Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?[1]

Silvia Fernández[1,3], Patricia Velázquez, Sonia Mandin[2], Eric SanJuan[1], Juan-Manuel Torres-Moreno[1,4]

[1]Laboratoire Informatique d'Avignon – BP 1228, 84 911 Avignon – France

[2]Laboratoire des Sciences de l'Education & IUFM – BP 47, 38 040 Grenoble – France

[3]Laboratoire de Physique de Matériaux – BP 239, 54 506 Vandœuvre-lès-Nancy – France

[4]Ecole Polytechnique de Montréal – CP 6079 – Montréal, Québec – CANADA H3C3A7

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 require linguistic resources. In this study, we confront them to the summaries made by students of different levels (middle school, high school and graduate school). Finally, the results of this study allow us to determine which summarization system best corresponds to a given school level.

Résumé

Nous avons développé trois systèmes de Résumé Automatique (Cortex et Enertex basés sur le modèle vectoriel, et un autre basé sur l'Analyse de la Sémantique Latente LSA). Ces systèmes reposent principalement sur des méthodes qui n'ont (presque) pas besoin de ressources linguistiques. Dans ce travail nous les confrontons avec les jugements d'élèves et d'étudiants de différents niveaux scolaires (collège, lycée et master). Cette expérience nous a permis de déterminer le niveau scolaire auquel correspondent le mieux les systèmes résumant automatiquement des textes.

Mots-clés : condensés de textes, résumés automatiques, modèle vectoriel, Cortex, Enertex, LSA.

1. Introduction

Depuis sa constitution comme un domaine scientifique au début des années 70, le Traitement Automatique des Langues (TAL) s'est nourri des recherches dans quatre domaines : la linguistique, l'informatique, les mathématiques et les sciences cognitives. Les objectifs de chacun diffèrent tant que TAL recourt à des approches parfois antagonistes. Par exemple, dans le domaine de l'intelligence artificielle, une approche cognitiviste s'intéresse au comportement des humains (*e.g.*, en modélisant leur compréhension de textes) ; l'approche informatique se concentre sur la construction de logiciels qui répondent aux demandes des utilisateurs (*e.g.*, en créant des outils facilitant la lecture / écriture). Nous retrouvons cette confrontation d'idées dans les applications les plus courantes du TAL. C'est le cas des applications de résumé automatique dont les approches adoptées peuvent être classées en

[1] Nous remercions Benoît Lemaire (IMAG) et Philippe Dessus (UPMF & IUFM) pour une partie des données fournies.

deux groupes : les approches par compréhension et les approches par extraction. La principale différence entre ces approches réside dans l'utilisation de méthodes différentes pour résumer. La première, celle du « *bon élève* », postule que pour réussir un résumé de qualité, représentant au mieux le contenu du texte et les intentions de l'auteur, il faut passer par une étape de compréhension. En revanche, la méthode du « *mauvais élève* » se contente de repérer les phrases importantes (*e.g.*, 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. La première produit des contractions de textes basées sur des reformulations en utilisant un lexique nouveau (*abstracts*) alors que l'autre produit des contractions de textes basées sur une extraction de quelques phrases (*extracts*). L'avantage de cette seconde méthode, bien que moins proche du fonctionnement humain, est qu'elle est plus facile à mettre en œuvre et plus prolifique. Mais, les approches par extraction de phrases simulent-elles forcément les « *mauvais élèves* » ? Afin de définir la pertinence des différents systèmes de résumés automatiques, nous avons mesuré la proximité de résumés générés automatiquement avec un nombre suffisant d'*abstracts* et *extracts* de collégiens, lycéens et étudiants universitaires. La Section 2 présente une brève synthèse sur les résumés automatiques par extraction. En Section 3, nous décrivons trois algorithmes : Cortex et Enertex, basés sur la représentation vectorielle des textes et un algorithme d'analyse sémantique latente conçu à l'origine pour étudier les processus cognitifs dans l'activité de résumer. En Section 4, nous détaillons la procédure expérimentale utilisée dans une étude mettant en commun des travaux de chercheurs grenoblois basés sur une approche cognitive du résumé et de travaux d'une équipe avignonnaise basés sur deux approches par extraction. En Section 5 nous présentons nos résultats et évaluations avant de conclure.

2. Le résumé par extraction de phrases

Le résumé par extraction (*extract*) est le nom générique recouvrant différentes approches ayant en commun le fait de proposer un résumé par sélection des phrases importantes (Ibekwe-SanJuan, 2007). Cette démarche se retrouve dans la majorité des travaux actuels, y compris ceux qui cherchent à produire des résumés par reformulation (*abstract*) (Minel, 2004 ; Yousfi-Monod *et al.*, 2006) ou qui étudient les processus cognitifs au moment de résumer un texte (Lemaire *et al.,* 2005 ; Mandin *et al.*, 2005 ; Fayol, 1985). Tous reconnaissent la pondération des phrases comme une étape importante dans la production de résumés. Des méthodes linguistiques, statistiques et probabilistes permettent d'évaluer l'information dans des segments de textes. Des techniques qui utilisent la position textuelle (Edmundson, 1969 ; Brandow *et al.*, 1995 ; Lin et *al.*, 1997), les modèles Bayésiens (Kupiec et *al.*, 1995), la pertinence marginale maximale (Goldstein *et al.*, 1999) ou la structure de discours ont été utilisés. La plupart des travaux sur le résumé par extraction appliquent les techniques statistiques (analyse de fréquence, recouvrement de mots, etc.) aux unités textuelles 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), la combinaison de l'extraction de l'information et de la génération de textes, l'utilisation des SVM (Mani *et al.*, 1999 ; Kupiec *et al.*, 1995) pour trouver des patrons dans les textes, des chaînes lexicales (Barzilay *et al.*, 1997) ou encore de la théorie de la structure rhétorique (Mann, 1987).

3. Trois modèles de génération automatique de résumés

3.1. Cortex

Cortex[2] (Torres-Moreno *et al.*, 2001) est un système de résumé automatique procédant par extraction. Il utilise deux algorithmes : une méthode de construction des métriques informationnelles indépendantes et un algorithme combinant l'information provenant des métriques. Ce dernier prend une décision sur la pertinence des segments selon une stratégie de vote. La figure 1 montre l'architecture modulaire de Cortex.

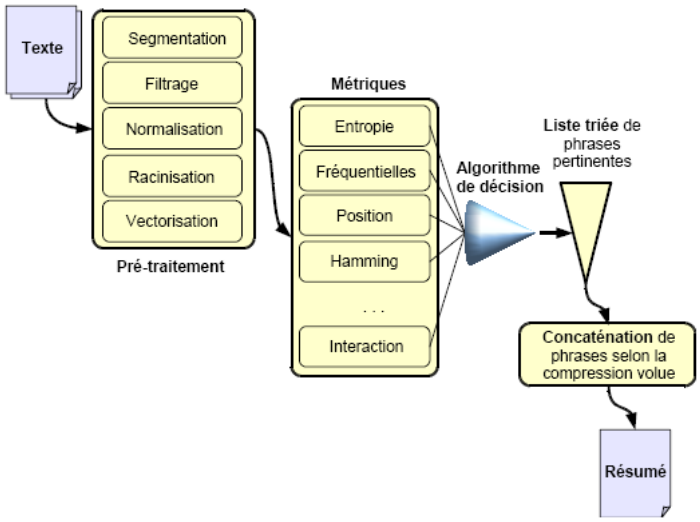

Figure 1. Architecture générale de Cortex.

L'idée consiste à représenter les textes dans un espace approprié et à leur appliquer des traitements vectoriels. Les documents sont prétraités avec des algorithmes classiques de filtrage de mots fonctionnels[3], de normalisation et de lemmatisation (Porter, 1980 ; Manning et al., 2000) afin de réduire la dimensionnalité. Une représentation en sac de mots produit une matrice $S[P \times N]$ de fréquences/absences composée de $\mu = 1, \ldots, P$ phrases (lignes) ; $\sigma_\mu = \{s^1_\mu, \ldots s^i_\mu, \ldots, s^N_\mu\}$ et un vocabulaire de $i = 1, \ldots, N$ termes (colonnes).

$$S = \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \\ \vdots & \vdots & \ddots & \vdots \\ s_P^1 & s_P^2 & \cdots & s_P^N \end{pmatrix} \qquad s_\mu^i = \begin{cases} TF^i \rightarrow terme - existe \\ 0 \rightarrow autrement \end{cases} \tag{1}$$

TF^i est la fréquence du terme i. Cette représentation sera aussi utilisée par le système Enertex. Cortex utilise jusqu'à $\Gamma=11$ métriques pour évaluer la pertinence des phrases. Quelques unes

[2] Cortex es Otro Resumidor de TEXtos.

[3] Nous avons effectué le filtrage de chiffres et l'utilisation d'anti-dictionnaires.

de ces métriques sont : l'angle entre le titre et chaque phrase, des calculs d'entropie, le poids fréquentiel des segments et des mots, ainsi que plusieurs mesures d'Hamming parmi d'autres. Le système pondère les phrases avec un algorithme de décision qui combine les sorties normalisées de toutes les métriques (entre *[0,1]*) dans une forme sophistiquée. Deux moyennes sont calculées : la tendance positive, où $\lambda_s > 0,5$, et la tendance négative, où $\lambda_s < 0,5$. Pour calculer cette moyenne, nous divisons par le nombre total de métriques Γ et non pas uniquement par le nombre d'éléments positifs ou négatifs (la moyenne réelle des tendances). En divisant par Γ, nous avons développé un algorithme plus décisif que la simple moyenne et plus réaliste que la moyenne réelle des tendances. L'algorithme de décision est :

$$\begin{aligned} \sum \alpha &= \sum_{\nu=1}^{\Gamma} (\left\|\lambda_s^\nu\right\| - 0.5); \left\|\lambda_s^\nu\right\| > 0.5 \\ \sum \beta &= \sum_{\nu=1}^{\Gamma} (0.5 - \left\|\lambda_s^\nu\right\|); \left\|\lambda_s^\nu\right\| < 0.5 \end{aligned} \tag{2}$$

Γ est le numéro totale de métriques et ν est l'indice de la métrique. La valeur attribuée à chaque phrase *s* est calculée de la façon suivante :

$$\begin{aligned} &Si \quad \left(\sum \alpha > \sum \beta\right) \quad alors \quad score_s = 0.5 + \frac{\sum \alpha}{\Gamma}: \quad retenir\ s \\ &Sinon \qquad\qquad\qquad\quad score_s = 0.5 + \frac{\sum \alpha}{\Gamma}: \quad éliminer\ s \end{aligned} \tag{3}$$

3.2. Enertex : l'énergie d'interaction entre phrases

Enertex est inspirée par la physique statistique de systèmes magnétiques. Ce système modèle les documents comme un réseau de neurones dont l'énergie textuelle est étudiée (Fernández et al., 2007). Un document peut être traité comme un ensemble d'unités (les mots) qui interagissent les unes avec les autres (figure 2).

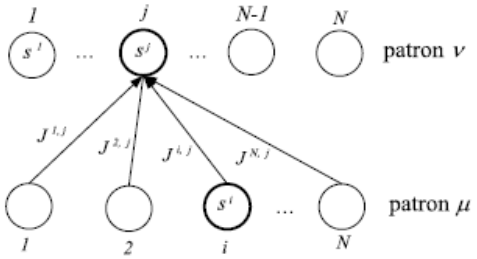

Figure 2. Le champ produit par les termes de la phrase de μ affecte les N termes de la phrase ν.

Le point de départ est la mémoire associative de Hopfield (Hopfield, 1982) qui se base sur le modèle magnétique d'Ising (modèle issu de la physique statistique décrivant un système avec des unités à deux états nommés *spins*). Ce réseau neuronal a des capacités d'apprentissage et de récupération de patrons et les unités correspondent aux neurones qui interagissent selon la règle d'apprentissage d'Hebb[4] :

$$J^{i,j} = \sum_{\mu=1}^{P} s_\mu^i s_\mu^j \tag{4}$$

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

s^i et s^j sont les états des neurones i et j. La somme porte sur les P patrons à stocker. Ce modèle est capable de stocker et de récupérer un certain nombre de configurations du système, car la règle d'Hebb transforme ces configurations en attracteurs (minimaux locaux) de la fonction d'énergie (Hopfield, 1982) :

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s^i J^{i,j} s^j \quad (5)$$

Les limitations de ce réseau ont été bien établies (Hopfield, 1982) : les patrons doivent être non corrélés afin que leur récupération s'effectue sans erreur, le système sature rapidement et seulement une fraction *(~0,14N*) des patrons peut être stockée. Cette situation restreint leurs applications. Mais, Enertex exploite ce comportement. Si on utilise le modèle vectoriel de textes (Salton *et al.*, 1983) pour transformer les phrases d'un document en vecteurs, ils peuvent être traités comme un réseau de neurones de type Hopfield. Les phrases sont représentées comme des chaînes (patrons) de N neurones actifs (termes présents) ou inactifs (termes absents) avec un vocabulaire de N termes par document. Un document de P phrases est formé de P chaînes dans un espace vectoriel de N dimensions. Ces vecteurs sont plus ou moins corrélés, selon les mots qu'ils partagent. Si les thématiques sont proches, le degré de corrélation sera élevé. Cependant, cela pose des problèmes pour le stockage et la récupération de ces représentations. Cependant, dans le cas du traitement automatique de la langue naturelle l'intérêt porte non pas sur la récupération, mais sur les interactions entre les mots et entre les phrases. Nous calculons l'interaction entre les termes en utilisant à la fois (4) et l'énergie textuelle entre phrases avec (5). Partant de l'hypothèse que l'énergie d'une phrase μ reflète son poids dans le document, nous avons appliqué cette méthode au résumé par extraction.

3.3. Simulation de l'activité de hiérarchisation de phrases par LSA

LSA (Landauer *et al.*, 1997) est un modèle computationnel qui permet de représenter la sémantique à partir des idées que :

- Deux mots sont proches s'ils apparaissent dans des contextes similaires et
- Deux contextes sont similaires s'ils contiennent des mots proches.

Pour opérationnaliser ces définitions, les mots d'un vaste corpus sont représentés dans une matrice d'occurrences. Cette matrice stocke, pour chaque mot du corpus, 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 entre 100 et 300. Chaque mot présent dans le corpus est représenté par un vecteur de tel façon que la proximité sémantique entre deux mots puisse, par exemple, se mesurer en estimant le cosinus de l'angle formé par ses deux vecteurs. 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 l'utilisation, en *input*, d'importants corpus. Dans nos traitements, nous en utilisons 3 différents qui seront réduits en un espace de 300 dimensions :

- *Le Monde (LSA_lemonde)* : il s'agit d'un corpus généraliste de 5 millions de mots. Il contient l'ensemble des articles parus, en 1999, dans le quotidien français Le Monde.
- *Enfants (LSA_enfants)* : il a été construit dans le but de représenter au mieux la mémoire sémantique d'enfants entre 9 et 11 ans (Lemaire *et al.*, 2006). Il possède

3,3 millions de mots issus de productions d'enfants, de contes, de manuels scolaires et d'encyclopédies pour enfants.

- *Adultes (LSA_adultes) :* ce corpus de 13 millions de mots rassemble une large partie des 2 corpus précédents en plus de romans pour adultes. Il est supposé représenter la mémoire sémantique d'un adulte.

LSA est fréquemment employé pour simuler des processus cognitifs (Foltz *et al.*, 1998). Le processus simulé qui nous intéresse présentement est celui de la hiérarchisation de l'importance des phrases d'un texte. Différents modèles basés sur des hypothèses cognitives différentes ont été testés (Lemaire *et al.*, 2005). Nous retiendrons celui dont l'hypothèse stipule que plus une phrase est importante, plus elle est sémantiquement proche de l'intégralité du texte source. Ce modèle produit des résultats (proximités sémantiques entre un texte et chacune de ses phrases) qui sont corrélés aux résultats d'élèves (sélection de 3 à 5 phrases les plus importantes d'un texte) et ne nécessite aucun calibrage particulier.

4. Procédure expérimentale

Nous allons à présent décrire d'une part le protocole par lequel nous avons recueilli des résumés humains et des résumés automatiques, puis d'autre part, le processus d'évaluation des systèmes de génération de résumés automatiques.

4.1. Recueil de résumés humains

Dans une étude précédente (Lemaire *et al.*, 2005), des sélections de phrases importantes (dont la concaténation produit des *extracts*) et des résumés (*abstracts*) ont été produits par des élèves allant de la 4e à la 1re et appartenant à différents établissements français. Nous avons soumis le même protocole à des étudiants en Master 2 (M2) d'informatique. Un feuillet d'exercices comprenant deux tâches a été distribué à chacun : souligner 3 à 5 phrases estimées les plus importantes dans un premier texte, et résumer un second texte. Les textes proposés sont un texte narratif, *Miguel de la faim*[5] *(Miguel*), et un texte expositif, *La pharmacie des éléphants*[6] *(Eléphants)*. Le texte narratif, en proposant une structure linéaire d'événements, est censé être plus aisément résumable que le texte explicatif, censé décrire des concepts plus abstraits (Brewer, 1980). Nous obtenons finalement 296 *extracts* et 372 *abstracts* distribués selon le tableau 1.

4.2. Génération des résumes automatiques

En plus de Cortex, Enertex et LSA, nous avons choisi d'inclure dans cette comparaison d'autres systèmes : Copernic[7], Pertinence[8] 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). Pour garder les mêmes conditions expérimentales, tous les systèmes ont produit deux types de résumés. Le premier représente 30% du texte source. L'objectif est de comparer les *extracts* générés automatiquement à des *abstracts* humains généralement proches de cette valeur. Le deuxième type de résumés se compose des 5 phrases

[5] Vidal, N. (1984). Miguel de la faim. Paris : Rageot.

[6] Pfeffer, P. (1989). Les pharmacies des éléphants. In Vie et mort d'un géant.

[7] http://www.copernic.com/fr/products/summarizer/index.html.

[8] http://www.pertinence.net.

identifiées comme les plus importantes par les systèmes informatiques. Cette fois, l'objectif est de comparer ces *extracts* générés automatiquement à des *extracts* d'humains.

Niveau scolaire		**Nombre de références**			
		Miguel *382 mots, 24 phrases*		*Eléphants* *523 mots, 18 phrases*	
		Abstacts	***Extracts***	***Abstratcs***	***Exstracts***
Collège (13-15 ans)	**4e**	29	23	29	22
	3e	39	24	40	34
Lycée (16-18 ans)	**2nde**	67	48	86	71
	1re	22	14	19	20
Lycée Professionnel (âge variable)	**CAP**	6	6	7	6
Université (22-23)	**M2**	14	14	14	14
	Total	*177*	*129*	*195*	*167*

Tableau 1. Corpus de référence.

4.3. Processus d'évaluation

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 par différents systèmes (résumés candidats) à ceux produits par un nombre de juges humains (résumés de référence). Pour des raisons évidentes de temps et de manque de disponibilité, les juges humains ne sont pas toujours très nombreux. Le fait d'avoir réuni 668 références offre un avantage considérable. Nous avons donc utilisé cet avantage pour mener les tests suivants :

a) *Utilisation d'abstracts humains comme références*. La qualité des juges humains a été évaluée par la lecture directe sur un échantillon de 10 résumés par niveau. Le groupe identifié comme « expert » (Master) a été utilisé comme référence pour mesurer les performances des systèmes automatiques. La comparaison a été réalisée avec le système *ROUGE* (Lin, 2004). Cet outil mesure l'intersection d'ensembles de n-grammes entre les résumés candidats et les résumés de référence. Les métriques utilisées sont ROUGE-2 (bigrammes) et ROUGE-SU4 (bigrammes séparés au maximum par un intervalle de 4 mots).

b) Utilisation *d'extracts humains comme références*. Nous avons fait une analyse sur deux niveaux de granularité :

- Un niveau dans lequel nous repérons les phrases estimées importantes à la fois par les systèmes et par les humains.
- Un niveau dans lequel nous tenons compte des n-grammes avec *ROUGE*. L'analyse porte alors sur les mots ou ensembles de mots du texte.

c) *Analyse discriminant les niveaux scolaires.* Le résumé produit pour chaque système a été comparé à l'ensemble des résumés humains groupés par niveau scolaire.

d) *Etude de la performance des systèmes face aux textes longs et composites.* Les mêmes systèmes de résumé automatique sont testés ici avec des textes plus longs et composés de fragments de plusieurs textes à thématiques différentes.

5. Résultats et évaluation

5.1. Les abstracts humains comme références

5.1.1. Evaluation qualitative d'abstracts humains

Un échantillon de 10 *abstracts* par niveau scolaire et par texte a été choisi au hasard. La lecture directe a permis de juger de la qualité des productions selon les critères de taille du résumé, pertinence du contenu, cohérence, fluidité et continuité des idées. Nous présentons les observations les plus caractéristiques de chaque groupe.

M2 : les résumés sont bien construits. Plusieurs phrases du texte ont été généralisées et utilisent des mots de liaison. Les phrases sont réduites par élimination de mots. Il y a une continuité d'une phrase à l'autre.

1re : les résumés sont courts et concis. Ils sont produits en utilisant des synonymes, en éliminant des mots ou en généralisant. La lecture est fluide. Les phrases sont reliées par des connecteurs ou utilisent des anaphores.

2nde : les résumés sont longs, composés d'un grand nombre de phrases juxtaposées. Ils contiennent beaucoup de détails et de descriptions.

3e : les résumés sont courts. Ils contiennent des généralisations mais aussi des inventions qui s'éloignent du contenu original. Les phrases n'ont pas de continuité.

4e : la rédaction adopte un style narratif composé d'expressions comme « ça parle de », « il était », « c'est une histoire ». Le contenu original est distordu.

CAP : les phrases sont paraphrasées et les mots remplacés par des synonymes. Ces résumés sont parfois incomplets, ils finissent brutalement sans aborder l'intégralité du contenu.

Il semblerait qu'il y ait une corrélation entre la qualité des résumés et le niveau scolaire. Dans les études expérimentales, les étudiants sont souvent considérés comme experts (Bianco *et al.*, 2005), cela semble être le cas ici pour les étudiants M2.

5.1.2. Evaluation ROUGE

Nous avons utilisé ROUGE (métriques 2 et SU4) pour mesurer la performance des systèmes en mesurant la proximité des résumés qu'ils produisent aux résumés du groupe d'experts. Pour *Miguel* (Figure 3), nous observons la présence d'écart types importants. Ceci peut s'expliquer par la présence de dialogues qui favorisent les reformulations et donc l'utilisation d'un lexique différent. Les *abstracts* du texte narratif sont davantage composés de mots non présents dans le texte alors que le ceux du texte explicatif utilise davantage le lexique des textes sources. Si les systèmes numériques Enertex et Cortex, sont performants pour les deux textes, c'est sans doute parce qu'ils sont moins sensibles aux variations lexicales.

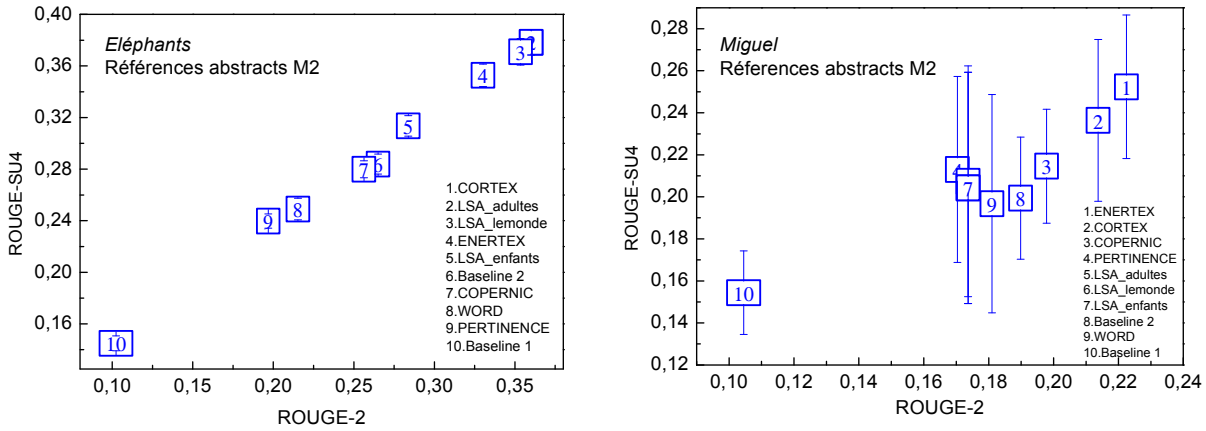

Figure 3. Proximité avec les résumés rédigés par les étudiants de M2.

5.2. Les extracts humains comme références

5.2.1. Evaluation quantitative de la sélection de phrases

Dans la section 5.1, nous avons conclu que les étudiants d'un niveau scolaire plus avancé ont une meilleure conception de la façon de rédiger un abstract. La réalisation de cette tâche inclut notamment un processus de sélection des phrases pertinentes. Aussi, est-ce que dans ce processus nous retrouvons également des différences importantes entre humains de niveaux scolaires différents ? En ce qui concerne le texte *Eléphants* (figure 4) de 18 phrases, les étudiants de M2, 1re, 2nde et 3e en sélectionnent quatre (phrases 1, 7, 11 et 18) comme les plus importantes. Les élèves de CAP s'accordent entre eux que pour trois phrases de cet ensemble (1, 7 et 18) et les 4es seulement sur la phrase 1. En ce qui concerne le texte *Miguel* (figure non présentée) de 24 phrases, les classes de 1res, de 3es et de 4es s'accordent aussi sur 4 d'entre elles (2, 4, 8 et 24). La classe de 2nde et les M2 s'accordent sur seulement trois (4, 8 et 24). Il semble que les humains ont une tendance générale à repérer les mêmes phrases pertinentes. Il est toutefois difficile de reconnaître les experts.

Nous avons déterminé la précision des systèmes en calculant le rapport entre le nombre de phrases pertinentes qu'ils ont repérées et les 4 phrases sélectionnées de façon consensuelle par les humains. Ceci donne un indicateur de la performance du système sur la sélection des phrases importantes. Pour *Eléphants,* les systèmes Cortex, LSA_adultes, LSA_lemonde et Enertex ont une précision de 0,75 ; LSA_enfants, Copernic et Word de 0,5 ; et 0 pour Pertinence. 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. Les systèmes Cortex, LSA_adultes et Enertex semble donc être les plus appropriés pour extraire les phrases pertinentes d'un texte.

5.2.2. Evaluation ROUGE

ROUGE utilise les *n-grammes* des mots pour mesurer l'intersection entre des résumés candidats et des références. Ce niveau d'analyse plus fin permet de mieux mesurer la proximité entre les *extracts* automatiques et les *extracts* humains. Etant donné que dans la sélection de phrases nous n'avons pas pu détecter d'experts, nous avons utilisé les sélections des phrases importantes de tous les individus, tous niveaux scolaires confondus. La figure 5 montre les valeurs des rappels moyens ROUGE-2 *vs* ROUGE-SU4 pour les 8 systèmes et les deux *baselines*. Les rappels moyens sont les plus élevés dans le cas du texte explicatif

Eléphants. Il semble que les systèmes concordent davantage avec les humains quand il s'agit de la présentation de concepts abstraits liés entre eux. Dans les textes narratifs, le thème évolue au fur et mesure de l'avancée du texte et peut même complètement changer. Le système qui semble le plus sensible au type de texte est finalement *Pertinence.* Il est très performant sur un texte narratif et médiocre sur un texte explicatif. Les algorithmes basés sur LSA et Cortex produisent, quant à eux, les meilleurs *extracts* du texte explicatif. La figure 6 présente la performance finale des systèmes telle que la somme normalisée de la précision et le produit moyen des rappels ROUGE-2 * ROUGE-SU4. Cette quantité a été moyennée sur les deux textes. Cortex, LSA_adultes et Enertex obtiennent les meilleurs résultats.

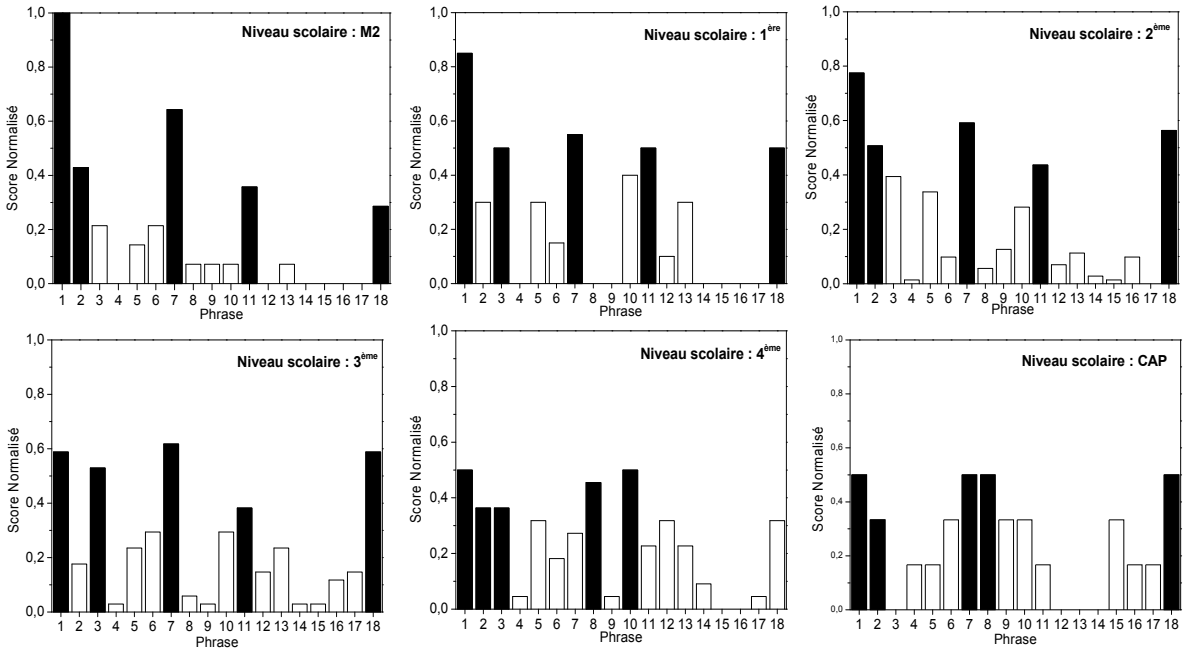

Figure 4. Phrases sélectionnées par les humains des différents niveaux scolaires pour le texte Eléphants (18 phrases). Les bars pleins représentent les 5 phrases avec les meilleurs scores.

5.4. Analyse discriminant les niveaux scolaires

Nous confrontons les systèmes aux jugements humains répartis en niveaux scolaires. Pour chaque système et chaque niveau, nous avons calculé le produit ROUGE-2 * ROUGE-SU4. Cette expérience nous a permis de déterminer (tableau 2) les niveaux scolaires auxquels correspondent les mieux les systèmes résumant automatiquement des textes. Le niveau scolaire attribué aux meilleurs systèmes est le niveau M2. Par contre, les systèmes les moins bons sont plus proches des collégiens (4es et 3es).

5.5. Traitement des textes longs et composites

Les textes que nous avons utilisés précédemment pour recueillir des résumés sont courts et mono-thématiques. Nous nous intéressons donc à présent aux performances des systèmes de résumé automatique face au traitement de textes longs et/ou composites. Un texte composite contient des thématiques différentes. Nous utilisons toujours *ROUGE* pour évaluer les systèmes. Toutefois, les références sont limitées aux *extracts* des étudiants en M2. Est-ce que les systèmes peuvent générer des résumés qui tiennent compte de toutes les thématiques d'un texte similairement aux étudiants ? Le tableau 3 montre la taille, le taux de compression, le nombre de références pour chaque texte utilisé, ainsi que les rappels moyens ROUGE-2 et

ROUGE-SU4 des 8 systèmes et du *baseline* aléatoire. Les systèmes Cortex et Enertex sont toujours bien positionnés. Ils ne semblent donc sensibles ni à la longueur ni à l'hétérogénéité thématique des textes. L'approche avec LSA donne des résultats moins satisfaisants qu'au préalable. Cependant, nous noterons que les corpus utilisés ne conviennent pas nécessairement (beaucoup de mots du texte « Puces » ne sont pas présents dans le corpus) et que l'algorithme avait été réalisé pour des textes courts (« J'accuse » est un texte très long) et à thématique unique (3-mélanges est une concaténation de 3 textes).

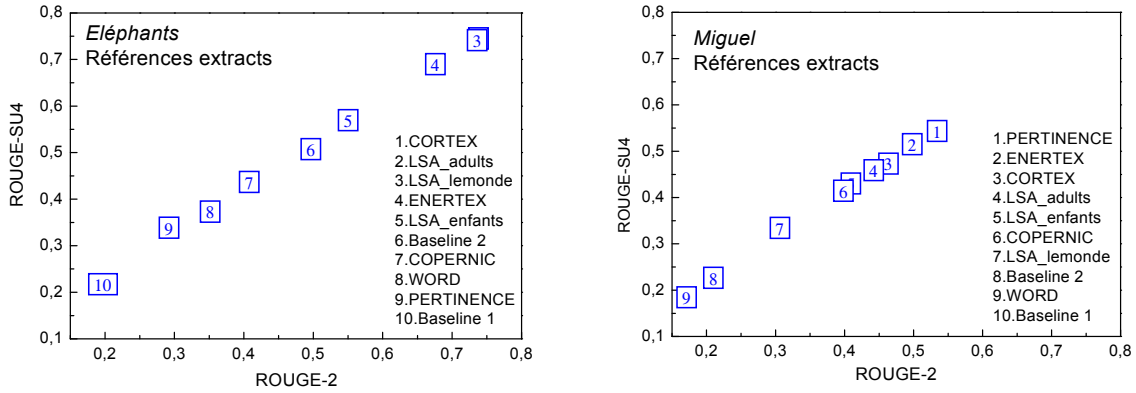

Figure 5. Rappel moyen ROUGE-2 et ROUGE-SU4 pour les 8 systèmes et deux baselines confrontés aux extracts des humains de tous les niveaux confondus.

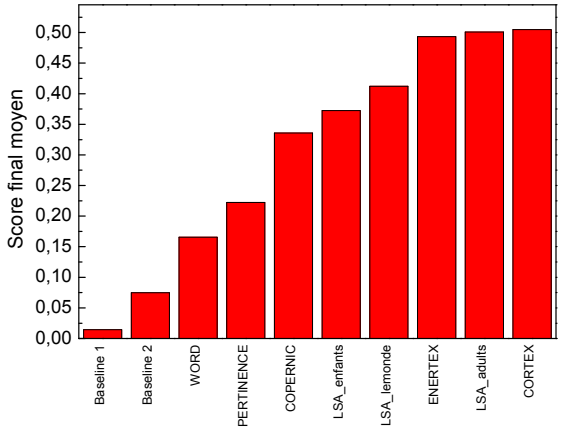

Figure 6. Score final : moyenne du produit des rappels moyens ROUGE-2 X SU4 et la précision sur les deux textes.

6. Conclusions

Nous avons présenté trois systèmes de Résumé Automatique par Extraction (Cortex et Enertex basés sur un modèle vectoriel, et LSA basé sur l'Analyse de la Sémantique Latente). Nous avons montré que ces systèmes, qui reposent sur des méthodes n'ayant (presque) pas besoin de ressources linguistiques, sont bien plus performants que d'autres systèmes testés. Dans cette étude, nous avons confronté tous ces systèmes aux productions de « novices » (collégiens et lycéens) et d'« experts » (étudiants). Ces tests ont confirmé que les résumés générés par les systèmes les plus performants sont davantage similaires à ceux d'« experts ». Nous pensons que cette plus grande proximité de certains systèmes avec les experts est le résultat d'une amélioration des algorithmes utilisés. Ils ne se limitent pas à de simples

analyses fréquentielles et capturent les relations pas toujours évidentes qu'entretiennent les différents termes du texte. De tels algorithmes produisent finalement des résumés qui rendent compte des informations les plus importantes d'un texte. Les systèmes de résumé automatique par extraction ne sont donc pas toujours aussi mauvais élèves que ce que l'on pourrait penser.

Extracts Eléphants				***Extracts Miguel***		
Système	**Niveau**	**Rouge-2XSU4**		**Système**	**Niveau**	**Rouge-2XSU4**
LSA_adultes	M2	0.67190779		Pertinence	M2	0.3080199
Cortex	M2	0.67190779		LSA_adultes	M2	0.2864887
LSA_lemonde	M2	0.66759349		Enertex	1re	0.2826907
Enertex	M2	0.63622629		Cortex	M2	0.2627781
LSA_enfants	M2	0.353795		LSA_enfants	M2	0.2370125
Copernic	M2	0.26738725		Copernic	M2	0.2056173
Baseline 2	M2	0.3027539		LSA_lemonde	4e	0.1578756
Word	3e	0.17862023		*Baseline* 2	3e	0.0719227
Pertinence	4e	0.16615629		Word	4e	0.0420054
Baseline 1	4e	0.09072123		*Baseline* 1	3e	0.0213965

Tableau 2. Niveau scolaire où chaque système maximise la valeur ROUGE-2 X SU4.

	3-mélanges (3 thèmes)[9] 27 phrases, 826 mots taux de compression 25% 8 références		**Puces (2 thèmes)**[8] 29 phrases, 653 mots taux de compression 25% 8 références		**J'accuse**[8] 206 phrases, 4 936 mots taux de compression 12% 6 références	
	R2	**SU4**	**R2**	**SU4**	**R2**	**SU4**
Baseline	0.3074	0.3294	0.3053	0.3272	0.2177	0.2615
Copernic	*0.4231*	*0.4348*	**0.5775**	**0.5896**	0.2235	0.2707
Cortex	**0.4968**	**0.5064**	*0.5360*	*0.5588*	**0.6316**	**0.6599**
Enertex	**0.4958**	**0.5064**	0.5204	0.5336	*0.6146*	*0.6419*
LSA_adultes	0.31010	0.33464	0.26389	0.30493	0.22725	0.26868
LSA_enfants	0.34047	0.35919	0.36766	0.39744	0.22536	0.27510
LSA_lemonde	0.42608	0.43670	0.30820	0.34649	0.19808	0.25138
Pertinenc	0.33101	0.35068	0.40964	0.43704	0.28308	0.31841
Word	0.43006	0.43753	0.16557	0.19552	0.31399	0.34407

Tableau 3. Rappels moyens ROUGE-2 et SU4 pour les 8 systèmes et une baseline aléatoire. En gras les premières places et en italique les deuxièmes.

Références

Barzilay R. and Elhadad M. (1997). Using lexical chains for Text Summarization. *Actes de ACL Intelligent Scalable Text Summarization*, p.10-17.

Bianco M., Dessus P., Lemaire B., Mandin S. and Mendelsohn P. (2005). Modélisation des processus de hiérarchisation et d'application de macrorègles et conception d'un prototype d'aide au résumé. In *Projet ACI École et sciences cognitives 2003-2005*.

Brandow R., Mitze K. and Rau L. (1995). Automatic condensation of electronic publications by sentence selection. Inf. *Proc. and Management.* Vol.(31): 675-685.

Brewer W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R.J. Spiro, B.C. Bruce and W.F. Brewer Eds., *Theoretical Issues in Reading Comprehension*, p.221-239.

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

Edmundson H. P. (1969). New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery* 16(2): 264-285.

Fayol M. (1985). Analyser et résumer des textes : Une revue des études développementales. *Etudes de Linguistique Appliquée*. Vol.(59) : 54-64.

Fernandez S., Sanjuan E. and Torres-Moreno J.-M. (2007). Energie textuelle des mémoires associatives. *Actes de TALN'07*. Vol.(1) : 25-34.

Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*. Vol. (25)(2-3): 285-307.

Goldstein J., Carbonell J., Kantrowitz M. and Mittal V. (1999). Summarizing text documents: sentence selection and evaluation metrics. *Actes de 22nd ACM SIGIR*, p.121-128.

Hopfield J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. of the National Academy of Sciences of the USA*, vol.(9): 2554-2558.

Ibekwe-SanJuan F. (2007). *Fouille de textes. Méthodes, outils et applications*. Hermès-Lavoisier.

Kupiec J., Pedersen J. and Chen F. (1995). A trainable document summarizer. *ACM SIGIR*, p.68-73.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104.

Lemaire B., Mandin S., Dessus P. and Denhière G. (2005). Computational cognitive models of summarization assessment skills. In B.G. Bara, L. Barsalou and M. Bucciarelli (Eds.), *Proc. of the 27th Annual Conference of the Cognitive Science Society,* p.1266-1271.

Lin C. and Hovy E. (1997). Identifying Topics by Position. *Actes de ACL Applied Natural Language Processing Conference*, p.283-290.

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

Mann W. and Thompson S. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. U. of Southern California, Information Sciences Institute.

Mandin S., Dessus P., Lemaire B. and Bianco M. (2005). Un EIAH d'aide à la production de résumés de textes. In P. Tchounikine, M. Joab, and L. Trouche (Eds.), *EIAH 2005,* p.69-80.

Mani I. and Maybury M. T. (1999). Advances *in Automatic Text Summarization.* MIT Press.

Manning C. D. and Schutze H. (2000). *Foundations of Statistical Natural Language Proc*. MIT Press.

Minel J.-L. (2004). Le résumé automatique de textes : solutions et perspectives. In *Proceedings of TAL, Résumé automatique de textes*. Vol.(45)/1 : 7-13.

Paice C. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*. Vol. (26) (1) : 171-186.

Porter M. (1980). An algorithm for suffix stripping. *Program*, vol(14) (3): 130-137.

Salton G. and Mcgill M. (1983). Introduction to modern information retrieval. *Computer Science Series,* McGraw Hill Publishing Company.

Torres-Moreno J.-M, Velázquez-Morales P. and Meunier J.-G. (2001). Cortex : un algorithme pour la condensation automatique des textes. *ARCo'01*. Vol.(2) : 365-366.

Yousfi-Monod M. and Prince V. (2006). Compression de phrases par élagage de l'arbre morpho-syntaxique. *Technique et Science Informatiques.* Vol.(25)(4) : 437-468.