Traitement Automatique de la Langue Naturelle
TALN Avignon, juin 2008

ENERTEX
un système basé sur l'énergie textuelle

Silvia FERNANDEZ

Juan Manuel TORRES, Eric SANJUAN
Laboratoire Informatique, Université d'Avignon

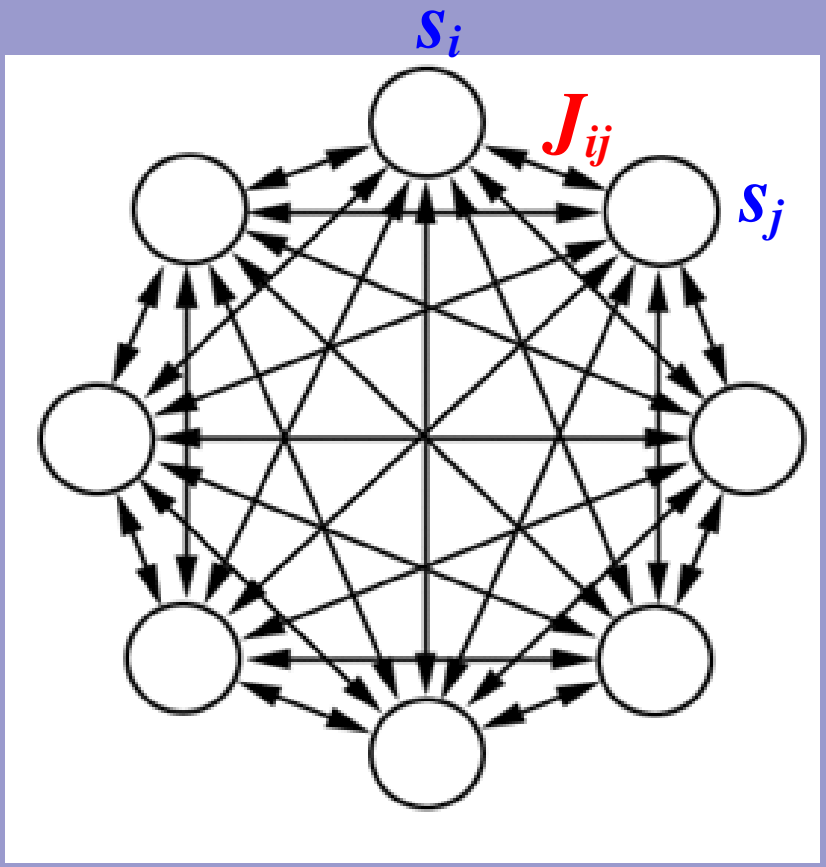
Comment visualiser un texte dès le point de vue de la physique ?

Texte : système composé d'unités en interaction (mots, phrases) dont les propriétés peuvent être étudiés

Départ :

- mémoire associative (réseaux de neurones)
- chaînes binaires en interaction
- garder et récupérer d'information

Mémoire associative de Hopfield
(Hopfield, 1982)



Neurone active/inactive= Spin 

Apprentissage

$J_{ij} = s_i\, s_j$ Règle d'Hebb

Récupération

$s_i = sgn(\Sigma\, J_{ij}\, s_j)$

$E = \Sigma\, J_{ij}\, s_i\, s_j$

Minimisation de l'énergie d'Ising

Limitations

- Patrons corrélés → erreur de récupération
- Capacité $\approx 0,14\, N$, étant N le nombre de neurones
- Etats : récupération → mélangés → amnesie totale

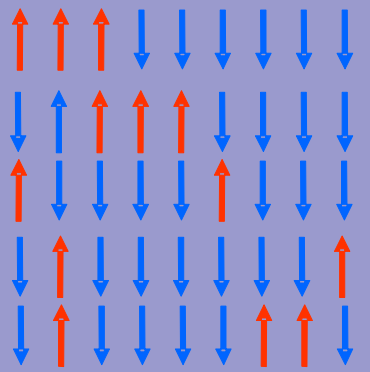
Codage des documents comme un système de spins

Les maisons bleues de ma tante.
Ma tante s'appelle Lulu.
J'adore sa maison.
Le bleu est ma couleur préférée.
J'ai une nouvelle paire de chaussures bleues.

Corrélation !!!

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

- Modèle vectoriel (sac-à-mots)
- Mots filtrés, normalisés et lemmatisés
  (Porter, 1980; Manning & Schutze, 2000)



Énergie textuelle

mot ~ spin s_i $[s_0 \quad s_1 \quad s_2 \ldots. \ s_N]$ Phrase ~ chaîne des spins

Document →

$$S = \begin{pmatrix} s_1^1 & s_2^1 & \cdots & s_N^1 \\ s_1^2 & s_2^2 & \cdots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^P & s_2^P & \cdots & s_N^P \end{pmatrix}$$

Matrice terme-segment

$$J = (S^T \times S) = \begin{pmatrix} j_{1,1}^{\mu} & j_{1,j}^{\mu} & \cdots & j_{1,N}^{\mu} \\ \vdots & \vdots & \ddots & \vdots \\ j_{i,1}^{\mu} & j_{i,j}^{\mu} & \cdots & j_{i,N}^{\mu} \\ \vdots & \vdots & \ddots & \vdots \\ j_{N,1}^{\mu} & j_{N,j}^{\mu} & \cdots & j_{N,N}^{\mu} \end{pmatrix}$$

Règle de Hebb

$$E = -S \times J \times S^T = \begin{pmatrix} e^{1,1} \ldots & e^{1,P} \\ \vdots & \vdots \\ e^{\mu,1} \ldots & e^{\mu,P} \\ \vdots & \vdots \\ e^{P,1} \ldots & e^{P,P} \end{pmatrix}$$

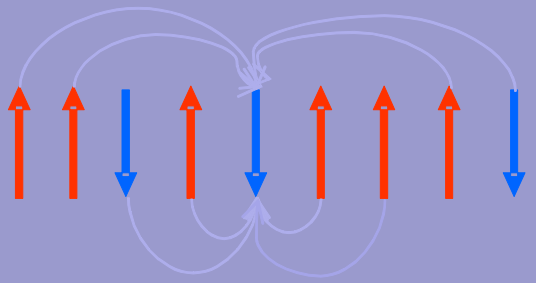
Energie textuelle

$e^{\mu,\nu}$ = énergie entre la phrase μ et la phrase ν

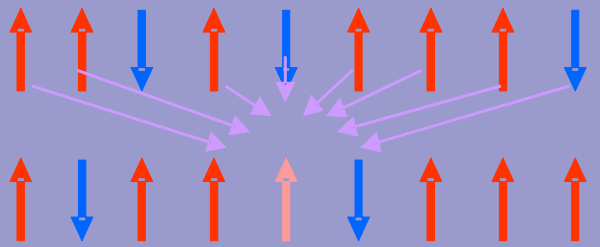
$E=$

$$\begin{pmatrix} e^{1,1} & \cdots & e^{1,P} \\ \vdots & & \vdots \\ e^{\mu,1} & \cdots & e^{\mu,P} \\ \vdots & & \vdots \\ e^{P,1} & \cdots & e^{P,P} \end{pmatrix}$$

diagonale

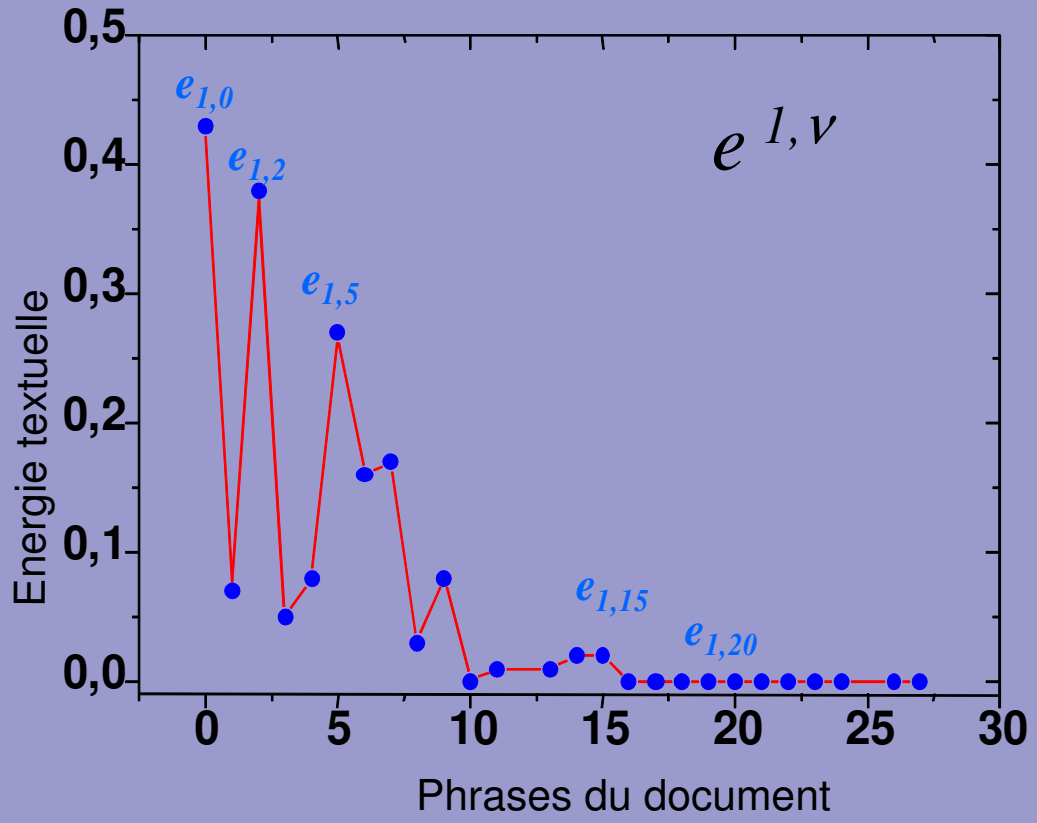


dehors de la diagonale



$$E = \begin{pmatrix} e^{1,1} & \dots & e^{1,P} \\ \vdots & & \vdots \\ e^{\mu,1} & \dots & e^{\mu,P} \\ \vdots & & \vdots \\ e^{P,1} & \dots & e^{P,P} \end{pmatrix}$$

$$E^{\mu,\,docto} = \sum |e^{\nu,\mu}|$$



les phrases plus énergétiques -> les plus représentatives du contenu

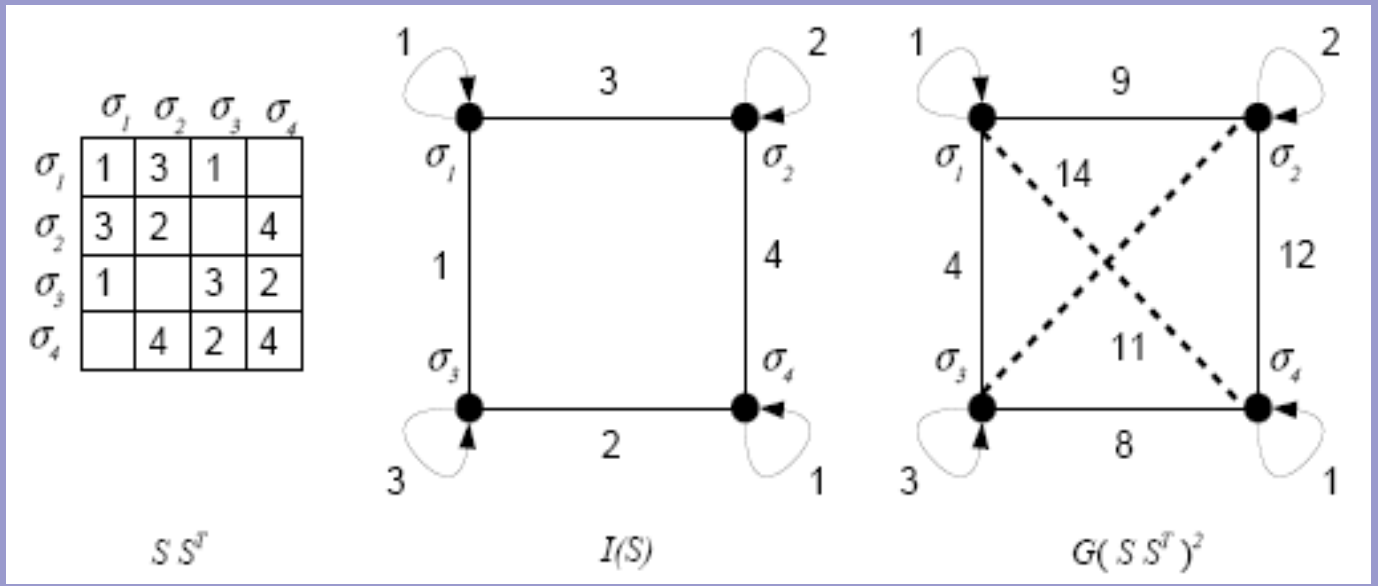
Résumé par extraction

- Identifier les phrases les plus importantes
- Les concaténer
- Les afficher comme un « extract » du document

Nous avons construit un système basé sur
l'énergie textuelle : Enertex

Interprétation (théorie de graphes)

$$E = -S \times J \times S^T = -S \times S^T \times S \times S^T = -(S \times S^T)^2$$

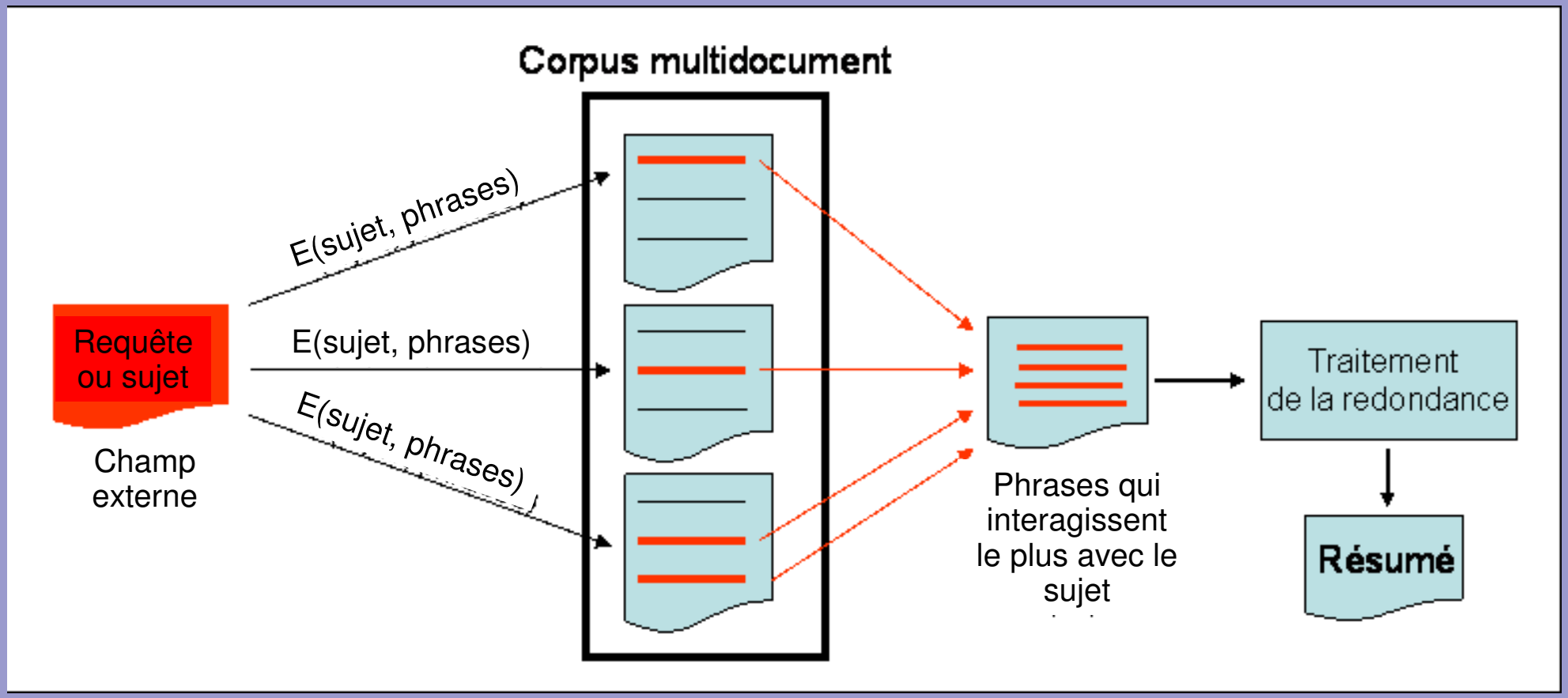


$\sigma_1 \cap \sigma_4 = \varnothing$ mais $\sigma_3 \cap \sigma_1 \neq \varnothing$ *et* $\sigma_3 \cap \sigma_4 \neq \varnothing$

$\Rightarrow$ l'énergie entre σ_1 et σ_4 n'est pas nulle

Complexité : de *P^2 à P logP;* *P= nb phrases*

Résumé multidocument guidé



Réduction de la redondance : $If\ |E_\mu - E_\eta| < \varepsilon \longrightarrow \sim même\ information$

Garder la phrase la plus courte

Évaluation

NIST-Document Understanding Conference (DUC)

Tâche : étant donnée une thématique et un ensemble de 25 documents pertinents (~1000 phrases/doc), générer un court résumé de 250 mots, qui répondra aux questions de la thématique*.

~50 thématiques

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

Réduction de la redondance : $If\ |E_{\mu} - E_{\eta}| < \varepsilon \longrightarrow \sim même\ information$

Diversifier le contenu : $If\ taille\ de\ phrase > k * M \longrightarrow omettre$

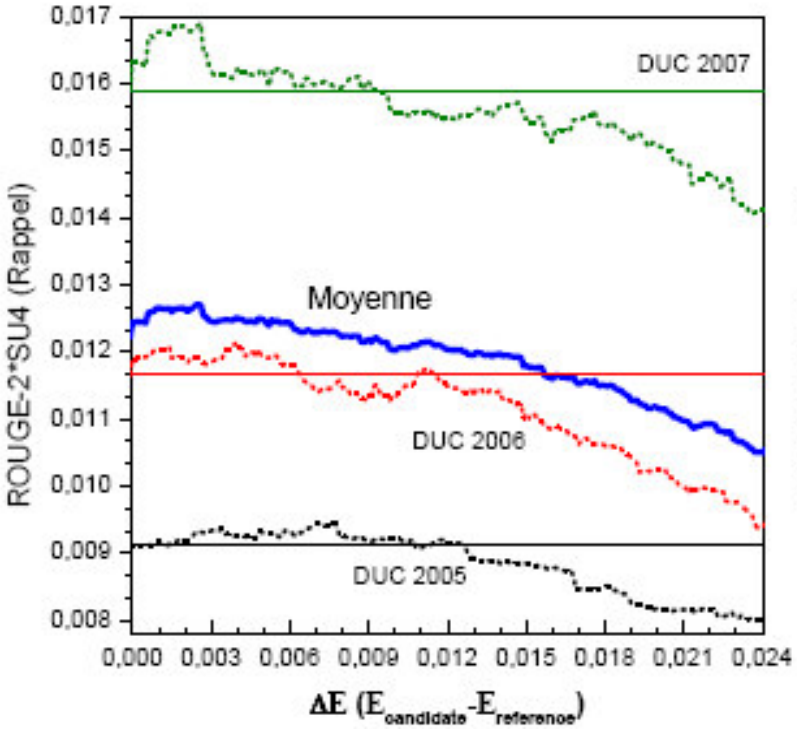
M = taille moyenne phrase (mots)

Corpus AQUAINT : articles d'Associated Press, New York Times (1998-2000) et Xinhua News Agency (1996-2000).

Détermination de ε et k

DUC : ~ 5 références/ thématique

ε = 0, 003
k = 1, 6



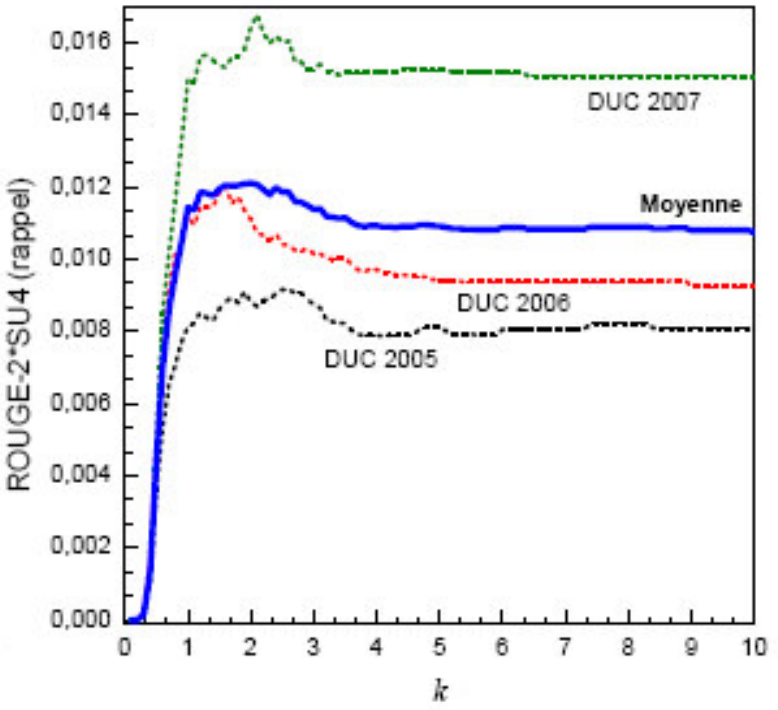
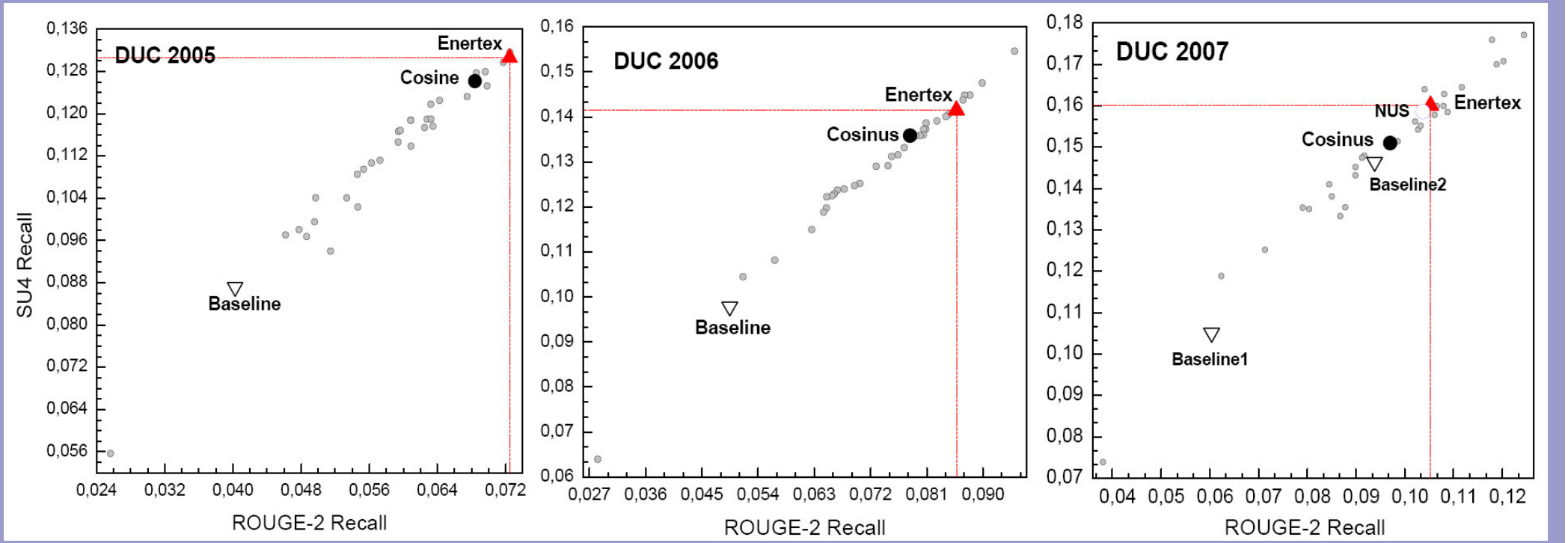


FIG. 2 – Diminution de la redondance : ΔE d'énergie des phrases et moyenne du longueur de phrases (en nombre de mots).

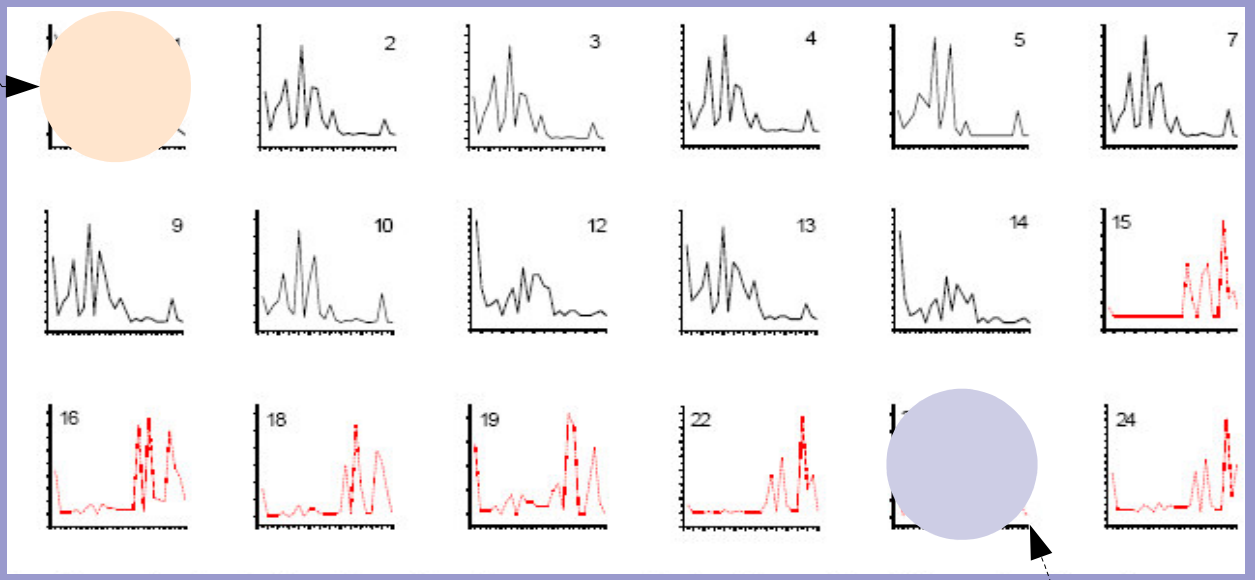
Résultats

Comparaison avec les participants DUC 2005-07

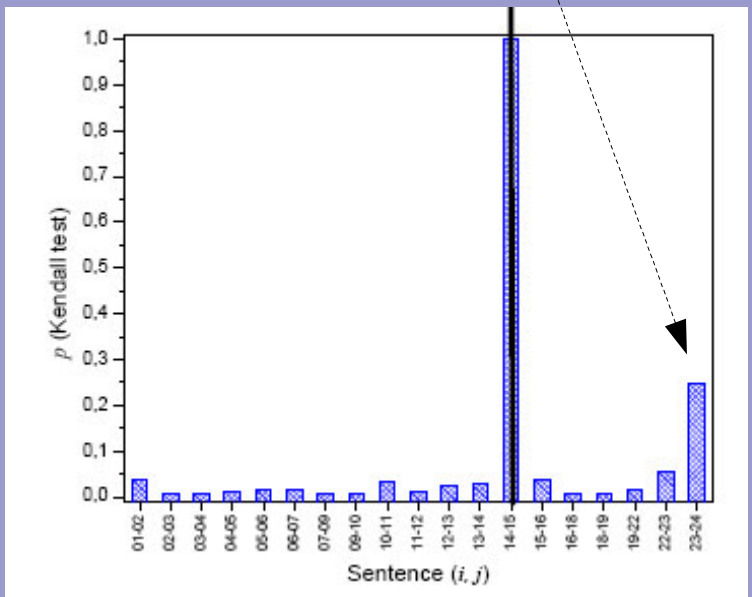


Segmentation thématique

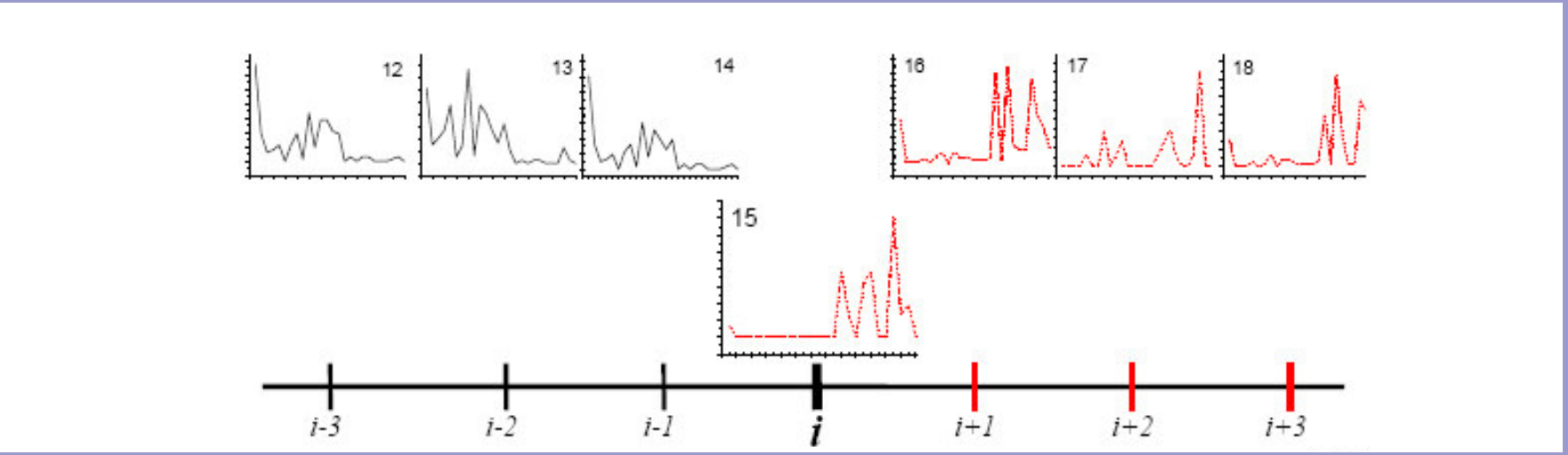
$$E = \begin{pmatrix} e^{1,1} \dots & e^{1,P} \\ \vdots & \vdots \\ e^{\mu,1} \dots & e^{\mu,P} \\ \vdots & \vdots \\ e^{P,1} \dots & e^{P,P} \end{pmatrix}$$



Test de concordance
de Kendall



Probabiliste

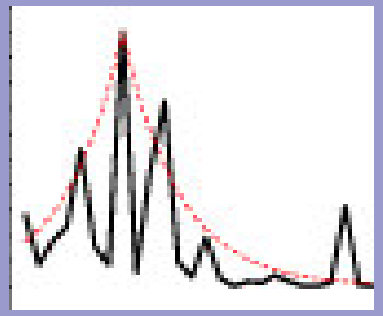


Test de Kendall avec fenêtre de taille 7

Il existe une rupture en *i* si:

$i \neq 0{,}5 * |gauche| \quad \& \quad i = 0{,}5 * |droite|$

« Physique »
Réduire le bruit des courbes



exp(-r/T)

r: distance au max plus proche
T: « température »

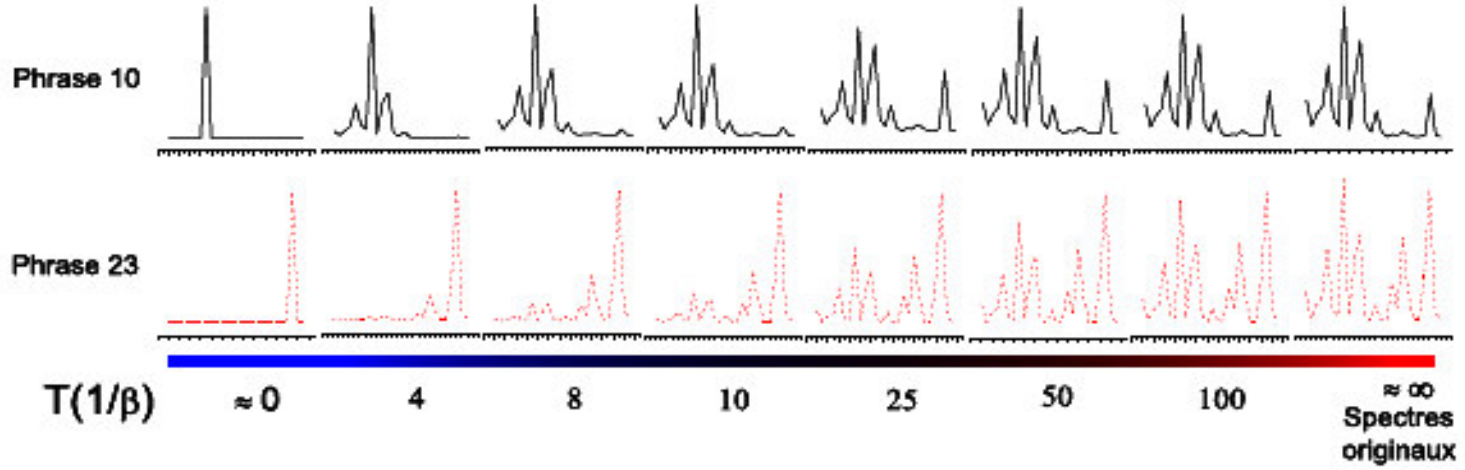
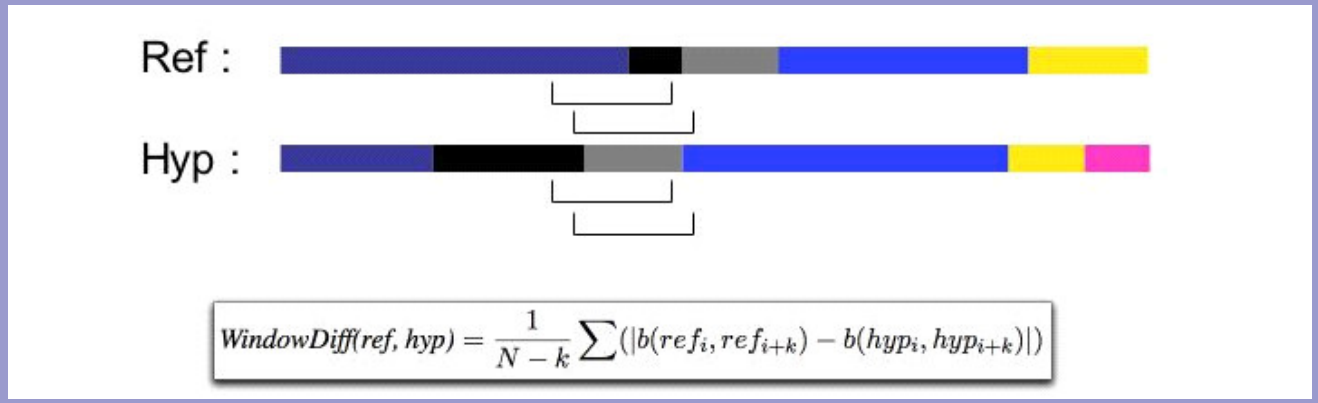


FIG. 5 – Lissage par $\exp^{-r/T}$ des spectres. En trait continu le spectre d'une phrase thématiquement bien définie et en pointillé celui d'une phrase inclassable. r est la distance au maximum et T la témpérature.

Mesure d'évaluation : WindowDiff



Plus petit est la valeur WD, meilleure performance en la détection de frontières.

Détermination de *T*

T=80

Corpus;
Français Le Monde
100 docs
10 seg/doc
seg 6-8 phrases

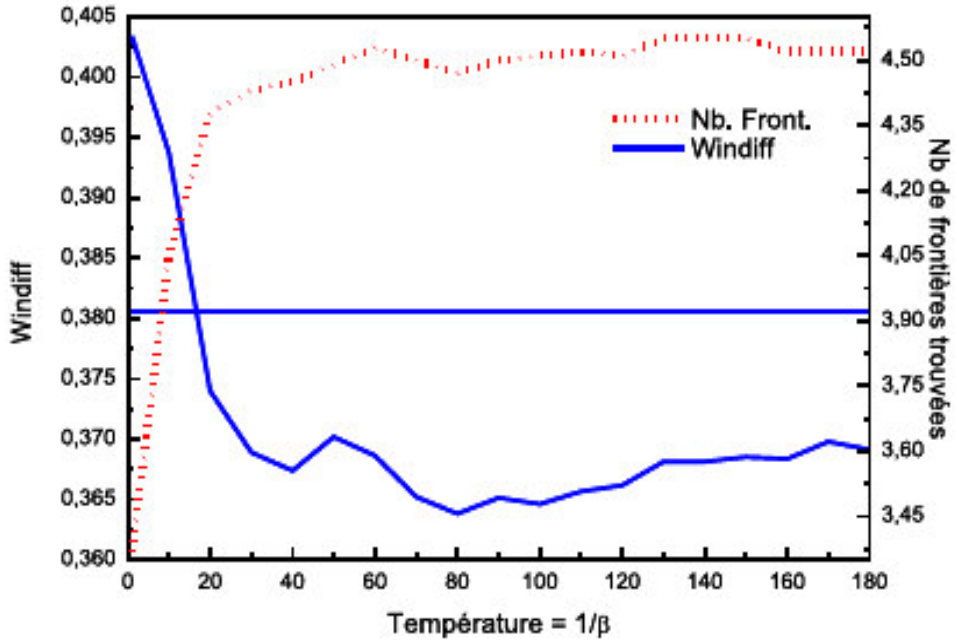


FIG. 6 – Évolution des moyennes de Windiff et du nombre de frontières trouvés en fonction de T pour le corpus en français. La ligne horizontale représente la valeur de Windiff à température infinie rapporté dans (Fernández *et al.*, 2007b). Taille des segment entre 6 et 8 phrases.

Corpus :

Espagnol : La Jornada
Anglais : Brown corpus
Français : Le Monde
Corpus : 400 documents pour langue
4 taille seg : 100 documents pour taille
Document : 10 segmentes thématiquement éloignées choisis au hasard

Segment size (sentences)	β	Enertex			LCseg	LIA_seg
		Spanish	English	French	French	French
9-11	≈ 0.005	0.3897	0.3925	0.4109	0.3272	(**0.3187**-0.4635)
6-8	≈ 0.010	0.3601	0.3804	0.3638	non-reported	non-reported
3-11	≈ 0.025	0.3646	0.3709	0.3885	0.3837	(**0.3685**-0.5105)
3-5	≈ 0.050	0.3598	0.3786	**0.3851**	0.4344	(0.4204-0.5856)

Table 1. Windiff for Enertex, LCseg and LIA_seg (for variable segment size).

Nouvelle mesure d'évaluation

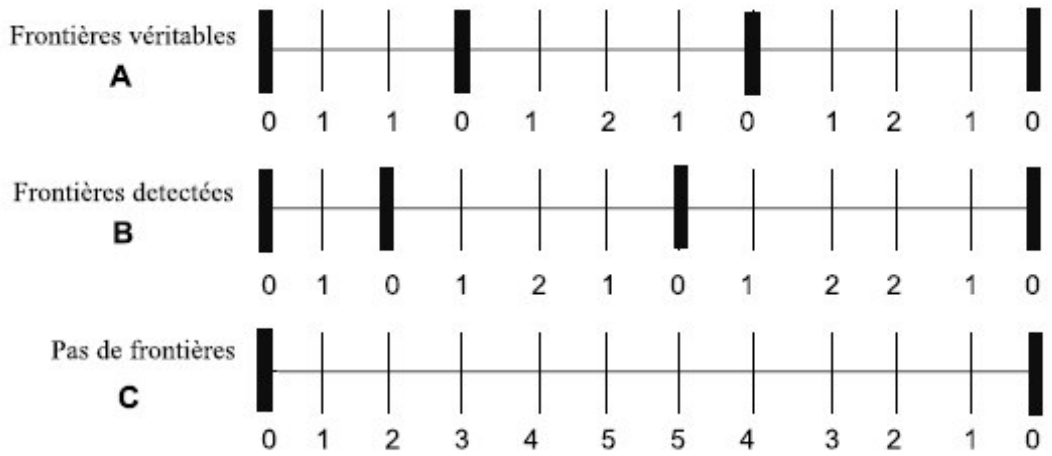


FIG. 7 – La mesure δ-Front.

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

Taille du segment	T	Français		Espagnol		Anglais		Nb. frontières trouvées
		WD	δ-Front	WD	δ-Front	WD	δ-Front	
9-11	120	0,4109	0,1817	0,3897	0,2069	0,3925	0,1524	$\approx$6/9
6-8	80	0,3638	0,1957	0,3601	0,2031	0,3804	0,1640	$\approx$5/9
3-11	40	0,3885	0,1974	0,3646	0,2043	0,3709	0,1634	$\approx$5/9
3-5	20	0,3851	0,4540	0,3598	0,3257	0,3786	0,3864	$\approx$3/9

TAB. 1 – Mesures WD et δ-Front pour des corpus en 3 langues et segments de tailles variables.

Introduction d'impuretés dans les réseaux textuels

Une collaboration avec le *College of Information Sciences, Drexel University Philadelphia*
Exploiting Semantic Annotations in Information Retrieval(ECIR'08)
Glasgow, Scotland, U.K, March 2008

1. Etiqueter les textes selon les renseignements portés par chaque sentence.
   - Génération automatique de patrons
   - Automates à états finis

2. Utiliser ces étiquettes dans l'analyse du texte.
   - Enertex

[OBJECTIVE] We investigate the expected correlation between the weak gravitational shear of distant galaxies and the orientation of foreground galaxies, through the use of numerical simulations. **[HYPOTHESIS]** This shear-ellipticity correlation can mimic a cosmological weak lensing signal, and is potentially the limiting physical systematic effect for cosmology with future high-precision weak lensing surveys. **[RESULT]** We find that the redshift dependence of the effect is proportional to the lensing efficiency of the foreground, and this offers prospects for removal to high precision, although with some model dependence.

Un exemple de texte etiquetté

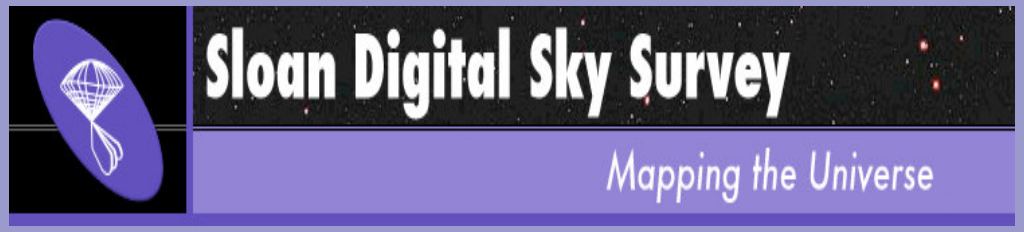
Premier test sur le calcul de l'energie textuelle

Requête : *"Randall-Sundrum"* (occurrence dans le corpus = 1)

Randall-Sundrum models imagine that the real world is a higher-dimensional Universe described by warped geometry. (Wikipedia)

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

One of the top ranked abstracts for the query "*Randall-Sundrum*".



Corpus Astronomie
1293 abstracts
ISI Web of Science (WoS)
"Sloan Digital Sky Survey" (SDSS)

Recherche d'information basé sur les deux :

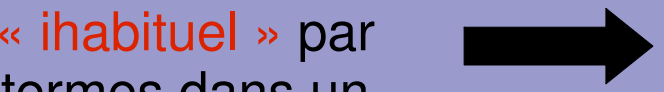
- termes scientifiques (atomes)
- étiquettes (impuretés)

Strategie : T + E → requête
Pretraitement (filtrage, lemmatisation + stemming)
Energie textuelle (requête, abstracts)

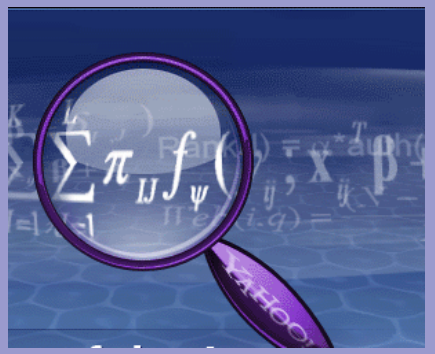
Exemples des premiers essais :

Requête 1 : *NEWTHING spectral classification of quasar*	Non pertinente
Requête 2 : *spectral classification of quasar*	Pertinente
Requête 3 : *NEWTHING*	Non pertinente

Ces impuretés ont une fréquence « ihabituel » par rapport au termes dans un texte filtré.

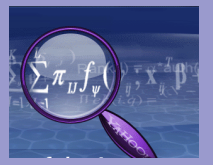


Il faut contrôler ses effets



Termes : $$f(w,s) = \log(trunc(\frac{f_{w,s}^2}{f_{w,\cdot} \times f_{.,s}} > 10^{-n} \times 10^{n}))$$

Impuretés : $$g(w,s) = \log\left(trunc\left(\frac{\left(f_{w,s} - \bar{f}_{w,\cdot} > 0\right)^2}{\sum_{t \in S} \left(f_{w,t} \times \bar{f}_{w,\cdot}\right)^2} \times f_{w,\cdot} \right) \right)$$



$$F = (f+1) * (g+1)$$

Requête : *NEWTHING spectral classification of quasar*

Termes pertinents attendus : *luminosity, quasar, redshift, quasar spectra, spectrum, optical, Balmer, eigen, Fe II emission, Baldwin*

[NEWTHING] Discovery of three new quasars and the spatial density of luminous quasars at z similar to 6. Quasars with intrinsically red (optically steep) power-law continua tend to have narrower Balmer lines and weaker C IV, C III], He II, and 3000 &ANGS; bump emission as compared with bluer (optically flatter) quasars.

Requête *pi*: *TAGi existence of the Gunn-Peterson*

Termes pertinents attendus : *Gunn-Peterson, neutral hydrogen, Intergalactic Medium, IGM, quasar spectra, Lyman Alpha, reionization*

Id	Query	Expected Terms	HYPO-THESIS	FINDING	CONCLUD	RESULT
p2	*HYPOTHESIS existence of the Gunn-Peterson*	93%	**43%**	17%	17%	35%
p3	*FINDING existence of the Gunn-Peterson*	84%	24%	**48%**	24%	56%
p4	*CONCLUSION existence of the Gunn-Peterson*	89%	18%	29%	**33%**	37%
p5	*RESULT existence of the Gunn-Peterson*	89%	22%	33%	22%	**44%**

Conclusions et perspectives

- Introduction de la notion d'énergie textuelle
- Inspiré sur le modèle magnétique d'Ising
- Construction du système Enertex
- Performant dans les différentes tâches du TALN
  - Résumé générique
  - Résumé multidocument guidé
  - Segmentation thématique
  - Recherche d'information
  - Classification de documents (DEFT 2008)
- Indépendante de la langue et de la thématique

- L'application de la physique dans l'analyse textuelle reste un chemin ouvert :
  - Autre type d'interaction entre termes
  - Analyse thermodynamique

Soumission : Silvia Fernandez, Eric SanJuan et Juan Manuel Torres-Moreno, "*Combining Statistical Physics and Graph Theory for Text Analysis*", 2008 PKDD 2008, 15-19 September 2008, Antwerp, Belgium.

Eric Charton, Nathalie Camelin, Rodrigo Acuna-Agost, Pierre Gotab, Remi Lavalley, Remy Kessler, Silvia Fernandez. *“Pré-traitements classiques ou par analyse distributionnelle : application aux méthodes de classification automatique déployées pour DEFT08”* TALN 2008, Avignon, France.

Silvia Fernandez, Eric SanJuan et Juan Manuel Torres-Moreno, *"Enertex: un système basé sur l'énergie textuelle"*, 2008 Traitement Automatique de la Langue Naturelle (TALN 2008), Avignon, France.

Ibekwe-SanJuan F., Fernandez S., SanJuan E., Charton E, "Annotation of Scientific Summaries for Information Retrieval", 2008 "ECIR-ESAIR 2008".

Silvia Fernández, Patricia Velázquez, Sonia Mandin, Eric SanJuan, Juan Torres-Moreno Manuel , *"Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves ?",* 2008 Journées internationales d’Analyse statistique des Données Textuelles JADT 2008.

Silvia Fernandez, Eric SanJuan, Juan Manuel Torres-Moreno, *"Energie textuelle de mémoires associatives"*, 2007 TALN 2007 Vol 1, pp 25-34.

Iria da Cunha, Silvia Fernandez, Patricia Velazquez Morales, Jorge Vivaldi, Eric SanJuan, Juan Manuel Torres Moreno, *"A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics"*, 2007 MICAI 2007, Aguascalientes, Mexique.

Silvia Fernandez, Eric SanJuan et Juan Manuel Torres-Moreno, *"Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation"*, 2007 MICAI 2007, Aguascalientes, Mexique.

Merci