 372 résumés distribuées selon le tableau B.1.
L'évaluation de la qualité des résumés automatiques reste un problème ouvert. Une façon de palier à ce problème, consiste à comparer les résumés produits automatiquement à ceux produits par un certain nombre de juges humains. Plus nombreuses sont les références, plus fiable est notre comparaison. Cependant, pour des raisons évidentes de temps et de manque de disponibilité, les juges humains ne sont pas toujours très nombreux (les conférences DUC en offrent de 2 à 5). Pour cette raison, le fait d'avoir réuni plus de 600 références offre une possibilité d'évaluation très intéressante. De plus, les références sont classées par niveau scolaire ce qui offre de possibilités d'analyse supplémentaires.

1. Vidal, N. (1984). Miguel de la faim. Paris : Rageot.
2. Pfeffer, P. (1989). Les pharmacies des éléphants. In Vie et mort d'un géant.

Niveau scolaire		Nombre de références			
		MIGUEL 382 mots, 24 phrases		ELÉPHANTS 523 mots, 18 phrases	
		Abstracts	**Extracts**	**Abstracts**	**Extracts**
Collège (13-15 ans)	4ème	29	23	29	22
	3ème	39	24	40	34
Lycée (16-18 ans)	2ème	67	48	86	71
	1ère	22	14	19	20
Lycée professionnel	CAP	6	6	7	6
Université	M2	14	14	14	14
Total		177	129	195	167

TABLE *B.1* – *Description du corpus produit par des élèves et étudiants.*

Jugements des élèves et étudiants

Issu d'une collaboration avec le Laboratoire des Sciences de l'Éducation de Grenoble (LSE) [3], ce travail a comme objectif principal de comparer notre approche pour extraction avec l'approche cognitive proposée pour l'équipe grenobloise. Cette dernière est basée sur un algorithme d'analyse sémantique latente (LSA) conçu à l'origine pour étudier les processus cognitifs de résumé. LSA (Landauer et Dumais, 1997) est un modèle qui permet de représenter la sémantique à partir de l'hypothèse que : i) deux mots sont proches s'ils apparaissent dans des contextes similaires et ii) deux contextes sont similaires s'ils contiennent des mots proches. Pour LSA, les mots d'un vaste corpus sont représentés dans une matrice d'occurrences. Cette matrice stocke, pour chaque mot, les contextes dans lesquels les mots apparaissent ainsi que leur fréquence d'apparition. La résolution d'un tel système est réalisée par une décomposition en valeurs singulières. Le nombre de dimensions de la matrice diagonale (SVD) est généralement fixé à 300 (Landauer et Dumais, 1997). La proximité sémantique entre mots peut se mesurer en estimant le cosinus. La proximité sémantique de deux phrases se mesurera alors en estimant le cosinus de l'angle formé par les vecteurs sommes des vecteurs des mots qui composent chaque phrase.

LSA nécessite en entrée un corpus de taille importante. On a utilisé trois corpus différents : LSA_lemonde, généraliste d'environ 5 millions de mots ; LSA_enfants avec 3,3 millions de mots (productions d'enfants, contes, manuels scolaires et encyclopédies pour enfants) ; et LSA_adultes, de 13 millions de mots (rassemblant une large partie des deux corpus précédents en plus de romans pour adultes). En plus de ENERTEX et LSA, nous avons inclut dans cette comparaison les systèmes : CORTEX (Torres-Moreno et al., 2001), COPERNIC [4], PERTINENCE [5] et MICROSOFT WORD [6]. Nous avons également ajouté deux *baselines* : *Baseline 1* (sélection aléatoire des phrases) et *Baseline 2* (phrases du début et de la fin du texte). Les systèmes ont produit des résumés de cinq phrases

3. http://web.upmf-grenoble.fr/sciedu
4. http://www.copernic.com/fr/products/summarizer/index.html
5. http://www.pertinence.net
6. http://www.microsoft.com

identifiées comme les plus importantes. Nous avons réalisée une évaluation à deux niveaux :

i) Au niveau du texte où l'on repère les phrases estimées importantes à la fois par les systèmes et par les humains ;

ii) au niveau de n-grammes de mots en utilisant ROUGE.

Avec une lecture échantillonnée des *abstracts* [7], nous avons constaté que les étudiants des niveaux scolaires supérieurs ont une idée plus claire de comment rédiger un résumé. Mais dans l'étape préalable de sélection des phrases pertinentes on n'a pas retrouvé des différences importantes en fonction des niveaux scolaires.

On observe dans la figure B.1 que pour le texte Éléphants, les étudiants M2, $1^{ère}$, $2^{ème}$ et $3^{ème}$ sélectionnent par consensus les phrases 1, 7, 11 et 18 comme les plus importantes. Les CAP sont d'accord uniquement pour trois phrases de cet ensemble (1, 7 et 18) et les 4ème seulement pour une (1). Pour le texte Miguel le consensus des classes est similaire. Les humains ont une tendance générale à repérer les mêmes phrases per-

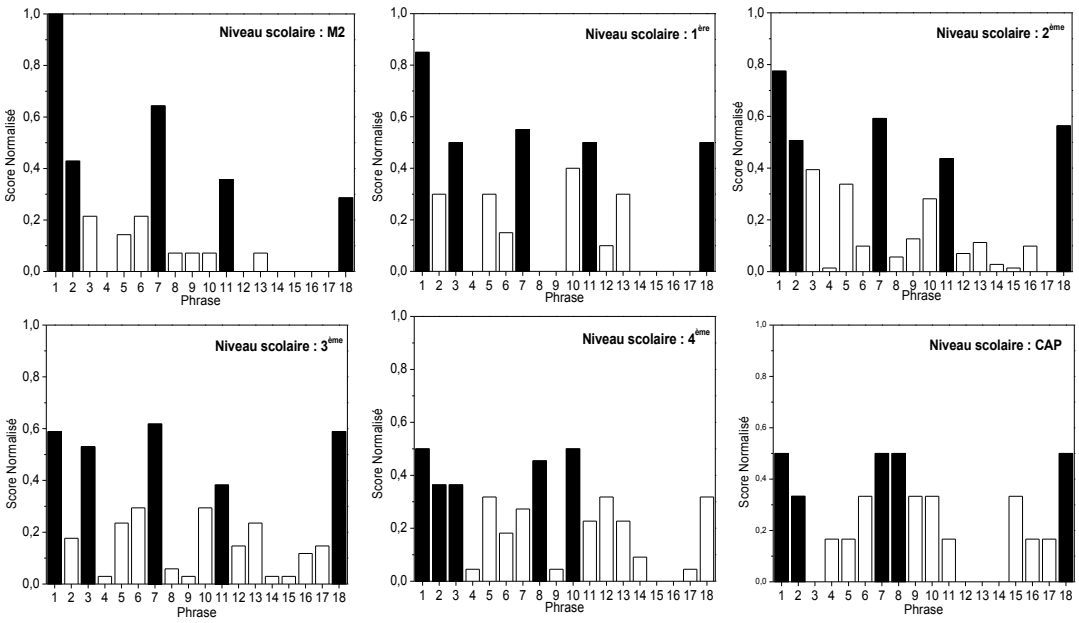

FIGURE *B.1* – *Phrases sélectionnées par les humains des différents niveaux scolaires pour le texte Eléphants (18 phrases). Les bars pleins représentent les cinq phrases avec les meilleurs scores.*

tinentes (quatre des cinq phrases demandés). Nous avons calculé la précision (équation

7. Les analyses manuelles ont été faites par Patricia Velázquez (Fernández et al., 2008b).

B-1) des systèmes par rapport aux quatre phrases sélectionnées par le consensus.

$$Précision = \frac{Nb.\ de\ phrases\ pertinentes\ retrouvées}{Nb.\ de\ phrases\ rapportées} \quad \text{(B-1)}$$

Ceci donne un indicateur de la performance du système au niveau de la phrase. Pour Éléphants, les systèmes CORTEX , LSA_adultes, LSA_lemonde et ENERTEX ont une précision de 0,75 ; LSA_enfants, COPERNIC et WORD de 0,5 ; et PERTINENCE de 0. Pour le texte Miguel, CORTEX , LSA_adultes et LSA_enfants, ENERTEX , COPERNIC et PERTINENCE ont une précision de 0,5 ; LSA_lemonde de 0,25 et WORD de 0. A cette échelle, LSA_adultes, CORTEX et ENERTEX sont les meilleurs systèmes.

Nous avons mené une évaluation plus fine avec ROUGE.Cela permet de relever les petites différences entre niveaux (les phrases qui ne font pas partie du consensus). La figure 4.2 présente la performance finale des systèmes comme la moyenne entre la précision normalisée et le produit moyen des rappels ROUGE-2×SU4. Cette quantité a été normalisée sur les deux textes. LSA_adultes, CORTEX et ENERTEX obtiennent les meilleures performances.

À quel niveau scolaire peuvent être situés les systèmes automatiques ?

Les approches par extraction se limitent à sélectionner et extraire des phrases. Elles sont souvent appelées « les mauvais élèves ». En revanche celles par compréhension sont de « bons élèves ». Mais, les systèmes par extraction de phrases sont-ils forcément de mauvais élèves ? Essayer de répondre à cette question a été le deuxième objectif de ce travail. Nous avons mesuré la proximité de résumés générés automatiquement avec les extraits séparés par niveaux scolaires (collégiens, lycéens et étudiants universitaires).

Pour chaque système et chaque niveau scolaire nous avons calculé le produit ROUGE-2 × SU4. On observe dans le tableau 4.5 que les résumés qui correspondent aux meilleurs systèmes sont plus similaires aux extraits des étudiants de 1ère et M2. Par contre, les systèmes avec des piètres performances sont plus proches des collégiens. Il est intéressant d'observer que PERTINENCE est sensible au type des documents. Il est le meilleur pour résumer le texte narratif Miguel et un des derniers pour le texte explicatif Éléphants. Ce résultat peut s'expliquer si l'on considère qu'un texte narratif, en proposant une structure linéaire d'événements, est censé être plus aisément résumable qu'un texte explicatif, qui contient des concepts plus abstraits (Brewer, 1980). Ainsi, on peut déduire que malgré les désavantages des résumés par extraction (manque de traitement des références et des anaphores), cette approche permet de condenser le texte en conservant les informations les plus importantes. ENERTEX utilisent un modèle physique pour sélectionner les phrases pertinentes des textes. Les évaluations ont montré que la performance du système cognitif LSA(adultes) est comparable à celle d'ENERTEX. Les résumés générés par les systèmes les plus performants, insensibles au type de document : LSA(adultes), CORTEX et ENERTEX . Ils sont les plus similaires à ceux des étudiants les plus avancés (1ère et M2). Les systèmes par extraction ne sont donc pas si mauvais élèves qu'on pourrait le croire.

B.2 Un résumeur hybride

Une collaboration avec l'IULA [8] (*Institut Universitari de Lingüística Aplicada*) de l'Université Pompeu Fabra, Barcelone Espagne avait pour objectif de combiner des méthodes des statistiques (ENERTEX et CORTEX) et linguistiques (DISICOSUM) pour résumer des articles médicaux en espagnol. Le système DISICOSUM , proposé par l'équipe espagnole, se fonde sur l'hypothèse que dans les domaines spécialisés, les professionnels utilisent des techniques concrètes pour résumer leurs documents. Après une analyse manuelle d'un ensemble d'articles médicaux, des règles ont été déduites et intégrées à DISICOSUM (da Cunha et al., 2007).

Description du système hybride

ENERTEX et CORTEX ont été combinés avec le système d'extracteur de termes YATE (da Cunha et al., 2007). Les termes extraits pour YATE censés représenter les concepts les plus importants des articles, ont reçu un poids plus important dans les calculs (par un facteur de 10). La figure B.2 montre l'architecture globale de l'approche proposé. D'abord, quelques règles linguistiques de DISICOSUM sont appliquées sur le do-

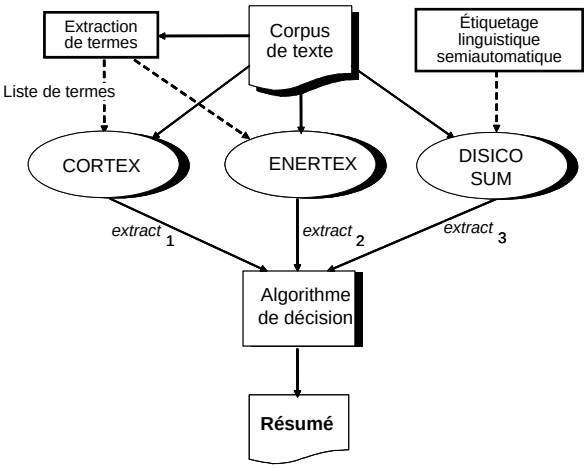

FIGURE *B.2* – *Système de résumé hybride combinant des approches numériques et linguistiques.*

cument original. La sortie est un document plus court qui contient uniquement des phrases principales libres d'information supplémentaire. Ce document a été résumé séparément par CORTEX , ENERTEX et DISICOSUM . La sortie est analysée par un algorithme de décision qui garde les phrases à extraire pour le résumé final :
i) D'abord, l'algorithme choisi les phrases sélectionnées par le consensus des trois systèmes ;
ii) si le consensus n'existe pas, on choisit les phrases sélectionnées par deux systèmes ;

8. http://www.iula.upf.edu

iii) enfin, s'il y a des phrases choisies par un seul système, la priorité est donnée à celles ayant le plus grand score.

Évaluation

Pour analyser la performance de l'approche hybride, nous avons utilisé un corpus de test de 10 articles médicaux fournis par l'IULA. Les articles ont été analysés Nous avons procédé à l'évaluation avec ROUGE en utilisant comme références les résumés rédigés par les auteurs. Deux *baselines* aléatoires ont aussi été inclues. La première a été extraite du document original et la deuxième du document réduit par l'élimination des phrases accessoires détectées par DISICOSUM . Les résultats, selon la médiane du score ROUGE sur les dix articles analysés, sont affichés au tableau 4.6.

Nous observons que le système hybride améliore la proximité avec l'auteur. Les performances individuelles des trois systèmes sont similaires mais inférieures à celle du système hybride. Il semblerait que la combinaison des techniques statistiques et linguistiques accumule les avantages des systèmes isolés.

B.3 Résumé en langues à structure éloignée

L'objectif de ces expériences est de montrer que les algorithmes numériques du TAL, en utilisant une représentation vectorielle, restent très indépendant de la langue. C'est pour cette raison qu'ils sont capables de gérer des langues à structure éloignée du français ou l'espagnol, comme le somali et le maya. Nous avons utilisé deux corpus contenant des textes parallèles. Dans le premier, chaque phrase en français est alignée avec sa traduction au somalienne, et dans le deuxième, des phrases en espagnol sont alignées aux correspondantes en maya. Nous avons résumé ces documents avec les systèmes CORTEX et ENERTEX.

B.3.1 Le français et le somali

Le corpus utilisé contient les documents en français PARLEMENT (7 phrases), ONU (14 phrases), PRINCE-1 (37 phrases) et leurs traductions correspondantes en somali. Pour résumer ces textes nous avons suivi la démarche suivante :

1. Les textes ont été filtrés et normalisés selon les procédures décrites en section 2.3.3. Les outils pour la traduction et le pretraitement des textes en somali (la liste de mots fonctionnels et les règles pour la normalisation du vocabulaire) ont été conçues par des chercheurs de l'Institut des Sciences et des Nouvelles Technologies de Djibouti (ISNT) [9].
2. Nous avons généré des résumés de ces textes avec CORTEX et ENERTEX. Le taux de compression a été du 25% du nombre de phrases.

Comme illustration, les condensés produits pour le texte PRINCE-1 en français et somali sont présentés dans les figures B.3, B.4, B.5 et B.6. On peut remarquer la pertinence des phrases retenues en français. Nous avons observé le même résultat pour tous les autres textes en langue française. Dans le cas du somali, la sélection des phrases a été évaluée par un chercheur de l'ISNT, spécialiste en cette langue. Il a constaté la pertinence des phrases choisies par les systèmes.

Comparaison des systèmes

Nous avons comparé statistiquement la performance des deux systèmes (en somali et en français) avec le test de concordance de Kendall (Siegel et Castellan, 1988) qui mesure le degré de concordance parmi k juges qui classent un ensemble de P objets [10].

9. L'ISNT est un des cinq instituts du Centre d'Étude et de Recherche de Djibouti (CERD), http://www.cerd.dj où sont développes des outils TAL pour des langues peu dotées de ressources informatisées. Le but est de sauvegarder et valoriser le patrimoine culturel africain. Des exemples de ces outils se trouvent en (Abdillahi et al., 2006).
10. Le test de Kendall a été expliquée en la section 5.3.1

Résumé automatique CORTEX**-FRANCAIS :** Lorsque j'avais six ans j'ai vu une image dans un livre sur la Forêt Vierge s'appelant Histoires Vécues. Ca représentait un serpent boa avalant un fauve. J'ai réfléchi sur les aventures de la jungle et j'ai réussi avec un crayon de couleur à tracer mon premier dessin. J'ai montré mon chef-d'oeuvre aux adultes et je leur ai demandé si mon dessin leur faisait peur. Mon dessin ne représentait pas un chapeau. J'ai dessiné l'intérieur du serpent boa afin que les adultes puissent comprendre. Les adultes m'ont conseillé d'abandonner les dessins de serpents boas ouverts ou fermés et de m'intéresser à la géographie à l'histoire au calcul et à la grammaire. J'ai abandonné à l'âge de six ans une carrière de peinture. Et je ne lui parlais ni de serpents boas ni de forêts vierges ni d'étoiles.

FIGURE *B.3* – *Résumé* CORTEX*-FRANCAIS pour le texte* PRINCE-1

Résumé automatique ENERTEX**-FRANCAIS :** Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Forêt Vierge s'appelant Histoires Vécues. Ca représentait un serpent boa qui avalait un fauve. On disait dans le livre : " Les serpents boas avalent leur proie tout entière, sans la mâcher. J'ai montré mon chef-d'oeuvre aux grandes personnes et je leur ai demandé si mon dessin leur faisait peur. Mon dessin ne représentait pas un chapeau. J'ai alors dessiné l'intérieur du serpent boa, afin que les grandes personnes puissent comprendre. Mon dessin numéro deux était comme cela : dessin numéro deux. Les grandes personnes m'ont conseillé de laisser de côté les dessins de serpents boas ouverts ou fermés, et de m'intéresser plutôt à la géographie, à l'histoire, au calcul et à la grammaire. J'avais été découragé par l'insuccès de mon dessin numéro un et de mon dessin numéro deux.

FIGURE *B.4* – *Résumé* ENERTEX*-FRANCAIS pour le texte* PRINCE-1

Résumé automatique CORTEX**-SOMALI :** Markii aan jiray lix sanadood waxaan arkay maalin sawir qurux badan oo ku yaala buug ka hadlaya kaynta hawdka oo magaciisu ahaa Taariikh Nololeed. Markaan sawirkaa arkay aad ayaan uga fikiray arimaha ka dhaca kaynta isla markiina waxaan ku guuleystay aniga oo isticmaalay qalin nashqad leh inaan soo saaro sawirkaygii ugu horeeyay. Waxaan tusay farshaxankaygii dadkii iga weyna oo waxaan weydiiyay inay ka cabsanayaan sawirkayga. Waxaan sawirkii aad ugu muujiyay gudaha jabisada oo furan si dadka waaweyni u fahmaan. Waxay markaa dadkii waaweynaayi igu dardaareen inaan iskaga hadho sawiradan jabisooyinka furan ama xidhan oo waxay igula taliyeen inaan isku taxaluujiyo jiqoraafiga taariikhda xisaabta iyo naxwaha. Dadka waaweyni waxba ma fahmaan keligood arintaasina way daalisa ubadka oo markasta macno bixiya. Markaan la kulmo qof weyn oo garaad leh waxaan ku tijaabin jiray sawirkaygii kowaad oo aan haystay. Aniguna Waan iska dhaafi jiray oo kama aan hadli jirin jabisooyinka iyo kaynta hawdka iyo xidigaha.

FIGURE *B.5* – *Résumé* CORTEX*-SOMALI pour le texte* PRINCE-1

Résumé automatique ENERTEX**-SOMALI :** Markii aan jiray lix sanadood waxaan arkay, maalin, sawir qurux badan, oo ku yaala buug, ka hadlaya kaynta hawdka, oo magaciisu ahaa « Taariikh Nololeed ». Markaan sawirkaa arkay, aad ayaan uga fikiray arimaha ka dhaca kaynta, isla markiina, waxaan ku guuleystay, aniga oo isticmaalay qalin nashqad leh, inaan soo saaro sawirkaygii ugu horeeyay. Waxaan sawirkii aad ugu muujiyay gudaha jabisada oo furan si dadka waaweyni u fahmaan. Sawirkaygii labaad wuxuu ahaa sidan : Sawirkii labaad. Waxay markaa dadkii waaweynaayi igu dardareen inaan iskaga hadho sawiradan jabisooyinka furan ama xidhan, oo waxay igula taliyeen inaan isku taxaluujiyo jiqoraafiga, taariikhda, xisaabta iyo naxwaha. Waxaa niyada iga dilay, fahamwaagi sawirkaygii kowaad iyo sawirkaygii labaad. Dadka waaweyni, waxba ma fahmaan keligood, arintaasina way daalisa ubadka oo markasta macno bixiya. Waxa i soo maray, noloshayda, kulamo badan oo ah dad mudakar ah. Aniguna, Waan iska dhaafi jiray, oo kama aan hadli jirin jabisooyinka, iyo kaynta hawdka iyo xidigaha.

FIGURE *B.6* – *Résumé* ENERTEX*-SOMALI pour le texte* PRINCE-1

Dans le tableau B.2, nous montrons pour les textes étudiés, la colonne Kendall avec l'accord entre les juges (CORTEX et ENERTEX) et la p valeur associée, qui étant très petite indique que les systèmes ont choisi presque toujours les mêmes phrases pertinentes.

Texte	Kendall	p-value
parlement-fr	0,96847	⊘
baarlamanka-som	1,00000	⊘
ONU-fr	0,95105	0,02506
ONU-somali	0,91958	0,04227
prince-1-fr	0,88879	0,01087
prince-1-somali	0,97832	0,00339

TABLE *B.2* – *Comparaison de résumés produits par ENERTEX et CORTEX. La colonne Kendall correspond au calcul de l'indice τ. Lorsqu'il est proche de 1 cela signifie que les classements sont presque les mêmes. La p-valeur donne la probabilité que les deux classements soient indépendants.*

B.3.2 L'espagnol et le maya

Un intérêt personnel, car je suis née à Yucatán [11] au Mexique, m'a motivé à essayer les outils du TAL sur la langue maya. Il s'agit d'utiliser le système numérique ENERTEX pour produire les résumés automatiques d'un corpus parallèle maya-espagnol contenant une centaine de lignes traitant sur les pratiques des sages femmes mayas dans le village de Kaua Yucatán [12]. Un extrait de ce document est montré dans la section A.5.

11. Yucatán est une des principales régions d'Amérique où la civilisation Maya s'est développée pendant les premiers siècles de notre ère ; une importante population maya habite encore cette région.

12. Le texte *Experiencias de las parteras de Kaua Yucatán* a été proportionné par le Profr. Miguel Güémez du Centre de Recherches Régionales (CIR) en Sciences Sociales de l'Université Autonome de Yucatán (UADY). La traduction de l'espagnol au maya a été réalisé par Feliciano Sánchez Chan.

Maya	Espagnol	Français
aktáan	*delante*	devant
ba'ax	*qué*	quoi
ba'ale'	*pero*	mais
beyo'	*así*	comme ça
hunp'éel	*uno*	un
ka'ap'éel	*dos*	deux
óoxp'éel	*tres*	trois
kanp'éel	*cuatro*	quatre
ho'p'éel	*cinco*	cinq
náach	*lejos*	loin
naats'	*cerca*	proche
te'elo'	*allí*	là-bas
ti'	*a*	à

TABLE *B.3* – *Quelques exemples des mots fonctionnels en maya et leurs traductions en espagnol et français.*

Pretraitement des textes

Étant donné qu'il n'existe pas des outils permettant le pretraitement des textes en langue maya (filtrage et normalisation), nous avons suivi la procédure suivante :

1. Construction d'une liste de mots fonctionnels (articles, prépositions, adverbes, verbes auxiliaires) à partir du dictionnaire maya-espagnol et du cours de langue maya du site web *Identidad y Cultura Maya* [13]. Cette liste contient les mots qui seront éliminés des textes dans la phrase de filtrage car ils sont considérés non porteurs d'information essentielle. Cela nous a permis une première réduction de la taille du vocabulaire. Des exemples des mots fonctionnels sont présentés dans le tableau B.3.
2. Normalisation du vocabulaire (pour réduire encore plus la taille du vocabulaire). Pour cette tâche, nous avons utilisé le logiciel de racinisation construit par Rachid Nabi et Jalal Bouhafer au LIA [14]. Il s'agit d'un système de racinisation par acquisition de familles morphologiques en utilisant des méthodes d'apprentissage non supervisé. L'approche consiste à former des familles par regroupements successifs. Les critères de regroupement reposent sur la similarité graphique des mots ainsi que sur des méthodes vectoriels. Nous pouvons observer des exemples de ces regroupements dans le tableau B.4.

13. Le site web http://www.mayas.uady.mx, sur l'identité et la culture du peuple maya, a été construit en avril du 2000 par initiative du Profr. Miguel Güémez du CIR Sociales de la UADY, projet dans lequel j'ai eu l'opportunité de collaborer de façon très étroite.

14. Le projet « Acquisition et racinisation statistique de familles morphologiques » a été développé par Rachid Nabi et Jalal Bouhafer pendant leur stage de Master en Informatique au LIA d'Avignon en 2008 sous la co-direction de Juan Manuel Torres et moi même (http://projets-gmi.iup.univ-avignon.fr/projets/proj0708/M1/p28/).

Famille morphologique en maya	Racine	Traduction espagnol/français
k'aas - k'aasi' - k'aasil	*k'aas*	*feo*/moche
k'iin - k'iino'ob - k'iine'	*k'iin*	*sol*/soleil
ko'olelo' - ko'olelilo'ob - ko'olelil - ko'olel	*ko'olel*	*dama*/dame
suuka'anile' - suuka'anil - suuk	*suuk*	*costumbre*/habitude
yotoche' - yotoch	*yotoch*	*casa*/maison

TABLE *B.4* – *Quelques exemples des regroupements morphologiques en maya. On prend comme racine commune le mot le plus court.*

Résumés par ENERTEX

Le but de cette expérience est de vérifier si un système numérique d'analyse textuel comme ENERTEX est capable de produire un résumé pertinent des textes en langue maya. La démarche a été la suivante :

1. Le corpus est composé des sept petits textes en maya traitant différents aspects liés à l'activité des sages femmes mayas. Nous avons aussi leur traduction en espagnol. Chacun de ces ensembles de textes ont été réuni en deux documents d'environ une centaine des phrases.
2. Nous avons filtré et normalisé les deux textes. Pour cette tâche nous avons utilisé les outils décrits dans la section précédente.
3. Nous avons utilisé ENERTEX pour produire les deux résumés, l'un en maya et l'autre en espagnol.

L'évaluation a été fait de façon manuelle. Un natif de la langue a constaté la pertinence de l'information extraite en espagnol. De plus, le condensé est équilibré : ENERTEX a choisi des phrases importantes venant des toutes les sept parties du document originale. Les pourcentages du résumé appartenant à chaque partie sont environ : 14%, 20%, 9%, 14%, 20%, 9% et 14%.

La qualité du résumé en maya a été évalué de façon indirecte au travers de la comparaison phrase à phrase avec le résumé en espagnol. D'une telle démarche, nous avons observé une concordance de 86% entre les résumés en maya et espagnol.

B.3.3 Conclusion

Les résultats obtenus dans les sections B.3.1 et B.3.2 ont mis en évidence la capacité des systèmes numériques d'analyser des textes dont la structure s'éloigne de celles du français, l'anglais ou l'espagnol. Pour cela, nous avons choisi des corpus en deux langues vivantes : le somali, parlé en Afrique ; et le maya, utilisé dans une vaste région d'Amérique. Un tel analyse est possible grâce au traitement vectoriel qui est à la base d'approches comme ENERTEX et CORTEX et qui permettent aux algorithmes une considérable indépendance de la langue.

Annexe C

Changement d'échelle et dopage du réseau textuel

C.1 La recherche d'information guidée par des annotations

Aider les utilisateurs à localiser une catégorie spécifique d'information a rarement été abordé dans la communauté dédiée à la Recherche d'Information (IR) en raison de la difficulté de cette tâche (Ibekwe-SanJuan et al., 2008b). Pour cela, il est commun d'utiliser des collections de documents annotées où les phrases, ou segments pertinents, ont été étiquetées selon le type d'information qu'elles portent : hypothèses, conclusions, résultats, etc. Cette démarche est particulièrement utile dans l'analyse de textes scientifiques dont la structure est bien déterminée. Une fois identifié le rôle de chaque segment, choisir les plus importants et les extraire devient plus facile (Teufel, 1999; Teufel et Moens, 2002; Orasan, 2001).

L'objectif de ce travail, réalisé en collaboration avec le *College of Information Sciences* [1] de l'Université de Drexel à Philadelphie est d'utiliser des annotations sémantiques pour aider un chercheur ou spécialiste à accéder à une catégorie particulière d'information dans les textes scientifiques. Nous avons utilisé le corpus SDSS du domaine de l'astronomie. Il est composé de 1293 *abstracts* issus de la base de données *Web of Science* (WoS) de l'ISI [2], tous contenant le terme « *Sloan Digital Sky Survey* » [3]. C'est important de signaler que, en raison du manque de références, l'évaluation de cette partie a été faite manuellement sur critères de pertinence et de portée de l'information trouvée.

1. http://www.ischool.drexel.edu
2. *Institute for Scientific Information*.
3. http://www.sdss.org

C.2 Des phrases aux *abstracts*

Dans les expériences qui font partie du chapitre 4, l'énergie textuelle a montré son efficacité pour capturer les relations pas toujours évidentes entre les phrases à l'intérieur d'un texte ou bien entre les phrases et une requête. Calculer leur énergie permet de donner un score de pertinence aux phrases. Si les documents à scorer, vis-à-vis d'une requête, ne sont pas des phrases mais des résumés (*abstracts*), est ce que l'énergie textuelle a encore un mot à dire ? La réponse est non. Nous avons constaté que les résumés plus énergétiques face à une requête ne sont pas nécessairement les plus pertinents. En fait, ENERTEX semble ignorer les concepts utilisés dans la requête et le résultat est bien plus proche d'un résumé générique.

L'explication de ce comportement se trouve dans la distribution fréquentielle des mots dans ces petits textes. Un *abstract* est constitué d'environ 250 mots où l'auteur essai d'exprimer le contexte, ses motivations, méthodes, résultats et conclusions. Ce type de document produit une distribution particulière : des termes rares mais importants, comme les noms de modèles ou les nouvelles techniques proposées ; ou des termes trop communs faisant partie du contexte. Pour adapter la mesure d'énergie à ce nouveau contexte, nous proposons une fonction de pondération f sur des paires {terme, *abstract*}. Cette fonction est basée sur l'indice d'équivalence, produit des probabilités conditionnelles $P(s/w)$ et $P(w/s)$ entre un *abstract* s et un terme w. Seules les valeurs au dessus d'un seuil de 10^{-n}, où n dépend de la taille du corpus, ont été considérées. Ainsi nous avons fixé :

$$f(w,s) = \log\left[trunc\left(\left(\frac{f_{w,s}^2}{f_{w,\cdot} \times f_{\cdot,s}} > 10^{-n}\right) \times 10^n\right)\right] \tag{C-1}$$

où $f_{w,s}$ est la fréquence du terme w dans s, $f_{w,\cdot}$ est la fréquence de w dans le corpus et $f_{\cdot,s}$ est le nombre de termes dans la phrase s. Pour optimiser la vitesse de l'algorithme de classement, nous avons tronqué les nombres réels pour travailler uniquement sur des valeurs entières. Nous avons limité les valeurs élevées avec la fonction logarithme. En appliquant l'équation C-1 sur les éléments de la matrice terme-segment, ENERTEX peut alors guider la recherche avec les termes importants, toujours en capturant leurs relations directes et indirectes. Ceci a produit des résultats intéressants. Dans les paragraphes suivants nous présentons quelques exemples.

Termes à occurrence unique dans le corpus

À titre d'exemple, considérons la requête *Randall-Sundrum*. Ce terme correspond au nom d'un modèle de géométrie de l'espace qui n'apparaît qu'une seule fois dans le corpus SDSS. En utilisant la fonction de pondération C-1, ENERTEX classe en première position le résumé contenant *Randall-Sundrum* puis ceux d'autres modèles géométriques. Il est intéressant de noter que l'algorithme a trouvé la relation entre les termes de la requête et son contexte. Des exemples de termes pertinents dans ce contexte sont : *geometry, space plat, dimension, inflation, expansion, brane, braneworld, DGP model*. Nous montrons dans la figure C.1 un de ces résumés, classé en $7^{ième}$ position. Les termes pertinents

ont été soulignés. Ceci est similaire à une procédure d'expansion de requêtes [4] dans les-

> Two new one-parameter tracking behavior dark energy representations omega=omega(0)/(1+z) and omega=omega(0)e(z/(1+z))/(1+z) are used to probe the geometry of the Universe and the property of dark energy. The combined [RESULT] type Ia supernova, Sloan Digital Sky Survey, and Wilkinson Microwave Anisotropy Probe data indicate that the Universe is almost spatially flat and that dark energy contributes about 72% of the matter content of the present universe. The observational data also tell us that omega(0)similar to-1. It is argued that [FINDING] the current observational data can hardly distinguish different dark energy models to the zeroth order. The transition redshift when the expansion of the Universe changed from deceleration phase to acceleration phase is around z(T)similar to 0.6 by using our one-parameter dark energy models.

FIGURE *C.1* – *Un des abstracts classés parmi les premières places pour la requête Randall-Sundrum à ocurrence unique dans le corpus SDSS. Les termes pertinents ont été soulignés.*

quelles les termes de haut rang sont utilisés pour élargir les termes de la requête. La différence ici est qu'ENERTEX sélectionne les résumés les mieux classés pour étendre la recherche sur la matrice d'énergie.

Les acronymes

Nous avons réalisé un deuxième type d'expérience qui consiste à construire des requêtes d'un seul mot avec des acronymes de termes d'astronomie tels que ΔCDM, AGB, AMIGA, LBG. Le but est d'observer si le système est capable d'extraire la définition de l'acronyme et l'information proche. Les termes considérés ont une fréquence très basse mais sont des abréviations de concepts importants du corpus. En nous basant sur la définition des acronymes, prises de différentes sources, l'évaluation a consisté à comparer les termes de telles définitions qui apparaissent dans les résumés sélectionnés par le système. Par exemple, en analysant le contenu des premiers 48 résumés sélectionnés où la requête contient l'acronyme AGB, nous observons :

- la présence du terme utilisé dans la requête (AGB) ;
- la présence des termes scientifiques présents dans la définition du terme de la requête AGB (*Asymptotic Giant Branch*) et de son contexte [5]. Des exemples sont *Asymptotic Giant Branch*, *Life*, *Core*, *Non-LTE*, *Convection atmosphere*, *stratosphere*, *chemical evolution*.

ENERTEX a trouvé des relations entre l'entité de la requête et des concepts reliés en se basant sur le contexte dans lequel on trouve cette entité. Cette même analyse a été réalisée systématiquement sur une vingtaine de requêtes. Au moins 93% des résumés classés dans les 50 premières places contiennent des termes importants, contenus aussi dans des glossaires spécialisés. Ceci montre que l'énergie textuelle est un moyen simple de détecter des relations sémantiques cachées entre les termes d'un corpus formé par des résumés scientifiques.

4. L'expansion d'une requête consiste à utiliser des mots ou des phrases reliés pour étendre la recherche. Il s'agit de considérer qu'un sens peut être porté par des mots différents (Ruch et al., 2006).

5. Prises du site web http://www.eso.org/projects/vlti/science/node8.html

Il faut dire que les termes pertinents trouvés par une requête particulière ne sont plus présents quand on utilise une autre requête. C'est-à-dire qu'il ne s'agit pas de termes généraux mais en accord avec les termes utilisés pour guider la recherche.

C.3 Introduction d'annotations sémantiques

Pour mieux aider les utilisateurs à localiser une catégorie spécifique d'information, nous avons introduit un deuxième facteur dans le système textuel. Il s'agit des étiquettes sémantiques qui guideront la recherche d'information. Les annotations sémantiques sont des étiquettes introduites dans le corpus pour indiquer le type de renseignement porté par les phrases. Nous travaillons avec sept catégories d'information : *RESULT, CONCLUSION, FUTURE WORK, NEWTHING, OBJECTIVE, RELATED WORK, HYPOTHESIS*. L'introduction de ces étiquettes a été réalisée au moyen d'un automate à états finis. Pour les détails de cette étape, nous renvoyons à (Ibekwe-SanJuan et al., 2008b). Dans la figure C.2 à gauche, nous montrons un exemple de texte annoté. Cela se traduit par une sorte de dopage de la matrice terme segment (figure C.2 à droite).

La présence de ces impuretés dans le réseau textuel entraîne-t-elle une modification des interactions entre les termes ? Dans notre cas, il se trouve que le dopage effectué sur

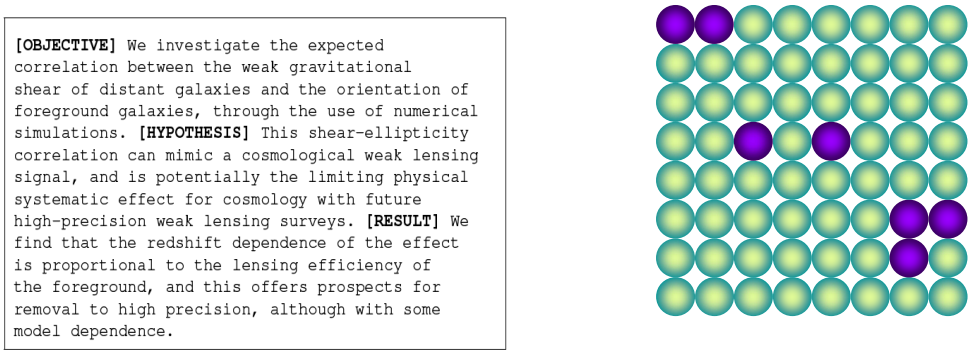

FIGURE C.2 – *À gauche, un exemple de texte annoté. On peut assimiler les étiquettes à des impuretés introduites dans le réseau textuel (à droite). Sa présence peut affecter les propriétés du matériau hôte.*

le réseau textuel (le corpus SDSS) est important. Les étiquettes constituent plus du 20% du vocabulaire du corpus. Nous avons observé que ces éléments étrangers perturbent le comportement habituel du système et empêchent de bien guider les recherches. Quand les étiquettes sont présentes dans les requêtes accompagnant des termes scientifiques, le système répond avec des informations non pertinentes. Il semble que les impuretés introduites sont trop nombreuses. On dirait que le dosage n'est pas correct et qu'il faut en retirer. Mais cela n'est pas approprié, car toutes les phrases ont la même priorité dans le processus d'étiquetage.

Requête : NEWTHING spectral classification of quasar.
Termes pertinents attendus : luminosity, quasar, redshift, quasar spectra, spectrum, optical, Balmer, eigen, Fe II emission, Baldwin.
Abstract identifié comme pertinent par le système :
[*NEWTHING*] Discovery of three new quasars and the spatial density of luminous quasars at z similar to 6. Quasars with intrinsically red (optically steep) power-law continua tend to have narrower Balmer lines and weaker C IV, C III], He II, and 3000 & ANGS; bump emission as compared with bluer (optically flatter) quasars.

FIGURE C.3 – *Recherche d'information guidé par des termes et étiquettes.*

Pour surmonter cette contrainte, nous avons implémenté un dosage alternatif basé sur la pondération des fréquences. La difficulté est que les étiquettes suivent un comportement fréquentiel différent de celui des termes courants. Par définition, les résumés scientifiques suivent l'hypothèse d'être bien formées [6] tandis que chaque étiquette de catégorie sémantique aura tendance à être répartie de manière uniforme dans l'ensemble du corpus et sera donc un événement de haute probabilité. Ainsi, quand on les considère comme des mots, les étiquettes sont tout simplement ignorées par la fonction de pondération introduite dans la section précédente. Nous proposons donc une deuxième fonction de pondération g mesurant combien la fréquence d'un mot est plus grande que prévu :

$$g(w,s) = \log\left[trunc\left(\frac{(f_{w,s} - \overline{f_{w,\cdot}} > 0)^2}{\sum_{t\in S}(f_{w,t} - \overline{f_{w,\cdot}} > 0)^2} \times f_{w,\cdot}\right) > 0\right] \tag{C-2}$$

Cette fonction $g(w,s)$ compare la fréquence $f_{w,s}$ du mot ou étiquette w avec la fréquence moyenne $\overline{f_{w,\cdot}}$ de ce terme dans les résumés. t est un résumé de l'ensemble total S. Á nouveau, nous avons optimisé l'algorithme en tronquant les nombres réels pour travailler sur valeurs entières et on limite les grandes valeurs avec la fonction logarithme. En raison de la taille du corpus, les tests statistiques complexes ont été écartés, et la plupart des calculs ont été fait sur des entiers. Finalement, nous avons combiné les fonctions f (equation C-1) et g (équation C-2) en prenant leur produit $(f(u)+1)\times(g(u)+1)$. Avec ce double effet, il est possible de guider la recherche d'information avec la combinaison de termes scientifiques et d'étiquettes sémantiques.

C.4 Expériences et discussion : requêtes à termes et étiquettes

Nous avons construit une vingtaine des requêtes avec différentes combinaisons de termes et d'étiquettes. ENERTEX a ordonné les résumés en accord avec leur pertinence par rapport à chaque requête.

L'évaluation a été la suivante : d'un côté, les résumés classés en premières positions doivent contenir les termes attendus. Ceci veut dire, ceux qui se trouvent dans les

6. Conformément aux règles grammaticales et syntaxiques.

sources scientifiques (articles, sites web) traitant le sujet. D'un autre côté, les étiquettes de la requête doivent aussi être présentes dans les résumés retenus.

En analysant les premiers 50 résumés obtenus pour chaque requête, nous avons observé qu'ENERTEX a toujours bien guidé la recherche. Un exemple est présenté dans la figure C.3 où la requête, les termes pertinents attendus et un des *abstracts* classé comme pertinent sont affichés. ENERTEX a été ainsi adapté pour prendre en compte les annotations sémantiques présentes dans le corpus et la requête. Ce travail constitue une nouvelle approche pour résoudre les problématiques de la recherche d'information avancée.

Annexe D

Le test de concordance τ de Kendall

D.1 Description

Le test de concordance τ de Kendall a été introduit par (Kendall, 1938) et permet de mesurer la concordance entre les classements, sous forme de rangs, établie par deux juges ou experts sur le même ensemble d'objets. Il appartient à un type de tests appelés non-paramétriques qui ne requièrent aucune supposition sur une éventuelle distribution des données (Siegel et Castellan, 1988). Nous allons illustrer son fonctionnement par une exemple inspiré par (Abdi, 2007). Soit S un ensemble de $N = 4$ vins :

$$S = \{a, b, c, d\} \tag{D-1}$$

Deux experts ont ordonné ces vins selon un certain critère commun. Un ensemble ordonné peut également être représenté par le rang donné aux objets. Ainsi, le premier expert a donné l'ordre $O_1 = [a, c, b, d]$ ce qui produit les rangs $R_1 = [1, 3, 2, 4]$ et le deuxième expert les a ordonné selon $O_2 = [a, c, d, b]$ qui correspond aux rangs $R_2 = [1, 3, 4, 2]$.

Le coefficient de corrélation τ permet d'évaluer la similarité entre les deux rangs R_1 et R_2 selon le nombre d'inversions de paires d'objets qui seraient nécessaires pour transformer un rang dans l'autre. Pour faciliter un tel analyse, les classements sont représentés par l'ensemble des paires d'objets et une valeur de 1 ou 0 est affecté à cette paire lors de son ordonnance correspond ou ne correspond pas à la façon dont ces deux objets ont été classés. Un ensemble ordonné O_i de N objets, peut être décomposé en $1/2N(N-1)$ paires ordonnés. Par exemple, O_1 est composé des 6 paires suivants :

$$P_1 = \{[a, c], [a, b], [a, d], [c, b], [c, d], [b, d]\}, \tag{D-2}$$

et O_2 des 6 paires :

$$P_2 = \{[a,c],[a,d],[a,b],[c,d],[c,b],[d,b]\}. \tag{D-3}$$

Afin de comparer les deux ensembles ordonnés, l'approche de Kendall compte le nombre de paires différents entre eux. On observe que entre P_1 et P_2 il y a deux paires différents :

$$\{[b,d],[d,b]\}. \tag{D-4}$$

Ce chiffre, notée d_Δ, est appelé la différence symétrique entre ensembles. Alors, entre P_1 et P_2 :

$$d_\Delta(P_1,P_2) = 2 \tag{D-5}$$

Le coefficient de corrélation τ de Kendall est obtenu en normalisant la différence symétrique en sorte d'avoir des valeurs comprises entre -1 et $+1$. La plus grande distance possible est -1 (obtenue quand un ordre est l'inverse de l'autre) et la plus petite est $+1$ (obtenue lorsque les deux ordres sont identiques). Si les classements sont totalement indépendants τ vaut 0. Le nombre maximum de paires qui peuvent varier entre deux séries de $1/2N(N-1)$ éléments est égal à $N(N-1)$ alors, le coefficient de corrélation des rangs de Kendall est égal à :

$$\tau = \frac{1/2N(N-1) - d_\Delta(P1,P2)}{1/2N(N-1)} = 1 - \frac{2 \times [d_\Delta(P1,P2)]}{N(N-1)} \tag{D-6}$$

que pour les deux experts en vin vaut :

$$\tau = 1 - \frac{2 \times 2}{12} = 1 - \frac{1}{3} = 0,67 \tag{D-7}$$

Étant donné que le coefficient de concordance τ est fondée sur le nombre de paires différentes entre deux ensembles classés, son interprétation peut être aussi formulée dans un contexte probabiliste. Pour une paire d'objets pris au hasard, τ est la différence entre la probabilité que les objets soient dans le même ordre ($P(\text{même})$) et la probabilité qu'ils soient dans un autre ordre ($P(\text{différents})$). En suivant notre exemple avec P_1 et P_2 :

$$\tau = P(\text{même}) - P(\text{différents}) = \frac{10}{12} - \frac{2}{12} = \frac{8}{12} = 0,67 \tag{D-8}$$

Cette grande valeur de τ semble indiquer que les deux experts sont fortement d'accord sur leurs évaluations sur les 4 vins. La question qui se pose maintenant est : un tel indicateur est-t-il la preuve suffisant pour établir l'accord entre les classements ? Pour y répondre, il faut calculer la p-valeur associée à cette valeur de τ.

D.2 La p-valeur et le test de signification

Pour juger la signifiance d'une valeur observée de τ, il est utile de savoir si une telle valeur aurait pu être obtenu par hasard à partir d'un univers dans lequel tous les rangs possibles des N objets occurrent le même nombre de fois. Pour cela, il est nécessaire de connaître la distribution de τ sous ces conditions. On appelé p-valeur à la probabilité de trouver, dans cet univers de référence, une valeur de τ égale ou supérieure à celle qu'on a mesuré sur les observations réeles. Si la p-valeur est plus grande qu'un certain seuil on peut conclure que le τ mesuré n'a pas de signification car qu'il peut être dû au hasard. Le seuil ou niveau de signification est couramment dénoté par la lettre grecque α. Dans la pratique des test statistiques les valeurs courantes de α sont $0,05$ (significatif), $0,01$ (très significatif), et $0,001$ (hautement significatif) (Saporta, 1990). Il est cependant indispensable de choisir une de ces valeurs *à priori*.

L'univers de référence

Étant donné que le coefficient τ dépend uniquement de l'ordre des paires, il peut toujours être calculé en supposant que l'un des classement sert de point de référence. Par exemple, pour $N = 4$ éléments, nous supposons arbitrairement que la référence est 1234. Par conséquent, avec des ordres de rang N sur les objets, il existe $N!$ différents résultats possibles (correspondant chacun à un ordre possible) à prendre en compte pour le calcul de la distribution d'échantillonnage de τ. À titre d'illustration, le tableau D.1 montre l'ensemble des $N! = 4 \times 3 \times 2 = 24$ possibles rangs pour un ensemble de $N = 4$ objets avec sa valeur τ d'écart à l'ordre canonique (c'est-à-dire, 1234).

De ce tableau, nous pouvons calculer la probabilité (p-valeur) associée à chaque valeur possible de τ. Par exemple, pour notre exemple des deux experts de vin, la p-valeur associée à $\tau = \frac{2}{3} = 0,67$ est égale à :

$$p = P(\tau \geq \frac{2}{3}) = \frac{\text{Nombre de } \tau \geq \frac{2}{3}}{\text{Nombre total de } \tau} = \frac{4}{24} = 0,17 \tag{D-9}$$

Ce résultat indique que dans 17 sur 100 des cas on pourrait obtenir une valeur de $\tau = \frac{2}{3}$ de manière entièrement aléatoire. La décision finale dépend de la valeur de signification α choisie.

Fixer α à $0,05$ indique qu'on est disposé à accepter un accord entre les experts seulement si la valeur de τ observée a moins de 5 sur 100 chances d'arriver par hasard. Or, on a obtenu qu'elle peut arriver dans 17 sur 100 des cas, alors, on ne peut pas accepter l'existence d'un accord entre les rangs assignés aux 4 vins par les deux experts. Et si on décide de l'accepter, on prend un risque de 17% de se tromper car ces événements ocurrent quand même dues au hasard dans les 17 sur 100 des cas.

On pourrait suivre le même raisonnement au travers d'un test par hypothèses.

	Rangs																							
1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4
2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	1	1	2	2	3	3	3	3
3	4	2	4	2	3	3	4	1	4	1	3	2	4	1	4	1	2	3	1	3	1	2	1	2
4	3	4	2	3	2	4	3	4	1	3	1	4	2	4	1	2	1	2	3	1	2	1	2	1
	2	2	1	1	1	2	1	1			1	1			1	1	2	1	1	2	2			
τ 1	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$-\frac{1}{3}$	$\frac{1}{3}$	0	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{2}{3}$		-1

TABLE *D.1* – *Les* 24 *ordres possibles pour* $N = 4$ *objets et les valeurs de* τ *calculées selon leurs corrélations avec l'ordre canonique* 1234.

Les tests par hypothèses

Il est habituel de traiter un test statistique comme un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon. On appelle H_0 et H_1 ces hypothèses dont une et une seule est vraie. La décision aboutira à choisir H_0 ou H_1. Dans notre exemple, on peut définir :

H_0 : les classements O_1 est O_2 sont indépendants (dues au hasard) ;
H_1 : il existe une corrélation entre les classements O_1 et O_2 (non due au hasard).

H_0 est appelée l'hypothèse nulle (ou d'indépendance) et H_1 l'hypothèse alternative (ou de corrélation). Cette dernière est généralement l'hypothèse qu'on cherche à tester.

La p-valeur mesure le risque d'établir une corrélation fausse entre les classements O_1 et O_2, c'est-à-dire, elle estime la probabilité de se tromper en choisissant H_1 alors que H_0 est vraie. Le seuil de signification α doit être fixée *a priori*. Il indique le niveau maximum de risque qu'on est prêt à prendre. Si la p-valeur est inférieure à α, on rejette l'hypothèse nulle c'est-à-dire on accepte que l'échantillon donne des éléments de preuve suffisants pour soutenir l'hypothèse alternative. Dans notre exemple, la p-valeur (0,17) est plus grande que α (0,05) et par conséquent, l'accord entre les deux experts ne peut pas être établie.

Quand le nombre d'objets classés est supérieur à 10, le calcul de la distribution d'échantillonnage de τ est toujours possible car elle est finie. Cependant, il comprend le calcul de $N!$ coefficients de corrélation, ce que, pour de grandes (et même pas trop grandes) valeurs de N, devient compliqué et dans certains cas, impossible.

Or ce problème n'est pas si drastique parce que pour $N > 10$, la distribution normal est une bonne approximation à la distribution d'échantillonnage de τ (Sheskin, 2004; Abdi, 2007). La demostration d'une telle convergence peut être consulte dans (Kendall, 1938). Dans ce cas, la moyenne est 0 et la variance :

$$\sigma_\tau^2 = \frac{2(2N+5)}{9N(N-1)} \tag{D-10}$$

La p-value peut donc être calculée en utilisant les tables correspondantes au niveau de signification désirée.Ces tables sont souvent intégrées dans les logiciels statistiques.

Liste des illustrations

Liste des tableaux

Liste de publications personnelles

Revues internationales avec comité de lecture

(Dumesnil et al., 2009)

K. Dumesnil, C. Dufour, S. Fernandez A. Avisou, M. Oudich, A. Rogalev, F. Wilhelm et E. Snoeck

Low temperature exchange bias in $[DyFe_2/YFe_2]$ superlattices : effect of the thermo-magnetic preparation

Publié dans *Journal of Physics : Condensed Matter*, 2009

Abstract : *The effect of the thermo-magnetic preparation on exchange-bias is investigated in an exchange coupled [3nm $DyFe_2$/12nm YFe_2]22 superlattice. X-ray Magnetic Circular Dichroïsm (XMCD) experiments at low temperature reveal that exchange bias originates from the quenched $DyFe_2$ magnetization, biasing the unpinned YFe_2 reversal. This quenched configuration might be tailored by changing the cooling field or the magnetic preparation at 300K before zero-field cooling. Changing the amplitude of the cooling field induces interface domains walls and tends to modify the orientation of the pinning moments at the interfaces. This results in the observation of single loops and in the continuous variation of the bias field as a function of the cooling field. A specific magnetic preparation (field cycling) at 300K may induce different remanent states with lateral domains in the pinning layer, which remain unchanged at low temperature after zero-field cooling, and behave independently. This gives rise to combined loops, whose shape reflects the domains populations.*

(Moukarzel et al., 2007)

Cristian F. Moukarzel, Silvia F. Fernández Sabido et J.C. Ruíz-Suárez

Phase transition in liquid drop fragmentation

Publié dans *Physical Review E (Vol.75, No.6)* , 2007

Abstract : A liquid droplet is fragmented by a sudden pressurized-gas blow, and the resulting droplets, adhered to the window of a flatbed scanner, are counted and sized by computerized means. The use of a scanner plus image recognition software enables us to automatically count and size up to tens of thousands of tiny droplets with a smallest detectable volume of approximately 0.02 nl. Upon varying the gas pressure, a critical value is found where the size distribution becomes a pure power law, a fact that is indicative of a phase transition. Away from this transition, the resulting size distributions are well described by Fisher's model at coexistence. It is found that the sign of the surface correction term changes sign, and the apparent power-law exponent τ has a steep minimum, at criticality, as previously reported in nuclear multifragmentation studies. We argue that the observed transition is not percolative, and introduce the concept of dominance in order to characterize it. The dominance probability is found to go to zero sharply at the transition. Simple arguments suggest that the correlation length exponent is $\nu = 1/2$. The sizes of the largest and average fragments, on the other hand, do not go to zero abruptly but behave in a way that appears to be consistent with recent predictions of Ashurst and Holian.

Communications internationales avec actes

(Fernández et Torres-Moreno, 2009)

Silvia Fernández et Juan Manuel Torres-Moreno

Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques

En attente de publication dans *Traitement Automatique des Langues Naturelles* (TALN) 2009

Abstract : *We present an exploratory approach based on thermodynamic concepts of Statistical Physics for sentence compression.We describe the magnetic model of spin glasses, well suited to our conception of problem. The Metropolis simulations allows to introduce thermal fluctuations to drive the compression. Interesting comparisons of our method were performed on a French text corpora.*

(Fernández et al., 2008a)

Enertex : un système basé sur l'énergie textuelle

Silvia Fernández, Eric SanJuan et Juan Manuel Torres-Moreno

Publié dans *Traitement Automatique des Langues Naturelles* (TALN) 2008

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 borders detection in three languages : English, French and Spanish.

(Fernández et al., 2008b)

Silvia Fernández, Patricia Velázquez, Sonia Mandin, Eric SanJuan et Juan Torres-Moreno Manuel

Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?

Publié dans *Actes de Journées internationales d'Analyse statistique des Données Textuelles* (JADT) 2008

Abstract : *We have developed three Automatic Summarization systems (Cortex and Enertex based on the vectorial model, and another based on Latent Semantic Analysis LSA). These systems use methods that do not need linguistic resources. In this work, we confront them to the summaries made by students of different levels (middle and graduated). Finally, it is possible to determine which school level fits better to each system.*

(Fernández et al., 2007b)

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

Silvia Fernández, Eric SanJuan et Juan Manuel Torres-Moreno

Publié dans *6th Mexican International Conference on Artificial Intelligence* (MICAI) 2007

Abstract : *In this paper we present a neural networks approach, inspired by statistical physics of magnetic systems, to study NLP fundamental problems. 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.*

(da Cunha et al., 2007)

Iria da Cunha, Silvia Fernández Patricia Velázquez Morales, Jorge Vivaldi, Eric SanJuan et Juan Manuel Torres Moreno

A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics

Publié dans *Lecture Notes in Artificial Intelligence* (LNAI) 4287, *6th Mexican International Conference on Artificial Intelligence* (MICAI) 2007

Abstract : *In this article we present a hybrid approach for automatic summarization of Spanish medical texts. There are a lot of systems for automatic summarization using statistics or*

linguistics, but only a few of them combining both techniques. Our idea is that to reach a good summary we need to use linguistic aspects of texts, but as well we should benefit of the advantages of statistical techniques. We have integrated the Cortex (Vector Space Model) and Enertex (statistical physics) systems coupled with the Yate term extractor, and the Disicosum system (linguistics). We have compared these systems and afterwards we have integrated them in a hybrid approach. Finally, we have applied this hybrid system over a corpora of medical articles and we have evaluated their performances obtaining good results.

(Fernández et al., 2007a)

Énergie textuelle de mémoires associatives

Silvia Fernández, Eric SanJuan et Juan Manuel Torres-Moreno

Publié dans *Traitement Automatique des Langues Naturelles* (TALN) 2007

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.*

Ateliers

(Ibekwe-SanJuan et al., 2008a)

Fidelia Ibekwe-SanJuan, Silvia F. Fernández Sabido, Eric SanJuan et Eric Charton

Annotation of Scientific Summaries for Information Retrieval

Publié dans les actes de l'atelier *Exploiting Semantic Annotations in Information Retrieval (ESAIR)* 2008

Abstract : *We present a methodology combining surface NLP and Machine Learning techniques for ranking asbtracts and generating summaries based on annotated corpora. The corpora were annotated with meta-semantic tags indicating the category of information a sentence is bearing (objective, findings, newthing, hypothesis, conclusion, future work, related work). The annotated corpus is fed into an automatic summarizer for query-oriented abstract ranking and multiabstract summarization. To adapt the summarizer to these two tasks, two novel weighting functions were devised in order to take into account the distribution of the tags in the corpus. Results, although still preliminary, are encouraging us to pursue this line of work and find better ways of building IR systems that can take into account semantic annotations in a corpus.*

(Charton et al., 2008)

Eric Charton, Nathalie Camelin, Rodrigo Acuna-Agost, Pierre Gotab, Remi Lavalley, Remy Kessler et Silvia Fernández

Pré-traitements classiques ou par analyse distributionnelle : application aux méthodes de classification automatique déployées pour DEFT08

Publié dans *les actes d'atelier DÉfi Fouille de Textes (DEFT)* 2008

Abstract : *In this paper we describe a set of automatic classification methods applied to the DEFT08 campaign. First, we evaluated and compared some of state-of-the-art classifiers like SVM, AdaBoost, probabilistic-based classifiers, and cosine-based classifiers. Subsequently, we developed a method to normalize classes using a distributional analysis of the text with the aim of improving the performance. Lastly, some additional results were obtained by two merging methods that showed to increase the scores of the individual classifiers.*

Soumissions

(Dumesnil et al., 2008)

K. Dumesnil, A. Avisou, S. Fernandez, C. Dufour, A. Rogalev, F. Wilhelm et E. Snoeck

Magnetization reversal in $DyFe_2/YFe_2$ exchange-coupled superlattices

Soumis dans *Magnetism and Magnetic Materials* en 2008

Abstract : *$[DyFe_2/YFe_2]$ superlattices, with a high single crystal quality and rather abrupt interfaces, have been grown by Molecular Beam Epitaxy. The magnetic properties of this hard/soft composite system, the components of which are exchange-coupled at the interfaces, have been investigated in the 10K-300K temperature range, with a specific attention paid to the influence of the soft and hard materials thicknesses. In order to unravel the very rich magnetization reversal processes, conventional susceptibility and magnetization measurements have been combined with element selective X-ray Magnetic Circular Dichroism analysis. The superlattice with thin individual thicknesses ([1nm DyFe2/4nm YFe2]70) reverses as a unique giant ferromagnetic block, the exchange-favored antiparallel arrangement between net magnetization being kept under magnetic field. In the superlattices with rather thick individual thicknesses ($[10nmDyFe_2/13nmYFe_2]18$ and $[10nmDyFe_2/20nmYFe_2]13$), the expected exchange spring behavior develops when the soft YFe_2 layers reverse for positive bias fields, followed by the irreversible switch of the hard $DyFe_2$ layers. In the case of intermediate thickness for the individual $DyFe_2$ layers ($[3nmDyFe_2/12nmYFe_2]22$, $[5nmDyFe_2/20nmYFe_2]13$, $[7nmDyFe_2/28nmYFe_2]10$), the magnetization reversal process strongly depends on temperature. In particular, an unusual magnetization reversal process occurs in the high temperature range where it becomes easier to reverse the hard $DyFe_2$ layers for positive fields, while keeping the dominant YFe_2 magnetization along the field. This phenomenon is attributed to the simultaneous thermal reduction of magnetization density and anisotropy in the $DyFe_2$ layers.*

Bibliographie

(Abdi, 2007) H. Abdi, 2007. *Kendall rank correlation In N.J. Salkind (Ed.) : Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA) : Sage.

(Abdillahi et al., 2006) N. Abdillahi, P. Nocera, et J.-M. Torres-Moreno, 2006. Boîtes à outils tal pour les langues peu informatisées : Le cas du somali. Dans les actes de *JADT 2006*.

(Amini et al., 2000) M.-R. Amini, H. Zaragoza, et P. Gallinari, 2000. Learning for sequence extraction tasks. Dans les actes de *RIAO 2000*, 476–489.

(Bartolozzi et al., 2006) M. Bartolozzi, D. B. Leinweber, et A. W. Thomas, 2006. Scale-free avalanche dynamics in the stock market. *Physica A : Statistical Mechanics and its Applications 370*(1), 132–139.

(Barzilay et Elhadad, 1997) R. Barzilay et M. Elhadad, 1997. Using lexical chains for text summarization. Dans les actes de *ACL Workshop on Intelligent Scalable Text Summarization*, 10–17.

(Bavaud et Xanthos, 2002) F. Bavaud et A. Xanthos, 2002. Thermodynamique et statistique textuelle : concepts et illustrations. *Actes des Journées d'Analyse des Données Textuelles JADT. 1*, 101–111.

(Beaujard et Jardino, 1999) C. Beaujard et M. Jardino, 1999. Classifications de mots non étiquetés par des méthodes statistiques. *Mathématiques et sciences humaines 147*, 7–23.

(Bechet et al., 2000) F. Bechet, A. Nasr, et F. Genet, 2000. Tagging unknown proper names using decision trees. Dans les actes de *ACL '00 : 38th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 77–84. Association for Computational Linguistics.

(Berger et al., 1996) A. L. Berger, S. D. Pietra, et V. J. D. Pietra, 1996. A maximum entropy approach to natural language processing. *Computational Linguistics 22*(1), 39–71.

(Biemann, 2007) C. Biemann, 2007. A Random Text Model for the Generation of Statistical Language Invariants. Dans les actes de *HLT-NAACL-07*, Volume 30, Rochester, NY, USA.

(Binder et al., 2008) K. Binder, W. Paul, T. Strauch, F. Rampf, V. Ivanov, et J. Luettmer-Strathmann, 2008. Phase transitions of single polymer chains and of polymer solutions : insights from Monte Carlo simulations. *J. Phys. : Condens. Matter 20*(49), 494215 (8pp).

(Boudin et Torres-Moreno, 2008) F. Boudin et J.-M. Torres-Moreno, 2008. Efficient organic chemistry summarization. Dans les actes de *GOTAL 08*, Gothenburg, Sweden, 89–99.

(Brants et al., 2002) T. Brants, F. Chen, et I. Tsochantaridis, 2002. Topic-based document segmentation with probabilistic Latent Semantic Analysis. Dans les actes de *CIKM'02*, McLean, Virginia, USA, 211–218.

(Caillet et al., 2004) M. Caillet, J.-F. Pessiot, M. Amini, et P. Gallinari, 2004. Unsupervised learning with term clustering for thematic segmentation of texts. Dans les actes de *RIAO'04*, France, 648–657.

(Castellano et al., 2000) C. Castellano, M. Marsili, et A. Vespignani, 2000. Nonequilibrium phase transition in a model for social influence. *Phys. Rev. Lett. 85*(16), 3536–3539.

(Chaboussant et al., 2004) G. Chaboussant, A. Sieber, S. Ochsenbein, H.-U. Güdel, M. Murrie, A. Honecker, N. Fukushima, et B. Normand, 2004. Exchange interactions and high-energy spin states in Mn12-acetate. *Phys. Rev. B 70*(10), 104422.

(Charton et al., 2008) E. Charton, N. Camelin, R. AcunaAgost, P. Gotab, R. Lavalley, R. Kessler, et S. Fernández, 2008. Pré-traitements classiques ou par analyse distributionnelle : application aux méthodes de classification automatique déployées pour deft08. Dans les actes de *TALN 2008-Atelier DEFT'08*, 161–170.

(Chuang et Chien, 2004) S.-L. Chuang et L.-F. Chien, 2004. A practical Web-based approach to generating Topic hierarchy for Text segments. Dans les actes de *Thirteenth ACM conference on Information and knowledge management*, Washington, D.C., USA, 127–136.

(Clarke et Lapata, 2007) J. Clarke et M. Lapata, 2007. Modelling compression with discourse constraints. Dans les actes de *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*, Prague, 1–11.

(Collobert et al., 2001) R. Collobert, S. Bengio, et C. Williamson, 2001. SVMTorch : Support vector machines for large-scale regression problems. *Journal of Machine Learning Research 1*, 143–160.

(Cortes et Vapnik, 1995) C. Cortes et V. Vapnik, 1995. Support vector networks. Dans les actes de *Machine Learning*, 273–297.

(da Cunha et al., 2007) I. da Cunha, S. Fernández, P. Velázquez Morales, J. Vivaldi, E. SanJuan, et J. M. Torres Moreno, 2007. A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. Dans les actes de *LNAI 4287, MICAI'07, Mexico*, 872–882.

(Dahui et al., 2005) W. Dahui, L. Menghui, et D. Zengru, 2005. True reason for zipf's law in language. *Physica A : Statistical Mechanics and its Applications 358*(2-4), 545–550.

(Dorr et al., 2003) B. Dorr, D. Zajic, et R. Schwartz, 2003. Hedge : A parse-and-trim approach to headline generation. Dans les actes de *HLT-NAACL DUC'03*, Edmonton, Canada, 1–8.

(Dreyfus et al., 1992) G. Dreyfus, J. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria, et L. Hérault, 1992. *Réseaux de neurones. Méthodologie et applications*. France : Eyrolles.

(Dumesnil et al., 2008) K. Dumesnil, A. Avisou, S. Fernandez, C. Dufour, A. Rogalev, F. Wilhelm, et E. Snoeck, 2008. Magnetization reversal in $dyfe_2/yfe_2$ exchange-coupled superlattices. *Magnetism and Magnetic Materials*.

(Dumesnil et al., 2009) K. Dumesnil, C. Dufour, S. Fernandez, A. Avisou, M. Oudich, A. Rogalev, F. Wilhelm, et E. Snoeck, 2009. Low temperature exchange bias in [dyfe2/yfe2] superlattices : effect of the thermo-magnetic preparation. *Journal of Physics : Condensed Matter*.

(Dumesnil et al., 2000) K. Dumesnil, M. Dutheil, C. Dufour, et P. Mangin, 2000. Spring magnet behavior in $dyfe2/yfe2$ laves phases superlattices. *Phys. Rev. B 62*(2), 1136–1140.

(Dupuis et al., 2005) V. Dupuis, F. Bert, J. P. Bouchaud, J. Hammann, F. Ladieu, D. Parker, et E. Vincent, 2005. Aging, rejuvenation and memory phenomena in spin glasses. *P.J. of Physics 64*, 1109.

(Edmundson, 1969) H. P. Edmundson, 1969. New methods in automatic extraction. *Journal of the Association for Computing Machinery 16*(2), 264–285.

(Egawa et al., 2008) S. Egawa, Y. Kato, et S. Matsubara, 2008. Sentence compression by removing recursive structure from parse tree. Dans les actes de *PRICAI'08 : Trends in AI*, Volume 5351, 115–127.

(Farmer, 1999) J. D. Farmer, 1999. Physicists attempt to scale the ivory towers of finance. *Computing in Science and Engineering 1*(6), 26–39.

(Fayol, 1985) M. Fayol, 1985. Analyser et résumer des textes : Une revue des études développementales. *Etudes de Linguistique Appliquée 59*, 54–64.

(Fernández et al., 2007a) S. Fernández, E. SanJuan, et J. M. Torres-Moreno, 2007a. Energie textuelle des mémories associatives. Dans N. H. et Philippe MULLER (Ed.), *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse, 25–34. ATALA : IRIT.

(Fernández et al., 2007b) S. Fernández, E. SanJuan, et J. M. Torres-Moreno, 2007b. Textual energy of associative memories : performants applications of enertex algorithm in text summarization and topic segmentation. Dans les actes de *MICAI '07, Aguascalientes (Mexico)*, 861–871.

(Fernández et al., 2008a) S. Fernández, E. SanJuan, et J. M. Torres-Moreno, 2008a. Enertex : un système basé sur l'énergie textuelle. Dans les actes de *TALN 2008*, 99–108.

(Fernández et Torres-Moreno, 2009) S. Fernández et J. M. Torres-Moreno, 2009. Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques. Dans les actes de *TALN 2009*.

(Fernández et al., 2008b) S. Fernández, P. Velázquez, S. Mandin, E. SanJuan, et J. M. Torres-Moreno, 2008b. Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ? Dans les actes de *Journées internationales d'Analyse statistique des Données Textuelles, JADT 2008*, 469–481.

(Ferrer i Cancho et Solé, 2003) R. Ferrer i Cancho et R. V. Solé, 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America 100*, 788–791.

(Ferret, 2007) O. Ferret, 2007. Finding document topics for improving topic segmentation. Dans les actes de *ACL'07*, 480–487.

(Freund et Schapire, 1996) Y. Freund et R. E. Schapire, 1996. Experiments with a new boosting algorithm. Dans les actes de *Thirteenth International Conference on Machine Learning*, 148–156. Morgan Kaufmann Ed.

(Gagnon et Sylva, 2006) M. Gagnon et L. D. Sylva, 2006. Text compression by syntactic pruning. Dans les actes de *19th Conference of the Canadian Society for Computational Studies of Intelligence (AI06)*, Volume 4013, 312–323. Springer Berlin / Heidelberg.

(Galley et al., 2003) M. Galley, K. R. McKeown, E. FoslerLussier, et H. Jing, 2003. Discourse segmentation of multi-party conversation. Dans les actes de *41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, Sapporo, Japan, 562–569.

(Harremöes et Topsoe, 2005) P. Harremöes et F. Topsoe, 2005. Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics 21*, 315–318.

(Hertz et al., 1991) J. Hertz, A. Krogh, et G. Palmer, 1991. *Introduction to the theorie of Neural Computation*. Redwood City, CA : Addison Wesley.

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

(Ibekwe-SanJuan, 2007) F. Ibekwe-SanJuan, 2007. *Fouille de textes : méthodes, outils et applications*. Paris, France : Paris , Hermes Science Publications, Lavoisier.

(Ibekwe-SanJuan et al., 2008a) F. Ibekwe-SanJuan, S. Fernández, E. SanJuan, et E. Charton, 2008a. Annotation of scientific summaries for information retrieval. Dans les actes de *ECIR-ESAIR 2008, 14p*.

(Ibekwe-SanJuan et al., 2008b) F. Ibekwe-SanJuan, S. Fernández, E. SanJuan, et E. Charton, 2008b. Annotation of Scientific Summaries for Information Retrieval. Dans les actes de *ECIR-ESAIR 2008, 14p*.

(Ibekwe-SanJuan et al., 2008c) F. Ibekwe-SanJuan, E. SanJuan, et M. Vogeley, 2008c. Graph decomposition of terminology graphs for domain knowledge acquisition. Dans les actes de *ACM 17th Conference on Information and Knowledge Management (CIKM 2008), Poster 630*.

(Inoue et Carlucci, 2001) J. Inoue et D. M. Carlucci, 2001. Image restoration using the q-ising spin glass. *Physical Review E 64*(036121), 1–18.

(Ising, 1925) E. Ising, 1925. Beitrag zur Theorie des Ferromagnetismus. *Z. Phys 31*(253-258), 3–4.

(Jaynes, 1957) E. T. Jaynes, 1957. Information theory and statistical mechanics. *Phys. Rev. 106*(4), 620–630.

(Jelinek et al., 1977) F. Jelinek, R. L. Mercer, L. R. Bahl, et J. K. Baker, 1977. Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America 62*, S63. Supplement 1.

(Jing, 2000) H. Jing, 2000. Sentence reduction for automatic text summarization. Dans les actes de *6th Applied Natural Language Processing Conference (ANLP'00)*, 310–315.

(Kendall, 1938) M. G. Kendall, 1938. A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

(Knight et Marcu, 2000) K. Knight et D. Marcu, 2000. Statistics-based summarization - step one : Sentence compression. Dans les actes de *AAAI/IAAI*, 703–710.

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

(Kuhn et De Mori, 1995) R. Kuhn et R. De Mori, 1995. The application of semantic classification trees to natural language understanding. *IEEE Trans. Pattern Anal. Mach. Intell. 17*(5), 449–460.

(Kupiec et al., 1995) J. Kupiec, J. Pedersen, et F. Chen, 1995. A trainable document summarizer. Dans les actes de *18th annual international ACM SIGIR conference on Research and development in information retrieval*, 68–73. ACM Press New York, NY, USA.

(Landauer et Dumais, 1997) T. Landauer et S. Dumais, 1997. A solution to plato's problem : the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review 104*(2), 211–240.

(Landsberg, 1984) P. Landsberg, 1984. Is equilibrium always an entropy maximum ? *Stat. Physics 35*(1-2), 159–169.

(Lebart et Salem, 1994) L. Lebart et A. Salem, 1994. *Statistique Textuelle*. Paris, France : Dunod.

(Lemaire et al., 2005) B. Lemaire, S. Mandin, P. Dessus, et G. Denhière, 2005. Computational cognitive models of summarization assessment skills. Dans les actes de *27th Annual Conference of the Cognitive Science Society. In B.G. Bara, L. Barsalou, & M. Bucciarelli (Eds.)*, 1266–1271.

(Lin et Hovy, 2003) C. Lin et E. Hovy, 2003. The potential and limitations of automatic sentence extraction for summarization. Dans les actes de *Human Language Technologies (HLT-NAACL). Association for Computational Linguistics Morristown, NJ, USA*, 73–80.

(Lin, 2004) C.-Y. Lin, 2004. ROUGE : A Package for Automatic Evaluation of Summaries. Dans S. S. Marie-Francine Moens (Ed.), *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, Barcelona, Spain, 74–81.

(Luhn, 1958) P. Luhn, 1958. Automatic creation of literature abstracts. *IBM Journal of Research and Development 2(2)*, 155–164.

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

(Mandelbrot, 1953) B. Mandelbrot, 1953. An informational theory of the statistical structure of languages. Dans les actes de *Communication Theory, Willis Jackson, Ed.*, New York : Academic Press., 486–502.

(Mandin et al., 2005) S. Mandin, P. Dessus, B. Lemaire, et M. Bianco, 2005. Un EIAH d'aide à la production de résumés de textes. Dans les actes de *Conférence EIAH 2005, P. Tchounikine, M. Joab, and L. Trouche (Eds.)*, 69–80.

(Mani et Mayburi, 1999) I. Mani et M. Mayburi, 1999. *Advances in automatic text summarization*. MIT Press, USA.

(Mann et Thompson, 1987) W. Mann et S. Thompson, 1987. *Rhetorical Structure Theory : A Theory of Text Organization*. University of Southern California, Information Sciences Institute.

(Marucco, 2004) J.-F. Marucco, 2004. *Chimie des solides*. Paris France : EDP Sciences.

(McKeown et Radev, 1995) K. McKeown et D. Radev, 1995. Generating summaries of multiple news articles. Dans les actes de *18^{th} ACM SIGIR*, 74–82.

(Mellet, 2002) S. Mellet, Novembre 2002. Corpus et recherches linguistiques. *Corpus, revues.org 1*, 1–12.

(Mihalcea, 2004) R. Mihalcea, 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Dans les actes de *ACL 2004 on Interactive poster and demonstration sessions*, Morristown, NJ, USA, 20. Association for Computational Linguistics.

(Minel, 2004) J. L. Minel, 2004. Le résumé automatique de textes : solutions et perspectives. Dans les actes de *TAL*, Volume 45/1, 7–13.

(Monod et Prince, 2006) M. Y. Monod et V. Prince, 2006. Compression de phrases par élagage de l'arbre morpho-syntaxique. *Technique et Science Informatiques 25*(4), 437–468.

(Moukarzel et al., 2007) C. Moukarzel, S. Fernández-Sabido, et J. Ruiz-Suárez, 2007. Phase transition in liquid drop fragmentation. *Physical Review E 75*(6), 061127 10p.

(Nadal et Gordon, 2005) J.-P. Nadal et M. B. Gordon, 2005. Physique statistique de phénomènes collectifs en sciences économiques et sociales. *Mathématiques et Sciences Humaines. N° spécial Modèles et méthodes mathématiques dans les sciences sociales : apports et limites 43*(172), 65–88.

(Nestler et al., 2005) B. Nestler, D. Danilov, et P. Galenko, 2005. Crystal growth of pure substances : phase-field simulations in comparison with analytical and experimental results. *J. Comput. Phys. 207*(1), 221–239.

(Newman et Barkema, 1999) M. E. J. Newman et G. T. Barkema, 1999. *Monte Carlo Methods in Statistical Physics*. Great Britain : Clarendon Press - Oxford University Press.

(Nguyen et al., 2004) M. L. Nguyen, S. Horiguchi, A. Shimazu, et B. T. Ho, 2004. Example-based sentence reduction using the hidden markov model. *ACM TALIP 3*(2), 146–158.

(Nowak et al., 2000) M. Nowak, J. Plotkin, et V. Jansen, 2000. The evolution of syntactic communication. *Nature 404*, 495–498.

(Onishi et al., 2003) T. Onishi, D. Yamaki, et K. Yamaguchi, 2003. Theoretical calculations of effective exchange integrals by spin projected and unprojected broken-symmetry methods. I. Cluster models of K2NiF4-type solids. *J. Chem. Phys. 118*(21), 9747–9761.

(Orasan, 2001) C. Orasan, 2001. Patterns in scientific abstracts. Dans L. University (Ed.), *Corpus Linguistics 2001 Conference*, Lancaster, UK, 433–443.

(Page et al., 1998) L. Page, S. Brin, R. Motwani, et T. Winograd, 1998. The pagerank citation ranking : Bringing order to the web. Rapport technique, Stanford Digital Library Technologies Project.

(Paice, 1990) C. D. Paice, 1990. Constructing literature abstracts by computer : techniques and prospects. *Inf. Process. Manage. 26*(1), 171–186.

(Papineni et al., 2001) K. Papineni, S. Roukos, T. Ward, et W. Zhu, 2001. Bleu : a method for automatic evaluation of machine translation.

(Pevzner et Hearst, 2002) L. Pevzner et M. Hearst, 2002. A critique and improvement of an evaluation metric for text segmentation. Dans les actes de *Computational Linguistic*, Volume 1, 19–36.

(Poibeau, 2003) T. Poibeau, 2003. *Extraction Automatique d'information*. Paris, France : Paris, Hermes Science Publications, Lavoisier.

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

(R Development Core Team, 2006) R Development Core Team, 2006. *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

(Ruch et al., 2006) P. Ruch, I. Tbahriti, J. Gobeill, et A. R. Aronson, 2006. Argumentative feedback : a linguistically-motivated term expansion for information retrieval. Dans les actes de *COLING/ACL on Main conference poster sessions*, Morristown, NJ, USA, 675–682.

(Sabah, 2006) G. Sabah, 2006. *Compréhension des langues en interaction (Traité IC2, Série Cognition et Traitement de l'Information)*. Paris, France : Paris, Hermes Science Publications, Lavoisier.

(Salton et McGill, 1983) G. Salton et M. McGill, 1983. *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company.

(Salton et al., 1975) G. Salton, A. Wong, et C. S. Yang, 1975. A vector space model for automatic indexing. *Commun. ACM 18*(11), 613–620.

(Salton et Yang, 1973) G. Salton et C. Yang, 1973. On the specification of term values in automatic indexing. *Journal of Documentation* (29), 351–372.

(Saporta, 1990) G. Saporta, 1990. *Probabilités, analyse des données et statistique*. Paris, France : Technip.

(Sébillot, 2005) P. Sébillot, 2005. Symbolic machine learning : A different answer to the problem of the acquisition of lexical knowledge from corpora. Dans les actes de *TripleC (Cognition, Communication, Co-operation, special issue : ECAP 2005 - European Computing and Philosophy Conference 2005)*, Sweden, 277–283.

(Schapire et Singer, 2000) R. E. Schapire et Y. Singer, 2000. BoosTexter : a boosting-based system for text categorization. Dans les actes de *Machine Learning*, 135–168.

(Schmid, 1994) H. Schmid, 1994. Probabilistic partofspeech tagging using decision trees. Dans les actes de *International Conference on New Methods in Language Processing*, Manchester, UK, 44–49.

(Schwartz et al., 2005) R. Schwartz, T. Phoenix, et B. d'Foy, 2005. *Learning Perl*. Paris, France : O'Reilly & Associates.

(Shannon, 1948) C. Shannon, 1948. A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423.

(Shannon, 1951) C. Shannon, 1951. Prediction and entropy of printed english. *Bell Systems Technical Journal 30*, 50–64.

(Sheskin, 2004) D. Sheskin, 2004. *Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition*. Great Britain : Taylor and Francis Group, CRC Press.

(Siegel et Castellan, 1988) S. Siegel et N. Castellan, 1988. *Nonparametric statistics for the behavioral sciences.* McGraw Hill.

(Sitbon et Bellot, 2004) L. Sitbon et P. Bellot, 2004. Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. Dans les actes de *TALN 2004*, 10–19.

(Sitbon et Bellot, 2005) L. Sitbon et P. Bellot, 2005. Segmentation thématique par chaînes lexicales pondérées. Dans les actes de *TALN 2005*, Volume 1, 505–510.

(Spärck Jones, 1972) K. Spärck Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*, 11–21.

(Stephens et Bialek, 2008) G. J. Stephens et W. Bialek, 2008. Toward a statistical mechanics of four letter words. *CoRR : A Computing Research Repository abs/0801.0253*.

(Swanson et al., 2006) D. R. Swanson, N. R. Smalheiser, et V. I. Torvik, 2006. Ranking indirect connections in literaturebased discovery : The role of medical subject headings : Research articles. *J. Am. Soc. Inf. Sci. Technol. 57*(11), 1427–1439.

(Szolnoki, 1999) A. Szolnoki, 1999. Stationary state in a two-temperature model with competing dynamics. *Phys. Rev. E 60*(2), 2425–2428.

(Taira et al., 2007) R. K. Taira, V. Bashyam, et H. Kangarloo, 2007. A field theoretical approach to medical natural language processing. *IEEE Transaction on Information Technology in Biomedicine 11*(4), 364–375.

(Takamura et al., 2005) H. Takamura, T. Inui, et O. Manabu, 2005. Extracting semantic orientations of words using spin model. Dans les actes de *ACL'05*, 133–140.

(Teufel, 1999) S. Teufel, 1999. *Argumentative Zoning : Information Extraction from Scientific Text*. Thèse de Doctorat, University of Edinburgh.

(Teufel et Moens, 2002) S. Teufel et M. Moens, 2002. Summarizing scientific articles : Experiments with relevance and rhetorical status. *Computational Linguistics 28*, 409–445.

(Torres-Moreno, 2007) J.-M. Torres-Moreno, 2007. *Rapport HDR. Du textuel au numérique : analyse et classification automatiques*. Avignon, France : Laboratoire Informatique d'Avignon.

(Torres-Moreno et al., 2001) J.-M. Torres-Moreno, P. Velázquez-Morales, et J. Meunier, 2001. Cortex : un algorithme pour la condensation automatique de textes. Dans les actes de *ARCo*, Volume 2, 365.

(Torres-Moreno et al., 2002) J.-M. Torres-Moreno, P. Velázquez-Morales, et J. Meunier, 2002. Condensés de textes par des méthodes numériques. Dans les actes de *JADT*, Volume 2, 723–734.

(Trémolet et al., 2000) E. Trémolet, M. Cyrot, et D. Michel, 2000. *Magnétisme, I - Fondaments*. Grenoble France : EDP Sciences.

(Waszak et Torres-Moreno, 2008) T. Waszak et J.-M. Torres-Moreno, 2008. Compression entropique de phrases controlée par un perceptron. *JADT 2*, 1163–1173.

(Winkler, 1999) W. Winkler, 1999. The state of record linkage and current research problems.

(Wong et Mooney, 2007) Y. W. Wong et R. J. Mooney, 2007. Learning synchronous grammars for semantique parsing with lambda calculus. Dans les actes de *45th Annual Meeting of the Association of Computational Linguistics, ACL 2007*, 960–967.

(Zener, 1951) C. Zener, 1951. Interaction between the *d*-shells in the transition metals. ii. ferromagnetic compounds of manganese with perovskite structure. *Phys. Rev. 82*(3), 403–405.

(Zipf, 1935) G. Zipf, 1935. *The psychobiology of language : An introduction to dynamic philology*. Boston, Mass., Houghton-Mifflin.

(Zipf, 1949) G. Zipf, 1949. *Human behavior and the principle of least effort : An introduction to human ecology*. Hafner Pub. Co.

Applications exploratoires des modèles de spins au Traitement Automatique de la Langue

Résumé

Dans cette thèse nous avons exploré la capacité des modèles magnétiques de la physique statistique à extraire l'information essentielle contenue dans les textes. Les documents ont été représentés comme des ensembles d'unités en interaction magnétique, l'intensité de telles interactions a été mesurée et utilisée pour calculer de quantités qui sont des indices de l'importance de l'information portée. Nous proposons deux nouvelles méthodes. Premièrement, nous avons étudié un modèle de spins qui nous a permis d'introduire l'énergie textuelle d'un document. Cette quantité a été utilisée comme indicatrice de pertinence et appliquée à une vaste palette de tâches telles que le résumé automatique, la recherche d'information, la classification de documents et la segmentation thématique. Par ailleurs, et de façon encore exploratoire, nous proposons un deuxième algorithme qui définie un couplage grammatical pour conserver les termes importants et produire des contractions. De cette façon, la compression d'une phrase est l'état fondamental de la chaîne de termes. Comme cette compression n'est pas forcement bonne, il a été intéressant de produire des variantes en permettant des fluctuations thermiques. Nous avons fait des simulations Métropolis Monte-Carlo avec le but de trouver l'état fondamental de ce système qui est analogue au verre de spin.

Mots clés : Énergie textuelle, Modèle de Hopfield, Résumé automatique, Frontière thématique, Verre textuel, Compression de phrases, Physique Statistique, Modèle de spin.

Exploratory applications of spin models in Natural Language Processing

Abstract

In this thesis we explored the ability of magnetic models of statistical physics to extract the essential information contained in texts. Documents are represented as sets of interacting magnetic units, the intensity of such interactions are measured and they are used to calculate quantities that are evidence of the importance of information scope. We propose two new methods. Firstly, we studied a spin model which allowed us to introduce the textual energy. This quantity was used as an indicator of information relevance. Several adaptations were necessary to adapt the energy calculation to a wide range of tasks such as summarisation, information retrieval, document classification and thematic segmentation. Furthermore, and even exploratory, we propose a second algorithm that defines a grammatical coupling between types of terms to retain the important terms and produce contractions. In this way, the compression of a sentence is the ground state of the chain of terms. As this compression is not necessarily good, it was interesting produce variants by thermal fluctuations. We have done simulations Metropolis Monte-Carlo with the aim of finding the ground state of this system that is analogous to spin glass.

Keywords : Textual Energy, Hopfield Model, Automatic Summarization, Thematic Boundary, Textual Glass, Sentence Compression, Statistical Physics, Spin Model.