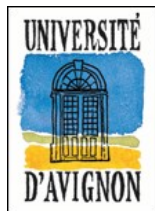




Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?

*Silvia Fernández, Patricia Velázquez, Sonia Mandin, Eric San Juan,
Juan Manuel Torres*



Deux méthodes pour résumer

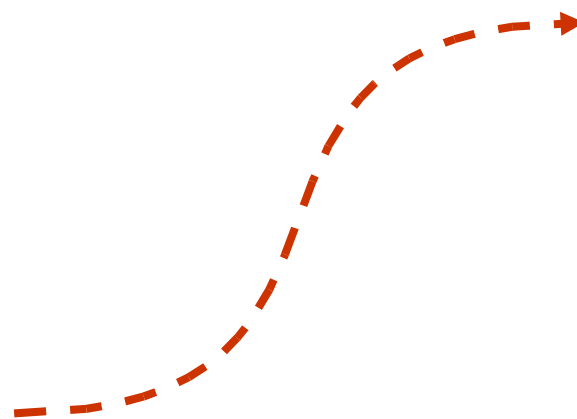
Bon élève

- Bonne compréhension
- Sélection des idées importantes
- Contractions
- Reformulations
- Lexique personnel
- *Abstracts*



Mauvais élève

- Sélection de phrases clefs
- Concaténation des phrases sélectionnées => *Extracts*
- Plus facile à mettre en œuvre !!



Système
automatique

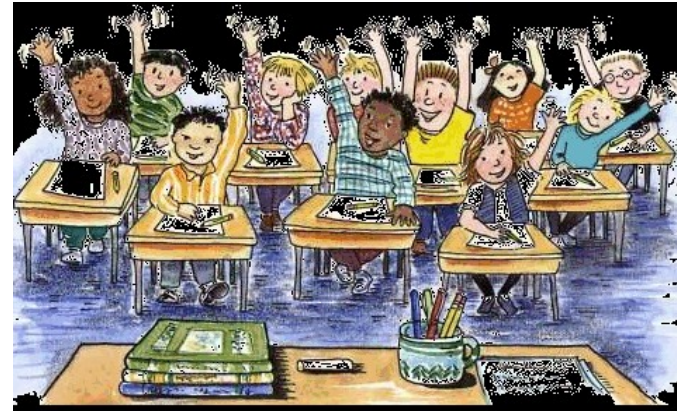


Stratégie :

Systèmes automatiques :
CORTEX, ENERTEX, LSA
Copernic, Pertinence, Word



~ 600 résumés d'élèves et étudiants :
Collège, Lycée, Master



**Comparaisons croisées
ROUGE**

**Déterminer le niveau scolaire auquel correspondent
le mieux les systèmes automatiques**

Résumés automatiques par extraction

Représentation vectorielle de textes

Pretraitement : mots filtrés,
normalisés et lemmatisés
(Porter, 1980; Manning & Schutze, 2000)

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 blues toutes neuves.

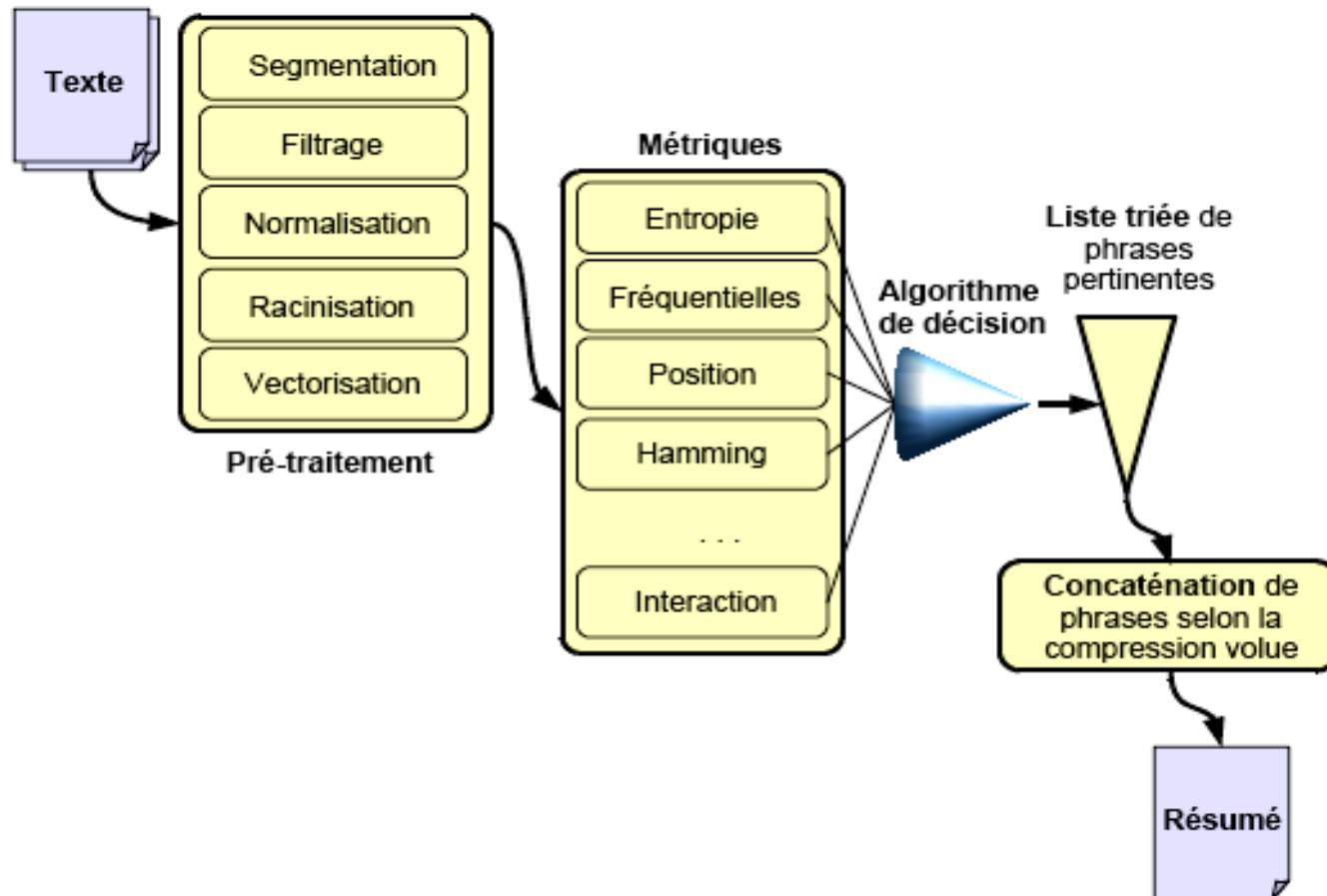
maison	bleu	tante	appeler	lulu	adorer	neuf	chaussure	couleur
TF	TF	TF	0	0	0	0	0	0
0	0	TF	TF	TF	0	0	0	0
TF	0	0	0	0	TF	0	0	0
0	TF	0	0	0	0	0	0	TF
0	TF	0	0	0	0	TF	TF	0

$$S = \begin{pmatrix} s_1^1 & s_1^2 & \dots & s_1^N \\ s_2^1 & s_2^2 & \dots & s_2^N \\ \vdots & \vdots & \ddots & \vdots \\ s_p^1 & s_p^2 & \dots & s_p^N \end{pmatrix} \quad s_{\mu}^i = \begin{cases} TF^i \rightarrow existe \\ 0 \rightarrow autrement \end{cases}$$

Système automatique 1 : Cortex

Cortex es Otro Resumidor de TEXtos

(Torres-Moreno *et al.*, 2001)



Système automatique 1 : Cortex...

L'algorithme de décision

$$\text{Si } \left(\sum \alpha > \sum \beta \right) \quad score_s = 0.5 + \frac{\sum \alpha}{\Gamma} : \text{retenir } s$$

$$\text{sinon} \quad score_s = 0.5 + \frac{\sum \alpha}{\Gamma} : \text{éliminer } s$$

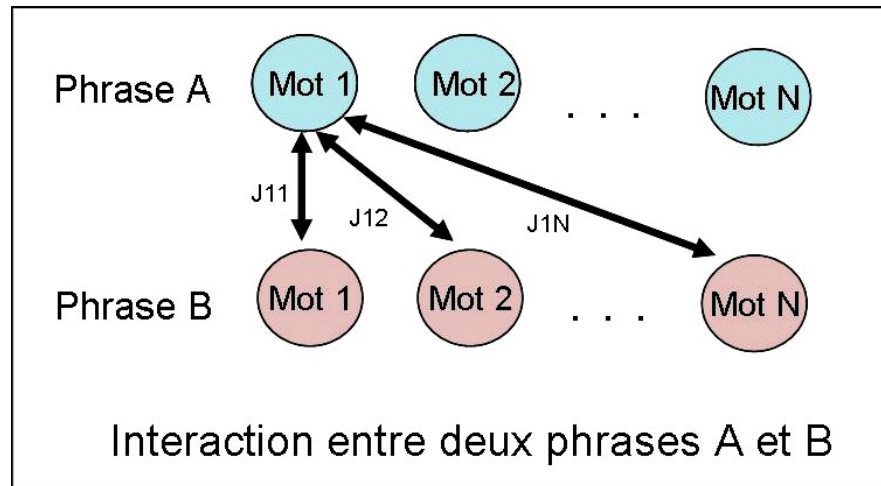
$$\sum \alpha = \sum_{v=1}^{\Gamma} (\|\lambda_s^v\| - 0.5); \|\lambda_s^v\| > 0.5$$

$$\sum \beta = \sum_{v=1}^{\Gamma} (0.5 - \|\lambda_s^v\|); \|\lambda_s^v\| < 0.5$$

Système automatique 2 : Enertex

Energie Textuelle entre phrases

(Fernández *et al.*, 2007)



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

L'énergie du modèle magnétique d'Ising

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

Règle d'Hebb (réseaux neuronaux)

Système automatique 2 : Enertex...

Energie Textuelle entre phrases

$$E = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N s^i J_{i,j} s^j \longrightarrow E = (S \times S^T)^2$$

Chaque élément de la matrice E représente l'énergie textuelle entre une couple des phrases.

Le résumé est formé avec les phrases ayant la plus grande énergie absolue.

Système automatique 3 : basé sur LSA

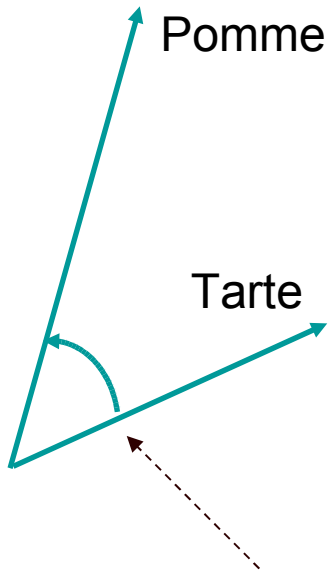
Principe (Landauer & Dumais 1997) :

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



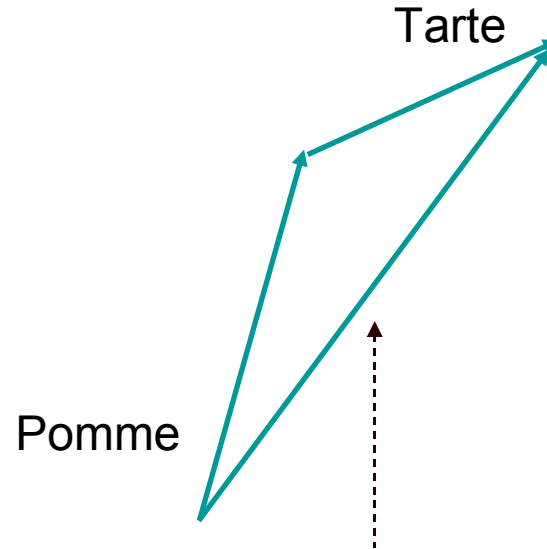
Systeme automatique 3 : basé sur LSA ...

Mot => vecteur



**Distance sémantique entre
deux éléments = Cosinus entre
les deux vecteurs**

**Paragraphe => vecteur somme
des vecteurs des mots**

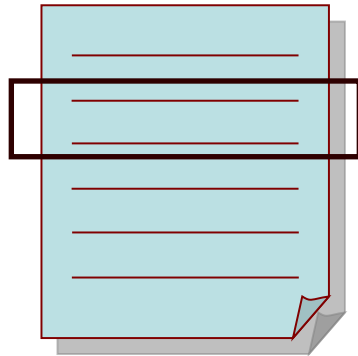


**Vecteur du paragraphe
« Tarte Pomme »**

LSA : modélisation de la sélection des phrases importantes d'un texte

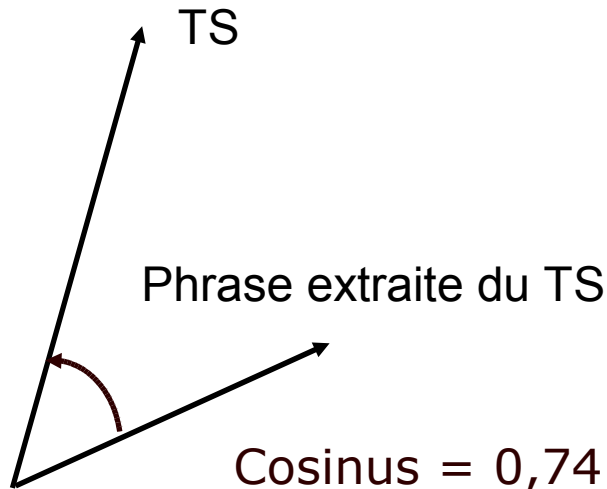
(Mandin, Dessus, Lemaire & Bianco, 2005)

Texte source



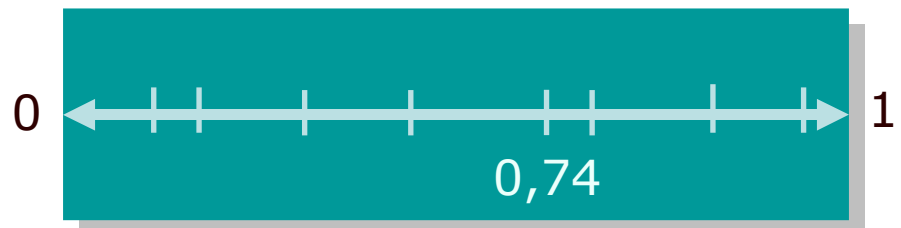
Phrase extraite du TS

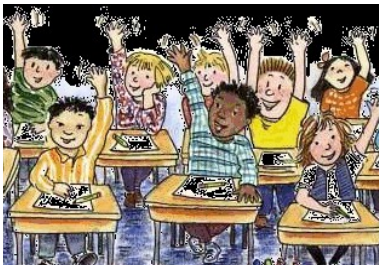
Comme tous les herbivores,
ou plus précisément...



**Phrases triviales
du TS**

**Phrases
importantes du TS**





Recueil de résumés humains

(Lemaire *et al.*, 2005)

Un feuillet d'exercices comprenant deux tâches a été distribué à chaque élève au étudiant :

1.

Texte original	Texte justifié
Le président américain a justifié mercredi soir le «programme» des prisons secrètes par les nécessités de la guerre contre le terrorisme. 14 prisonniers détenus par la CIA ont été transférés à Guantanamo dans l'attente de leur jugement. Parmi eux, Khalid Cheik Mohamed, le «cerveau» des attentats du 11 septembre. Changement de cap. Lors d'un discours mercredi devant plusieurs familles de victimes des attentats du 11 septembre, George Bush a annoncé plusieurs modifications majeures de sa politique concernant les prisonniers capturés dans le cadre de la «guerre contre le terrorisme». Pour la première fois, le président américain a reconnu que certains suspects ont été détenus et interrogés à l'étranger par la CIA. L'information, révélée par la presse américaine, était jusqu'ici interdite par la Maison Blanche. Bush a justifié ce «programme», réservé	Le président américain a justifié mercredi soir le «programme» des prisons secrètes par les nécessités de la guerre contre le terrorisme. 14 prisonniers détenus par la CIA ont été transférés à Guantanamo dans l'attente de leur jugement. Parmi eux, Khalid Cheik Mohamed, le «cerveau» des attentats du 11 septembre. Changement de cap. Lors d'un discours mercredi devant plusieurs familles de victimes des attentats du 11 septembre, George Bush a annoncé plusieurs modifications majeures de sa politique concernant les prisonniers capturés dans le cadre de la «guerre contre le terrorisme». Pour la première fois, le président américain a reconnu que certains suspects ont été détenus et interrogés à l'étranger par la CIA. L'information, révélée par la presse américaine, était jusqu'ici interdite par la Maison Blanche. Bush a justifié ce «programme», réservé

résumer un premier texte



ABSTRACT

2.

Notes (Ms.10.338, folios 266-291, Carnets d'enquêtes, op. cit., p. 38) La marée arrive sur des camions de chemin de fer. On jette du allées. Les moules sont dans des sacs bruns, mouillés, qui ressemblent de charbon. J'ai vu aussi arriver des campes encore vivantes dans de vide dans les viviers de la criée. Quant aux poissons de mer, ils sont paniers, bas et en pointe, plus ou moins grands. En général, petits. Il bouches longues. Le tout est garni de paille. Ces paniers sont mis grandes cages à grande claire-voie, qu'on doit charger dans les wagons arrivées on charge sur les camions sans toucher aux paniers. C'est une porte derrière permet de prendre les paniers jusqu'au fond. Chacuns deux ou trois cages. On décharge dans l'encointe des criées et peu à peu sur le cas Les sacs de moules sont remplis, et laissent couler leur eau. Les pa augmentent toujours. Avant la criée, il y a quelques poissonnières qui rôdent, des empl Les hommes de l'expéditeur, en blouse, avec des sacs de cuir, atten assis contre le bureau de vente. Les éperians arrivent dans de petits paniers carrés, pareils à c fraises. Il y a des merlans dans des boîtes. Cependant les compteurs-verseurs débaltent le poisson et le r paniers plats, marnes plates, superbes natures mortes. Anguilles de ou trois, en rond, gris perle bleuâtre ; avec le ventre plus blanc, de p Elles glissent l'une sur l'autre, elles semblent vivantes. On les prend crochets. Brochets dorés et grisâtres. Maquereaux à dos strié de bru et dorée, à ventre aux reflets de nacre rose. Soles par paire, grises. Crevettes grises dans des paniers. Crevettes roses, grosses, sur des Grondins rose très vif, avec les ventres blancs, fleurs à panachures blanches ; les tilas sont réunies au milieu. Grandes raies, à ventre t
--

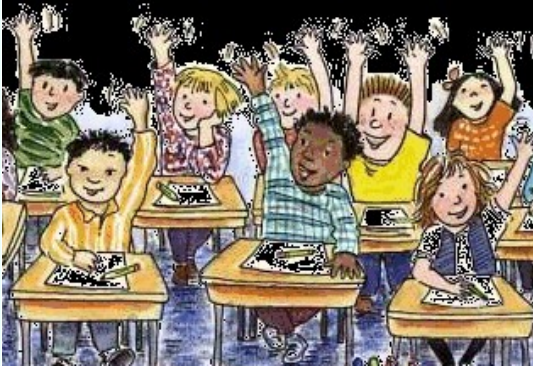
souligner 3 à 5 phrases estimées les plus importantes dans un second texte

Notes (Ms.10.338, folios 266-291, Carnets d'enquêtes, op. cit., p. 38) La marée arrive sur des camions de chemin de fer. On jette du allées. Les moules sont dans des sacs bruns, mouillés, qui ressemblent de charbon. J'ai vu aussi arriver des campes encore vivantes dans de vide dans les viviers de la criée. Quant aux poissons de mer, ils sont paniers, bas et en pointe, plus ou moins grands. En général, petits. Il bouches longues. Le tout est garni de paille. Ces paniers sont mis grandes cages à grande claire-voie, qu'on doit charger dans les wagons arrivées on charge sur les camions sans toucher aux paniers. C'est une porte derrière permet de prendre les paniers jusqu'au fond. Chacuns deux ou trois cages. On décharge dans l'encointe des criées et peu à peu sur le cas Les sacs de moules sont remplis, et laissent couler leur eau. Les pa augmentent toujours. Avant la criée, il y a quelques poissonnières qui rôdent, des empl Les hommes de l'expéditeur, en blouse, avec des sacs de cuir, atten assis contre le bureau de vente. Les éperians arrivent dans de petits paniers carrés, pareils à c fraises. Il y a des merlans dans des boîtes. Cependant les compteurs-verseurs débaltent le poisson et le r paniers plats, marnes plates, superbes natures mortes. Anguilles de ou trois, en rond, gris perle bleuâtre ; avec le ventre plus blanc, de p Elles glissent l'une sur l'autre, elles semblent vivantes. On les prend crochets. Brochets dorés et grisâtres. Maquereaux à dos strié de bru et dorée, à ventre aux reflets de nacre rose. Soles par paire, grises. Crevettes grises dans des paniers. Crevettes roses, grosses, sur des Grondins rose très vif, avec les ventres blancs, fleurs à panachures blanches ; les tilas sont réunies au milieu. Grandes raies, à ventre t
--

EXTRACT

grandes cages à grande claire-voie, qu'on doit charger dans les wagons arrivées on charge sur les camions sans toucher aux paniers. C'est une porte derrière permet de prendre les paniers jusqu'au fond. Chacuns deux ou trois cages. On décharge dans l'encointe des criées et peu à peu sur le cas Les sacs de moules sont remplis, et laissent couler leur eau. Les pa augmentent toujours. Avant la criée, il y a quelques poissonnières qui rôdent, des empl Les hommes de l'expéditeur, en blouse, avec des sacs de cuir, atten assis contre le bureau de vente. Les éperians arrivent dans de petits paniers carrés, pareils à c fraises. Il y a des merlans dans des boîtes. Cependant les compteurs-verseurs débaltent le poisson et le r paniers plats, marnes plates, superbes natures mortes. Anguilles de ou trois, en rond, gris perle bleuâtre ; avec le ventre plus blanc, de p Elles glissent l'une sur l'autre, elles semblent vivantes. On les prend crochets. Brochets dorés et grisâtres. Maquereaux à dos strié de bru et dorée, à ventre aux reflets de nacre rose. Soles par paire, grises. Crevettes grises dans des paniers. Crevettes roses, grosses, sur des Grondins rose très vif, avec les ventres blancs, fleurs à panachures blanches ; les tilas sont réunies au milieu. Grandes raies, à ventre t
--

Corpus de référence

		Nombre de références			
		<i>Texte 1 (Miguel)</i> 382 mots, 24 phrases		<i>Texte 2 (Éléphants)</i> 523 mots, 18 phrases	
		<i>Abstracts</i>	<i>Extracts</i>	<i>Abstracts</i>	<i>Extracts</i>
Collège (13-15 ans)	4 ^{ème}	29	23	29	22
	3 ^{ème}	39	24	40	34
Lycée (16-18 ans)	2 ^{nde}	67	48	86	71
	1 ^{ère}	22	14	19	20
Lycée Professionnel (âge variable)	CAP	6	6	7	6
Université (22-23)	M2	14	14	14	14
<i>Sous-total</i>		<i>177</i>	<i>129</i>	<i>195</i>	<i>167</i>
Total		668			



Les ABSTRACTS humains comme références

- Évaluation qualitative par lecture directe échantillonnée
- Identification des experts
- Comparaison entre résumés humains des experts vs résumés automatiques (ROUGE-2 et SU4)

La qualité des *abstracts* humains



M2 – Très bons résumés : Bien construits, phrases généralisées ou réduites, mots de liaison, cohésion

EXPERTS

1ère – Très bons résumés : Courts et concis, synonymes, élimination, généralisation. Connecteurs, anaphores. Lecture fluide

2nde – Résumés moyens : Longs, beaucoup de détails et de descriptions

3ème – Résumés insatisfaisants : Généralisations mais aussi des inventions. Pas de cohésion

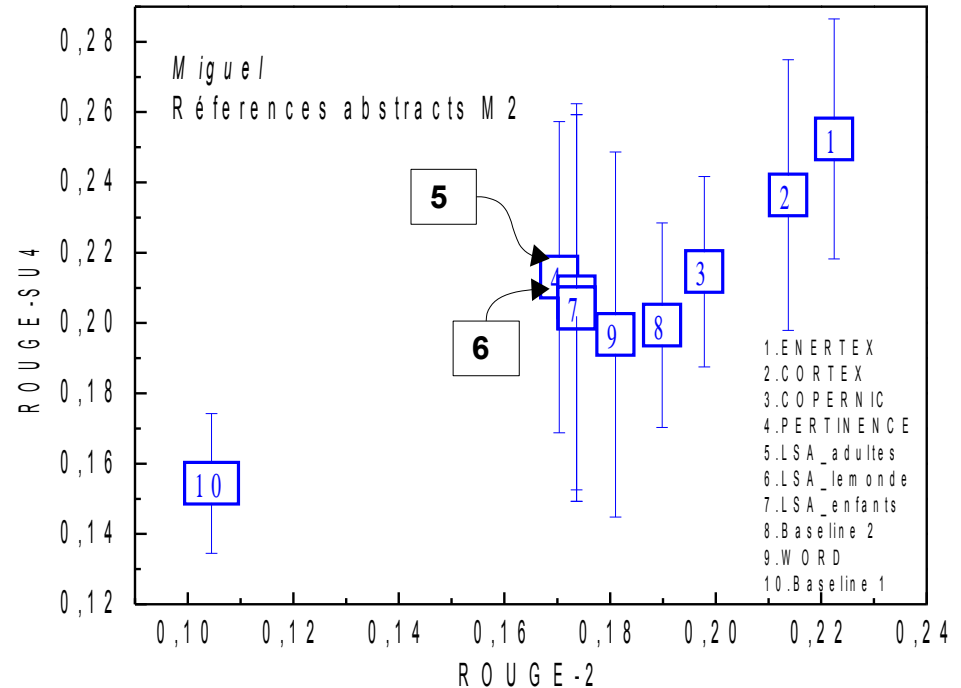
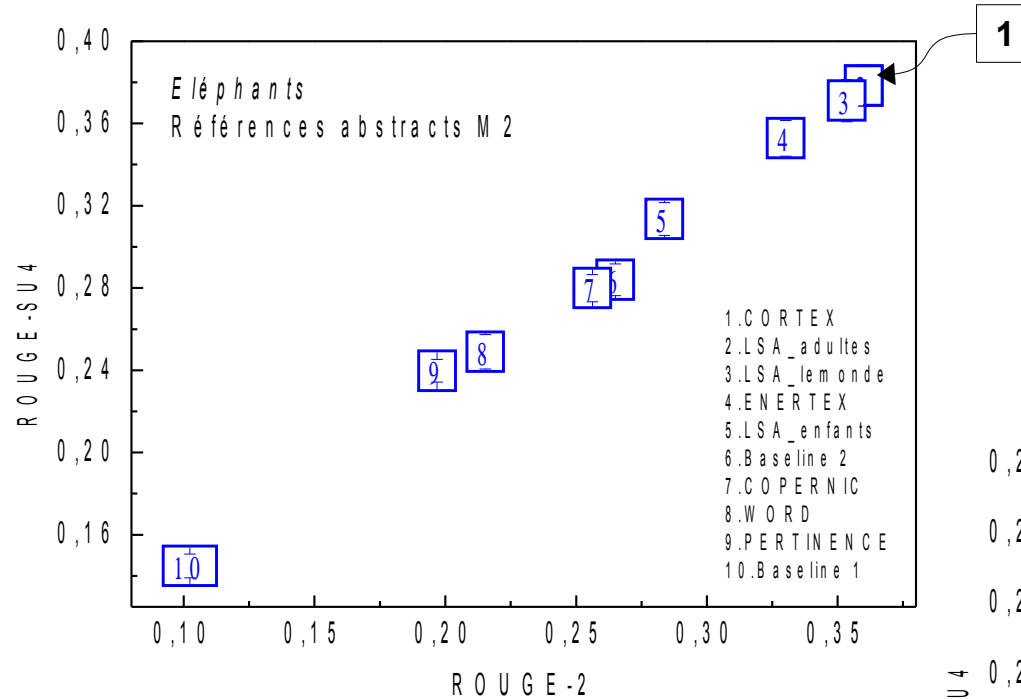
4ème – Très mauvais résumés : Style narratif. Le contenu original est déformé

CAP - Très mauvais résumés : Paraphrases, recours uniquement à des synonymes

Corrélation entre la qualité des résumés et le niveau scolaire



La performance des systèmes selon les «experts»



grands cages à grande cage-voie, qu'on doit charger dans les wagons.
l'arrivée on charge sur les camions sans toucher aux paniers. C'est q
Une porte derrière permet de prendre les paniers jusqu'au fond. Cha
deux ou trois cages.
On décharge dans l'enceinte des criées et peu à peu sur le r
Les sacs de moules sont remplis, et laissent couler leur eau. Les pa
augmentent toujours.
...Avant la criée, il y a quelques poissonnières qui rôdent, des empl
Les hommes de l'expéditeur, en blouse, avec des sacs de cuir, atten
assis contre le bureau de vente.
Les éperlans arrivent dans de petits paniers carrés, pareils à
traises. Il y a des merlans dans des boîtes.
Cependant les compteurs-verseurs débaltent le poisson et le r
paniers plats, manges plats, superbes natures mortes. Anouilles de

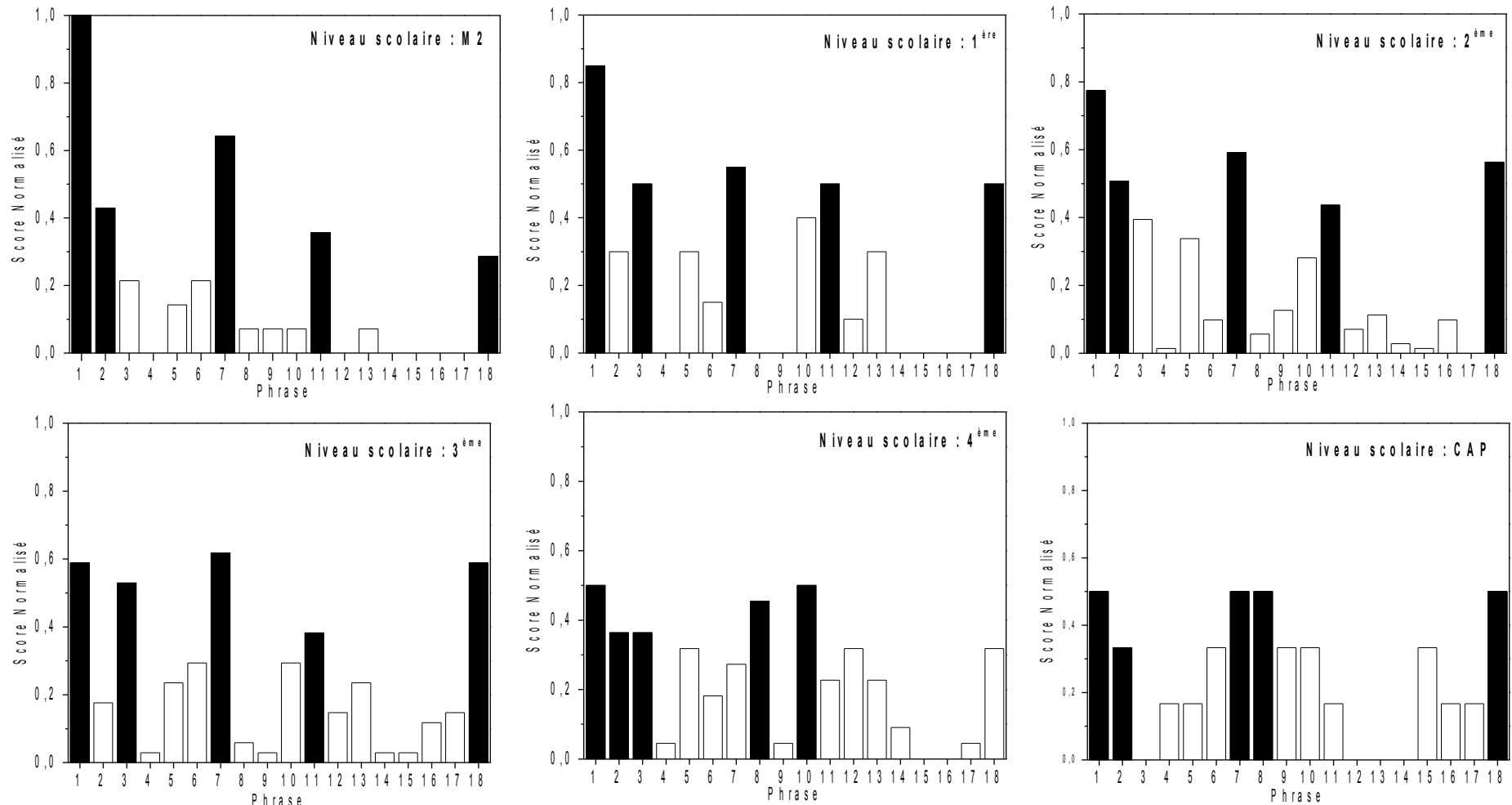
Les extracts humains comme références

1. Analyse de la sélection de phrases chez les humains
2. Peut-on caractériser des expertes dans el processus de sélection des phrases pertinentes ?
3. Évaluation au niveau de phrases
4. Évaluation au niveau de n-grammes (ROUGE)
5. Analyse discriminant les niveaux scolaires

Sélection des phrases chez les humains

Est-ce qu'il y a des expertes ?

Texte Éléphants



Les humains identifient presque les mêmes phrases importantes.

Il est difficile de reconnaître les experts

Précision de systèmes

$$P = \frac{\text{Nb. Phrs. pertinentes}}{\text{Nb. Phrs. rapportées}}$$

Éléphants

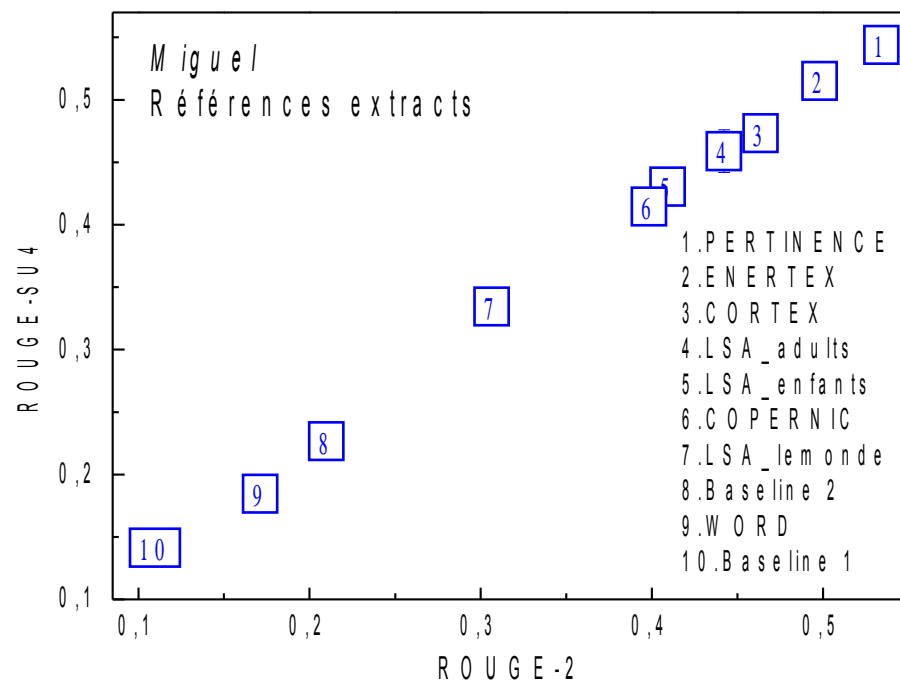
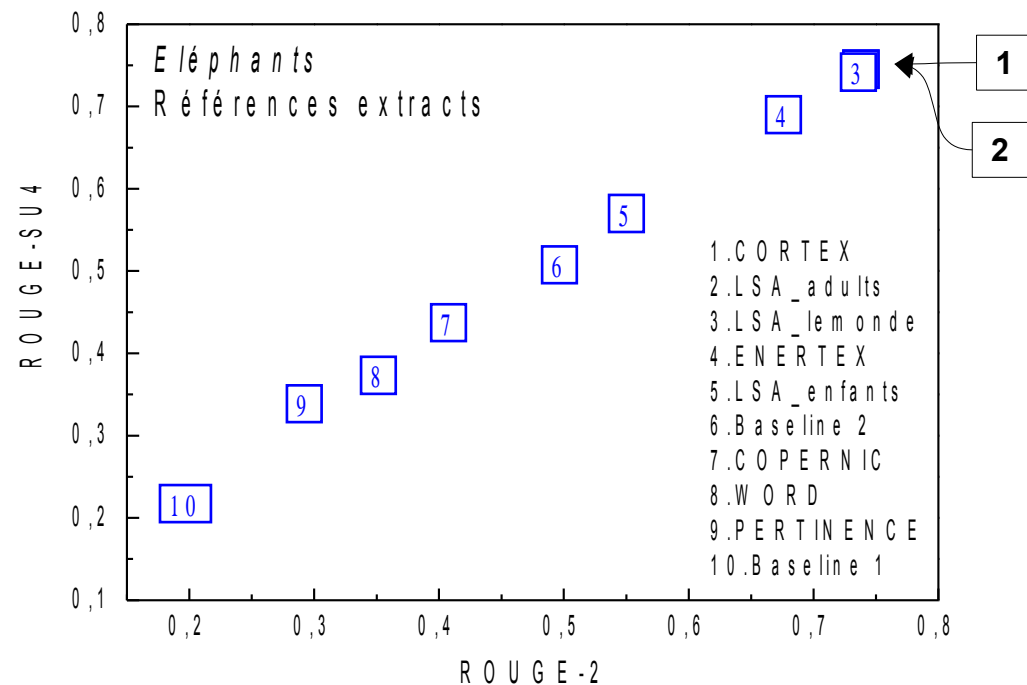
Systèmes	P
Cortex, LSA_adultes, LSA_lemonde et Enertex	0,75
LSA_enfants, Copernic et Word	0,50
Pertinence	0,0

Miguel

Systèmes	P
Cortex, LSA_adultes, LSA_enfants, Enertex, Copernic, Pertinence	0,50
LSA_lemonde	0,25
Word	0,0

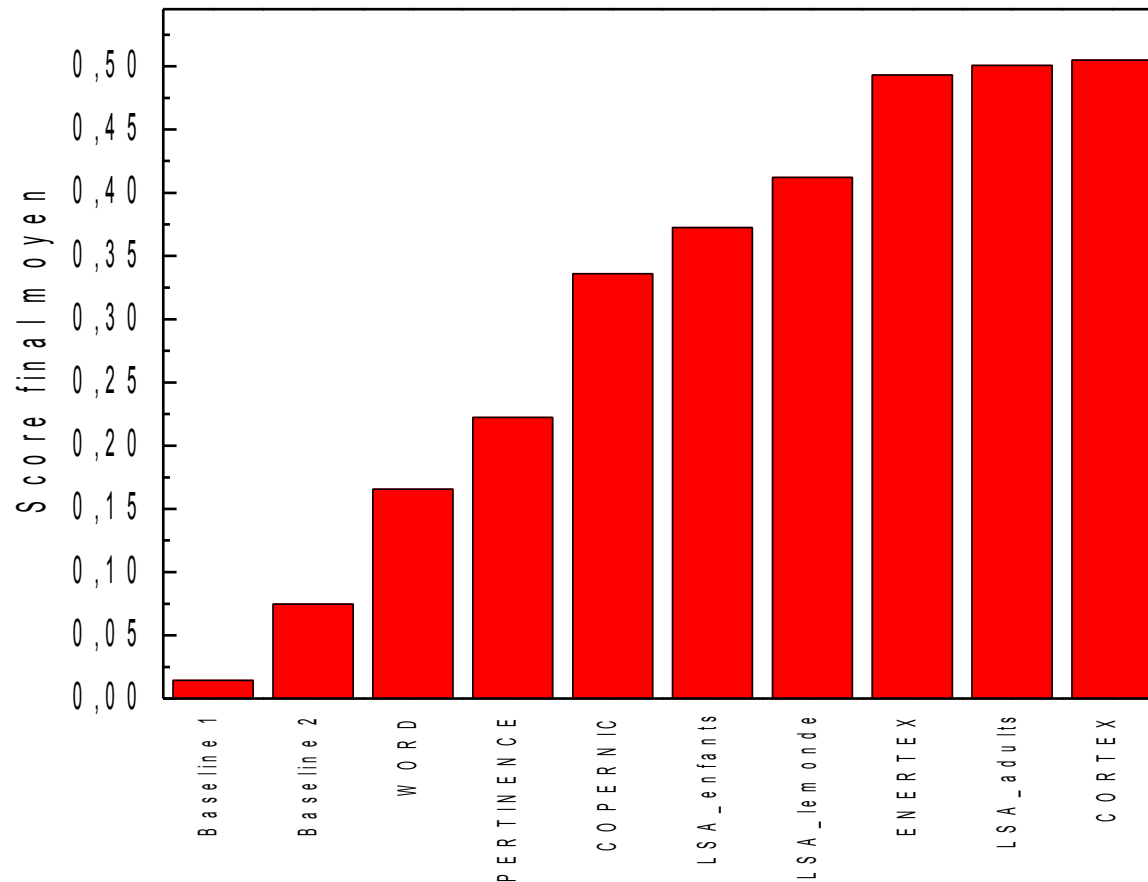
Évaluation plus fine avec ROUGE

Tous les niveaux scolaires confondus



Performance moyenne des systèmes

Sur les deux textes



Discriminant les niveaux scolaires

Mais, en fin...à quel niveau scolaire correspondent le mieux les systèmes de résumé automatique ?

<i>Texte Eléphants</i>		
Système	Niveau	Rouge-2XSU4
LSA_adultes	M 2	0.6719
Cortex	M 2	0.6719
LSA_le monde	M 2	0.6676
Energex	M 2	0.6362
LSA_enfants	M 2	0.3538
Copernic	M 2	0.2674
Baseline 2	M 2	0.3028
Word	3 ^{ème}	0.1786
Pertinence	4 ^{ème}	0.1662
Baseline 1	4 ^{ème}	0.0907

<i>Texte Miguel</i>		
Système	Niveau	Rouge-2XSU4
Pertinence	M 2	0.3080
LSA_adultes	M 2	0.2865
Energex	1 ^{ère}	0.2827
Cortex	M 2	0.2628
LSA_enfants	M 2	0.2370
Copernic	M 2	0.2056
LSA_le monde	4 ^{ème}	0.1579
Baseline 2	3 ^{ème}	0.0719
Word	4 ^{ème}	0.0420
Baseline 1	3 ^{ème}	0.0214

Textes longs et composites

Est-ce que les systèmes génèrent des résumés qui tiennent compte de toutes les thématiques similairement aux « experts » ?

	3-m é l a n g e s (3 t h è m e s) 27 phrases, 826 mots taux de compression 25 % 8 références		P u c e s (2 t h è m e s) 29 phrases, 653 mots taux de compression 25 % 8 références		J ' a c c u s e 206 phrases, 4 936 mots taux de compression 12 % 6 références	
	R 2	S U 4	R 2	S U 4	R 2	S U 4
C o p e r n i c	0.4231	0.4348	0.5775	0.5896	0.2235	0.2707
C o r t e x	0.4968	0.5064	0.5360	0.5588	0.6316	0.6599
E n e r t e x	0.4958	0.5064	0.5204	0.5336	0.6146	0.6419
L S A _ a d u l t e s	0.31010	0.33464	0.26389	0.30493	0.22725	0.26868
L S A _ e n f a n t s	0.34047	0.35919	0.36766	0.39744	0.22536	0.27510
L S A _ l e m o n d e	0.42608	0.43670	0.30820	0.34649	0.19808	0.25138
P e r t i n e n c e	0.33101	0.35068	0.40964	0.43704	0.28308	0.31841
W o r d	0.43006	0.43753	0.16557	0.19552	0.31399	0.34407
B a s e l i n e	0.3074	0.3294	0.3053	0.3272	0.2177	0.2615

R a p p e l s m o y e n s R O U G E - 2 e t S U 4 . E n g r a s l e s p r e m i è r e s p l a c e s e t e n i t a l i q u e l e s d e u x i è m e s .

Conclusions

- Nous avons présenté trois systèmes de Résumé Automatique par Extraction : Cortex, Enertex et un modèle basé sur LSA
- Ces systèmes n'ont (presque) pas besoin de ressources linguistiques
- Leur performance est supérieure à celle d'autres systèmes testés

Conclusions...

- Confronter les systèmes aux productions de « novices » (collégiens et lycéens) et d'« experts » (étudiants)
- Les résumés générés par les systèmes les plus performants sont davantage similaires à ceux d'« experts » :
 - Les algorithmes ne se limitent pas à de simples analyses fréquentielles et capturent les relations pas toujours évidentes qu'entretiennent les différents termes du texte



Les systèmes de résumé
automatique par extraction ne sont
pas toujours aussi mauvais élèves
que ce que l'on pourrait penser !