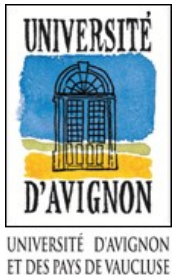


Compresion automatica de frases: ¿Cómo decir algo en menos palabras y aún así decirlo bien?

Juan-Manuel Torres

Alejandro Molina

juan-manuel.torres@univ-avignon.fr

Laboratoire Informatique d'Avignon

Université d'Avignon et des Pays de Vaucluse

Ecole Poytechnique de Montréal

Compresión automática de frases

Tarea subjetiva

Trabajos pioneros de Marcu (Canal ruidoso) y arboles sintacticos

- Tarea bien planteada ?

Se deben eliminar palabras ?

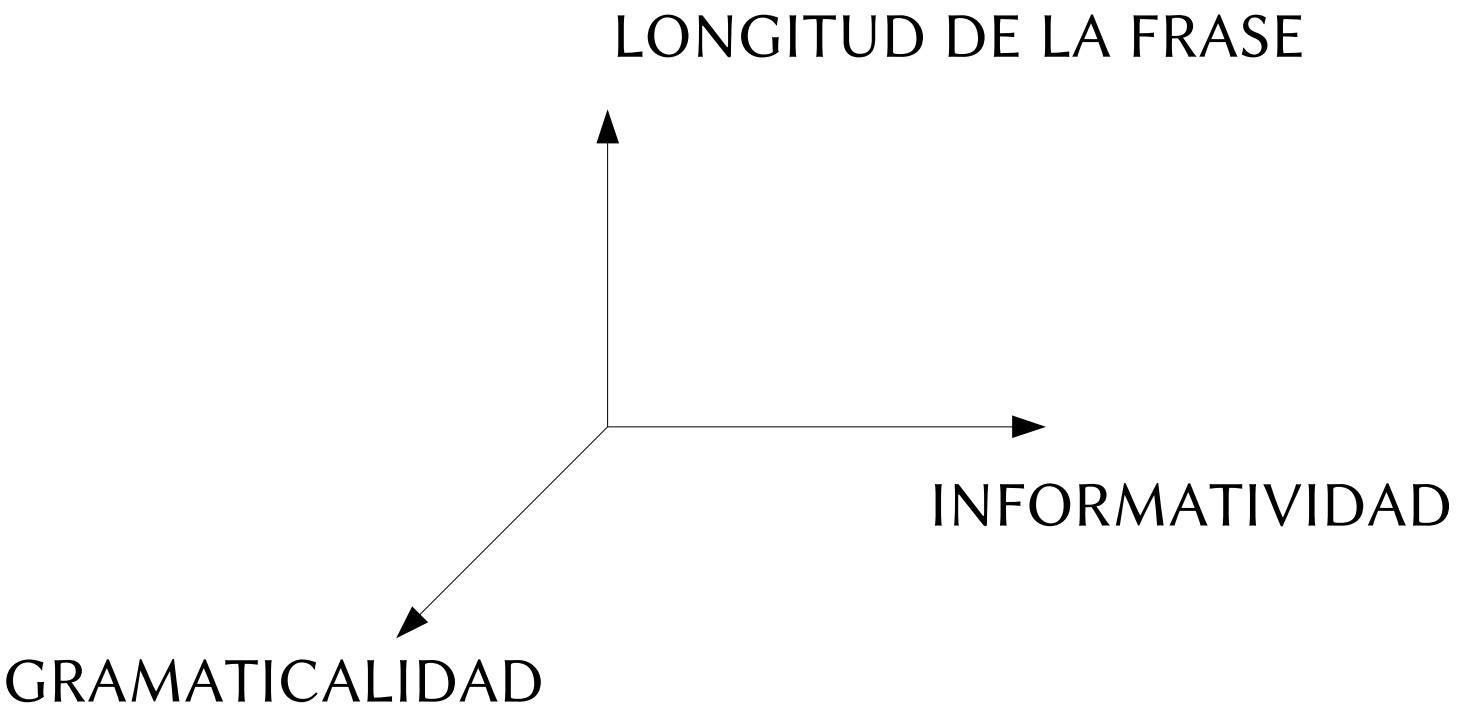
- Las palabras dependen del contexto

Por tanto, eliminar mejor segmentos...

- Qué segmentos ?

Compresión automática de frases

Problema a triple dimension



Compresiones gramaticales... pero qué tan informativas son?

φ : Juliette prépare un gâteau pour le manger bien qu'elle n'ait pas faim.

$\widetilde{\varphi_1^\star}$: Juliette prépare un gâteau bien qu'elle n'ait pas faim.

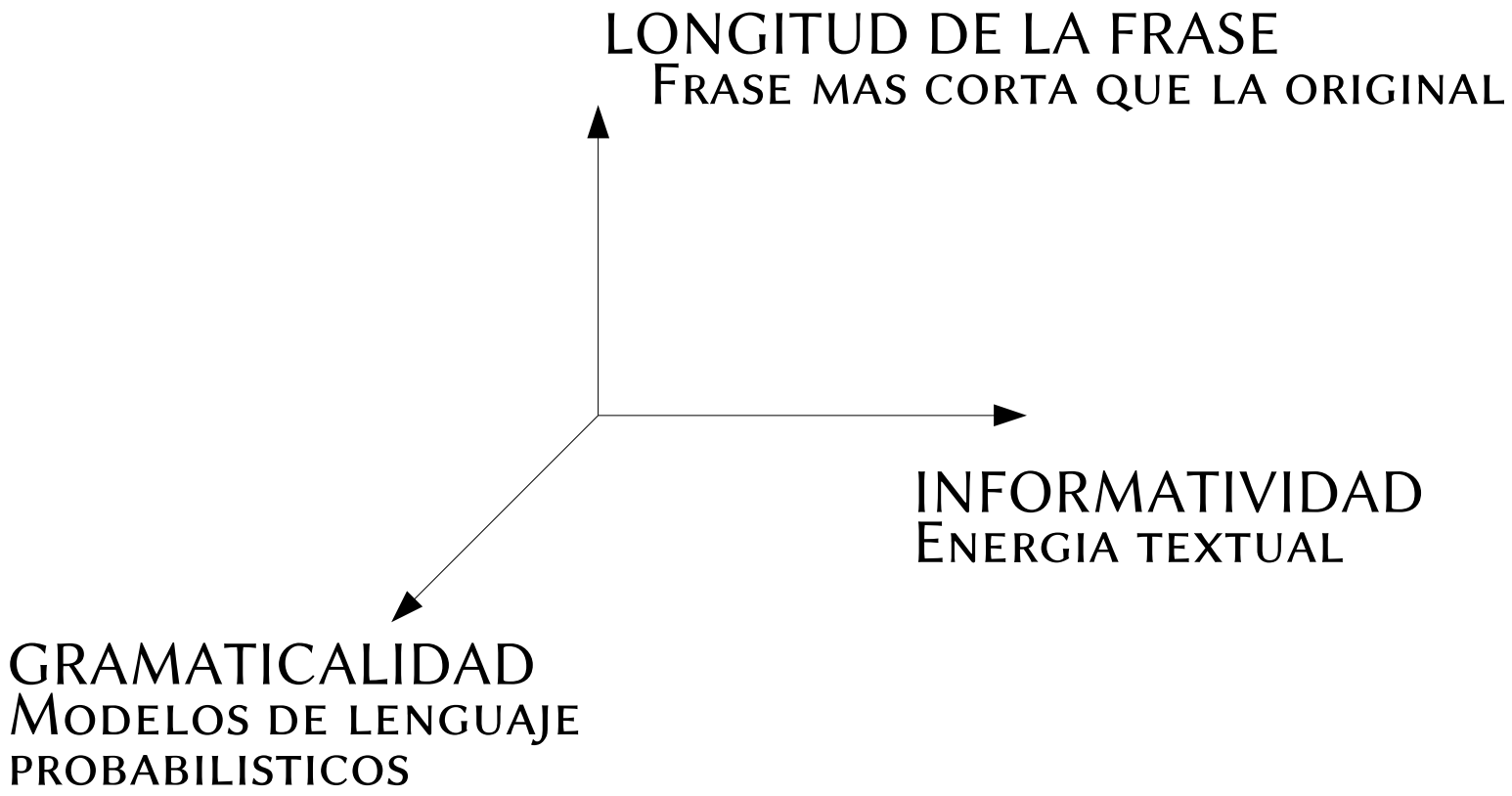
$\widetilde{\varphi_2^\star}$: Juliette prépare un gâteau pour le manger.

$\widetilde{\varphi_3^\star}$: Juliette prépare un gâteau.

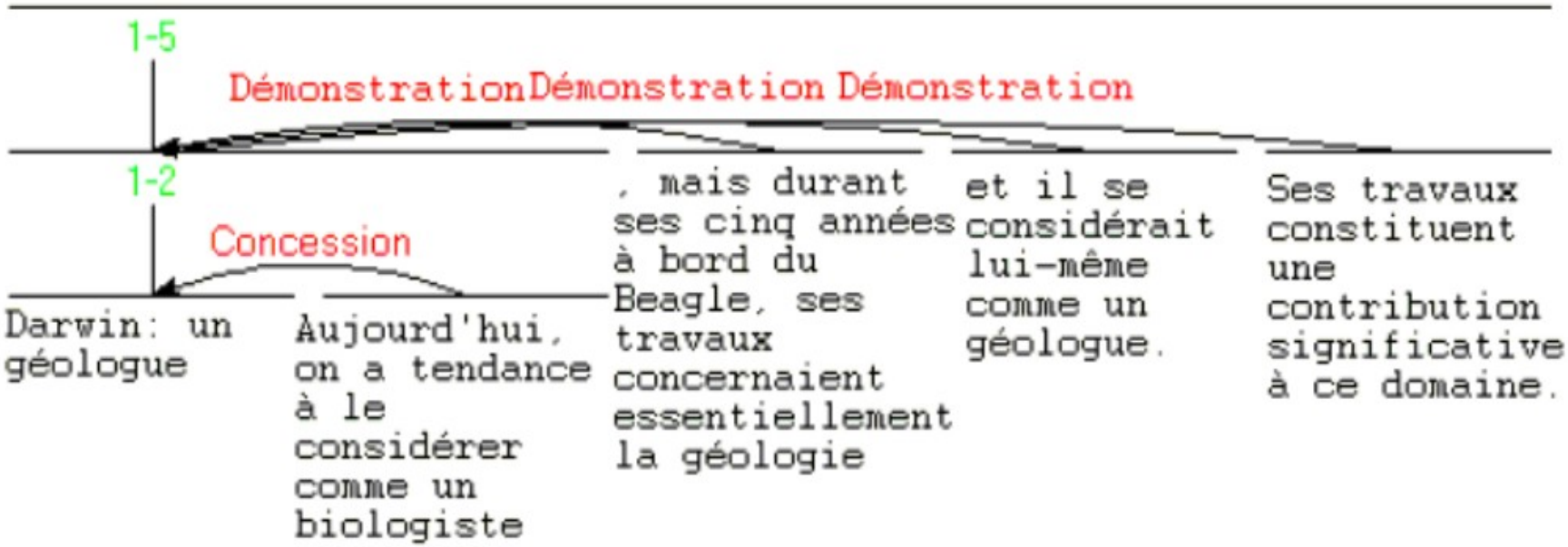
Dimensiones antagonistas (ortogonales)

Compresión automática de frases

Problema a triple dimension



Informatividad : informacion discursiva



Informatividad : segmentos discursivos

S0. Darwin : un géologue.

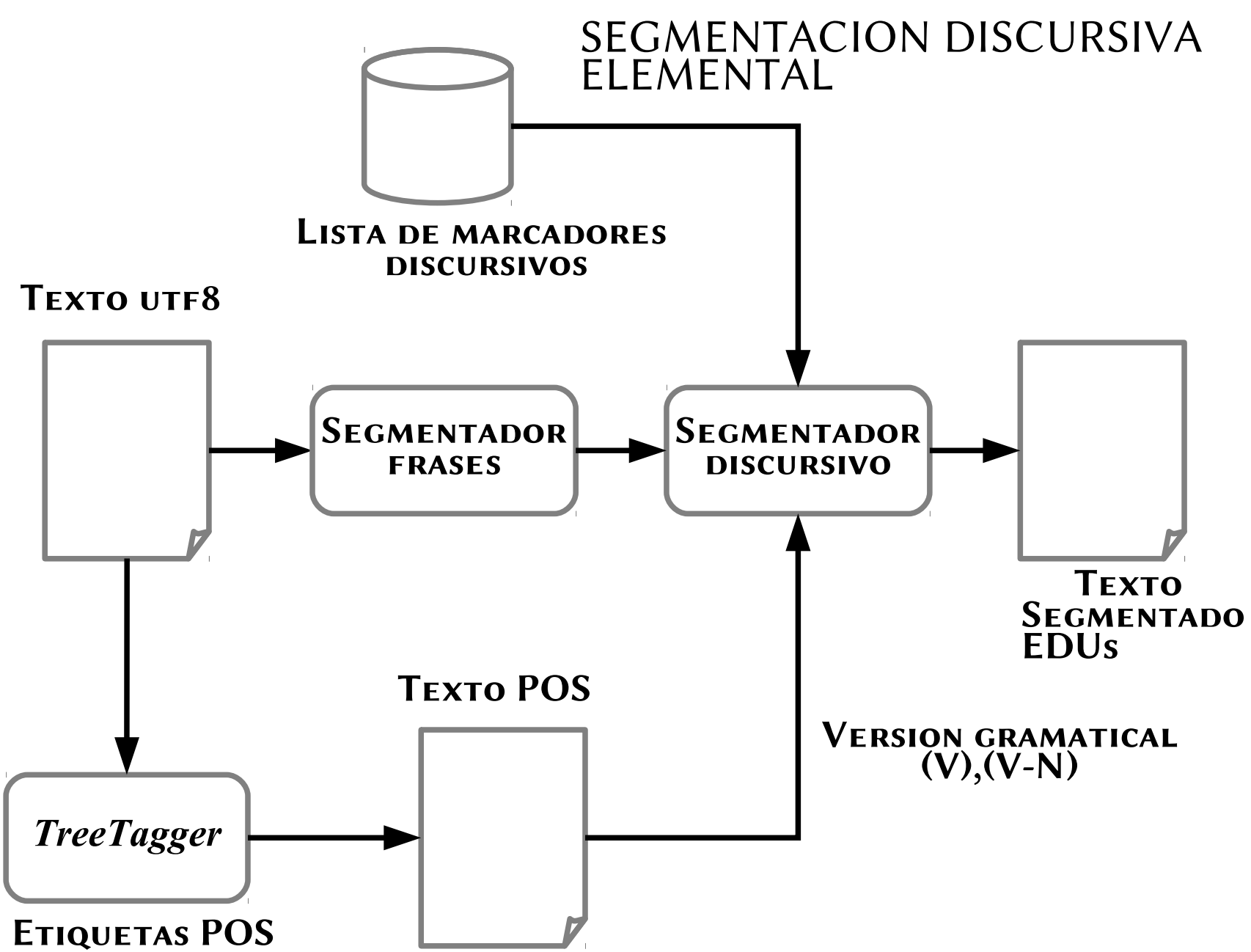
S1. Aujourd'hui on a tendance a le considérer comme un biologiste

S2. mais ses 5 années à bord du Beegle, ses travaux concernant essentiellement la Géologie

S3. et il se considérait lui-même comme un Géologue.

S4. Ses travaux constituent une contribution significative à ce domaine.

Segmentados automaticamente con herramientas Hechas en « casa » (FRANCIA/MEXICO)


SEGMENTACION DISCURSIVA ELEMENTAL
LISTA DE MARCADORES DISCURSIVOS
TEXTO UTF8
SEGMENTADOR FRASES
SEGMENTADOR DISCURSIVO
TEXTO SEGMENTADO EDUs
TEXTO POS
VERSION GRAMATICAL (V),(V-N)
TreeTagger
ETIQUETAS POS

Modelos de lenguaje: probabilidad de existencia de la frase

$$\begin{aligned}\mathbf{P}(\text{Juliette, prépare, un, gâteau, pour, le, manger}) = \mathbf{P}(\text{Juliette}) \times \\ \mathbf{P}(\text{prépare} \mid \text{Juliette}) \times \\ \mathbf{P}(\text{un} \mid \text{Juliette prépare}) \times \\ \mathbf{P}(\text{gâteau} \mid \text{Juliette prépare un}) \times \\ \mathbf{P}(\text{pour} \mid \text{Juliette prépare un gâteau}) \times \\ \mathbf{P}(\text{le} \mid \text{Juliette prépare un gâteau pour}) \times \\ \mathbf{P}(\text{manger} \mid \text{Juliette prépare un gâteau pour le})\end{aligned}$$

$$\mathbf{P}(w_1^n) = \mathbf{P}(w_1) \times \mathbf{P}(w_2|w_1) \times \mathbf{P}(w_3|w_1^2) \times \ldots \times \mathbf{P}(w_n|w_1^{n-1}).$$

Probabilidades estimadas sobre un corpus representativo : GOOGLE 5-grammes En Fr Es

Compresión de frases: Modelo lineal combinando parámetros

$$\mathbf{P_{elim}}(s,\varphi) = \mathbf{Ener}(s,\varphi) + \mathbf{Gram}(s,\varphi) + \mathbf{Seg}(s,\varphi) + \mathbf{Lon}(s,\varphi)$$

Ener ~ informatividad

Gram ~ Gramaticalidad

Seg ~ Segmentador

Lon ~ Longitud

GRAN INDEPENDENCIA DEL IDIOMA

Evaluación de compresiones

Métodos clasicos

- de traduccion: BLEU
- de resumen semi-automaticos: ROUGE (Lin 2007)
- de resumen automaticos: FRESA (Torres et al 2010, Saggion et al. 2011)

Test de Turing revisitado (Molina, SanJuan & Torres, 2013)

Test de Turing (*The Imitation Game*)

Alan Turing et son test de éponyme

Descubrimiento de mamut emociona a científicos

(Documento fuente)

El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacer se con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra para ver si pueden determinar qué tanto de los restos del mamut siguen enterrados. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales para proteger el sitio.

Descubrimiento de mamut emociona a científicos

(Documento fuente)

El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos cientificos que podrían hacer se con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra para ver si pueden determinar qué tanto de los restos del mamut siguen enterrados. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales para proteger el sitio.

Descubrimiento de mamut emociona a científicos

(Resumen por compresion)

Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

20 % del tamano original, segmentador DISEG
Resumidor ENERTEX

Test de Turing

Réponse du juge	Origine du résumé Humain	Machine
Humain	a	b
Machine	c	d

Juge id	Contingence		p	H_0
Juge 1	4	0	0.030	faux
	2	6		
Juge 4	1	5	0.998	vrai
	5	1		
Juge 7	5	5	0.772	vrai
	1	1		

H0 : Independencia : no hay asociacion entre el origen del resumen y las respuestas
H1 : El juez identifica el origen del resumen

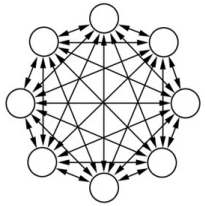
Resultados : sobre 54 humanos, 53 fueron incapaces de Descubrir el origen artificial de los resumenes

(p-value = 0.496 > 0.05 se acepta H0)

Los documentos son informativos,
pero...
poseen propiedades fisicas...?
Volumen? Masa? Longitud?
Energia...?

Probablemente puedan usarse una
transposicion de ideas
de otras disciplinas...

Como estudiar el lenguaje humano?



Física Estadística

Modelo de spins de Ising

Linguistica

Aprendizaje automatico

Computacion

cognicion

PLN

Comprension De texto

Clasificacion

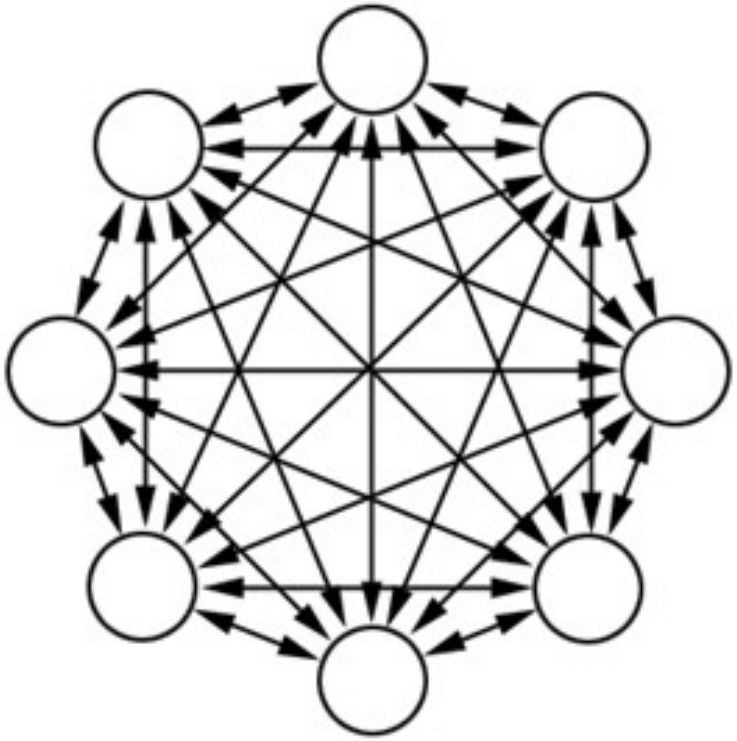
...

Generacion De texto

Resumen automático

Compresion de frases

Ingredientes basicos...
Mecanica estadistica :
energia textual

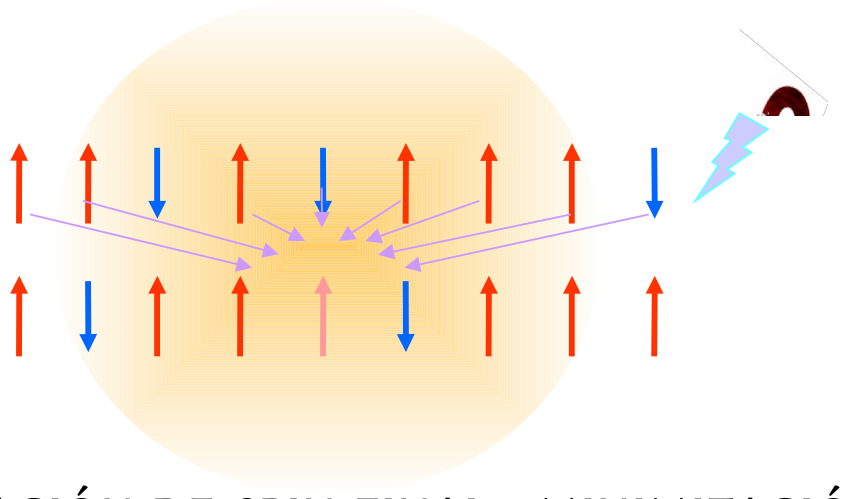


Energía del sistema

$E = E(\text{interacciones}) + E(\text{campo})$

$$E_{ij} = J_{ij}\, s_i\, s_j \qquad + \qquad E_i = H\, s_i$$

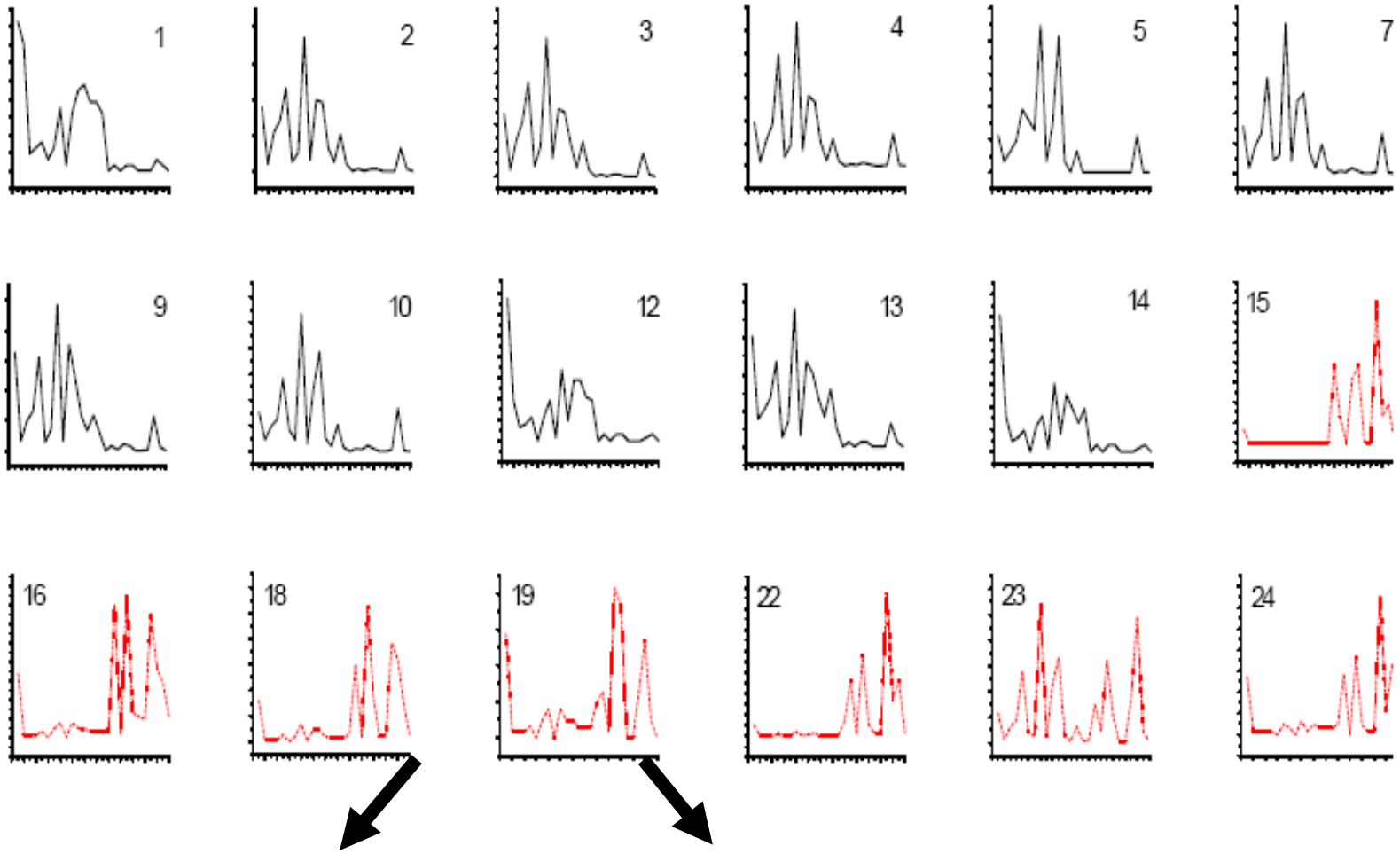
$$J_{ij} = J_{ji}$$



Configuración de spin final : minimización de E

p(estado del sistema) = $f(E, T, Z)$; Z=función de partición ;
T = temperatura

Energía textual



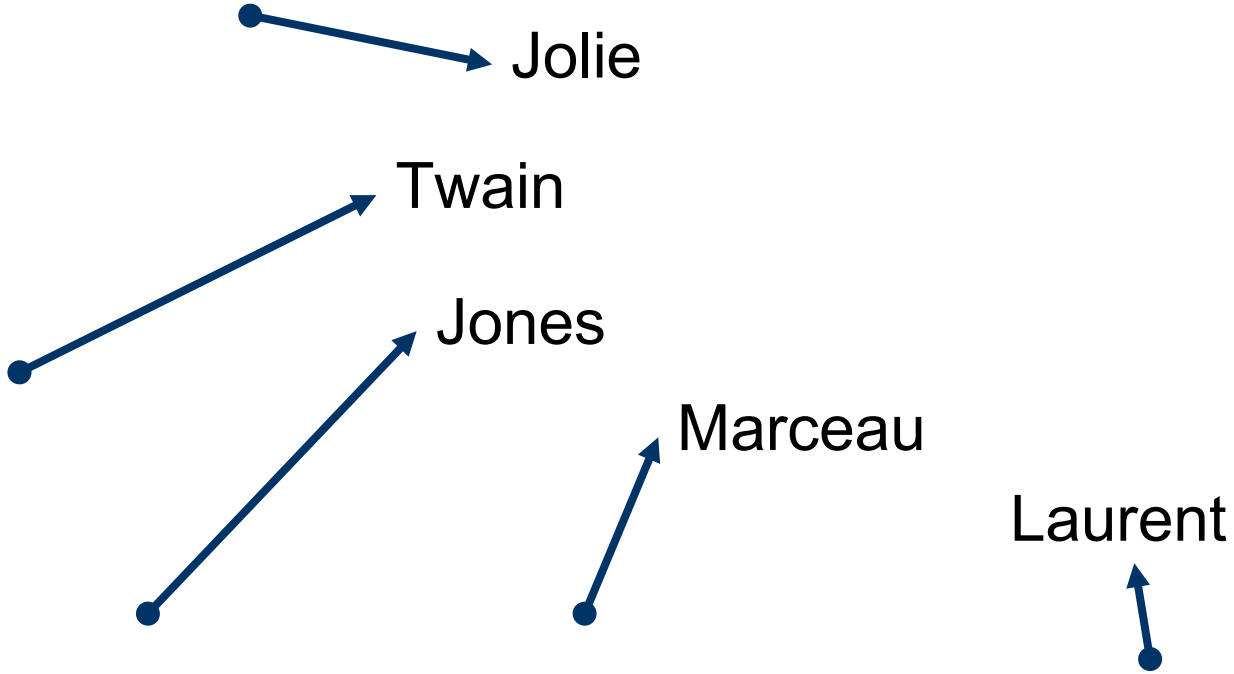
| E^{μ} | de frases :
Resumen automatico

Concordancia de curvas :
Segmentacion tematica

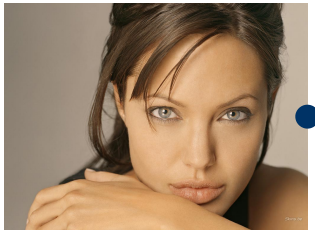
Ingredientes basicos:

Ideas de cognicion...

memorias asociativas

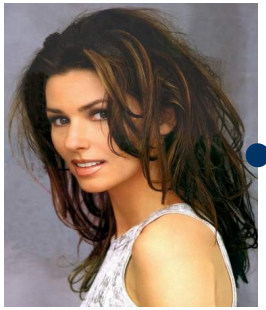


Memorias asociativas



Angelina Jolie

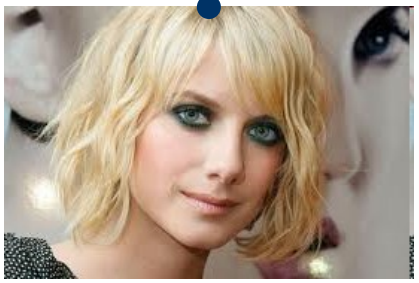
Shania Twain



Catherine Z Jones

Sophie Marceau

Mélanie Laurent



Ingredientes basicos

Linguistica

Analisis discursivo

Corpus

Computacion

Matematicas

Ingenieria...

Aplicaciones: Detección de fronteras temáticas

- Separar temáticamente documentos
- Corpus
  - Politica | Ciencia | Arte | Deportes | Cultura
- Tarea clásica de PLN
- Independiente del idioma (trilingüe en/fr/es)

Aplicaciones: Detección de similitud textual

La energia textual puede servir para detectar similitud textual...

En particular parafrasis, independiente del idioma, tematica y contenido

Por mi raza hablara el espiritu

El espiritu va a hablar por mi raza

Quien hablara por mi raza sera el espiritu

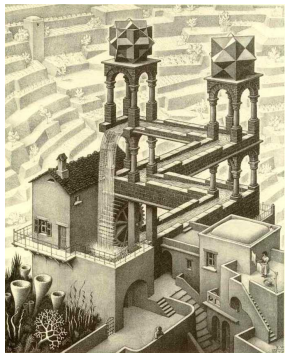
El espiritu sera quien hable por nosotros

...

Mas aplicaciones

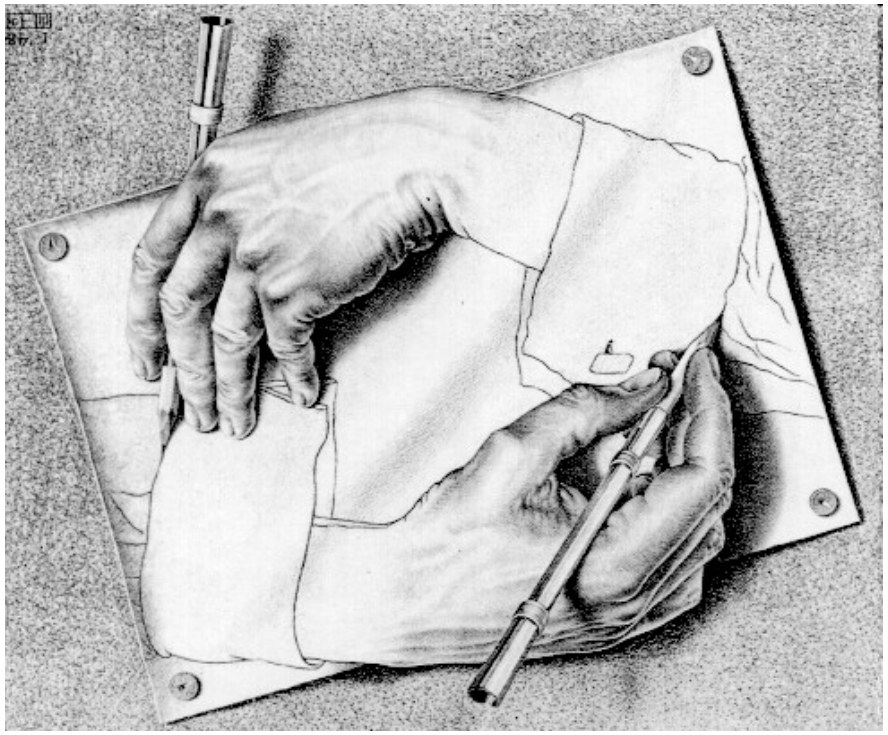
Resumen automatico,
Generacion de texto,
Clasificacion (tweets, blogs, documentos,...),
Identificacion de estilos,
Analisis automatico de CV,
...

Conclusiones...



No sabemos escribir programas que comprendan el texto como lo hace un humano...

Probablemente no necesitamos (o no podemos) escribir programas que verdaderamente comprendan el texto
Necesitamos únicamente escribir programas que razonablemente procesen masas de documentos en lugar de las personas... y que lo hagan bien y rápidamente



Merci beaucoup!

Avez-vous des questions?

juan-manuel.torres@univ-avignon.fr
http://juanmanuel.torres.free.fr/