



Du textuel au numérique

*Habilitation à Diriger des Recherches
Informatique*

Juan Manuel Torres Moreno

LIA Avignon 12.12.2007

Plan

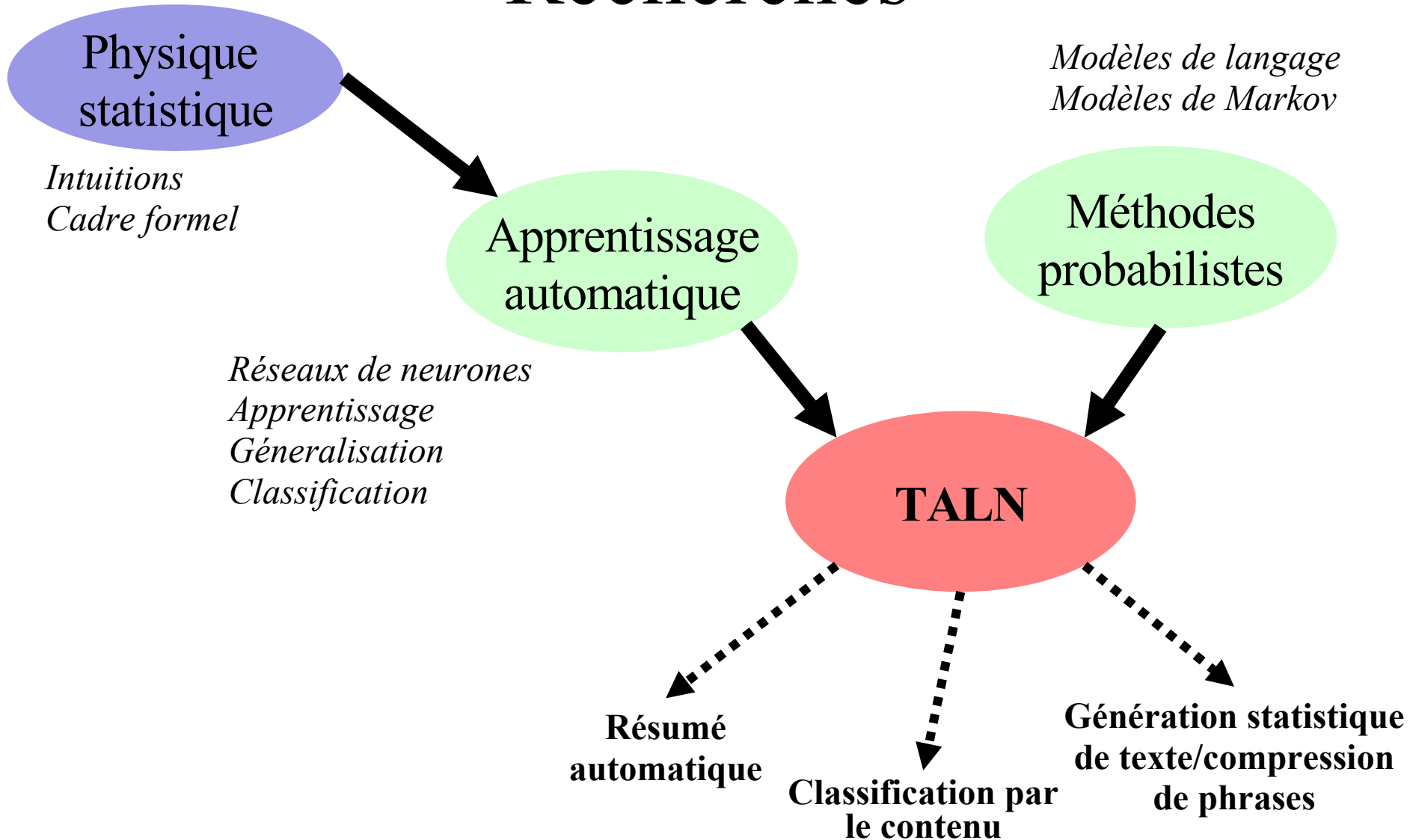
- Recherche
 - Approches
 - Résumé automatique
 - Classification par le contenu
- Projets
 - Encadrements
 - Enseignement
 - Responsabilités
- Perspectives : programme de recherche
- Conclusions

Parcours

- 1997 Doctorat (INPG/CEA Grenoble, France)
- 1998-99 Chercheur (LANIA, Veracruz, Mexique)
- 2000 Postdoctorat (LANCI/UQAM, Canada)
- 2001 Professeur (UQAC, Canada)
- 2001-03 Professeur (École Polytechnique, Canada)
- 2003- MdC (LIA Avignon, France) / Professeur associé
(École Polytechnique de Montréal)



Recherches



Approches

Linguistique

- Analyse syntaxique
- Analyse sémantique
- Rhétorique
- Structures
- ...

- Analyse fine... mais difficile
- Dépendant de la langue

Numérique

- Méthodes probabilistes
- Apprentissage supervisé/
non supervisé
- Traits superficiels
- ...

- Analyse grossière... facile
- Moins dépendant de la langue

Résumé automatique

Cortex et Enertex

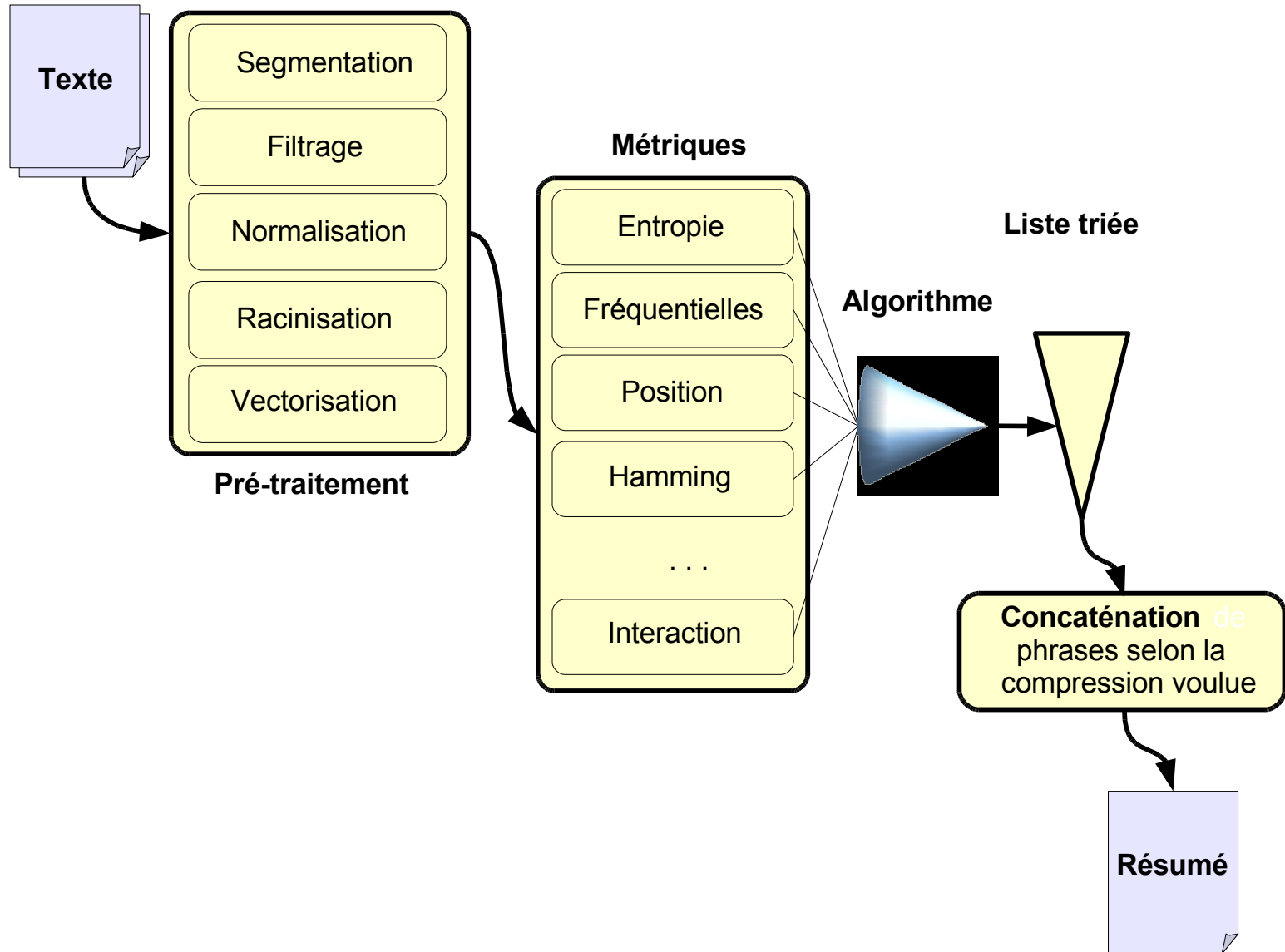
Résumé générique

Guidé par une thématique

Cortex : *Otro Resumidor de TEXTos*

- Basé sur les modèles de RI
- Métriques indépendantes
- Combinaison par un algorithme de décision
- Pas d'apprentissage
- Pas de connaissances
- Indépendant de la langue
- Indépendant de la thématique

Cortex : *Otro Resumidor de TEXTos*



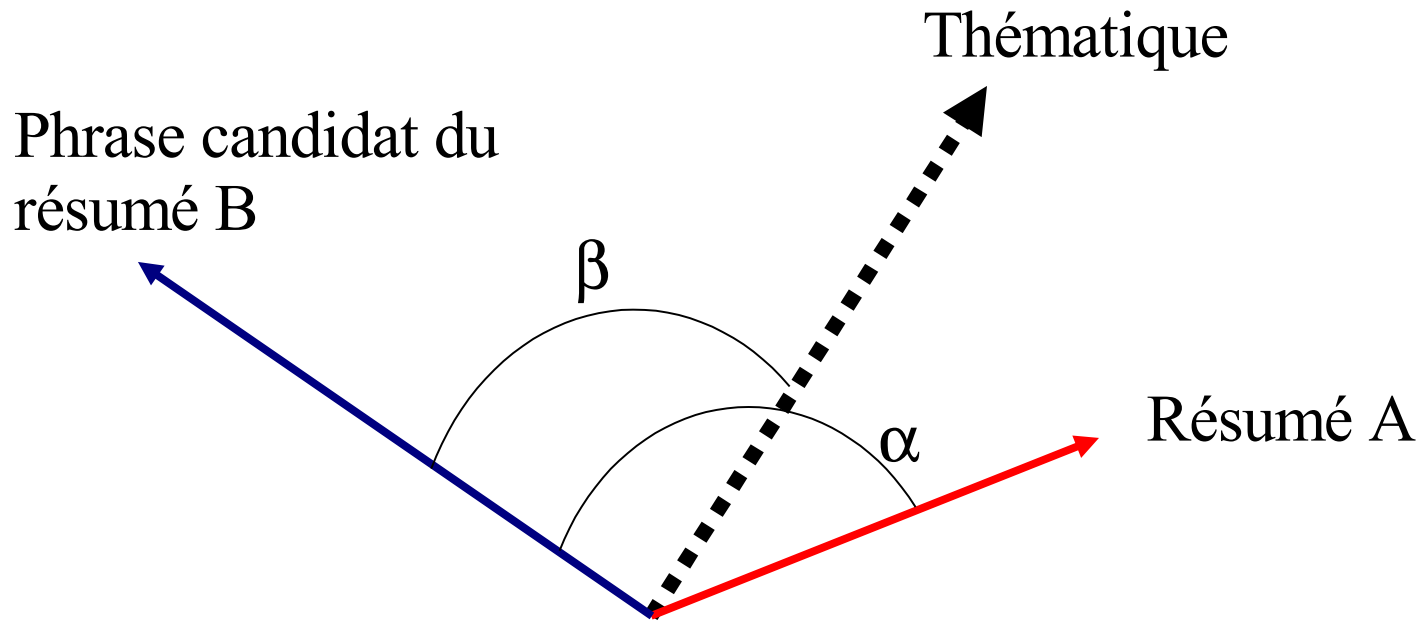
Variantes

- Neo Cortex :
 - Résumé multi-documents
 - Guidé par une thématique
- Evaluation en multi-documents
 - NIST : *Document Understanding Conferences* (DUC'06/07)

Variantes

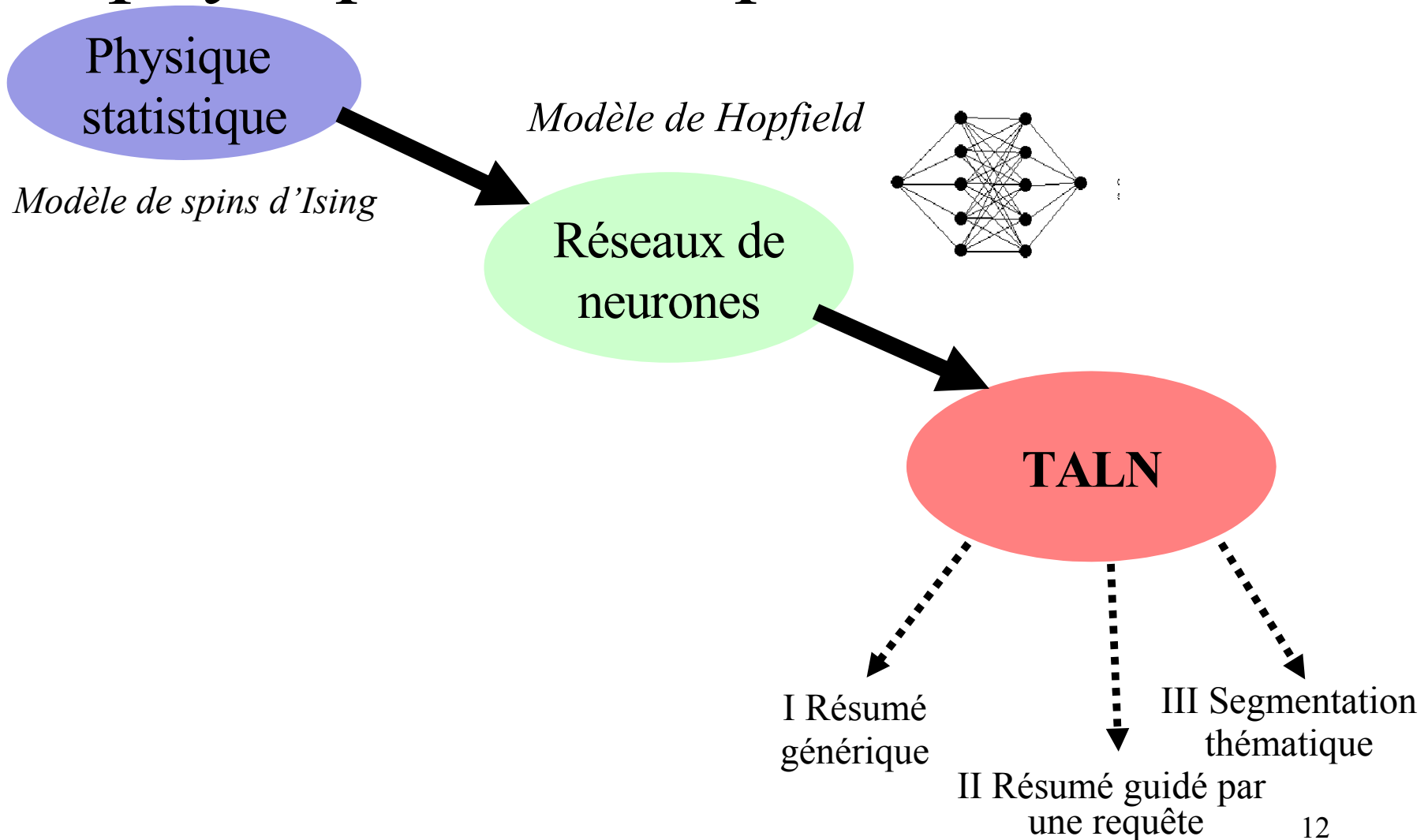
- Détection de la nouveauté
 - Tâche pilote de DUC 2007
 - Résumés guidés par une thématique
 - Résumés qui montrent des événements nouveaux
 - ~100 mots
 - Groupes de documents (temporisés)
 - $\text{temp}(A) < \text{temp}(B) < \text{temp}(C)$

Détection de la nouveauté



Max(α) et Min (β)

Applications exotiques de la physique statistique dans le TALN



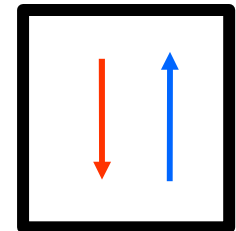
Energie des systèmes magnétiques

$$E = \sum_{ij} J_{ij} s_i s_j + H \sum_i s_i$$

Interaction de spins *champ externe*

$s_i s_j$ état des spins i et j , J_{ij} interaction entre eux

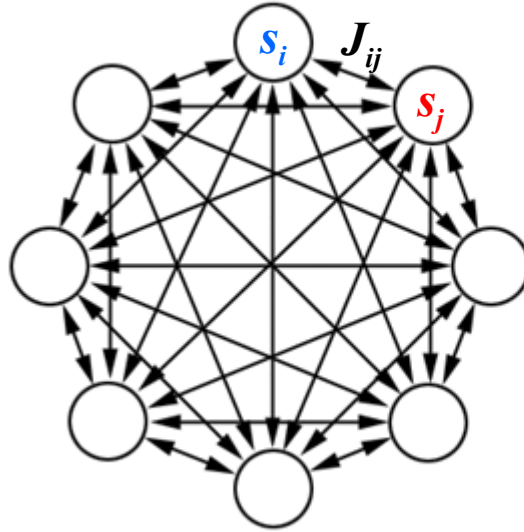
Le modèle le plus simple de la Physique
Statistique : modèle d'Ising à deux
orientations



Mémoire associative (Hopfield, 1982)

Modèle de Spin d'Ising

neurone = spin $\downarrow \uparrow$



Réseau de neurones

Règle d'Hebb

$$J_{ij} = s_i s_j$$

N unités binaires, 2^N configuration ou patrons

Stockage de P patrons

Apprentissage $\longrightarrow J_{ij} = \sum s_i s_j$

Rappel \longrightarrow minimisation ^{P} d'énergie $E_{ij} = - \sum \sum J_{ij} s_i s_j$

Limitations :

- Capacité $\approx 0,14 N$
- Patrons corrélés \rightarrow erreurs de récupération !

Document codé

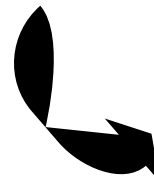
Comme un système de spins

The blue house of my aunt.
 My aunt's name is Lulu.
 I like her house.
 The blue is my favorite colour.
 I have a new pair of blue shoes.

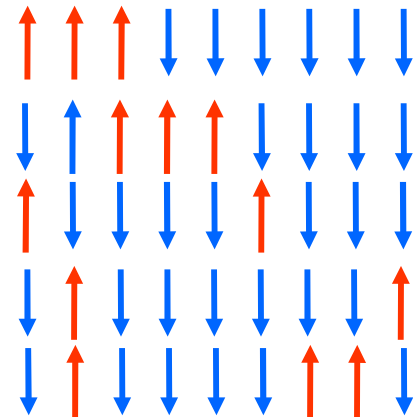


Corrélés !

house	blue	aunt	name	lulu	like	new	shoes	colour
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 de mots)
- Normalisation, filtrage et lemmatisation
(Porter, 1980; Manning & Schutze, 2000)

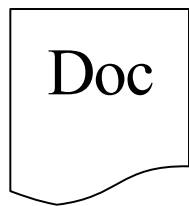


Calcul d'interactions

mot \sim spin s_i

$[s_0 \ s_1 \ s_2 \ \dots \ s_N]$ Phrase \sim chaîne de spins

P phrases, N mots



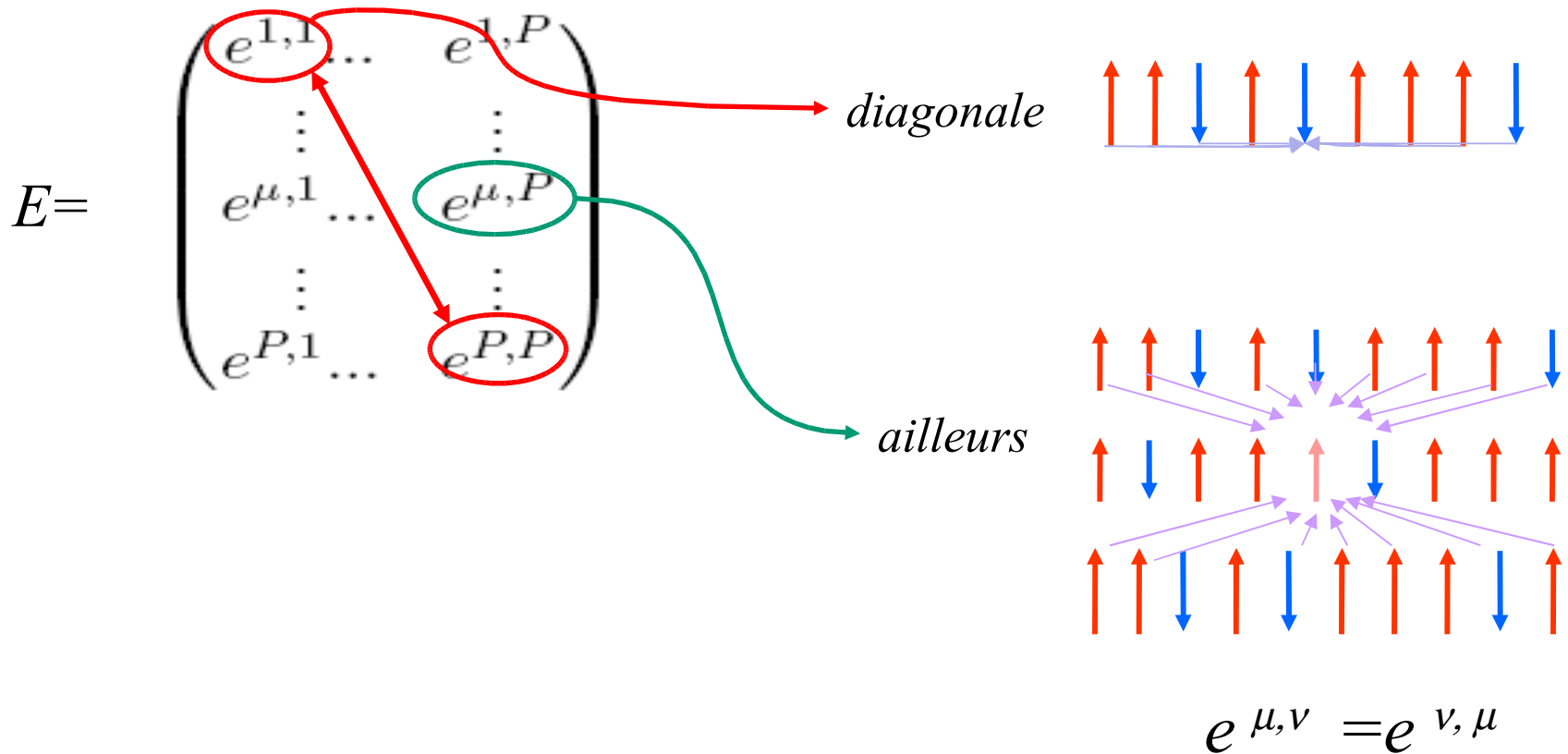
$$S = \begin{pmatrix} s_1^1 & s_2^1 & \dots & s_N^1 \\ s_1^2 & s_2^2 & \dots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^P & s_2^P & \dots & s_N^P \end{pmatrix} \quad \text{Phrases} \times \text{mots}$$

$$J = (S^T \times S) = \begin{pmatrix} j_{1,1}^\mu & j_{1,j}^\mu & \dots & j_{1,N}^\mu \\ \vdots & \vdots & \ddots & \vdots \\ j_{i,1}^\mu & j_{i,j}^\mu & \dots & j_{i,N}^\mu \\ \vdots & \vdots & \ddots & \vdots \\ j_{N,1}^\mu & j_{N,j}^\mu & \dots & j_{N,N}^\mu \end{pmatrix} \quad \text{Règle de Hebb} \rightarrow \text{interactions mots}$$

$$E = - S \times J \times S^T = \begin{pmatrix} e^{1,1} \dots & e^{1,P} \\ \vdots & \vdots \\ e^{\mu,1} \dots & e^{\mu,P} \\ \vdots & \vdots \\ e^{P,1} \dots & e^{P,P} \end{pmatrix} \quad \text{Energie textuelle} \rightarrow \text{interactions entre phrases}$$

L'énergie textuelle

$e^{\mu,\nu}$ = énergie d'interaction
entre les phrases μ et ν



Interprétation (Théorie de graphes)

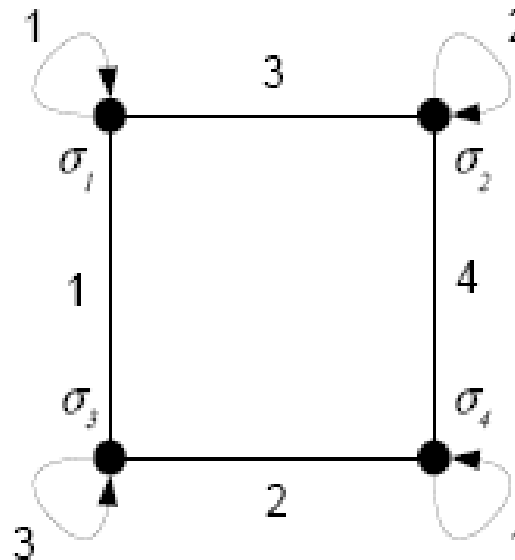
Energie textuelle sous forme matricielle :

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

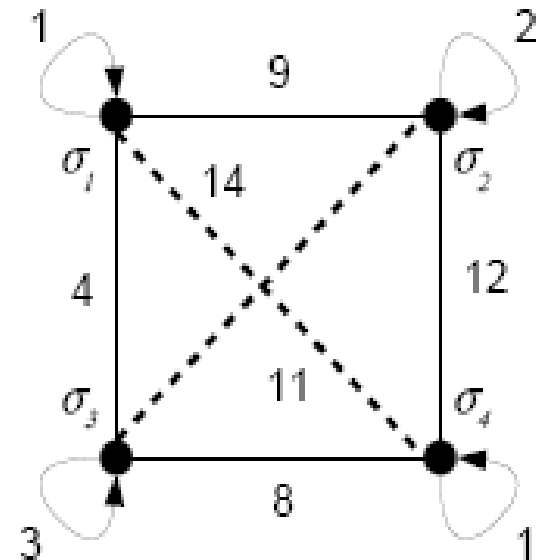
Exemple

	σ_1	σ_2	σ_3	σ_4
σ_1	1	3	1	
σ_2	3	2		4
σ_3	1		3	2
σ_4		4	2	4

$S S^T$



$I(S)$



$G(S S^T)^2$

$\sigma_1 \cap \sigma_4 = \emptyset$ mais $\sigma_1 \cap \sigma_3 \neq \emptyset$ et $\sigma_3 \cap \sigma_4 \neq \emptyset$

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

L'énergie textuelle

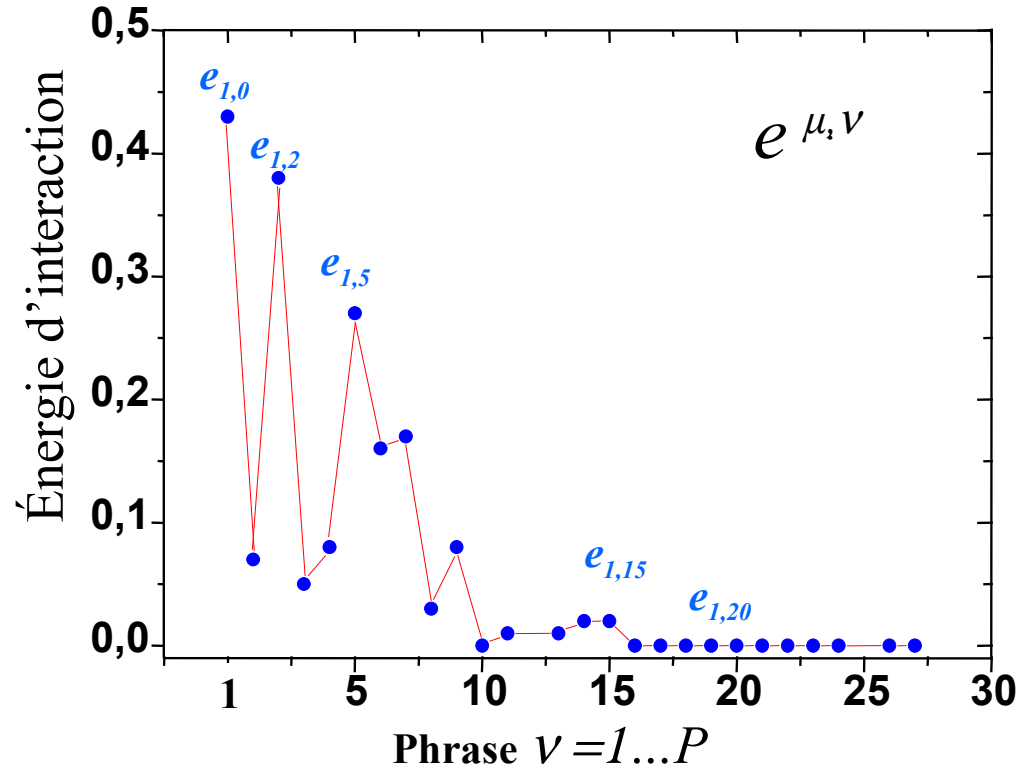
- Une nouvelle mesure de similarité entre phrases
- Interactions locales et globales
- TF IDF au niveau de la phrase
- Applications
 - Résumé automatique | mono/multi-document
 - Segmentation thématique

Résumé générique

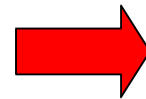
$$E = \begin{pmatrix} e^{1,1} & \dots & e^{1,P} \\ \vdots & & \vdots \\ e^{\mu,1} & \dots & e^{\mu,P} \\ \vdots & & \vdots \\ e^{P,1} & \dots & e^{P,P} \end{pmatrix}$$

$$E^{\mu, doc} = \sum e^{v, \mu}$$

Énergie totale de la phrase μ



Phrases les plus énergétiques



Résumé automatique

Résumé guidé par une thématique

The blue house of my old aunt.

My aunt's name is Lulu.

I like her house !

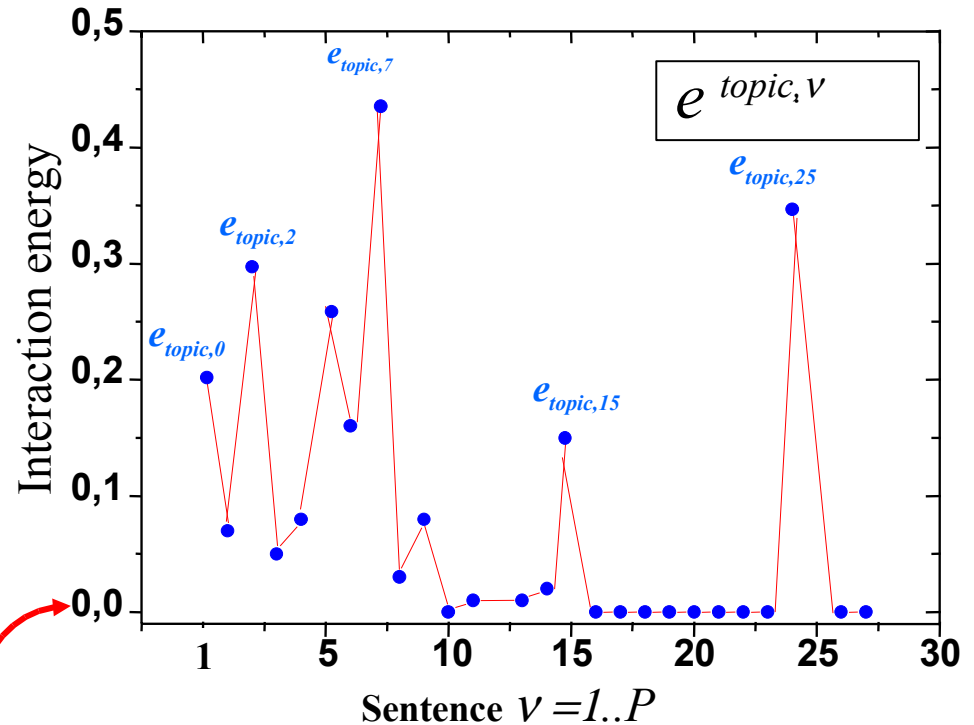
The bleu is my favorite colour.

I have a new pair of blue shoes...

Topic: the colour of my aunt's house

$$S = \begin{pmatrix} s_1^1 & s_2^1 & \dots & s_N^1 \\ s_1^2 & s_2^2 & \dots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ \text{topic} & \text{topic} & \dots & \text{topic} \\ s_1 & s_2 & \dots & s_N \end{pmatrix}$$

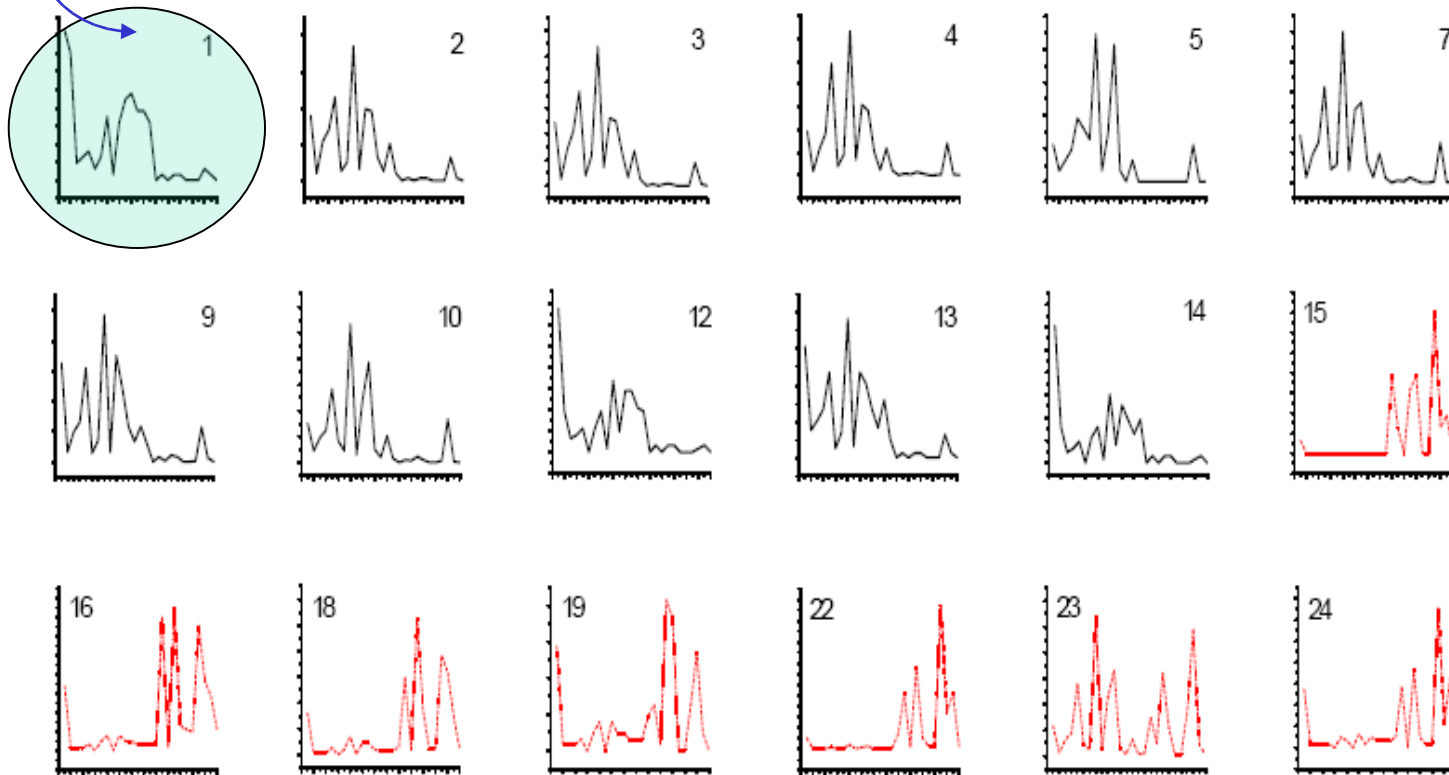
$$E = \begin{pmatrix} e^{1,1} & \dots & e^{1,P} \\ \vdots & & \vdots \\ e^{\mu,1} & \dots & e^{\mu,P} \\ \vdots & & \vdots \\ e^{\text{topic},1} & \dots & e^{\text{topic},P} \end{pmatrix}$$



Résumé : phrases avec la plus forte interaction avec la thématique (*topic*)

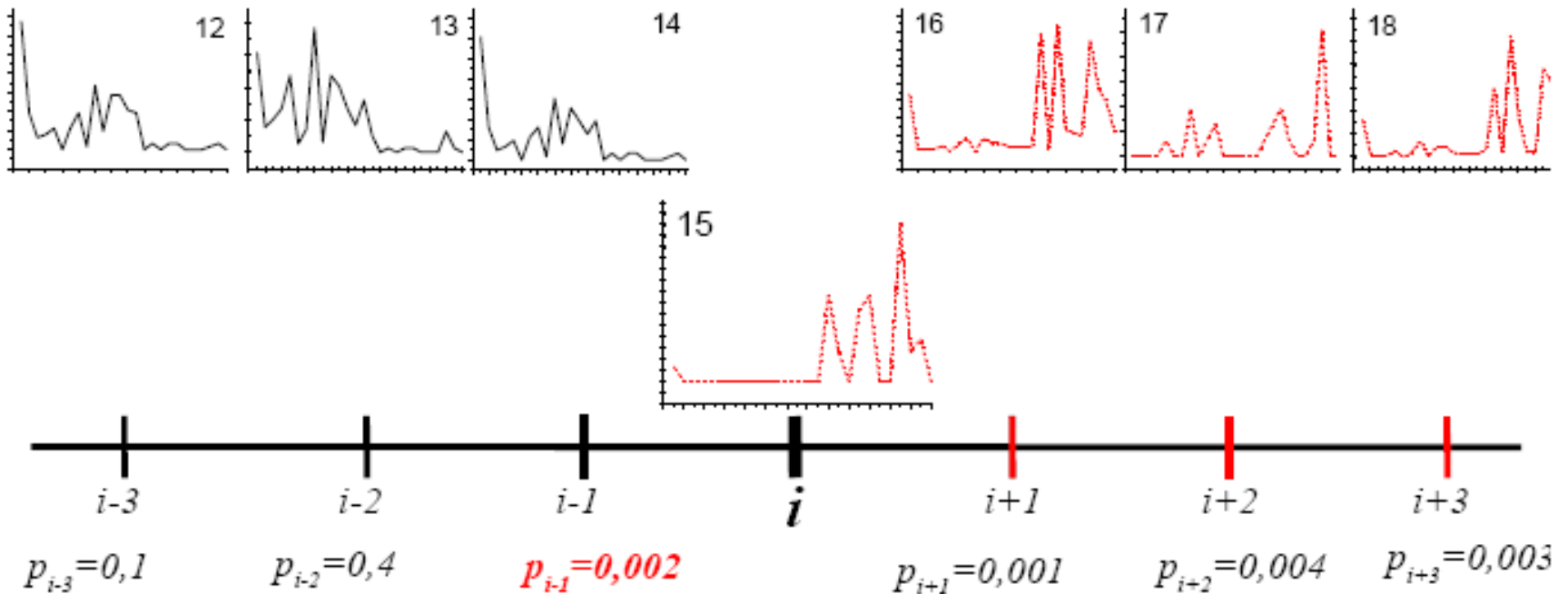
Segmentation thématique

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



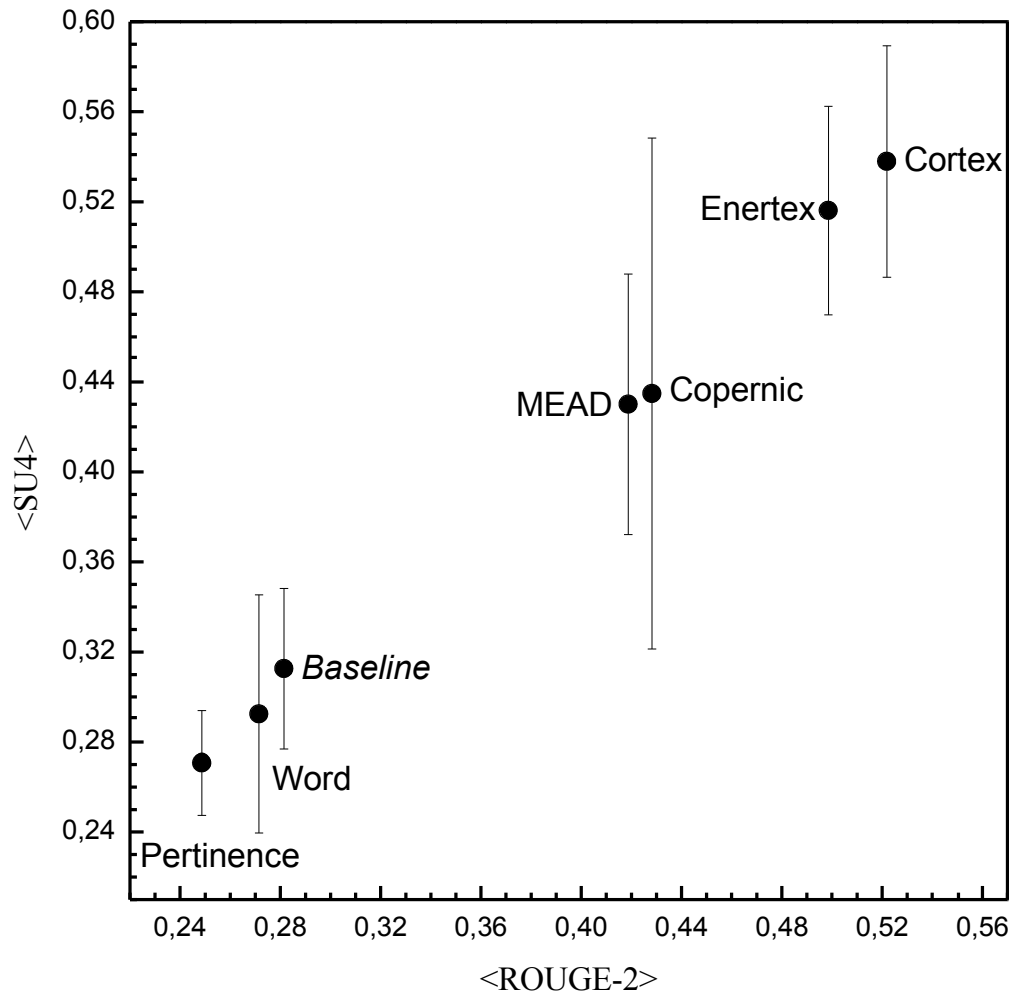
Texte à deux thématiques

Test statistique du τ de Kendall



- Test de Kendall en fenêtre glissante de taille k
- $p_{i\pm k}$ = probabilité de concordance entre $i\pm k$ et i

Résultats : résumé générique



Résumé en anglais,
français et espagnol

Textes composites

Évaluation multi-document

NIST - DUC

45 thématiques / 25 groupes de documents : générer des résumés de 250 mots répondant des questions dans la thématique

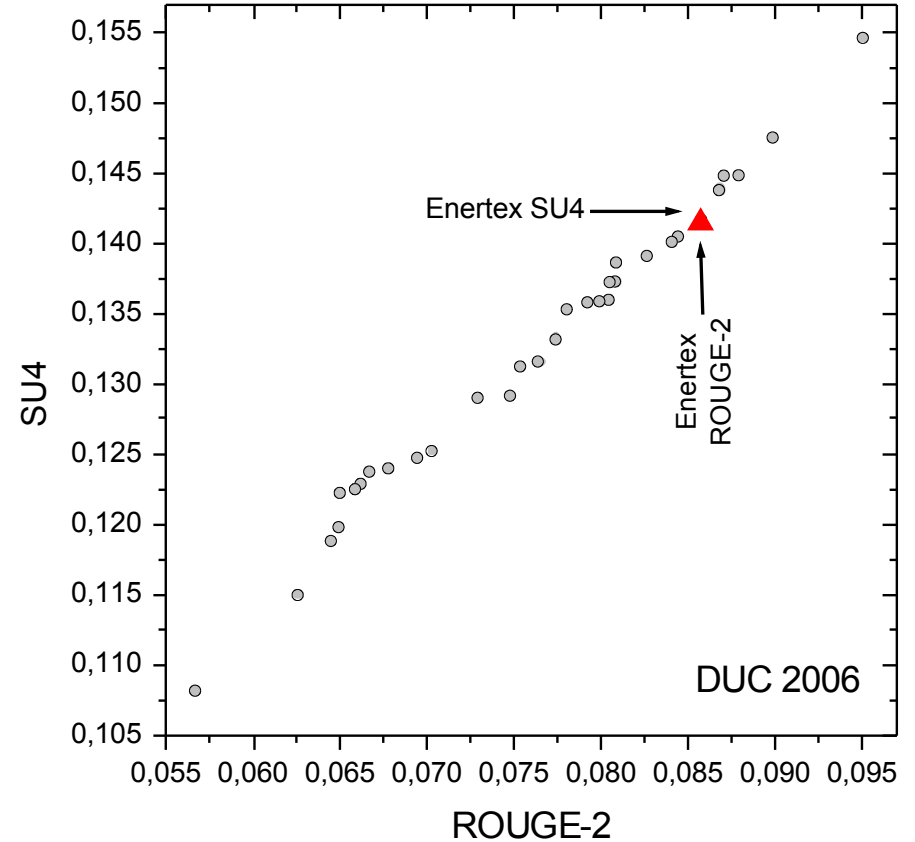
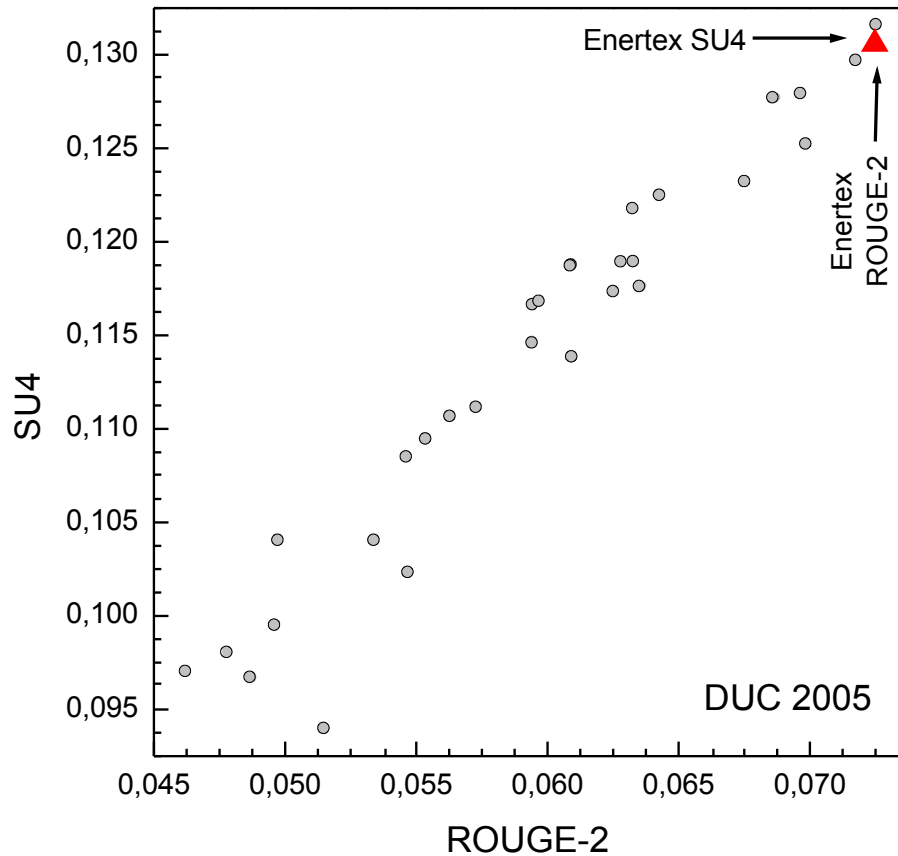
$$E^{topic,\mu} = - \sum_i^N \sum_j^N S_i^{topic} S_j^\mu J_{i,j}$$

Deux stratégies supplémentaires (long corpus):

Élimination de la redondance SI $|E_\mu - E_\eta| < \varepsilon$ \longrightarrow \sim même phrase

Diversifier le contenu SI *taille phrase* $>$ 2 *taille moyenne* \longrightarrow *Utiliser une autre*

Résultats : résumé guidé par une thématique



Classification par le contenu

DEFT 05

Problème

- Classification de textes selon leurs opinions
- Corpus limités
 - Jeux video
 - Relectures d'articles scientifiques
 - www.avoir-alire.com
 - Débats senat (2 classes)
- Evaluation : Fscore

Stratégie

- 11 approches numériques : apprentissage automatique
 - reproduire règles d'association à partir d'un corpus étiqueté
- Méthode
 - Déployer de nombreux systèmes avec de multiples représentations des données
 - plutôt que de régler très précisément un seul système (risque de sur-apprentissage) entraîner un grand nombre de systèmes (certains sans adaptation)
 - Apprentissage « 5-fold cross validation »
 - Fusionner les résultats des différents systèmes : vote simple

Systemes utilisés

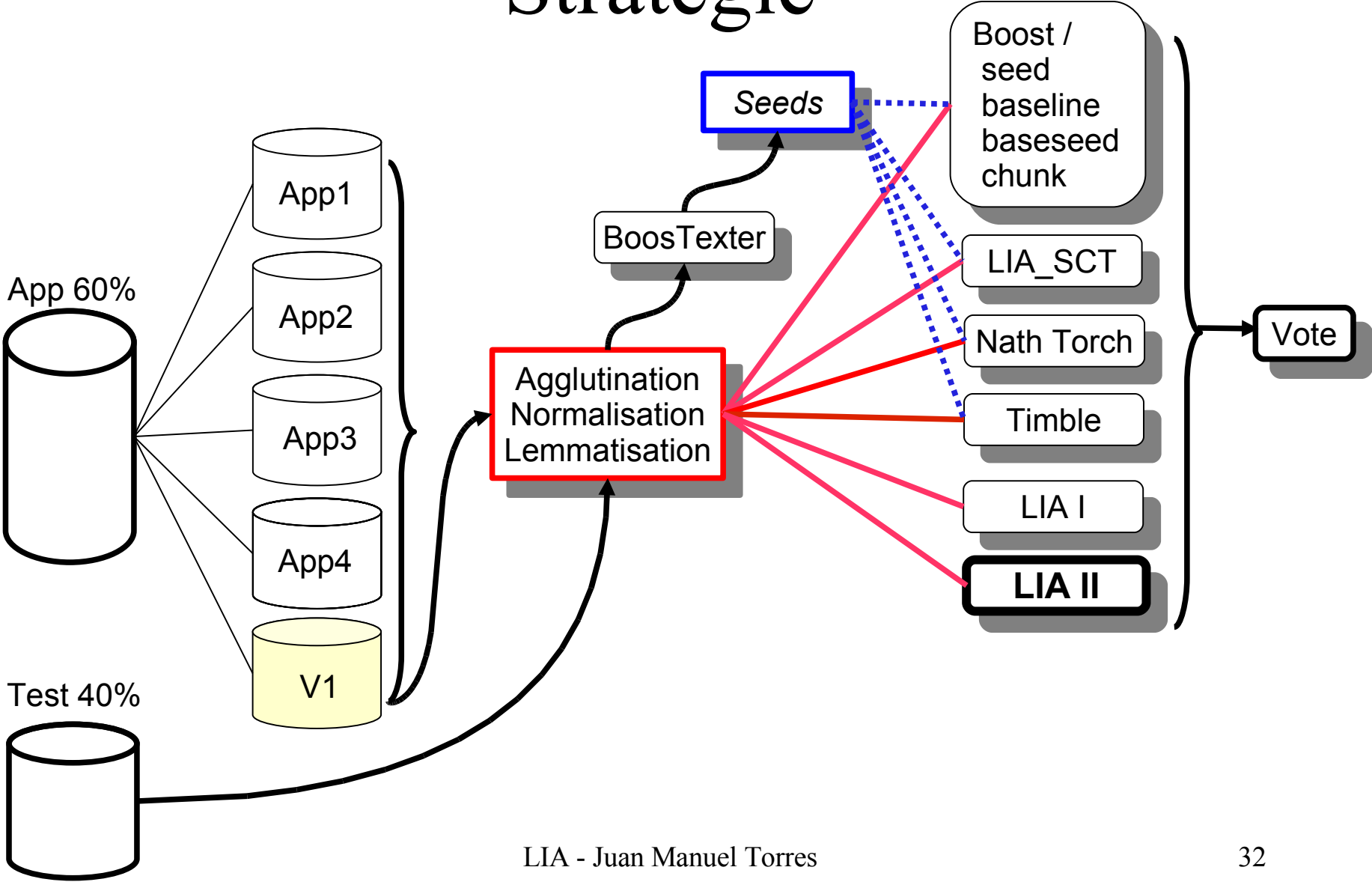
- Classifieurs « *sortis de leur boîte* »
 - BoosTexter : AdaBoost (Schapire & Singer)
 - SVM-Torch : Support-Vector-Machines (Vapnick)
 - Timble : K-plus proches voisins (Daelemans & Van den Bosch)
 - LIA_SCT : Arbres de classification sémantiques (Kuhn & de Mori)
- Classifieurs adaptés pour DEFT'07
 - LIA_I : Modélisation probabiliste discriminante
 - **LIA_II : Modèle de probabilités n-grammes avec/sans lemmatisation**

LIA_II (Torres et al 2007)

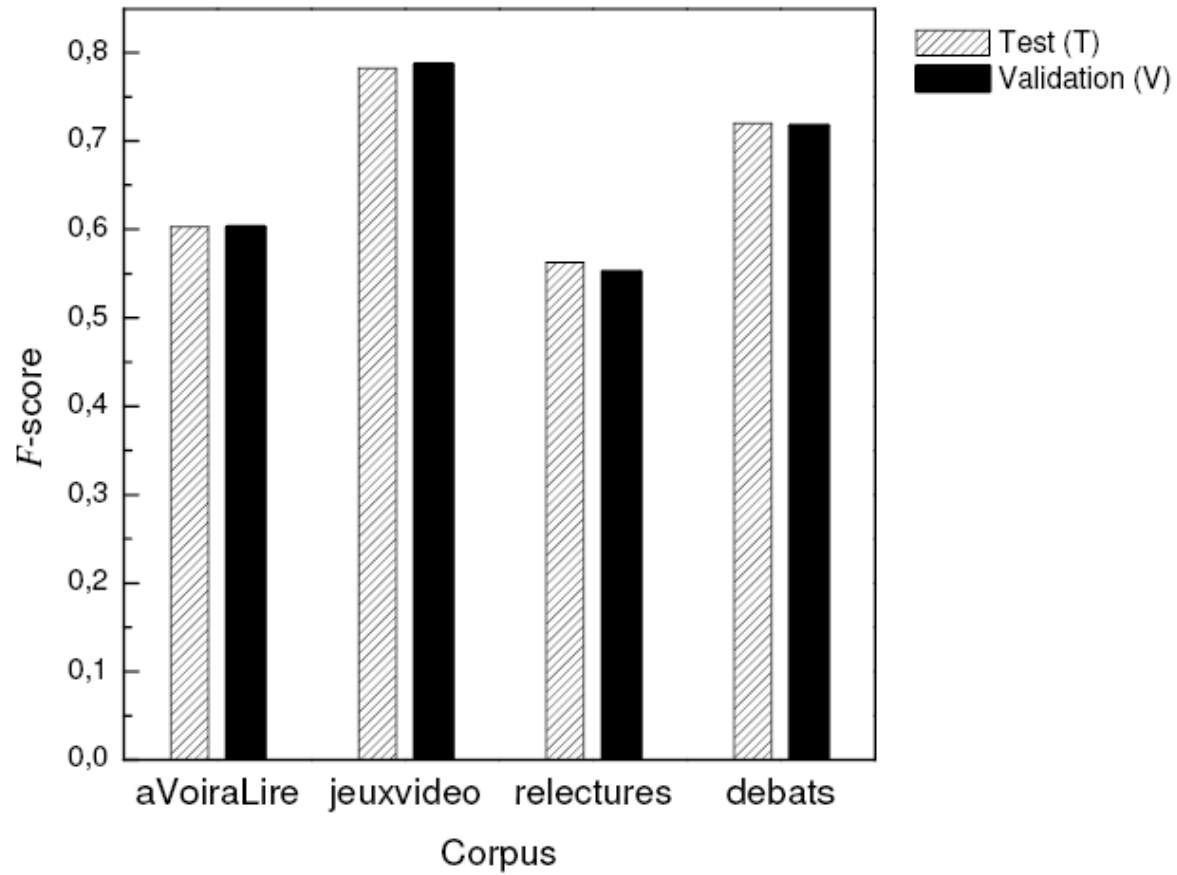
- Filtrage très léger afin de capturer des tournures
 - Voix passive, formes interrogatives et exclamatives
- Aggregation de mots dans la même famille
 - Collocations + morphologique
- Utilisation d'un modèle d'uni-lemmes
- Probabilité d'appartenance à une classe:

$$P_t(w) \approx \prod_i \lambda_1 P_t(w_i) + \lambda_0 U_0$$

Stratégie













Résultats



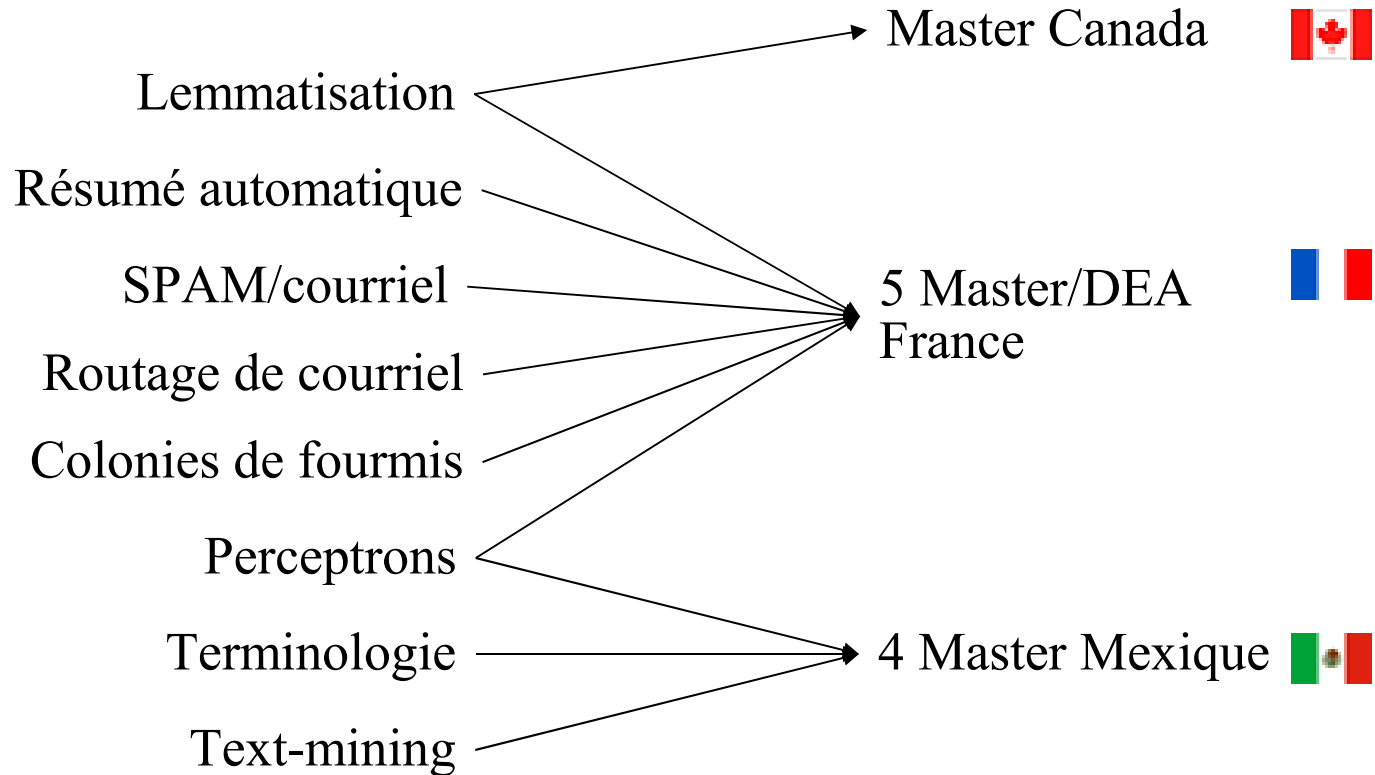
Projets

Encadrements
Publications
Responsabilités

Projets financés

- France 2007-10 Projet **ANR** avec Thématique Parole 
 - Post-doc Résumé pluridocument multimédia
- France 2005-2008 Thèses **CIFRE** 
 - Génération automatique de patrons ; Gestion automatique de CV/offres d'emploi
- Espagne 2007- (IULA) **UAPV**:
 - Résumés hybrides linguistique-numériques 
- Belgique 2006-2008 Thèse **FUNDP**
 - Résumé automatique dans un domaine spécialisé 
- Mexique 2004-2008
 - Energie textuelle, Thèse **Conacyt** avec LPM Nancy 
 - Financement **PACA/Nuevo Leon** 
 - 2 Master: terminologie campagne TREC ; Text-mining en ressources humaines
- Canada 2001-2003
 - **CRSNG** : Résumé automatique 
 - **CRSH** : LANCI Forage de texte 
 - **PAIR-UQAC** 
 - **PIED-Ecole Polytechnique** 

Encadrements master



Encadrements de thèses/post-doc

 2 Thèses CIFRE

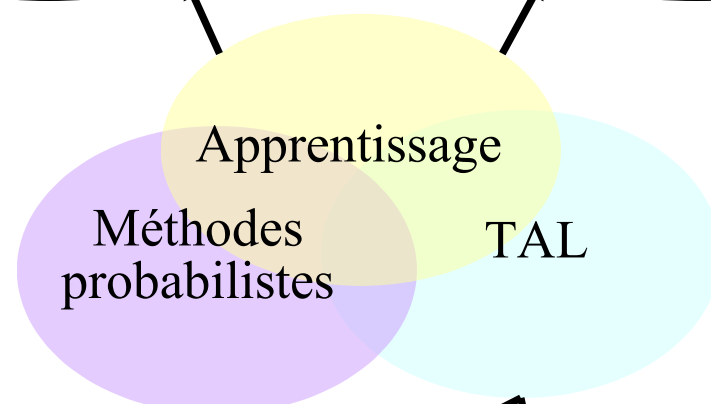


Classification

Thèse financée
Mexique



Energie textuelle




Génération de texte



Thèse financée
Belgique

Résumé automatique

 Thèse financée
Parole /Mistral

Post-doc financé
ANR



Publications

- **Publications**

- 6 publications en revues scientifiques
- 25 publications en congrès internationaux scientifiques
- 4 publications en revues *littéraires*

- **Responsabilités**

- Conseil de Direction du LIA (élu)
- Responsable de la Thématique LIA-TALNE
- Arbitre des conférences (IASTED, ACFAS, MICAI, JADT, CICLing, Recital)
- Arbitre de revues (Neural Processing Letters, Neural Computation, TAL, Computer & Systems, ISI)

Enseignement

Cours



Mexique (Master) : *Intelligence artificielle appliquée au TAL*



Canada (Licence) : Systèmes d'exploitation, Systèmes digitaux, Architecture d'ordinateurs, Programmation assembleur



France (Licence) : Programmation *Perl/C/C++*, Structure des ordinateurs, Bases théoriques de l'informatique

(DEA/Master) : *Mathématiques pour le TALN, Analyse et compréhension de la langue naturelle, Informatique décisionnelle, Résumé automatique, Apprentissage automatique*

Participation aux jurys

- Thèse
 - 1 Mexique, 3 au Canada, 1 en France
- Master
 - 5 Mexique, 5 Canada, ~10 France

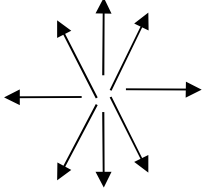
Séminaires et articles de vulgarisation

Perspectives

Programme de recherche
Applications

Programme de recherche :

Spins

- Modèle à spins autre qu'Ising : Potts 
- Etude des chemins de longueur > 2
- Introduction d'une *température* dans le modèle de spins
 - Compression automatique de phrases
 - Génération de texte
- Lissage du paysage d'énergie
 - Segmentation thématique

Programme de recherche :

Perceptrons et apprentissage automatique

- Compression automatique de phrases
 - Guidée par un perceptron : candidates à la compression
 - Compression
 - Systèmes hybrides
 - Systèmes statistiques
 - Énergie textuelle

Programme de recherche :

Résumé automatique

- Résumé : Modèles hybrides
 - numériques + linguistiques (peu profondes)
- Résumés conceptuels ?
 - N'existent pas encore... même pas au niveau de systèmes de jouet
 - *Prédiction* : utilisation de méthodes statistiques pour apprendre les transformations sémantiques. Grand besoin de ressources sémantiques cachées et profondes
- Résumé : évaluation
 - Par le contenu du texte source

Programme de recherche : Application industrielle

Projet RP2M

Résumé pluridocument, multimédia guidé par opinion

- Collaboration académique-universitaire (Synequa, LIA,...) financée ANR 2008-10
- Thématiques LIA : Parole + TALN
- Tâches
 - Détection de **nouveauté**
 - **Résumé** multi-document
 - Classification d'**opinions**
 - Support (**texte**, oral, vidéo)



Conclusions

- Le langage est trop complexe pour que les humains fassent des règles...
 les systèmes doivent les apprendre
- La recherche en TALN n'apporte pas de solutions générales, mais elle apporte de solutions spécifiques *performantes*
- Le TALN numérique suffit à un certain niveau de granularité
- Les méthodes d'apprentissage automatique : adéquates pour traiter des vastes corpus documentaires

Conclusions

- Problématiques TALN posées comme problèmes d'apprentissage
 - Classification d'opinions, Identification d'auteur, Classification de courriel,...
- Des tâches cognitives difficiles (telles que le résumé automatique) ont été approchées efficacement par des méthodes numériques
- Energie textuelle : nouvelle mesure de similarité à explorer
 - Résumé automatique, segmentation, terminologie
 - \sim TF.IDF au niveau de la phrase
- Programme de recherche pour approfondir mes recherches TAL

*Lorsque **moi** j'emploie un mot --déclara Humpty
Dumpty-- il veut dire exactement ce que **je** veux qu'il
dise - ni plus ni moins*

Lewis Carrol, *Through the Looking-Glass* 1871