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 sur des notions thermodynamiques de la Physique statistique pour la compression de phrases. Nous décrivons le modèle magnétique des verres de spins, adapté à notre conception de la problématique. Des simulations Métropolis Monte-Carlo 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 Monte-Carlo simulations allow 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. 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. Nous présentons une approche statistique-thermodynamique pour la compression de phrases en français. 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). Un tel système possède 2^N configurations où seulement un petit sous-ensemble correspond aux compressions acceptables de la phrase initiale. 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 voisins pour contrôler leurs retournements et réduire l'espace des configurations. Ces couplages seront mesurés sur un corpus aligné

de phrases complètes/compressées.

En Section 2 nous faisons un parcours des méthodes statistiques de compression de phrases. En Section 3 nous décrivons le modèle magnétique des verres de spins. Nous présentons en Section 4 notre stratégie pour calculer le couplage entre les termes. Les simulations Monte-Carlo, qui nous ont permis d'introduire des fluctuations thermiques, seront aussi décrites. Enfin, avant de conclure, nous présentons en Section 5 l'évaluation de notre algorithme.

2 La compression statistique de phrases

Le modèle du canal bruité (Knight & Marcu, 2000) considère que la compression c est la phrase originale, à laquelle 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. La méthode des arbres de décision (Knight & Marcu, 2000) part d'un arbre représentant la structure d'une phrase en produisant un autre plus petit, correspondant à la compression. (Jing, 2000) utilise plusieurs sources de connaissance : la syntaxe, le contexte et l'analyse statistique d'un corpus. L'idée est de supprimer les éléments qui ne sont pas porteurs du 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. Il existe des études qui se sont passées des arbres syntaxiques et qui ont obtenu des résultats comparables. On trouve 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. 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.

3 Les verres de spin

Les verres de spins sont des matériaux constitués d'unités magnétiques dont leurs interactions sont soit positives soit négatives, de façon aléatoire. 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). Il existe donc une compétition locale entre ces forces. Les spins peuvent se révéler incapables de satisfaire simultanément les interactions contradictoires auxquelles ils sont soumis par leurs voisins. Ce comportement peut donner lieu à ce qu'on appelle la frustration.

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 (Fernández *et al.*, 2007). 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. 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 aux couples de voisins proches. La dimension de chaque vecteur est donc le nombre de termes de la phrase représentée. Le système de compression de phrases que nous proposons utilise un corpus aligné de phrases complètes/compressées en français [1]. Nous avons mesuré les couplages entre termes adjacents. 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 $J_{maison,rouge} = -y$. La variété en valeur et en signe des interactions entre termes produit des compétitions internes dans la phrase. Cette situation fait du système une sorte de verre de spins ou verre textuel.

4 Calcul des règles d'interaction

Le corpus MYRIAM est composé de 219 phrases issues de sources journalistiques. Pour chaque phrase, une version compressée a été produite manuellement. Nous avons éclaté le corpus en deux ensembles : 80% pour l'apprentissage des couplages et 20% pour les tests de compression. 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 dans la phrase suivante, quelques termes ont disparu pendant la compression (leurs états ont changé 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. Ces règles indiquent à chaque terme de suivre ou pas l'orientation de ses voisins. Or, en raison de la petite taille du corpus, une grande partie du vocabulaire des phrases de test n'existe pas dans le corpus d'apprentissage et, par conséquent, aucune règle ne les concerne. Pour surmonter les problèmes liés au manque de termes, nous avons décidé de grouper les termes selon leur catégorie grammaticale. Cette stratégie nous a permis d'élargir la validité des règles obtenues. TreeTagger (Schmid, 1994) étiquette les termes selon leurs catégories grammaticales pour produire des relations du type :

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

Ainsi, on regroupe plusieurs règles en une seule qui représente le couplage entre deux types de termes. Nous avons choisi d'en utiliser leur valeur moyenne. Par exemple, $J_{\text{une,fois}} = +1$ et

1. Le corpus MYRIAM développé par Michel Gagnon à l'École Polytechnique de Montréal.

$J_{\text{la,pénurie}} = +2$, alors $J_{ART,NOM} = +1.5$. Nous avons ainsi obtenu 400 relations entre étiquettes grammaticales. Nous avons appliqué les règles apprises pour compresser les phrases du corpus de test. Le tableau 1 montre un exemple de cette application. On observe les 7 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

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

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

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. Les simulations du type Métropolis Monte-Carlo nous ont permis d'introduire des fluctuations thermiques qui apporteront de la flexibilité à l'application des règles et d'utiliser le recuit simulé pour faire face à la frustration des verres textuels.

4.1 Simulations Métropolis Monte-Carlo

L'idée principale d'une simulation Monte-Carlo est d'imiter les fluctuations thermiques 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 : $p_\mu \propto exp(-E_\mu/kT)$ 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. 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 ;
2. 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 $0 \leq r \leq 1$; et on le compare avec le facteur $exp(-\Delta E/kT)$;
   - si $r < exp(-\Delta E/kT)$ on accepte le *flip*, autrement, on reste dans le même état μ.
3. répéter la simulation un nombre suffisant de fois, pour permettre au système d'atteindre l'équilibre à une température établie.
4. quand le système est déjà à l'équilibre on pourrait commence la mesure des valeurs d'*expectation* (moyennes temporales) de quantités comme l'énergie, la magnétisation, la chaleur spécifique. On aura une valeur pour chaque température.

Être en équilibre signifie que le système ne fera plus de transitions importantes et la valeur de l'énergie devient quasi constante. Les conditions des simulations ont été les suivantes.

État initial : ferromagnétique où tous les termes sont présents.

Spins fixés : afin de ne pas confondre une configuration donnée avec sa configuration symétrique de même énergie, nous avons fixé dans l'état $\uparrow$ le symbole de ponctuation final qui ne disparaît jamais. Nous avons fixé aussi 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) :$IS(terme_{j,i}) = \sum_{i=1}^{P} \frac{ns(terme_{j,i})}{|phr_i|}$ 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. Le spin qui sera fixé $\downarrow$ dans la configuration initiale correspond au terme d'IS plus élevé.

Température : elle prend des valeurs allant de 1 à 0 en pas de 0,01.

Frustration : pour en faire face, nous avons utilisé la technique du recuit simulé.

Nombre d'itérations : $1\ 000$ itérations par température.

Les configurations retenues : 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 $E = -\frac{1}{2}\sum_{i,j} s_i s_j J_{i,j}$ et la magnétisation $M = \sum_i s_i$, où s_i est l'état du spin i.

5 Évaluation de la compression

Le système *Bilingual Language Evaluation Understudy* (BLEU) (Papineni *et al.*, 2001) mesure la concordance entre une phrase candidate (faite par un système) et une référence (faite par un humain). Les tableaux 2 et 3 montrent les scores BLEU obtenus par notre système sans et avec recuit simulé. Nous utilisons des n-grammes ($n = 3, 4$) tel que suggéré par (Papineni *et al.*, 2001). Pour la plupart des simulations, le critère de magnétisation maximale obtient des scores plus élevés que celui d'énergie minimale. Nous avons réalisé trois simulations (s1, s2 et s3 aux tableaux). Nous comparons nos résultats avec ceux produits par ENTROPIE de (Waszak & 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. Les scores obtenus par notre système sont légèrement supérieurs à ceux obtenus par ENTROPIE. Le recuit simulé ne semble pas avoir un effet significatif dans le résultat. BLEU est d'avantage une mesure de la

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 2 – 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 & Torres-Moreno, 2008) et une *baseline* où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.

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 3 – 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 & Torres-Moreno, 2008) et une *baseline* où les valeurs des couplages $J_{i,j}$ sont des valeurs aléatoires entre -1 et +1.

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. 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 compressés correctes, le nombre de phrases compressés incorrectes et le nombre de phrases non compressés pour chaque système. Les résultats sont présentés au tableau 4. Des exemples

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

TABLE 4 – Pourcentages de phrases du corpus qui ont été compressées par le système ENTROPIE et VERRE TEXTUEL pendant une simulation.

de compressions sont montrées dans le tableau 5. Il est important de signaler que le corpus utilisé contient des expressions agrammaticales dû à l'éclatement de certains contractions (par exemple *à les* à la place de *aux*, *de le* au lieu de *du*). Cette séparation de termes se trouve sur le corpus d'origine où ses auteurs ont eu l'idée de permettre aux méthodes de compresser au maximum les phrases. Notre processus de compression exploite cette propriété afin de capturer plus finement les interactions des tous les termes. D'où la présence de, par exemple, *de* et *ici* à la place de *d'ici* dans les exemples.

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 .**
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 .

TABLE 5 – Compressions générées par le système VERRE TEXTUEL. Nous montrons la phrase originale, la compression faite par un humain et celle produite par ENTROPIE. En gras, la meilleure compression. 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 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 sommes intéressés uniquement à l ?exploration des interactions des termes isolés (unigrammes).

6 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 les étiquettes grammaticales des termes ont été calculés sur un corpus d'apprentissage. Les phrases de test ont été 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 pour 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. Même si le groupement par catégorie grammaticale génère des règles d'échange générales qui produisent des compressions acceptables, nous pensons que le calcul des couplages avec des termes à longueur variable (1, 2 ou 3 mots) pourrait améliorer la qualité de nos compressions. Des études sont actuellement en cours.

Remerciements

Ce travail a été realisé en partie grâce au financement du CONACYT (Mexique), bourse 175225.

Références

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.

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.

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.

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