MPGI UC
MAGISTER EN PROCESAMIENTO Y GESTIÓN DE LA INFORMACIÓN

Curso de procesamiento del lenguaje natural

César Antonio Aguilar
Facultad de Lenguas y Letras
12/08/2020

Cesar.Aguilar72@gmail.com

PLN: ¿qué es?

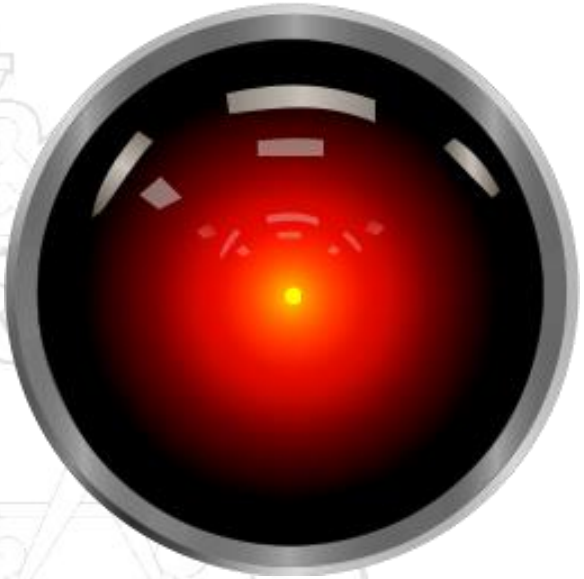
PLN es un área emergente de investigación, la cual tuvo su origen dentro del campo de la inteligencia artificial.

Si estableciéramos un año fundacional, éste podría ser 1950, cuando Alan Turing escribió un artículo titulado “Computer Machinery and Intelligence”.

En este artículo, se plantea que si una máquina es capaz de comprender y generar lenguaje humano, entonces se puede considerar que se trata de una máquina inteligente.

Alan Turing (1912-1954)

El Dr. Chandra platica con Sal, la hermana de Hal



Uno de los ejemplos más creíbles que podemos ver en el cine sobre lo que es la interacción de tipo humano-máquina lo podemos ver en una película hecha por Stanley Kubrick en 1968: *2001: A Space Odyssey*. En ella, vemos como el Dr. Sivasubramanian Chandrasegarampillai (o el Dr. Chandra) desarrolla en computador más avanzado de su tiempo, Hall 9000, capaz de realizar varias tareas **complejas, entre ellas dialogar con astronautas.**

Nueve años después, Hall tiene una hermana, Sal 9000, la cual en esencia tiene los mismos atributos que su hermano mayor. Un ejemplo de su funcionamiento lo pueden ver en la liga que anoté en la página del curso.

Cosas más reales (1)

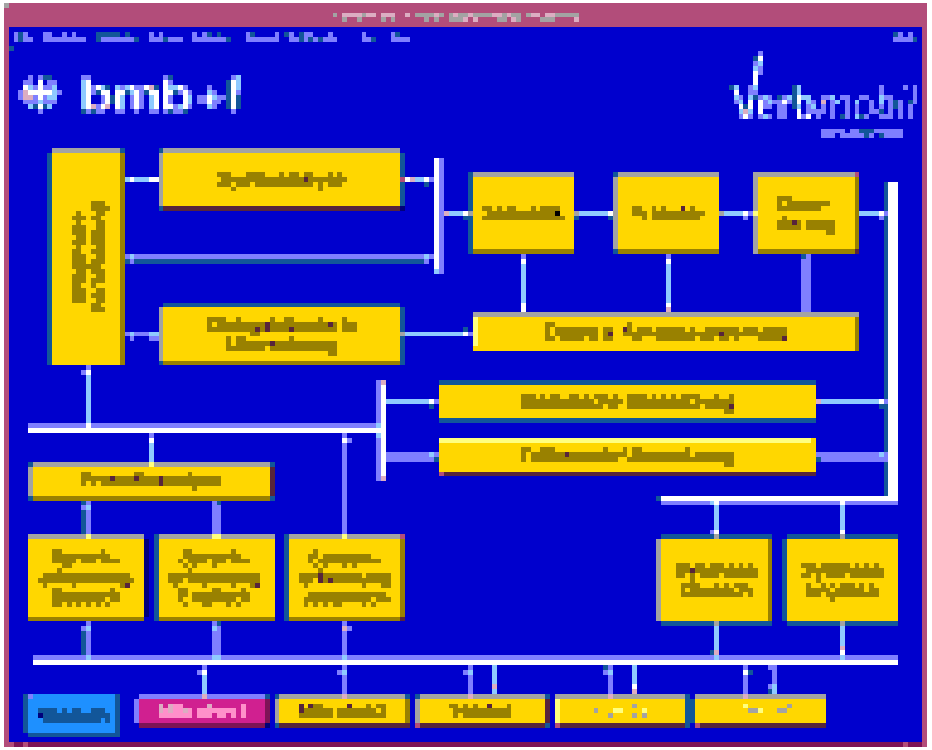
Threepio, a pesar del avance tecnológico actual, es todavía parte de la ciencia-ficción. ¿Qué podemos hacer hoy en día con las tecnologías del lenguaje?

Para tener una mejor idea de lo que es el PLN, enfocado en este caso su relación con la ingeniería, veamos la explicación que nos da el Dr. Jesús Cardeñosa Lera, quien es un profesor/investigador de la Universidad Politécnica de Madrid (UPM), en una entrevista que le hacen para el canal *Tendencias 21*:

www.youtube.com/watch?v=WOXbdVV2qMI

Cosas más reales (2)

Igualmente, un ejemplo de aplicación es el sistema de traducción automática Verbmobil, desarrollado bajo el auspicio del Ministerio Federal Alemán para la Educación, la Ciencia, la Investigación y la Tecnología en 1996.

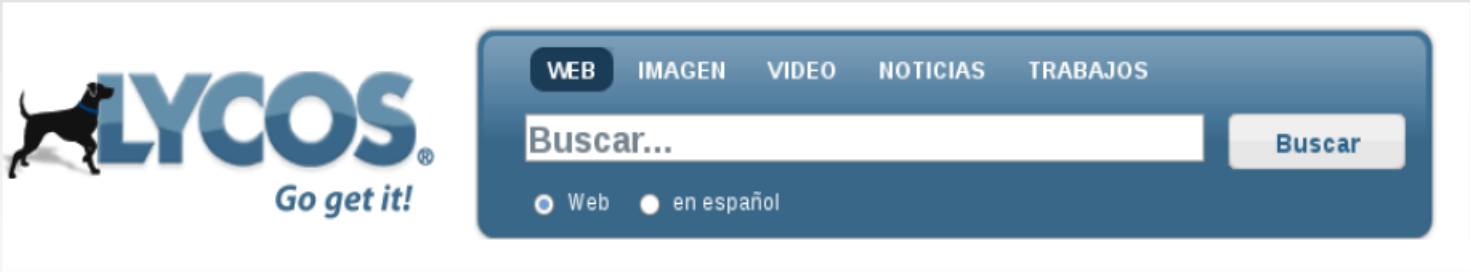


Liga al video: www.youtube.com/watch?v=noZBab-Lmss

Mayor información sobre Verbmobil: http://verbmobil.dfki.de/

Cosas más reales (3)

De acuerdo con a Martin, hoy en día las líneas de investigación más atractivas del PLN tienen que ver con procesamiento de información. Revisemos algunas:

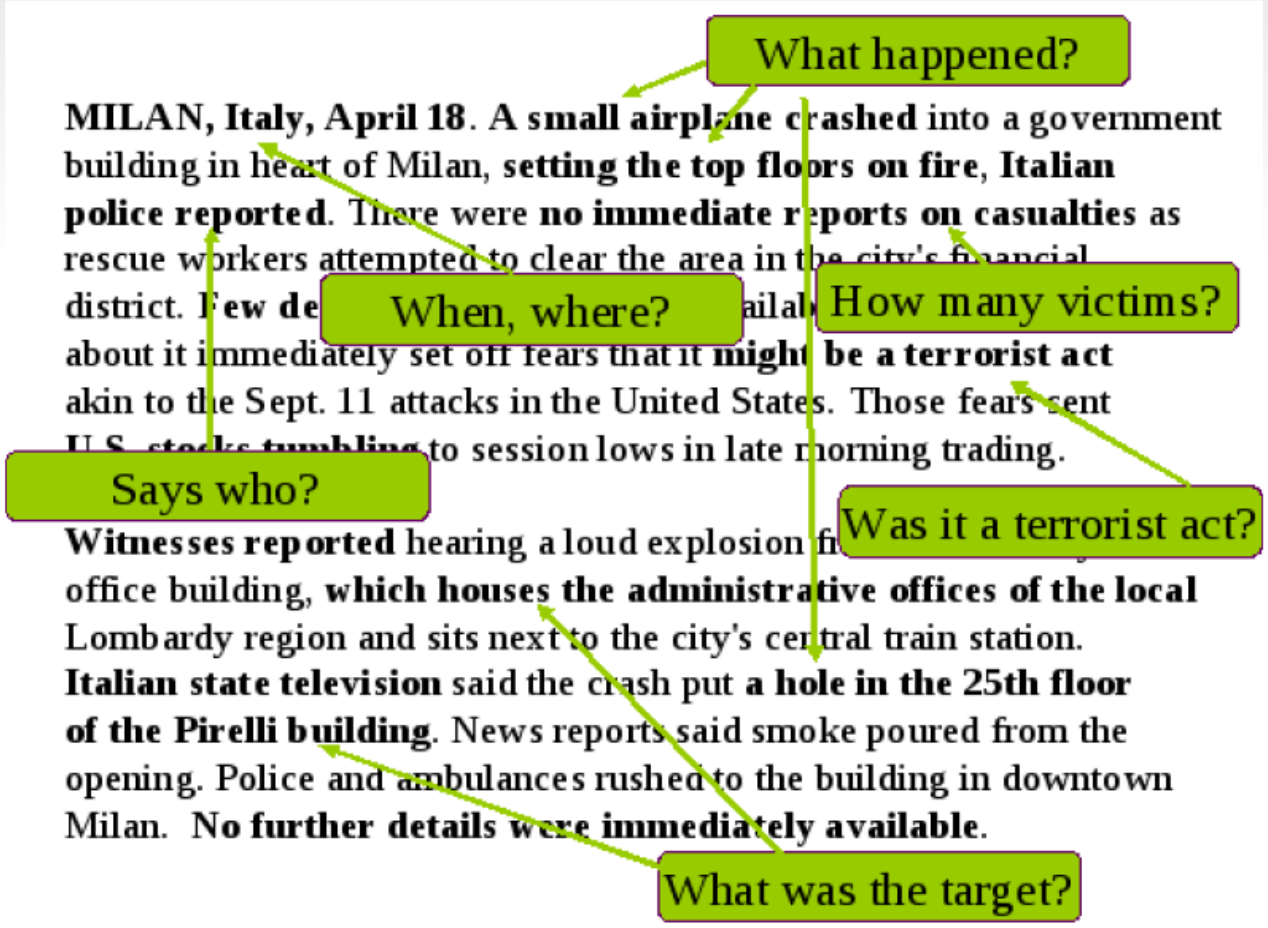


Recuperación de información: es un área de investigación dedicada a la localización de datos e información en documentos.

- Búsquedas de documentos localizados en Internet o en bases de datos.
- Esto se hace considerando datos y/o metadatos que describan, en síntesis, un documento: título, autores, palabras claves, etc.
- En general, el proceso consiste en generar listas de documentos asociados con alguna consulta.

Cosas más reales (4)

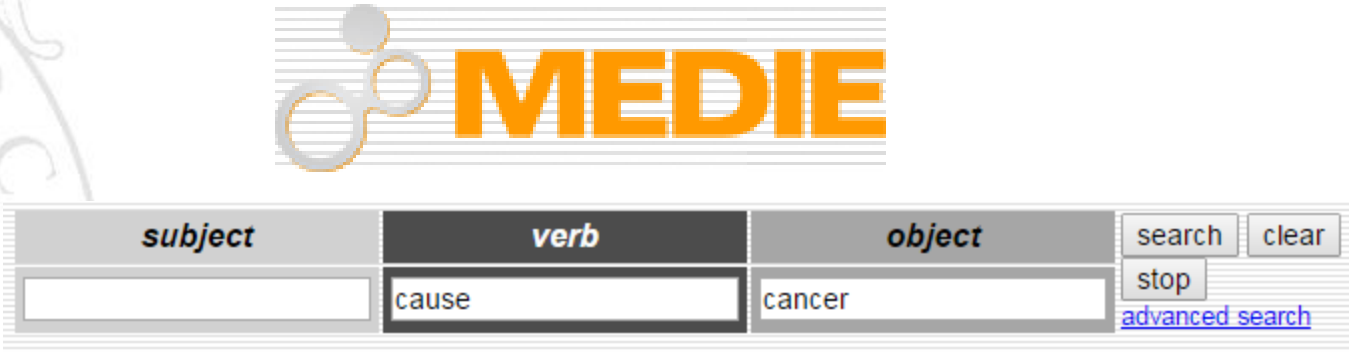
Extracción de información: con un mayor grado de complejidad, es un área que trata de reconocer, de forma (semi-)automática, fragmentos específicos dentro de un documento que contengan información concreta sobre una entidad o un evento dado.



Cosas más reales (5)

Minería de textos: esta línea de investigación está enfocada en la obtención de información textual sumamente precisa, combinando métodos probabilísticos y formales, particularmente técnicas de **aprendizaje automático** y **reconocimiento de patrones**.

Un ejemplo de un sistema para minería de textos es *Medie*, un motor de búsqueda de información médica desarrollado por el **National Center of Text Mining** (www.nactem.ac.uk):



www.nactem.ac.uk/medie/search.cgi

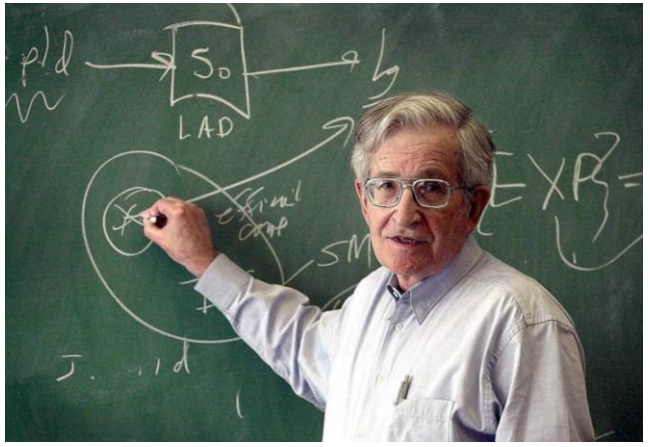
El impacto del PLN en nuestra tecnología (1)

Existen muchas expectativas acertadas e incorrectas sobre lo que se puede hacer hoy en día con las líneas de investigación y los sistemas que hoy creamos y usamos en PLN.

Ahora, lo cierto es que estas expectativas están atrayendo a una gran cantidad de personas de distintas ramas académicas y empresariales que quieren hacer cosas al respecto.

Noam Chomsky, el padre de la lingüística generativa, considera que uno de los aportes importantes del PLN es que ofrece un escenario interesante para validar (o invalidar) teorías lingüísticas, ya que no es pertinente fiarse del todo del análisis estocástico que hacen los computadores de un texto.

MIT Linguistics



www.youtube.com/watch?v=EjoywQq4V44

El impacto del PLN en nuestra tecnología (2)

En contraste, la opinión de Richard Wolniewicz (director de la división de 3M dedicada a sistemas de información en salud) es más positiva que la de Chomsky: el PLN ha cambiado nuestra manera de procesar información en línea, y gracias al uso de modelos estocásticos, hoy podemos diseñar prototipos de inteligenica artificial tan avanzados como **Watson**, desarrollado por IBM, por citar un ejemplo.

3M Science. Applied to Life.™

www.youtube.com/watch?v=O9kjGtQfdAs

El giro paradigmático de Watson (1)

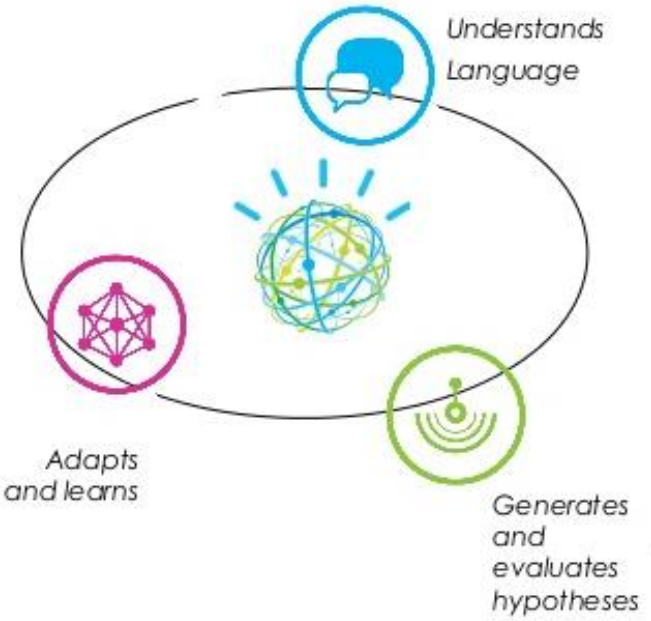
En el año 2011, IBM participó en el programa televisivo *Jeorpardy!*, en una serie de sesiones en donde sus competidores eran campeones de sorteos anteriores. La idea, a grandes rasgos, era demostrar que en ese momento existía la tecnología necesaria para desarrollar un cerebro electrónico capaz de responder preguntas y respuestas como lo hacen los humanos. Lo que ocurrió en ese concurso ya es parte de los anales de la historia de la inteligencia artificial:

www.youtube.com/watch?v=P18EdAKuC1U

El giro paradigmático de Watson (2)

Muchos expertos en inteligencia artificial, tanto científicos académicos como industriales, consideran que Watson ha inaugurado un nuevo paradgima en las ciencias computacionales, el cual denominan **computación cognitiva**, el cual se entiende del siguiente modo:

Cognitive Computing can fuel digital transformation



Dimensions we need

Perception:
understand the world as we do: it interprets sensory input beyond traditional data

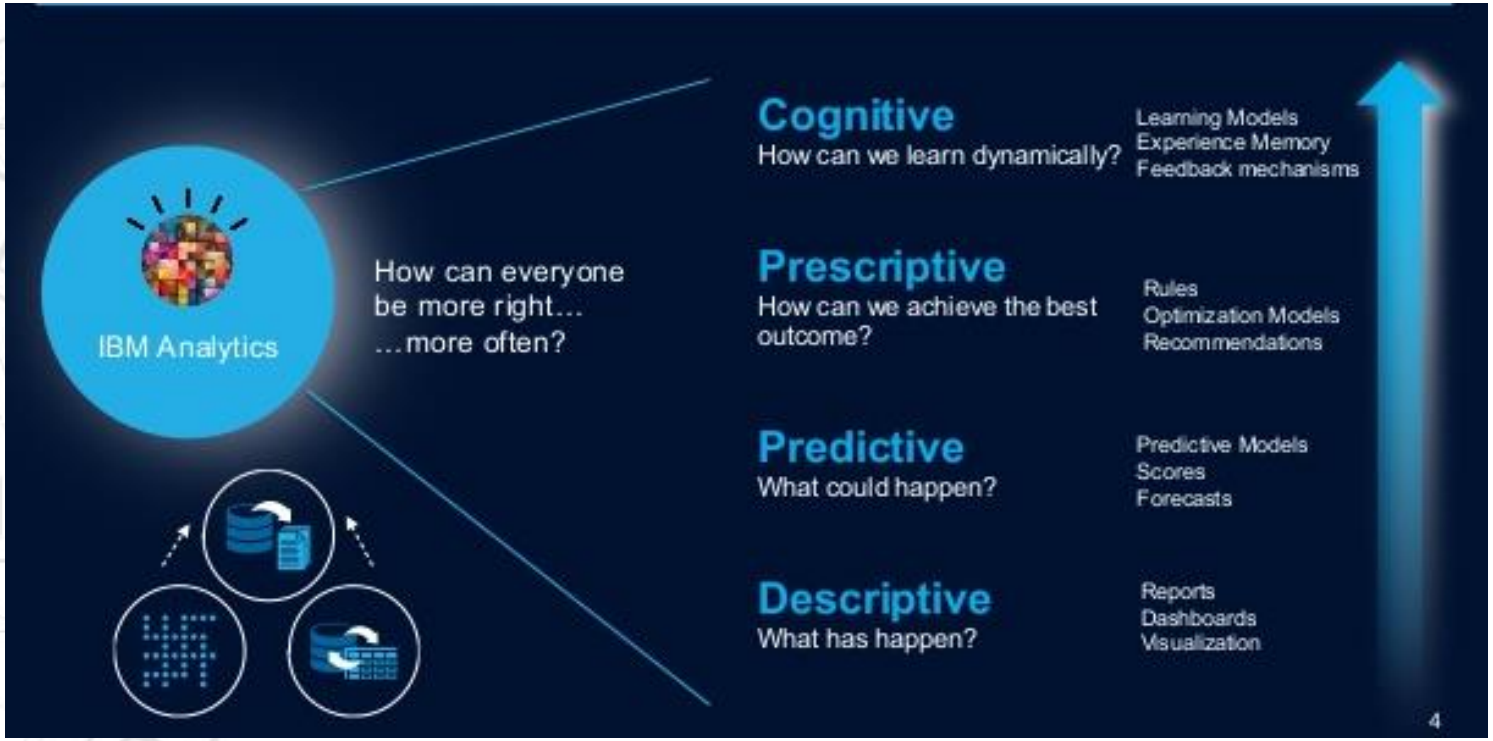
Reasoning:
think through complex problems: it deepens our analysis and inspires creativity

Relating:
understand how we communicate, and personalizes its interactions with each of us

Learning:
learn from every interaction, scaling our ability to build experience

El giro paradigmático de Watson (3)

Muchos expertos en inteligencia artificial, tanto científicos académicos como industriales, consideran que Watson ha inaugurado un nuevo paradgima en las ciencias computacionales, el cual denominan **computación cognitiva**, el cual se entiende del siguiente modo:



www.youtube.com/watch?v=_Xcmh1LQB9I

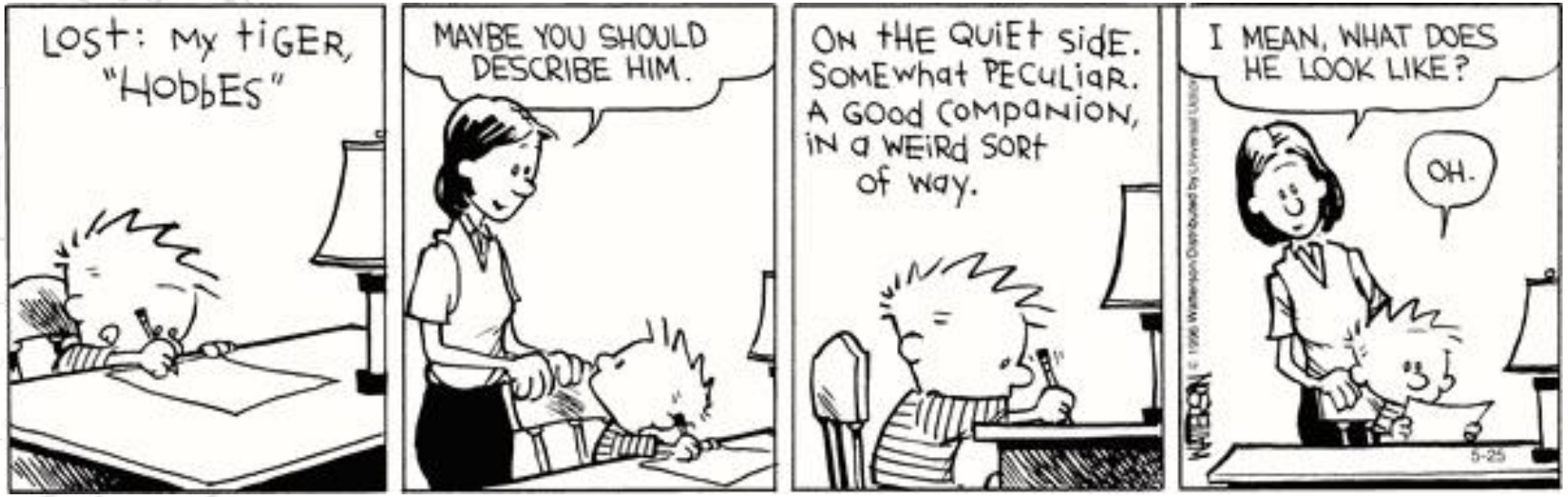
El giro paradigmático de Watson (4)

Claro está, la computación cognitiva también está teniendo presencia en América Latina, por lo pronto en un plano de difusión:

www.youtube.com/watch?v=cTcjpNokA5I

PLN: ¿la próxima revolución en IA? (1)

Todas estos avances generan discusiones y expectativas respecto a los aportes que puede brindar el PLN para el futuro. Hay personas (p. e., lingüistas como Chomsky) que desconfían mucho de los logros alcanzados por sistemas como IBM Watson, y los ven como *fuegos de artificio*, pues les generan muchas dudas pensar que realmente una máquina, algún día, pueda ser realmente capaz de comprender y generar lenguaje humano de manera autónoma, p. e.:



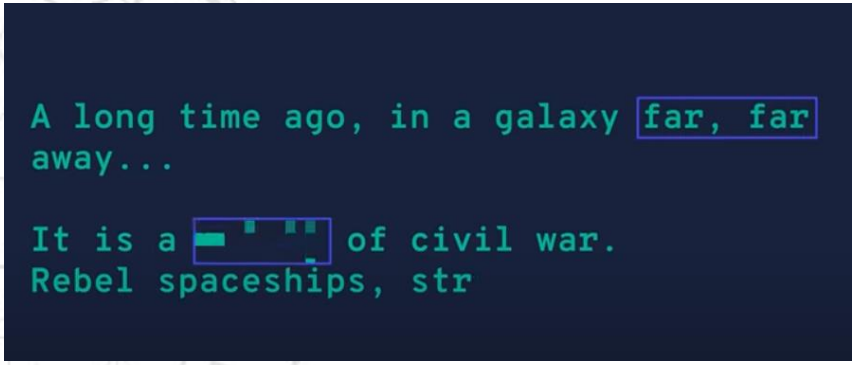
PLN: ¿la próxima revolución en IA? (2)

Por otro lado, muchos de nosotros tenemos poca información sobre las teorías, métodos y los efectos que conlleva el desarrollo de la IA en nuestras sociedades.

De hecho, una gran mayoría de personas en América Latina, África y Asia ni siquiera sabe de esto, pues son parte de la **brecha digital**: el nulo acceso a cualquier tipo de tecnología:

PLN: ¿la próxima revolución en IA? (3)

www.linkedin.com/in/carlossantanavega/

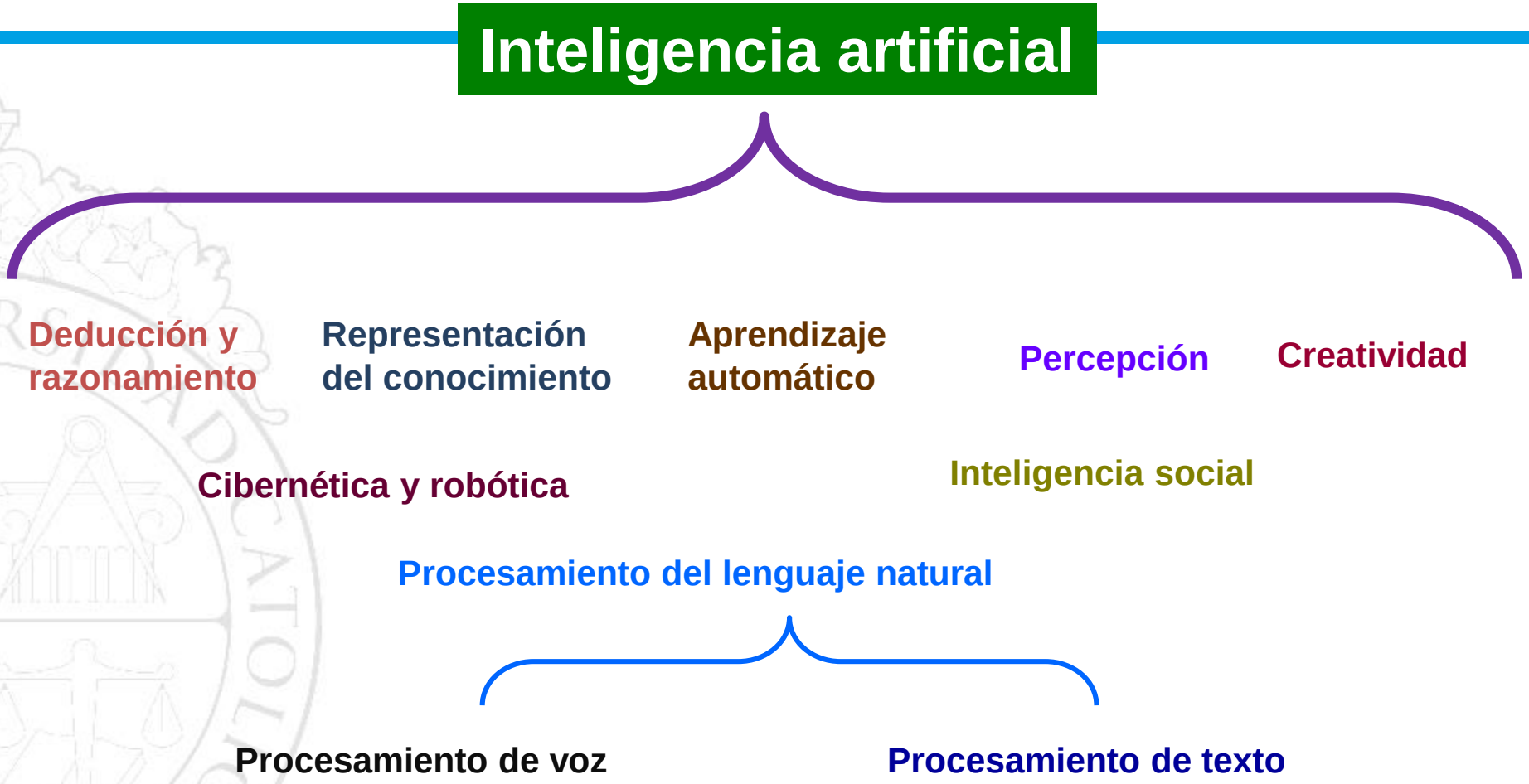


www.youtube.com/watch?v=cTQiN9dewIg

Finalmente, hay expertos que no únicamente crean nuevo *software* o *hardware* capaz de procesar lenguaje, sino que vislumbran cambios radicales en nuestra forma de comprender e interactuar con el mundo.

En este sentido, PLN es una de las áreas que, conforme a la opinión de **Carlos Santana Vega**, va a definir el futuro de la IA a partir de esta década:

Estableciendo equivalencias (1)



Estableciendo equivalencias (2)

Lingüística teórica

Fonética/ Fonología

Morfología

Lexicología

Sintaxis

Semántica

Análisis del discurso

Pragmática

Lingüística computacional

Estableciendo equivalencias (3)

Lingüística aplicada

Psicolingüística

Lexicografía

Traducción

Sociolingüística

Neurolingüística

Adquisición y desarrollo del lenguaje

Terminología

Enseñanza de lenguas

Terapia del lenguaje

Análisis del discurso

Ingeniería lingüística

Estableciendo equivalencias (4)

Lingüística computacional

Procesamiento de voz

Lematización /tokenización

Lexicones

Chunking /parsing

Relaciones semánticas

Representación del discurso

Inferencias pragmáticas

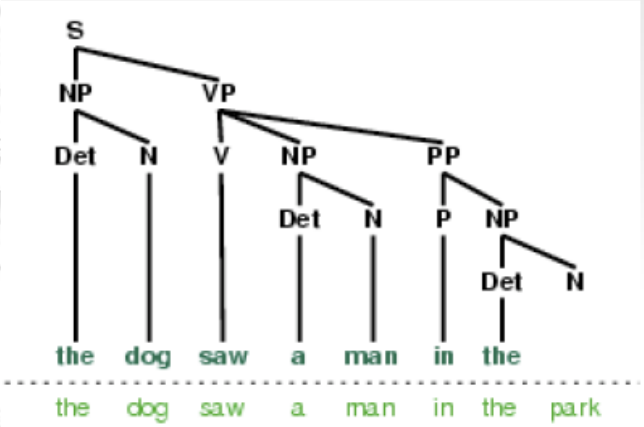
Estableciendo equivalencias (5)

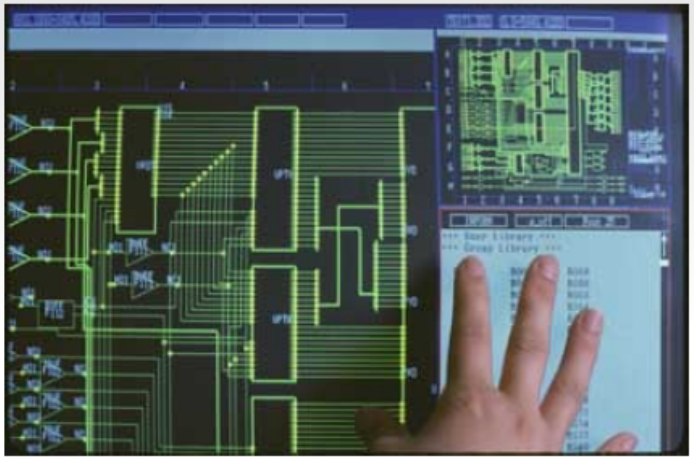
Ingeniería lingüística

- Reconocimiento/ generación de voz
- Recuperación y extracción de información
- Clasificación de documentos
- Lexicografía computacional
- Extracción terminológica
- Minería de textos
- Traducción automática
- Sistemas pregunta-respuesta
- Reconocimiento de entidades nombradas
- Corpus lingüísticos
- Resumen automático
- Ontologías

Estableciendo equivalencias (6)

Con todo, no siempre es fácil ni pertinente poner límites entre lingüistas computacionales e ingenieros lingüistas:

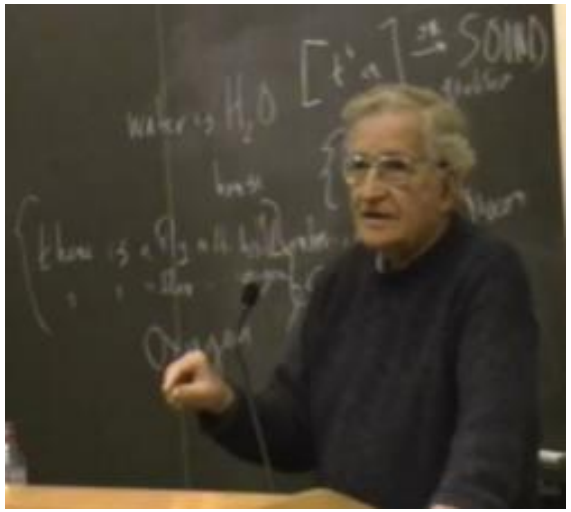




Digamos que tenemos entonces dos caras de la misma moneda: por un lado, un área enfocada en el estudio y modelado matemático (basado en reglas formales y métodos estadísticos) del lenguaje natural; y por el otro lado una rama que se encarga de desarrollar sistemas y herramientas capaces de procesar lenguaje natural.

¿Y qué hacemos? (1)

Como todo buen lingüista, nos interesa estudiar, describir y modelar las formas y las funciones que operan en toda lengua humana.



Como buenos computólogos, siguiendo la propuesta de Turing o Von Neumann, queremos ocupar ese conocimiento lingüístico para crear máquinas inteligentes.

¿Y qué hacemos? (2)

```
((Takes NP SBar) (Type 2))

((Takes NP NP Inf) (Type 2 ORaising))

(or ((Takes NP NP NP) (Type 2 ORaising))
    ((Takes NP NP AuxInf) (Type 2 ORaising)))

(or ((Takes NP NP AP) (Type 2 ORaising))
    ((Takes NP NP AuxInf) (Type 2 ORaising)))
```

Métodos basados en reglas lingüísticas: el lenguaje natural sigue una serie de reglas que sirven para construir estructuras con patrones regulares.

Métodos estadísticos: el lenguaje natural, al presentar cierta regularidad, ofrece un margen de probabilidad estadística para determinar si es viable que se genere o no cierta estructura lingüística.

$a_{i,j}::B_{i,j}$	forms	$p(b_{k,j} \mid a_{i,j})$	-p*log(p)
p::a	1365	0.189425	0.31518
p::b	1	0.000138773	0.00123268
p::c	2	0.000277546	0.00227297
p::d	1	0.000138773	0.00123262
p::e	1396	0.193727	0.317965

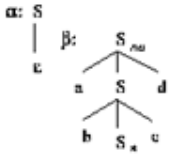
¿Y qué hacemos? (3)

Herramientas

Etiquetado

XMLWriter

EAGLES
Expert Advisory Group on Language Engineering Standards



The XTAG Project

TIGER Corpus

Annotation

Plataformas

Microsoft

Sun microsystems

suse
A NOVELL BUSINESS

Lenguajes y desarrollos

python

Java

perl

Natural Language Toolkit

Herramientas de análisis

WordSmith
OXFORD
UNIVERSITY PRESS

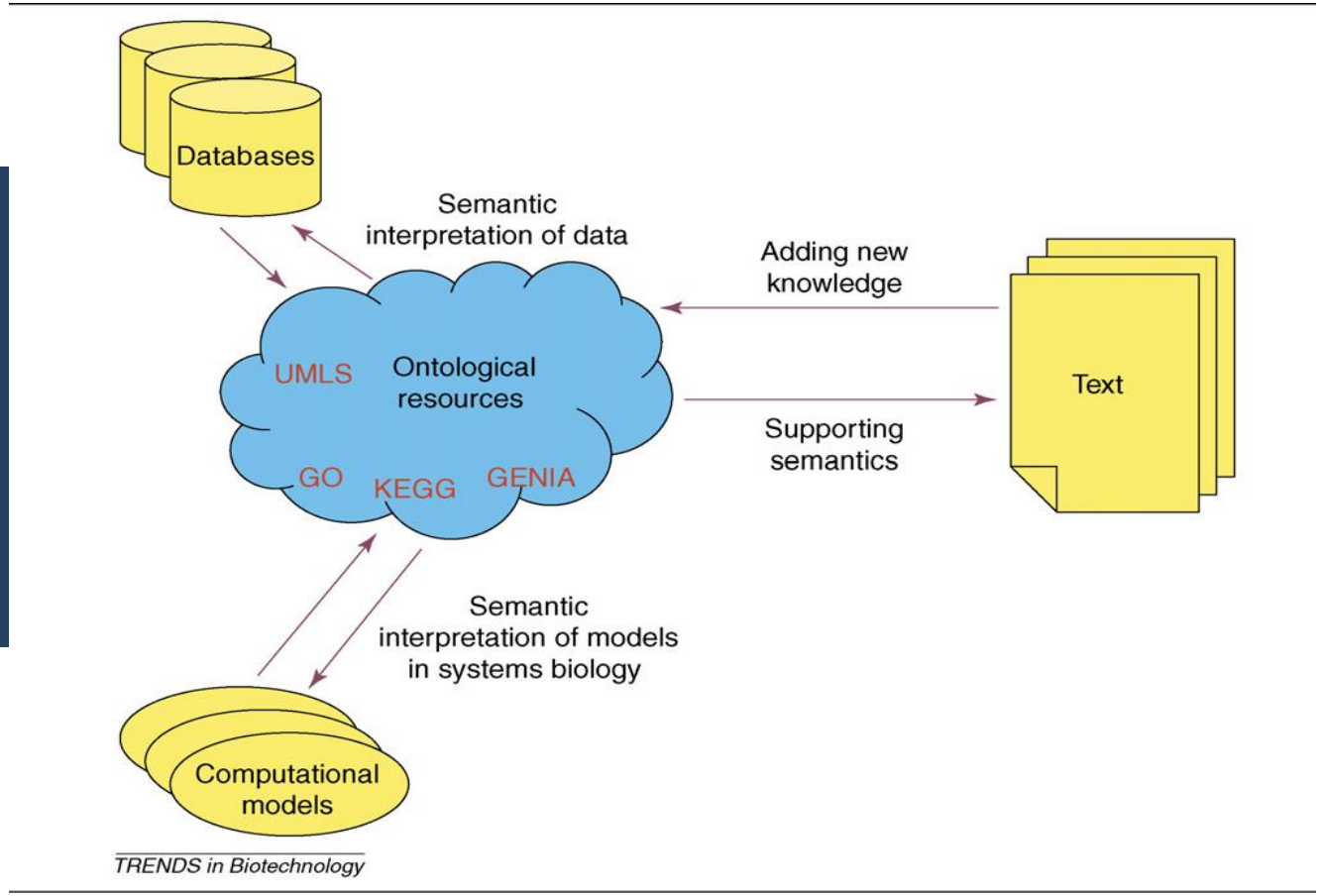
HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR

WordNet

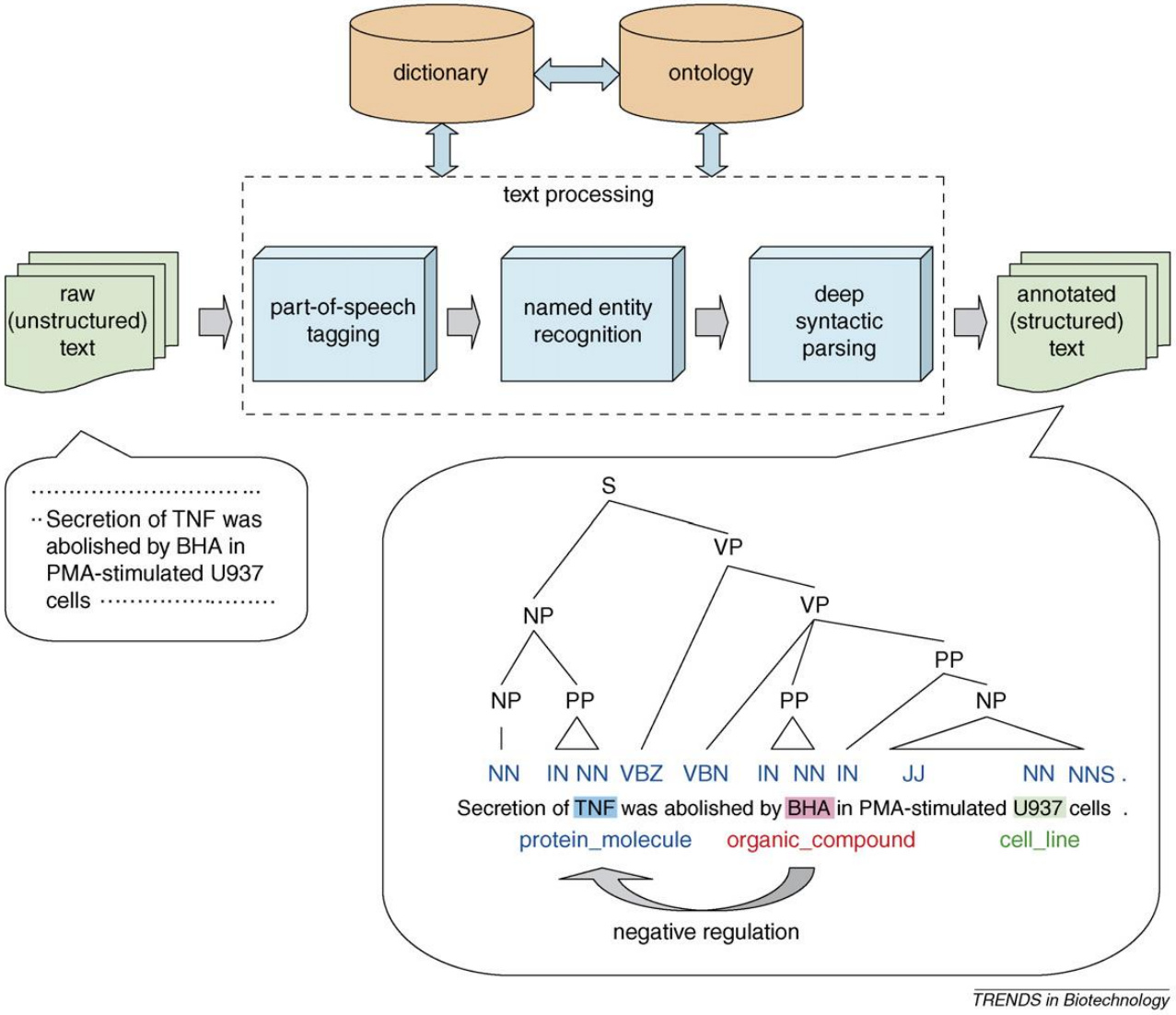
Propuesta de trabajo (1)

Siguiendo el ejemplo del NacTeM y su herramienta *Medie*, un tema de trabajo que podemos explorar durante el curso es el procesamiento de textos médicos, almacenados en repositorios electrónicos como MedLine.

Esta línea de investigación ha priorizado el análisis de documentos escritos en inglés, dada la gran producción que se tiene hoy en día a nivel mundial.



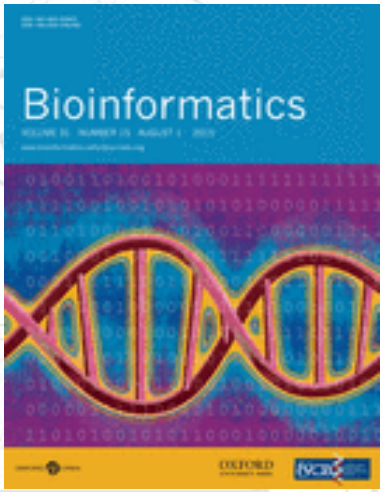
Propuesta de trabajo (2)



Estos trabajos siguen una secuencia de procesos que les permiten obtener información relevante de tales textos.

Propuesta de trabajo (3)

Básicamente, esta línea de investigación se ha orientado hacia tres áreas concretas:



- Creación de taxonomías y ontologías
- Minería de textos
- Detección de similitudes semánticas

Propuesta de trabajo (4)

En 2005 Paul Buitelaar, Philipp Cimiano y Bernardo Magnini desarrollaron el siguiente esquema, el cual resume una metodología para desarrollar ontologías (e igualmente taxonomías) basada en textos:

$\forall x, y$ (procesar (computadora,datos) $\rightarrow$ HACER (x,y))

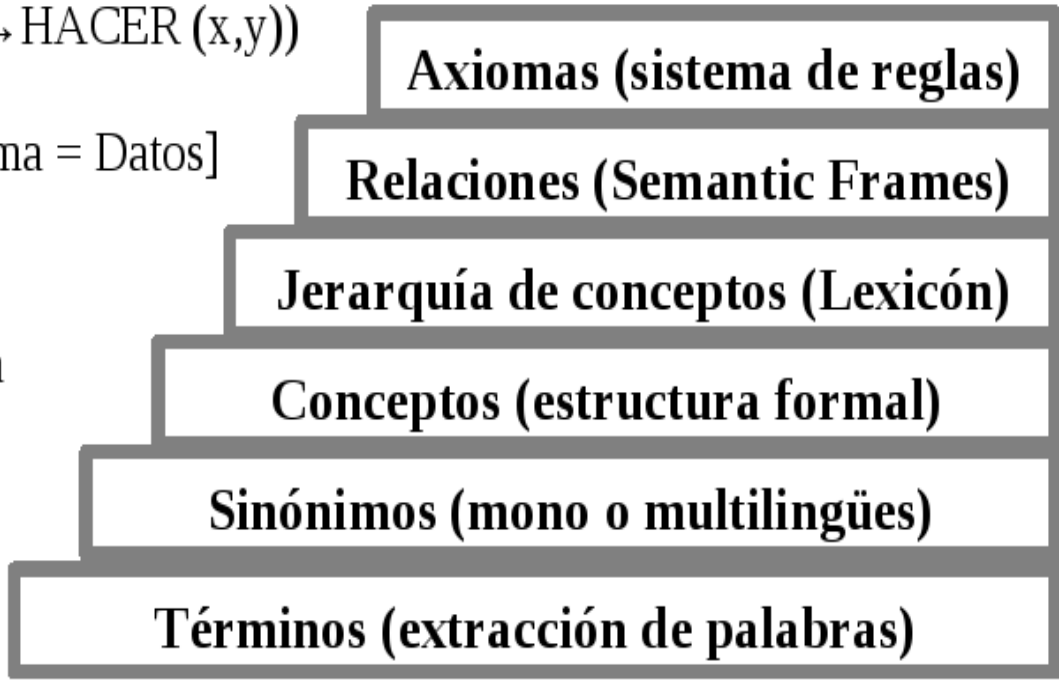
Procesar [Agente = Computadora; Tema = Datos]

Computadora *ES-UN* tipo de máquina

Computadora = "máquina que procesa datos"

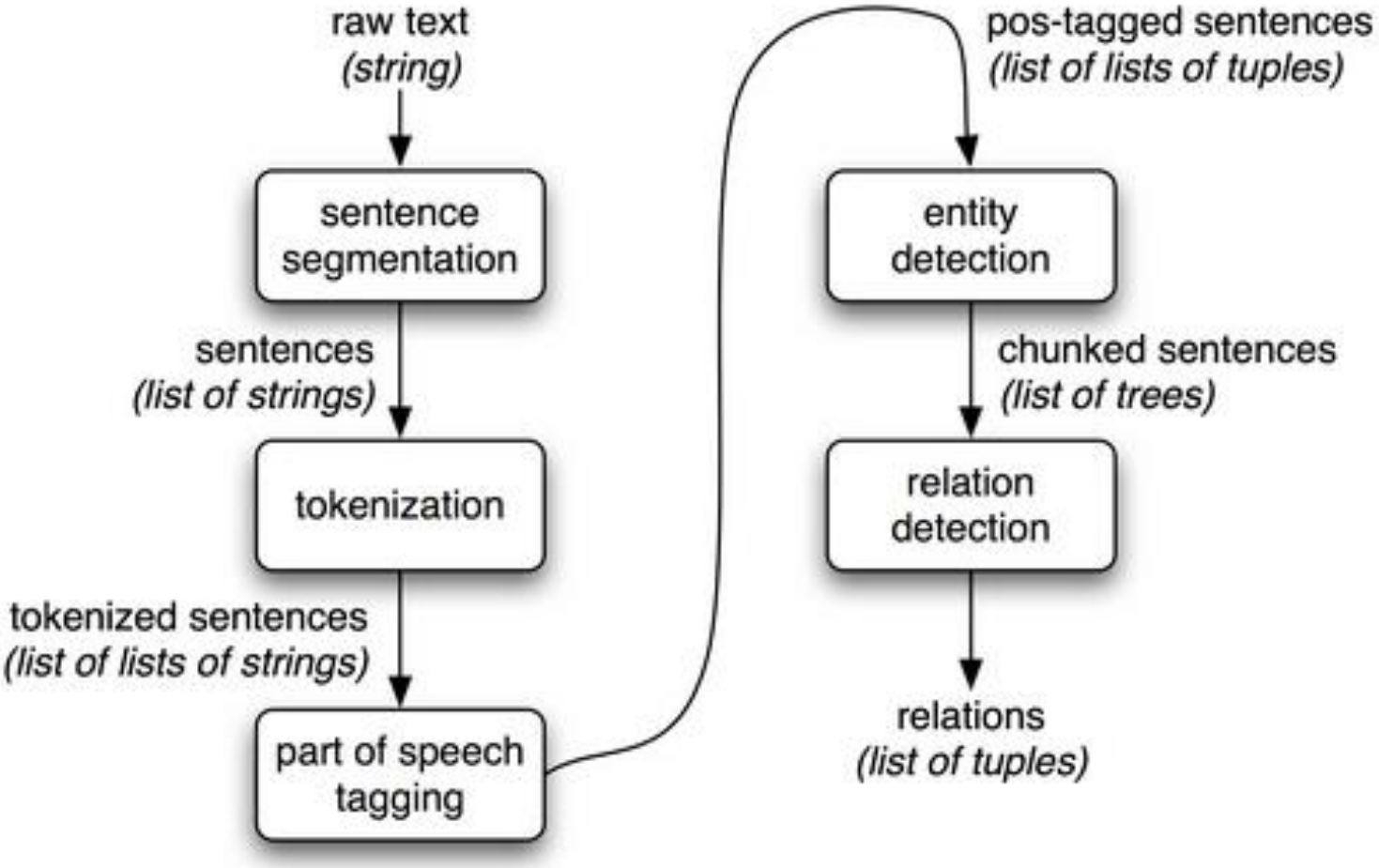
Ordenador = Computadora

Computadora, datos



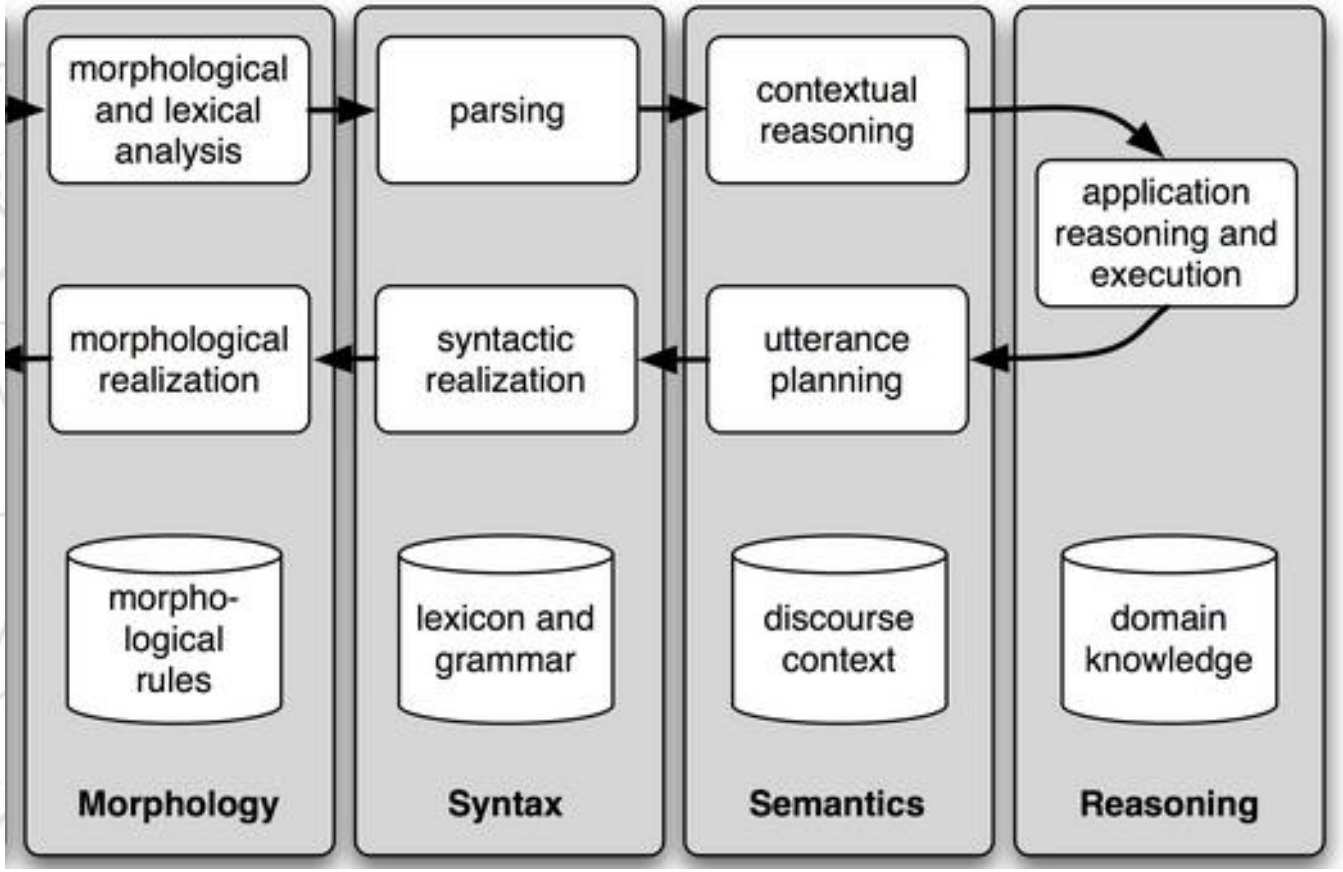
Propuesta de trabajo (5)

Partiendo de esta idea, podemos desarrollar una **cadena de procesamiento**, esto es, una serie de procesos organizados de manera modular, en donde el resultado de uno es el *input* de otro. Veamos:



Propuesta de trabajo (6)

Hacer esto, nos permite hacernos un modelo sobre cómo funciona el lenguaje natural, y particularmente también nos ayuda a entender cómo se configura un texto, tomando en cuenta las relaciones que se dan entre los niveles lingüísticos involucrados en tal configuración, esto es:



Punto de partida (1)

Dado que en PLN trabajamos con criterios computacionales y lingüísticos, vamos a establecer un punto de partida común:

Lenguas humanas $\in$ sistema de signos

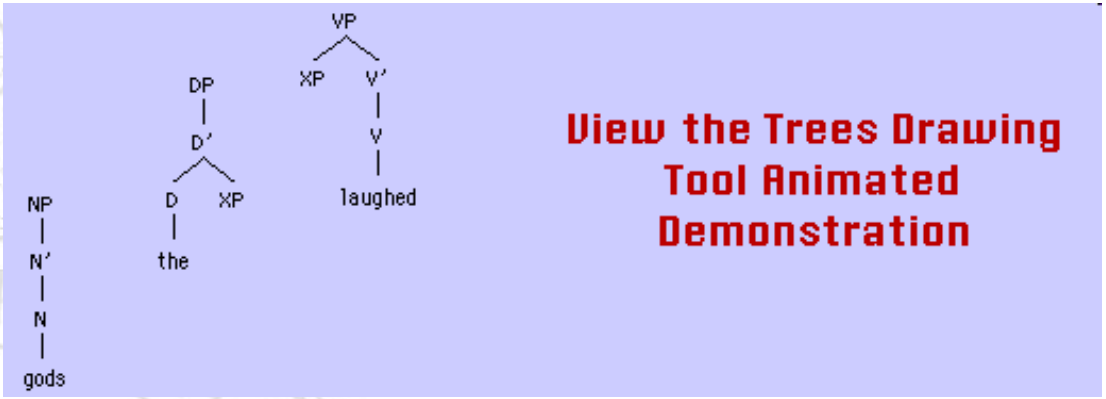
Permiten expresar información sobre el mundo y las personas

Se codifican siguiendo reglas y patrones regulares

Tal regularidad es el fruto de un uso rutinario por parte de sus hablantes. Dicha rutina, con el tiempo, da lugar a una gramática

Puntos en común (2)

¿Qué podemos hacer con una lengua viéndola como un sistema de signos?

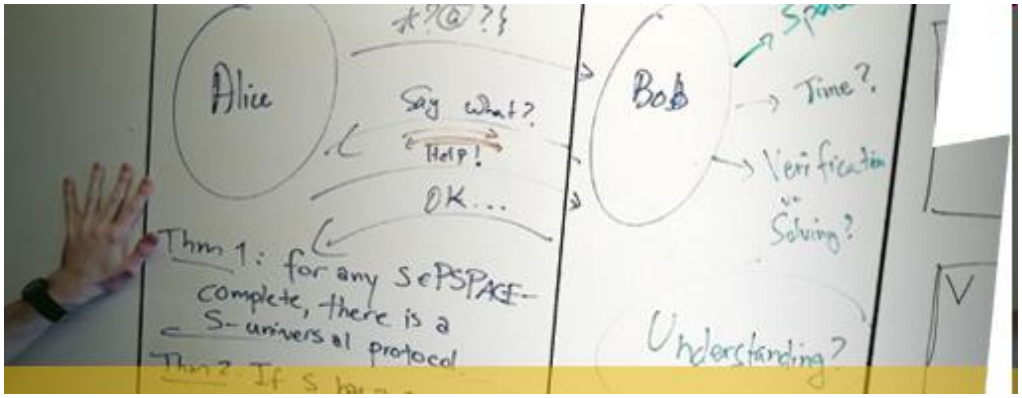


Hacer comparaciones entre distintas lenguas humanas

Deducir reglas y patrones

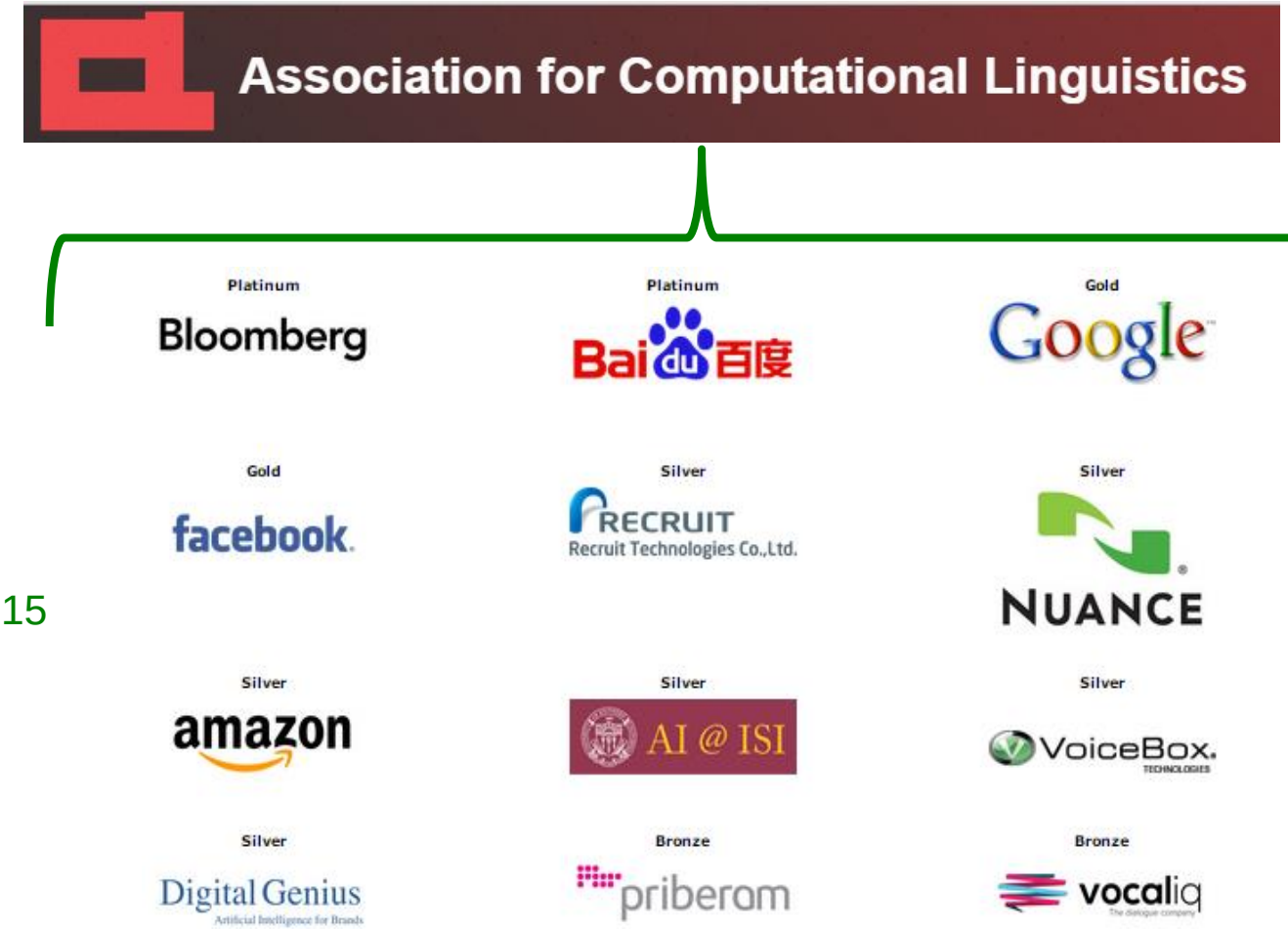
Establecer modelos gramaticales

Aprovechar esos modelos para desarrollar sistemas de cómputo capaces de generar y comprender una lengua



Punto de partida (2)

Una forma de ver las cosas, y que ha dejado mucho dinero a algunos, es utilizar el lenguaje natural para hacer búsquedas de información.

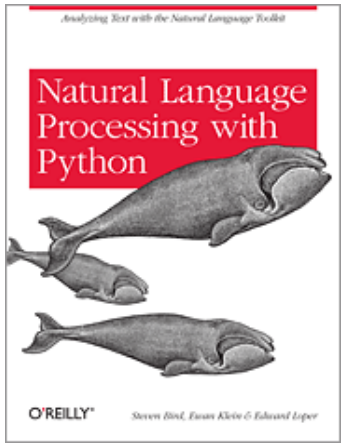


Patrocinadores de la conferencia Empirical Methods in Natural Language Processing 2015

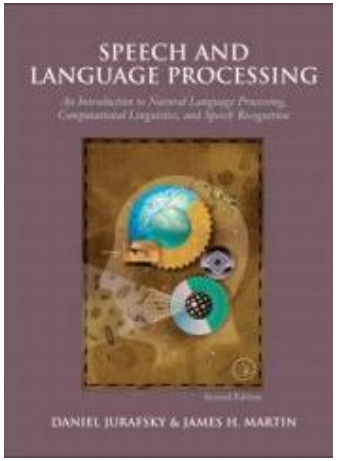
(www.emnlp2015.org)

Referencias de consulta (1)

Libros básicos:



1. Bird, Steven, Klein, Ewan y Loper, Edward (2010): *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*, Sebastropol, Cal., O'Reilly Media.



2. Jurafsky, Daniel y James H. Martin (2007): *Speech and Language Processing*, New York, Prentice-Hall.

Referencias de consulta (2)

Sitios WEB:

1. **Página oficial de Python: www.python.org.**
2. **Página oficial de NLTK: www.nltk.org.**
3. **Página académica de Dan Jurafsky: www.stanford.edu/~jurafsky, Stanford University.**
4. **Página académica de James Martin: www.cs.colorado.edu/~martin/csci5832/, University of Colorado.**
5. **Curso de procesamiento del lenguaje natural con Python: http://www.corpus.unam.mx/cursopln/, IINGEN-UNAM (México).**

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/curso-de-procesamiento-del-lenguaje-natural.html