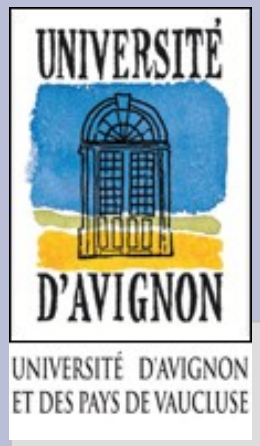


Semana del Procesamiento de Lenguaje Natural LIA, UAPV Optimización combinatoria y algoritmica, UAM-Azcapotzalco

Juan-Manuel Torres Moreno

LABORATOIRE
INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

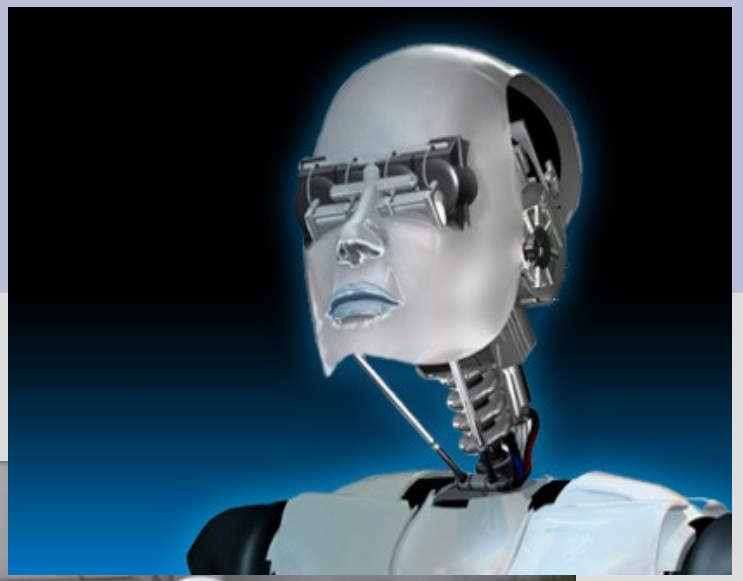
LABORATOIRE INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

¿Puede hacerse procesamiento automático de lenguaje natural sin usar lingüística?

Juan-Manuel Torres Moreno

juan-manuel.torres@univ-avignon.fr


AKIHABARANEWS
.com



¿Puede imaginarse una plática donde humanos y robots discuten sobre un tema del entorno en que se desenvuelven?

¿Puede imaginarse una plática donde humanos y robots discuten sobre un tema del entorno en que se desenvuelven?

- ¡Claro que eso es sólo ficción!
- ¡O mas aún: ciencia ficción!
- Pero aun siendo ficción, se trata de un objetivo atractivo que está mereciendo atención en el campo de la **inteligencia artificial**

Procesamiento de Lenguaje Natural (PLN)

- El problema se centra en el procesamiento de lenguajes naturales: espanol, francés, inglés, chino, alemán, arabe...
- A él se hace referencia en general, en términos de “lingüística computacional”
- Nace por motivos militares :
  - Desciframiento de los mensajes captados de los ejércitos enemigos (Turing 1954)
  - Resolver el problema de la traducción automática

PLN

- Problema **multidisciplinario** del que se ocupa:
  - la lingüística
  - la lógica
  - la informática
  - la psicología cognitiva
  - la ingeniería
  - las matemáticas
  - la epistemología...

¿Qué hace el PLN ?

- Area de investigación en continuo desarrollo
- Se aplica en diferentes actividades:
  - Traducción automática
  - Recuperación/Busqueda de información
  - Elaboración automática de resúmenes
  - Interfaces hombre/maquina
  - Generación de texto
  - Detección y clasificacion de opiniones/emociones
  - Diálogo...

¿Qué hace el PLN ?

- Un sueño desde hace mucho tiempo
- ¿Por qué tantas dificultades ?
  - Cantidad de lenguas humanas
  - Complejidad de las lenguas
  - Ambigüedad a varios niveles
  - Incertitumbre
  - Fenómenos culturales
  - ...

¿Qué tipo de textos trata el PLN?

- ¿Textos litérarios…?

  Bergère ô tour Eiffel le troupeau des ponts bêle ce matin

  (Pastor oh Torre Eiffel el rebaño de puentes bala esta mañana)

  Apollinaire, Alcools 1913

  À la nue accablante tu *S. Mallarmé*

  (A la pesadisima nube/desnuda tu)

- ¿Documentos (cientificos, periodisticos,…)?
  - vocabulario técnico
  - normalizado
  - formas sintácticas suficientemente simples

¿Qué tipo de textos trata el PLN?

- Texto literario
  - Expresa voluntariamente de forma *ambigua* un mensaje *complejo.* No podria ser resumido o reformulado en otros términos
  - La *litérariedad* soporte mal la traducción, los sinónimos…
- Documento (cientifico, periodistico,…)
  - El texto se desvanece frente a su contenido
  - El documento envejece rápidamente
  - Los documentos son géneralmente textos recientes que deben ser procesados rápidamente para difundir su contenido

¿Cómo hacer artefactos tecnológicos de PLN *reales*?

Ideas intuitivas:

- Si el problema es la lengua... ¡hagamos uso de la **lingüística**!
- La lingüística debe estar en el **centro de la problemática** de PLN
- La lingüística *debe aportar* soluciones a los problemas reales del PLN
- La lingüística *tiene un marco teórico* adecuado
- El análisis lingüístico es fino...

La lingüística estudia las lenguas humanas

- **Paradigma de la lingüística**
  - **Una lengua humana puede ser representada por una gramática formal**
- **Gramática formal**
  - Conjunto de reglas que deciden si una frase pertenece (*gramaticalmente correcta*) o no (*agramatical*) a una lengua

¿La lingüística resuelve los problemas del PLN?

- **Vi al hombre con el telescopio**
  - ¿Usé un telescopio para ver al hombre?
  - ¿El hombre tenía un telescopio ?
- **Me gusta acariciar los gatos. A tu novia también**
  - **Ah bon ??**
- ***J'amerais descendre un avocat***

Quisiera *(descendre=bajar/matar) un (avocat=aguacate/abogado)*

- **Mmmmmh ¿Tienes problemas ? No importa... ¿Cuánto pagas?**

¿La lingüística resuelve los problemas del PLN?

- **Los humanos son muy buenos para**
  - **Resolver la ambigüedad:** elegir el sentido pertinente de una frase
  - **Robustez:** los humanos detectan y no aceptan pequeñas desviaciones de las frases
  - **Desempeño:** los humanos son capaces de procesar frases complejas eficazmente *: ellas* representan desafíos enormes a la lingüística

Ambigüedad

- **Varios niveles**
  - **Significación:** *Puce* (pulga) vs *Puce* (circuito electronico)
  - **Etiquetas gramaticales (**Análisis sintáctico**):**

La belle ferme le voile

(a) La bella granja lo oculta

(b) La bella cierra el velo

(a) Art Adj Nom Pro Ver ?

(b) Art Nom Ver Art Nom

Ambigüedad

- **Las lenguas humanas son ambiguas e inciertas... ¡eso hace su riqueza!**
- Eso es al mismo tiempo su **dificultad** para ser **analizadas automaticamente**
- Las gramaticas formales pueden producir un numero de analisis de frases, *exponencial* respecto al numero de palabras de la frase

Ambigüedad

- **Un sistema PLN basado en lingüística debe resolver entre un gran número de posibles caminos**
- Para un humano es diferente: poseemos vastos recursos extra lingüísticos:
  - Conocimiento del mundo
  - Preferencias culturales
  - Experiencia...

Ambigüedad

- Una gramatica formal no tiene acceso a conocimientos extra-linguisticos
- Incapaz de tratar la ambiegüedad : no puede resolver la incertitud
- Determinista

Robustez

- **Frases no-gramaticales :**
  - John not home **(falta el verbo)**
- **No gramaticales y ambiguas**
  - L'artiste peins la nuit **(Mal acuerdo con la 3a persona:** L'artiste ***peint*** la nuit **; “pinta la noche” o “pinta de noche” ?)**
- **Uso no comun de la lengua**
  - “pie niche haut, oie niche bas”
  - **La** pie **niche** en **haut,** l'**oie niche** en **bas**
  - *El gorrion anida en lo alto, el ganso anida abajo*

Desempeño

- *Remarcablemente, el **linguista** (computacional o no) **pueden** llegar a pensar que el **modulo comun** de todas las aplicaciones de lenguaje natural debe ser un modelo de procesamiento **linguistico**; y más **aun**, que **deberia** jugar un rol central en la **interaccion** (vocal o escrita) hombre-**maquina**. Sin embargo la **pratica** en vigor para **esas tareas** no incluye absolutamente los modelos reales del procesamiento **linguistico.***

Desempeño

- *De hecho reportes **claros** salen constantemente **a la luz** sobre el uso de la **tecnologia** vocal y del lenguaje escrito: el uso de la estructura lingüística en **ellos** apenas mejora el procesamiento. Los reportes vienen **acompanados dela conclusion** que los modelos lingüísticamente fundamentados no son precisos, ni robustos y que, en cambio son **terriblement** ineficientes.*

¿La lingüística resuelve los problemas del PLN?

- Frases largas, estilo
- Uso de anáforas, sinónimos, antónimos
- Gramatica compleja
- Sintaxis pobre: palabras sin acentos o mal acentuadas
- **¿Y la semántica ?**

¿Los modelos lingüísticamente motivados para el PLN pueden aun jugar un rol en las tecnologias del lenguaje natural?

¿Y la semántica?

- ¿Se necesta comprender un texto para *procesarlo* adecuadamente?
- El caso de **FRUMP**: un sistema que genera resumenes automáticos por *comprensión* del texto
  - Análisis lingüístico
  - Interpretación usando conocimientos
  - Generación de texto: activación de *scripts* por palabras claves « comprendidas »

FRUMP

a small eartquake shook several Southern Illinois counties Monday night, the National Earthquake Information Service in Golden, Colo., reported. Spokesman Don Finley said the quake measured 3.2 on the Richter scale, "probably not enough to do any damage or cause any injuries." The quake occurred iabout 7:48 p.m. CST and was centered about 30 miles east of Mount Vernon, Finlay said. It was felt in Richland, Clay, Jasper, Effington, and Marion Counties.

(TEXTO)

There was an earthquake in Illinois with a 3.2 richter scale. (RESUMEN)

¡RESUMEN REMARCABLE!

FRUMP : resultados fuera de su dominio de comprensión …

- ¿Los 50 scripts de FRUMP son suficientes para interpretar al mundo?
- Los conocimientos son *codificados* manualmente
- ¿Es fácil aprender scripts de nuevos dominios ?

Ciudad del Vaticano. La noticia de la muerte del Papa *sacude* al mundo. Murió el martes pasado de forma… (TEXTO)

Sismo en el Vaticano : un muerto. (RESUMEN)

¿La lingüística resuelve los problemas del PLN?

- *Las gramáticas formales limitan las fronteras del lenguaje*
  - *las frases no-gramaticales son clasificadas como no pertenecientes a la langua*
- *Les frases extrañas, mal escritas o con sintáxis pobre (chat, SMS, e-mail...) no son un problema insoluble para un humano*
  - *La interpretación es elegida en función de su* ***uso corriente***

¿La lingüística puede procesar esto?

from <juan.manuel.torres@univ-avignon.fr>

to <ana.lilia@uam.mx>

Bonjour!!! el vuelo af estuvo ufffffff!! (mas de 13).
grcias por el e-ticket eh? :-)

A 1/2 noche no dieron de comer un chingo y casi nos da indi gestion... Vi n+1 pelis, ya ni se. La n++ je rien compris y hizo que me pusiera super >-(

confirmo @ pasado: doy 4 sem. uama y 1 IMASS... Salu2! y 1 :-*

JM

¿La lingüística puede procesar esto? (2/3 palabras cambiadas)

from <juan.manuel.torres@univ-avignon.fr>

to <ana.lilia@uam.mx>

¡Hola! El vuelo de Air France estuvo ¡cansadísimo! (Fueron más de 13 horas). Gracias por el boleto electrónico. Me dio mucho gusto.

A media noche nos dieron de comer muchisimo y casi nos indigestamos... Vi muchas películas. No sé cuántas. La última no la entendí e hizo que me pusiera muy enojado.

Confirmo tu correo pasado: doy cuatro seminarios en la UAM-Azcapotzalco y uno en el IIMAS. ¡Saludos! y un beso.

Juan Manuel

¿La lingüística resuelve los problemas del PLN?

- *El desempeño sale de **su** campo*
- *La robustez también*
- *Las gramaticas formales son indecuadas para resolver la ambigüedad porque no pueden resolver la incertidumbre*

- ***¿Qué hacemos entonces? ¿Qué queda?***

¡Los métodos probabilistas!

- Abordan la incertitud formalizando un ***modelo de lenguaje*** como una *distribución de probabilidad*
- Ambigüedad: cada análisis tiene asociado un valor ***0 ≤ p ≤ 1*** : indica el **grado de pertenencia** de la frase a la lengua
- Los valores son más que números: probabilidades fundadas en la **Teoria de Probabilidades**

Teoría de probabilidades

- Modelización adecuada de la noción intuitiva de la probabilidad de eventos (independientes o no)
- Camino empírico para estimar las probabilidades : la estadistica
- Interpretación de la teoría de probabilidades

Ejemplo : etiquetas gramaticales

- *La belle ferme le voile*

AMBIGUEDAD...

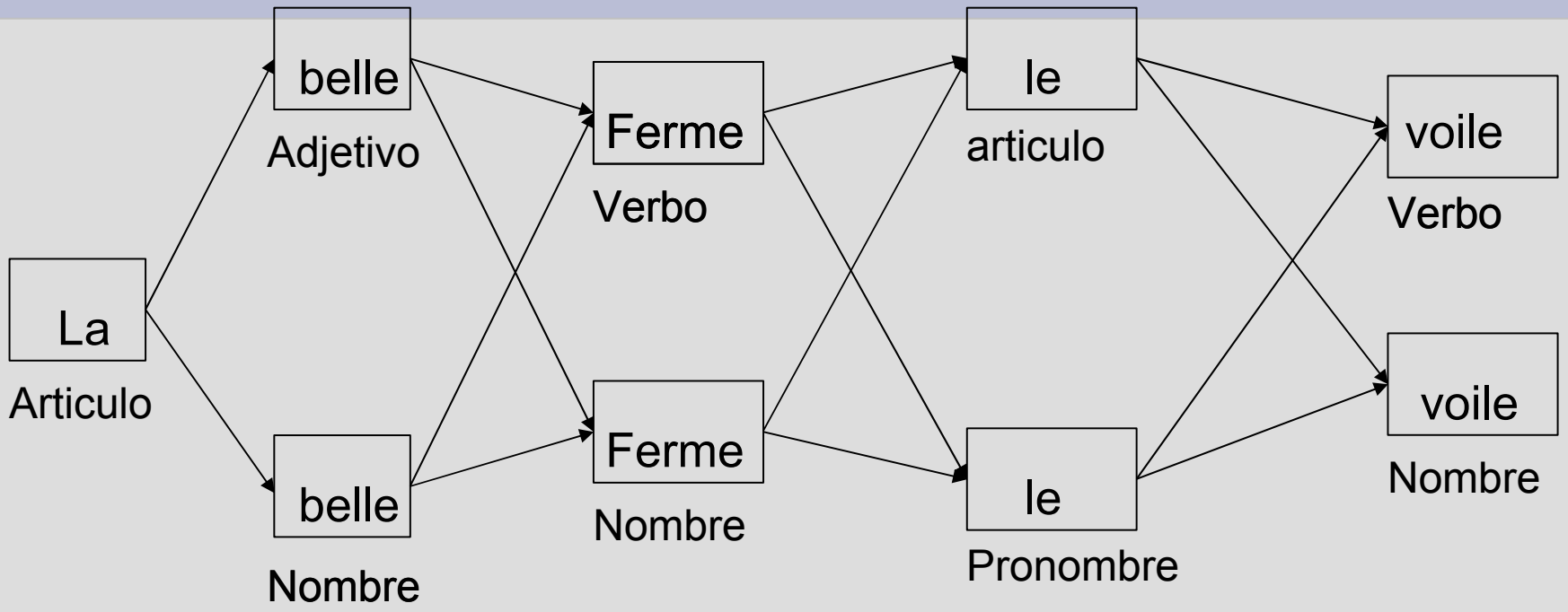
- (a) *La bella cierra el velo*
- (b) *La bella granja lo oculta*

(a) Art Adj Nom Pro Ver ? **o bien :**

(b) Art Nom Ver Art Nom ?

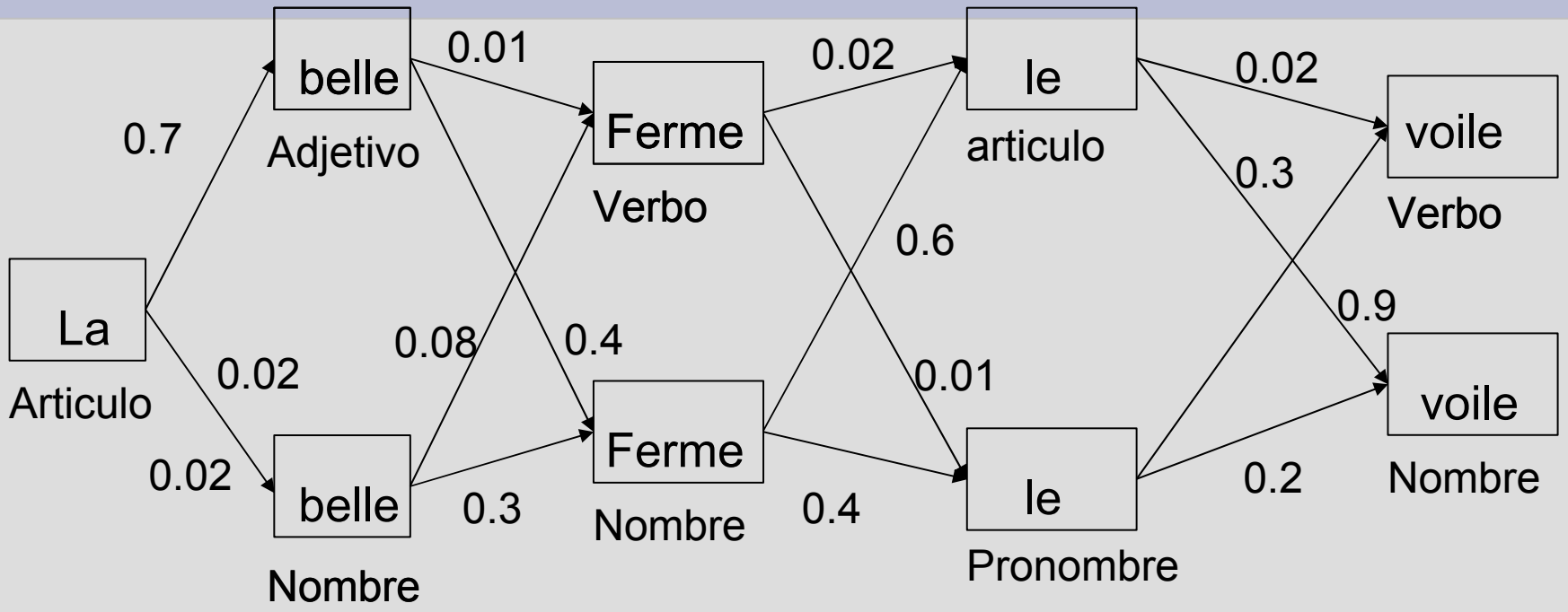
La linguistica no puede resolver esta ambiguedad...

Ejemplo : etiquetas gramaticales

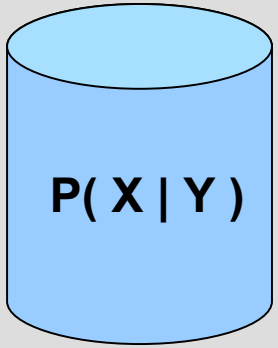


¿Qué camino tomar ?

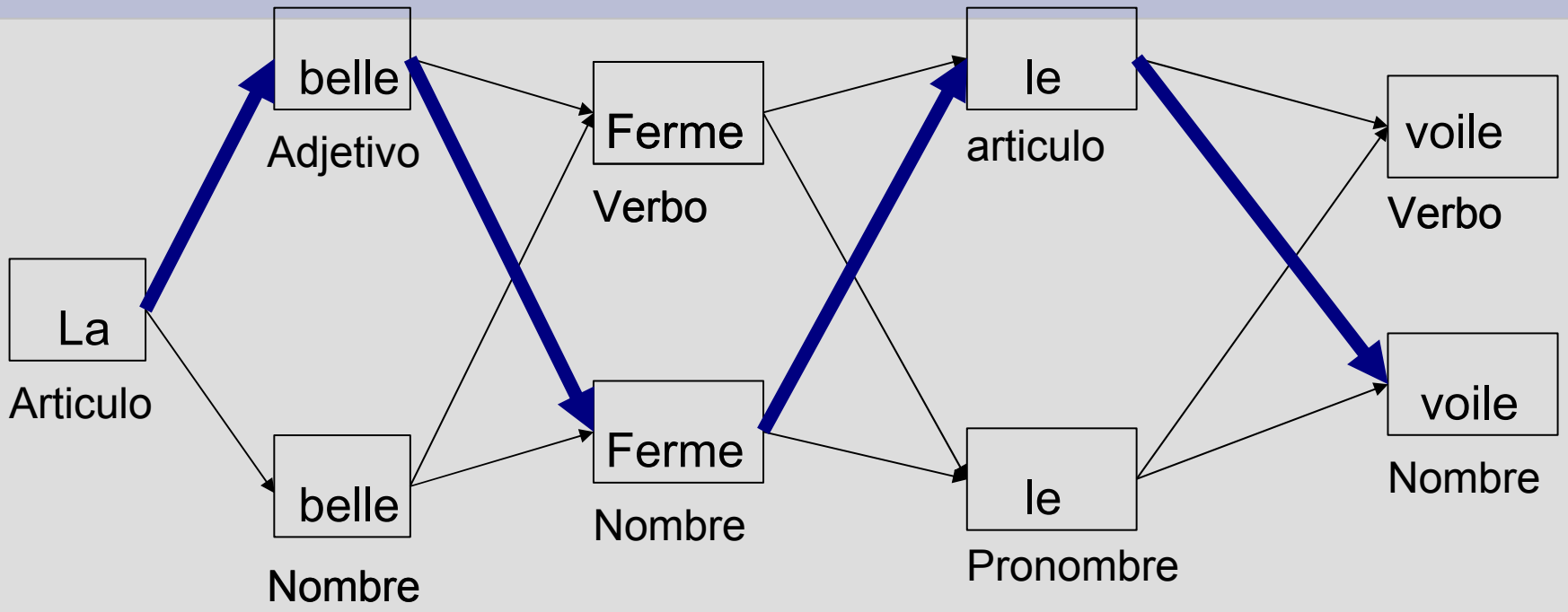
Etiquetas gramaticales : calculadas



Probabilidades a partir de un Corpus (miles de palabras)



Etiquetas gramaticales asociadas



LA / **Art** BELLE / **Adj** FERME / **Ver** LE / **Art** VOILE / **Nom**

Resúmen por métodos numéricos

Ciudad del Vaticano. La noticia de la muerte del Papa *sacude* al mundo. Murió el martes pasado de forma … (TEXTO)

La noticia de la muerte del Papa *sacude* al mundo. (RESUMEN)

NO HAY COMPRENSION. NO HAY ANALISIS LEXICO NI SEMANTICO. NO NECESITA SCRIPTS NI RECURSOS EXTRA LINGUISTICOS... PERO EL RESUMEN ES *ACEPTABLE*

Ventajas

- Extiende naturalmente la teoría de conjuntos para **tratar la incertitumbre** (implicaciones directas en ambigüedad, robustez y el desempeño)
- Es una **interpretación directa y empírica** a partir de la estadistica
- Posee un enlace directo con la **teoría del aprendizaje automático**
- Ventajas metodológicas importantes: la **optimización** y la **descomposición** modular

Abordar el problema pragmaticamente

- Todavía no sabemos escribir programas que comprendan realmente el texto como lo hace un humano
- La lengua es demasiado compleja para modelarla con reglas: ¡que la máquina la aprenda por ejemplos!
- Incluso para un individuo es difícil explicar como llega a una conclusión a partir de una lectura *entre líneas*

Reflexiones

- Un programa no puede reproducir la manera en que las personas leen y producen documentos.
- Son caminos muy diferentes: un abismo de experiencias, de percepciones, de emociones...
- El camino de la máquina es inhumano... ahí reside su fuerza

Conclusion...

- El enfoque numérico del PLN es *comprensible*: los números dicen cosas (a quien los sabe escuchar):
  - Siempre se puede saber por qué el sistema siguió un camino y no otro
- Probablemente no necesitamos escribir programas que verdaderamente comprendan el texto
  - Ejemplo: los resumidores automáticos

Conclusion

- Se necesita únicamente escribir programas que *razonablemente* procesen masas de documentos en lugar de las personas... y que lo hagan bien y rápidamente
- En tanto que el enfoque sea eficaz, poco importa si es inhumano... siempre y cuando los resultados sean satisfactorios

La ciencia se acerca cada vez más a la ficción

Ultima reflexión: por qué no combinar ambos caminos adecuadamente?

Ultima reflexión: por qué no combinar ambos caminos adecuadamente?

Base numérica para procesar la masa de documentos

+

Análisis lingüístico fino para mejorar resultados

=

Sistemas híbridos