TALN 2009, Senlis, 24–26 juin 2009

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

Silvia Fernández Sabido[1,2] Juan-Manuel Torres-Moreno[1]
(1) Laboratoire Informatique d'Avignon, BP 1228 84911 Avignon
(2) Laboratoire de Physique de Matériaux, UHP-Nancy, 54506 Vandœuvre
{silvia.fernandez, juan-manuel.torres}@univ-avignon.fr

Résumé. Nous présentons une approche exploratoire basée en notions thermodynamiques de la Physique statistique pour la compression de phrases. Nous décrivons le modèle magnétique des verres de spins, bien adapté à notre conception de la problématique. Des simulations Métropolis permettent d'introduire des fluctuations thermiques pour piloter la compression. Des comparaisons intéressantes de notre méthode ont été réalisées sur un corpus en français.

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.

Mots-clés : Compression de phrases, Résumé automatique, Résumé par extraction, Enertex, Mécanique statistique.

Keywords: Sentence Compression, Automatic Summarization, Extraction Summarization, Enertex, Statistical Mechanics.

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 en conservant le sens et la grammaticalité. Dans le résumé automatique par extraction, où les phrases les plus importantes ont été concaténées pour produire le condensé, aucun traitement n'est effectué au niveau intra-phrase. Ainsi, une phrase est soit conservée dans son intégralité, soit totalement supprimée. La compression peut alors combler ce manque afin d'en supprimer les constituants les moins pertinents (Monod & 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 utilise un corpus pour détecter des régularités afin de produire automatiquement les règles. Pour cette dernière, il est nécessaire de disposer d'un corpus d'apprentissage. Il doit ainsi contenir des phrases et une version acceptable de leurs compressions.

Nous présentons une approche statistique-thermodynamique pour la compression 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 où seulement un petit sous-ensemble correspond aux compressions acceptables de la phrase initiale. Pour une phrase de 25 termes il y a $2^{25} = 33\,554\,432$ 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 au préalable sur un corpus aligné de phrases complètes/compressées. Nous consacrons cette étude exclusivement à la langue française. Des exemples de systèmes qui compressent des phrases en français sont le modèle linguistique de (Monod & Prince, 2006) et l'approche statistique de (Waszak & Torres-Moreno, 2008). En modélisant la phrase comme un système thermodynamique sujet aux contraintes d'interaction entre unités, notre objectif vise l'étude du problème de la compression de phrases dans un nouveau cadre qui puisse donner des pistes pour des recherches futures.

En Section 1 nous faisons un parcours des méthodes statistiques de compression de phrases. En Sections 2 et 3 nous décrivons le modèle magnétique des verres de spins. Nous présentons en Sections 4 et 5 une stratégie pour calculer le couplage entre les termes des textes et entre leurs étiquettes grammaticales. Enfin en Section 6, des simulations Métropolis Monte-Carlo nous ont permis d'utiliser les couplages et d'introduire des fluctuations thermiques. La combinaison de ces deux facteurs nous permet de compresser automatiquement des phrases.

2 La compression statistique de phrases

Le modèle du canal bruité (Knight & Marcu, 2000) considère que la compression c est la phrase originale, et il lui a été ajouté du bruit pour produire une 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. Du fait que ces probabilités sont pondérées selon la longueur de la phrase compressée, le taux de compression n'est pas élevé. La méthode des arbres de décision (Knight & Marcu, 2000) utilise aussi des arbres syntaxiques. Il part d'un arbre représentant la structure d'une phrase en produisant un autre plus petit, correspondant à la compression. L'ordre des termes est conservé, mais les catégories syntaxiques peuvent changer. Ces travaux ont servi de référence à beaucoup d'autres, comme celui de (Turner & Charniak, 2005). (Clarke & Lapata, 2007) proposent une méthode qui utilise des arbres de décision et une autre qui évalue l'importance d'un terme (selon des critères sémantiques et fonctionnels) pour décider s'il doit être effacé. (Jing, 2000) utilise plusieurs sources de connaissance pour la compression : la syntaxe, le contexte et l'analyse statistique d'un corpus. L'idée est de supprimer les éléments qui ne sont pas portants au sujet principal du document.

L'analyse syntaxique a été privilégiée pour déterminer les éléments dont leur disparition affecteront 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 (Waszak & Torres-Moreno, 2008). De plus, les analyseurs syntaxiques ne sont pas toujours disponibles pour toutes les langues. Il existe des études qui se sont passés des arbres syntaxiques et qui ont obtenu des résultats comparables. On trouve par exemple le travail de (Nguyen et al., 2004) basé sur des *templates* de traduction. Il considère que les phrases non compressées sont

écrites dans une langue source et les phrases compressées dans une 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, dû à la grande quantité de règles possibles, le temps du calcul peut être exponentiel. Récemment, le système ENTROPIE (Waszak & Torres-Moreno, 2008) utilise aussi un corpus aligné pour apprendre un modèle de langage qui sert à déterminer quels termes ont une forte probabilité d'être supprimés. Ce choix est réalisé en utilisant des critères entropiques. Un perceptron [1] détermine si la phrase est suffisamment compressée. Ce système fonctionne pour des phrases en l'anglais et en français.

3 Les verres de spin

Les verres de spins [2] sont des matériaux constitués d'unités magnétiques entre lesquelles les interactions, dites d'échange, sont aléatoirement positives ou négatives. Si le couplage d'échange 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. Dans un triangle constitué par trois spins, si les trois interactions sont négatives, elles ne peuvent jamais être satisfaites au même temps (Trémolet et al., 2000).

3.1 Le texte vu comme un verre textuel

Un terme peut être vu comme un spin à deux états : $\uparrow$ (+1) indiquant sa présence dans une phrase ou $\downarrow$ (-1) son absence. Une phrase de N termes sera donc codée comme une chaîne de N spins orientés vers le haut, et sa compression correspond à la même chaîne où quelques spins ont changé d'orientation vers le bas (Fernández et al., 2007). Le calcul d'énergie textuelle (Fernández et al., 2008) dans un tel système établi une connectivité totale entre couples de termes. Or, dans notre modèle nous limitons les interactions d'échange 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 mots où la dimension vectorielle correspond à la taille du vocabulaire total du document. La dimension de chaque vecteur est donc le nombre de termes de la phrase représentée. Nous n'appliquons aucun prétraitement : les mots et les signes de ponctuation sont considérés comme de termes [3].

Le système de compression de phrases que nous proposons utilise un corpus aligné de phrases complètes/compressées en français [4]. Il s'agit du corpus utilisé par (Waszak & Torres-Moreno, 2008), ce qui nous permet de faire des comparaisons. Nous avons mesuré les couplages entre termes adjacents (voisins proches). Telles règles sont assimilées aux couplages entre les spins magnétiques qui interagissent dans un matériau. Nous sommes intéressés à établir des règles d'interaction qu'à partir de la phrase originale, amènent à une compression correcte. Il est clair

1. Le perceptron est un modèle de réseaux neuronaux (Hertz et al., 1991).
2. L'appellation « verre » vient du fait qu'ils présentent un comportement similaire aux verres structuraux.
3. En effet, (Bechet et al., 2008) proposent une méthode où la ponctuation est pertinents pour classer les textes.
4. Le corpus MYRIAM développé par Michel Gagnon à l'École Polytechnique de Montréal.

que pour supprimer les termes accessoires tout en gardant ceux pertinents, il faut avoir des interactions positives et négatives. Par exemple, si l'on veut que *la maison rouge* soit compressée 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 d'échange entre termes produit des compétitions internes dans la phrase. En Physique statistique, en absence d'autre facteur affectant le système (température, champ externe), obéir aux règles d'échange amène à une configuration appelé état fondamental où l'énergie est minimale. Par contre, si on applique des règles contradictores, on sera dans un cas de « frustration de termes » ou de « phrase frustrée » ayant par conséquence la production d'états méta-stables. Cette situation fait du système une sorte de verre de spins ou verre textuel.

4 Calcul des règles d'échange

Le corpus MYRIAM est composé de 219 phrases issues de sources journalistiques. Pour chaque phrase, une version compressée a été produite manuellement. Un échantillon est montré au tableau 1. Nous avons éclaté au hasard le corpus en deux ensembles : 80% pour l'apprentissage des couplages et 20% pour 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.

TABLE 1 – Exemples de phrases parallèles complètes/compressées du corpus MYRIAM.

4.1 Le couplage entre termes

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, les termes de la 1ère phrase du tableau 1, disparus pendant le processus de compression ont changé leur état de ↑ à ↓ :

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

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.

Nous établissons les huit couplages entre voisins présentés au tableau 2. Ces règles indiquent à chaque terme de suivre ou pas l'orientation de ses voisins. À partir de la phrase complète comme configuration initiale (les spins sont ↑), la satisfaction de toutes les règles amène à l'état fondamental d'énergie minimale qui, dans ce cas, correspond à une compression correcte. Nous avons suivi la même démarche avec toutes les phrases du corpus d'apprentissage. Deux

Couplage ferromagnétique ↑↑ ou ↓↓	**Couplage antiferromagnétique** ↑↓ ou ↓↑
$J_{je\ ,\ souhaite} = +1$ $J_{\grave{a}\ ,\ notre} = +1$ $J_{notre\ ,\ terre} = +1$ $J_{la\ ,\ paix} = +1$ $J_{paix\ ,\ .} = +1$	$J_{Enfin\ ,\ je} = -1$ $J_{souhaite\ ,\ \grave{a}} = -1$ $J_{terre\ ,\ la} = -1$

TABLE 2 – Couplages entre termes voisins.

termes peuvent être voisins proches dans plusieurs phrases. Pour avoir une valeur unique pour chaque couple de termes, nous avons fait la somme d'occurrences. Ce processus a produit $\approx 6\,000$ règles, quelques unes affichées au tableau 3. Nous avons appliqué les règles apprises pour

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

TABLE 3 – Exemples des règles d'échange entre termes apprises sur le corpus d'apprentissage.

compresser les phrases du corpus de test. Cependant, les résultats obtenus ont été mitigés. En voici les raisons : *i)* une grande partie du vocabulaire des phrases à compresser n'existe pas dans le corpus d'apprentissage, donc aucune règle ne les concerne. *ii)* même si 2 termes sont présents dans les corpus, il n'est pas sûr qu'ils soient voisins adjacents, donc leur règle d'échange est inexistante. En effet, pour un grand ensemble des phrases de test, aucune règle n'existe et pour d'autres phrases, les règles sont peu nombreuses. Un exemple est montré au tableau 4, où nous avons seulement deux règles pour une phrase de huit termes. Pour surmonter ces problèmes liés au manque de termes, nous avons groupé les termes selon leur catégorie grammaticale.

$J_{\text{la,pénurie}} = +1$
$J_{\text{fait,sentir}} = +1$

Configuration initiale	↑ Mais	↑ partout	↑ la	↑ pénurie	↑ se	↑ fait	↑ sentir	↑ .
État fondamental	?	?	↑ la	↑ pénurie	?	↑ fait	↑ sentir	?

TABLE 4 – Exemple d'application du couplage entre termes pour une phrase du corpus de test. Seulement 2/7 des couplages possibles entre termes voisins ont été déterminés pendant le processus d'apprentissage. Cette manque d'information peut être une conséquence de la taille réduite du corpus.

4.2 Le couplage grammatical

TreeTagger (Schmid, 1994) étiquette les termes selon leurs catégories grammaticales pour produire des relations du type : $J_{\text{mais,partout}} \rightarrow J_{\text{PREP , ADV}}$; $J_{\text{fait,sentir}} \rightarrow J_{\text{VERB :PRE , VERB :INF}}$. 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 $J_{ART,NOM} = +1.5$. Nous avons ainsi obtenu 400 relations entre étiquettes grammaticales à partir des $6\,000$ règles entre termes. Les distributions des valeurs des couplages de termes et étiquettes sont montrées dans la figure 1. On observe que dans le cas des termes (trait pointillé), environ 80% correspond aux valeurs +1. Elles sont produites en grande partie par des occurrences uniques des termes voisins avec la même orientation. En revanche, on observe pour les étiquettes (trait continu) un effet lissé. Le pic concerne environ 48% des couples. Nous observons dans les deux cas, une prédominance des valeurs positives. Il semble que les termes voisins

qui restent ou disparaissent ensemble pendant le processus de compression sont plus nombreux que ceux qui le font séparement.

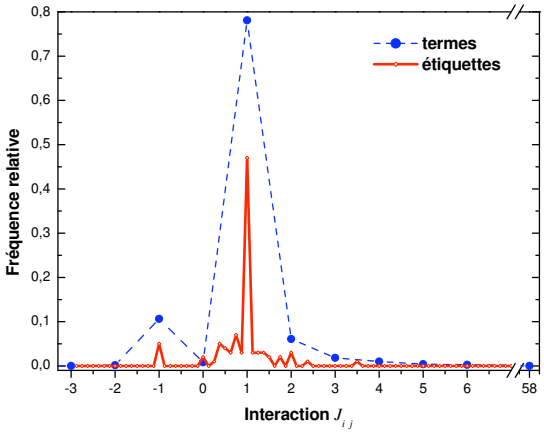

FIGURE 1 – Fréquences relatives des couplages entre termes et entre étiquettes du corpus MYRIAM.

5 Application des règles à la compression de phrases

5.1 Les états fondamentaux de la chaîne des spins

Nous avons appliqué l'ensemble réduit des règles générales sur les phrases du corpus de test. Le tableau 5 montre un exemple de cette application. On observe que pour la même phrase du tableau 4, on a les 7 valeurs de couplages entre voisins proches. Appliqués sur la phrase originale, ces couplages produisent une compression acceptable. Malheureusement ce n'est pas

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

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 5 – Application du couplage entre étiquettes grammaticales pour la phrase du tableau 4. Nous avons les sept valeurs des couplages qui produisent une compression bien acceptable de la phrase.

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 : *i)* Les sous-phrases obtenues avec les états fondamentaux ne sont pas systématiquement des bonnes compressions. Cet effet peut être lié à la petite taille du corpus d'apprentissage qui produit des règles rigides. *ii)* La frustration (impossibilité de satisfaire toutes les règles d'échange) est présente dans 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 une certaine souplesse dans l'application des règles et au même temps qui puisse traiter les phrases frustrées. Les simulations du type Métropolis Monte-Carlo nous ont permis d'introduire des fluctuations thermiques qu'apporteront flexibilité à l'application des règles et d'utiliser le recuit simulé pour faire face à la frustration des verres de termes.

5.2 Simulations Métropolis Monte-Carlo

L'idée principal est d'imiter les fluctuations thermiques aléatoires d'un système qui parcourt plusieurs états. La probabilité p_μ de trouver le système dans un état μ est donné par la distribution de Gibbs-Boltzmann :

$$p_\mu \propto exp(-E_\mu/kT) \tag{1}$$

où E_μ est l'énergie du système à 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 (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 le *flip* car l'état produit est plus stable que le précédant ;
   - si l'énergie augmente ($\Delta E > 0$), on génère un numéro r aléatoire tel que $0 \leq r \leq 1$. Si $r < exp(-\Delta E/kT)$ on accepte le *flip*, autrement, on reste dans l'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 l'énergie devient quasi constante. Nous sommes intéressés par récupérer les états dans lesquels le système s'stabilise à chaque température, car ils représentent des variantes potentielles de la compression. La température est une perturbation qui fait varier l'énergie du système. Deux facteurs sont en concurrence : l'interaction entre spins (couplage d'échange) et la température. À basse température l'échange domine et les spins gèlent dans la configuration dictée pour celui-ci. À haute température les fluctuations thermiques favorisent les états aléatoires qui ne répondent pas forcement au 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 ainsi des variantes pour en choisir la meilleure. Les conditions de simulations sont les suivantes :

L'**état initial** est ferromagnétique : tous les termes sont présents.

Spins fixés $\uparrow$: Pour ne pas confondre une configuration avec sa symétrique (même énergie où les spins ont l'état opposé) nous avons fixé quelques spins de la phrase. Nous avons fixé le premier substantif et/ou verbe (dans l'ordre original du texte) présents. En générale, ces éléments restent $\uparrow$ dans la phrase compressée. Nous fixons aussi le point final car il ne disparaît jamais.

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

$$IS(terme_j)_i = \sum_i \frac{ns(terme_{j,i})}{|phr_i|} \tag{2}$$

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. 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 , touts neufs~~ .	IS(bleu)$_3$=2/20=0,10
		IS(bleu)$_{\text{corpus}}$=0,24

Le spin fixé $\downarrow$ dans la configuration initiale correspond au terme d'IS le plus élevé.

La **Température** : T varie de 1 à 0 en pas de 0,01. On adapte 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.

Frustration : Pour éviter que le système reste piegé dans des états méta-stables, nous avons utilisé le recuit simulé. Il consiste à faire monter et descendre la température plusieurs fois dans un rang de températures suffisamment basses.

Configurations retenues : Après $1\,000$ itérations on récupère les états à chaque température. Cela produit un ensemble de variantes de compression de la phrase initiale. Nous utilisons 2 critères pour choisir la compression : l'énergie minimale et la magnétisation maximale (qui correspond au taux de compression minimum).

6 Évaluation de la compression

BLEU (*Bilingual Language Evaluation Understudy*) (Papineni et al., 2001) conçu pour juger la précision de la traduction automatique, est aussi utilisé dans la compression de phrases (Dorr et al., 2003; Egawa et al., 2008). Il mesure la concordance entre une phrase candidate (faite par un système) et une référence (faite par un humain). BLEU mesure la précision (% de n-grammes de la phrase candidate dans la référence) ce qui montre une forte corrélation avec les jugements humains sur la qualité de la compression. Le tableau 6 montre les scores obtenus avec et sans recuit. Les unités de comparaison sont les $3-4$-grammes, tel que suggéré par (Papineni et al., 2001). Les critères de sélection sont l'énergie minimale et la magnétisation maximale. Nous avons réalisé 3 simulations s_1, s_2 et s_3, en comparant nos résultats avec ceux du système ENTROPIE (Waszak & Torres-Moreno, 2008) et à une *baseline* où les couplages $J_{i,j}$ ont des valeurs aléatoires $\in [-1, +1]$. Plus la valeur BLEU est élevée, plus proche est la compression candidate de la référence. On observe que les deux critères, énergie et magnétisation, sont proches. 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. BLEU étant une mesure de la pertinence de l'information plus que de qualité grammaticale, on peut dire que notre système produit des compressions où l'information essentielle est conservée mais pour vérifier la

	Unité BLEU	*Baseline*	**ENTROPIE**	**VERRE TEXTUEL**			**VERRE TEXTUEL** simulations avec recuit		
				s_1	s_2	s_3	s_1	s_2	s_3
Énergie	**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
Magnet.	**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 – En haut scores BLEU pour le système VERRE TEXTUEL avec le critère d'énergie minimale, en bas avec le critère de magnétisation maximale. On montre les résultats d'ENTROPIE et d'une *baseline*.

qualité, une évaluation manuelle s'avère nécessaire. Nous montrons des compressions grammaticalement correctes (tableau 7) et incorrectes (tableau 8). Le système ENTROPIE semble un

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	**certains fabricants adoptent des mesures.**
Originale	Le déficit actuel pourrait doubler de ici le an 2000.
Humain	Le déficit pourrait doubler de ici 2000.
ENTROPIE	Le déficit de le an 2000.
VERRE TEXTUEL	**Le déficit pourrait doubler**

TABLE 7 – Exemples de compressions grammaticalement correctes. En gras, la meilleure compression.

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	, 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	souhaite une meilleure santé économique .

TABLE 8 – Exemples de compressions grammaticalement incorrectes. En gras, la meilleure compression.

peu plus robuste grammaticalement, possiblement grâce à l'utilisation de n-grammes comme unité de base. Dans notre cas, nous sommes intéressés à explorer les interactions des termes isolés (unigrammes). Le tableau 9 donne un aperçu des performances des systèmes.

Système	**% phrases non compressées**	**% phrases compressées correctes**	**% phrases compressées incorrectes**
ENTROPIE	$\approx 30\%$	$\approx 30\%$	$\approx 40\%$
VERRE TEXTUEL	$\approx 40\%$	$\approx 20\%$	$\approx 40\%$
Baseline	$\approx 5\%$	$\approx 5\%$	$\approx 90\%$

TABLE 9 – Pourcentages de phrases compressées par les systèmes.

7 Conclusion

Un système thermodynamique de compression de phrases en français a été proposé. Les phrases sont codées comme des chaînes de verres de spins. Les couplages entre termes, et entre leurs étiquettes grammaticales ont été calculés sur un corpus d'apprentissage. Les phrases de test sont compressées en appliquant les couplages appris avec une dynamique thermique de Métropolis. Pour chaque phrase et chaque température cette approche génère un ensemble de choix. Cet ensemble est différente en chaque simulation car le système n'est pas déterministe. Ce comportement est en accord avec la tâche de compression de texte, qui n'a pas une solution unique. Deux personnes ne compressent pas une phrase de la même façon, et plus encore, la même personne peut faire des compressions différentes à des moments différents. Les compressions, évaluées par rapport à celles faites par des humains, ont des scores BLEU comparables à ceux du système ENTROPIE. Le groupement par catégorie grammaticale génère des règles d'échange générales qui produisent des compressions acceptables. Or, la précision perdue impacte dans la qualité grammaticale des phrases. Nous pensons que il faut un corpus plus large pour avoir 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 & Barkema, 1999), bien adapté aux systèmes type verre de spins.

Références

BECHET F., EL-BÈZE M. & TORRES-MORENO J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes. In TALN 2008-Atelier DEFT'08, p. 161–170.

CLARKE J. & LAPATA M. (2007). Modelling compression with discourse constraints. In Empirical Methods in NLP and Computational Natural Language Learning, p. 1–11, Prague.

DORR B., ZAJIC D. & SCHWARTZ R. (2003). Hedge : A parse-and-trim approach to headline generation. In HLT-NAACL DUC'03, p. 1–8, Edmonton, Canada.

EGAWA S., KATO Y. & MATSUBARA S. (2008). Sentence compression by removing recursive structure from parse tree. In PRICAI'08 : Trends in AI, volume 5351, p. 115–127.

FERNÁNDEZ S., SANJUAN E. & TORRES-MORENO J. M. (2007). Textual energy of associative memories : performants applications of enertex algorithm in text summarization and topic segmentation. In MICAI '07, Aguascalientes (Mexico), p. 861–871.

FERNÁNDEZ S., SANJUAN E. & TORRES-MORENO J. M. (2008). Enertex : un système basé sur l'énergie textuelle. In TALN 2008, p. 99–108.

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

JING H. (2000). Sentence reduction for automatic text summarization. In 6th Applied Natural Language Processing Conference (ANLP'00), p. 310–315.

KNIGHT K. & MARCU D. (2000). Statistics-based summarization - step one : Sentence compression. In AAAI/IAAI, p. 703–710.

MONOD M. Y. & PRINCE V. (2006). Compression de phrases par élagage de l'arbre morpho-syntaxique. Technique et Science Informatiques, **25**(4), 437–468.

NEWMAN M. E. J. & BARKEMA G. T. (1999). Monte Carlo Methods in Statistical Physics. Great Britain : Clarendon Press - Oxford University Press.

NGUYEN M. L., HORIGUCHI S., SHIMAZU A. & HO B. T. (2004). Example-based sentence reduction using the hidden markov model. ACM TALIP, **3**(2), 146–158.

PAPINENI K., ROUKOS S., WARD T. & ZHU W. (2001). Bleu : a method for automatic evaluation of machine translation.

SCHMID H. (1994). Probabilistic partofspeech tagging using decision trees. In International Conference on New Methods in Language Processing, p. 44–49, Manchester, UK.

TRÉMOLET E., CYROT M. & MICHEL D. (2000). Magnétisme, I - Fondaments. Grenoble France : EDP Sciences.

TURNER J. & CHARNIAK E. (2005). Supervised and unsupervised learning for sentence compression. In ACL'05, p. 290–297, Morristown, NJ, USA : ACL.

WASZAK T. & TORRES-MORENO J.-M. (2008). Compression entropique de phrases controlée par un perceptron. JADT, **2**, 1163–1173.