

Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN – InCo

Introducción

¿Qué es el PLN?

Introducción

¿Qué es el PLN?

*El **Procesamiento de Lenguaje Natural (PLN)** es una subdisciplina de la Inteligencia Artificial que intenta resolver con **computadoras** tareas vinculadas al **lenguaje humano**, permitiendo la comunicación entre el humano y la computadora a través del lenguaje natural o resolviendo diferentes tareas que implican algún tipo de **procesamiento de texto o habla**. (Jurafsky & Martin, 2008)*

Introducción

– Recuperación de información



Busca en Google o escribe una URL 

– Traducción automática

receta de merengue italiano → 意大利蛋白酥皮食譜

– Respuestas a preguntas



– Análisis de sentimiento

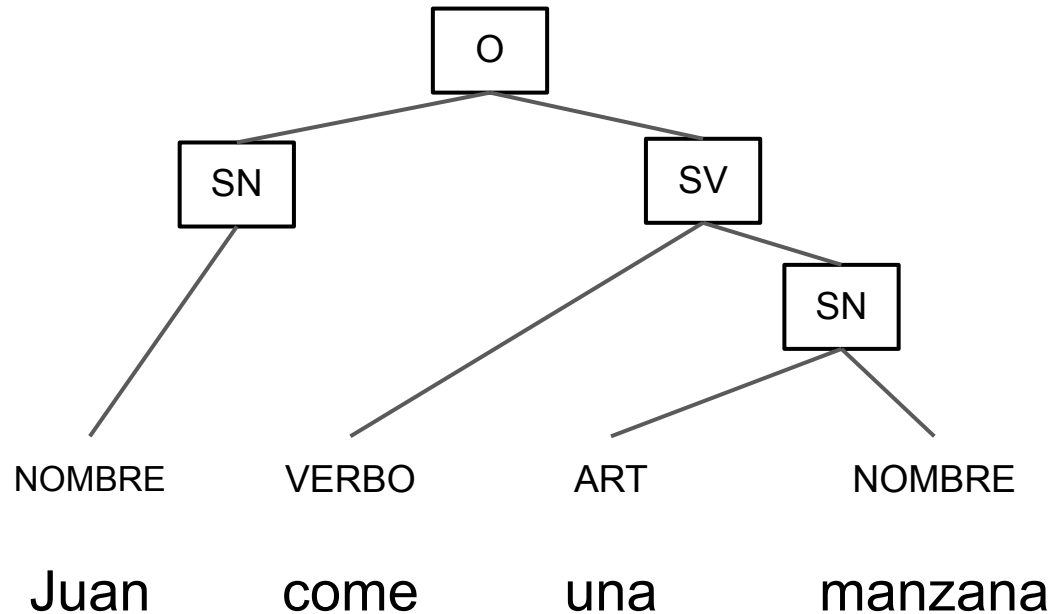


Introducción

– Parsing

– POS Tagging

– mucho más ...



Introducción

HAL 9000




HAL 9000

Habilidades de HAL

- ❖ comprensión de humanos vía:
 - reconocimiento del habla
 - comprensión de lenguaje natural

 - ❖ comunicación con humanos vía:
 - síntesis del habla
 - generación de lenguaje natural
-

HAL 9000

Señal sonora  Secuencia de palabras

Reconocer/Generar

- **Conocimientos de:**
 - **Fonética:** naturaleza física de los sonidos.
 - **Fonología:** cómo los sonidos funcionan en una lengua.
-

HAL 9000

- Debe saber, por ejemplo:
 - que los sustantivos tienen género y número:
 - Perr-o, Perr-o-s, Perr-a, Perr-a-s.
 - Pero:
 - Cas-a no es el femenino de Cas-o.
 - Ni Luz-s ni Luz-es son plurales de Luz.
 - que se pueden formar palabras agregando prefijos y sufijos a palabras existentes:
 - in-creíble (in- denota negación)
 - calmada-mente (-mente transforma adjetivo en adverbio)
 - Conocimientos de **Morfología**: estudio de la estructura interna de las palabras
-

HAL 9000

“Lo siento, Dave. Me temo que no puedo hacerlo.”

- Debe conocer el orden correcto en el que las palabras deben decirse para que la respuesta tenga sentido.
 - Por ejemplo: (*) *Lo puedo Dave siento que no temo me hacerlo.*
 - Sin embargo: (?) *Dave, lo siento. Que no puedo hacerlo, me temo.*
 - Conocimientos de **Sintaxis**: estudio de la estructuración (orden y agrupamiento) de las palabras en unidades mayores.
-

HAL 9000

- La sintaxis no es suficiente:
 - Abre las compuertas, HAL ($V + ART + N + SP + N$)
 - Baja las persianas, HAL
 - Saca los dados, HAL
 - Es necesario comprender el **significado** de lo que Dave está diciendo:
 - significado de cada palabra: **Semántica Léxica**
 - significado de la combinación de palabras para obtener significados mayores: **Semántica Composicional**
-

HAL 9000

- HAL retoma elementos mencionados anteriormente y va elaborando un discurso coherente y cohesivo:

Lo siento, Dave. Me temo que no puedo hacerlo.

- Significa: “no puedo abrir las compuertas”
 - Conocimientos de **Discurso**: estudio de las unidades mayores a la oración.
-

HAL 9000

- Adicionalmente, HAL presenta una forma de expresarse educada en el lenguaje utilizado:

Lo siento, Dave. Me temo que no puedo hacerlo.

- Pero en realidad: (1) no lo siente y
(2) puede abrir las compuertas
 - Conocimientos de **Pragmática**: estudio del lenguaje en un contexto, relaciones entre el significado y objetivos e intenciones de los hablantes.
-

Etapas clásicas en el Procesamiento de Lenguaje Natural

Etapas clásicas para el PLN

- ***Fonética y Fonología***: estudio de los sonidos lingüísticos (usados para la comunicación humana)
 - ***Morfología***: estudio de la estructura interna de las palabras
 - ***Sintaxis***: estudio de la estructuración (orden y agrupamiento) de las palabras en unidades mayores
 - ***Semántica***: estudio del significado
 - ***Discurso***: estudio de las unidades mayores a la oración
 - ***Pragmática***: estudio de cómo el lenguaje se utiliza para cumplir objetivos
-

Un poco de historia...



Década del '50

Warren Weaver



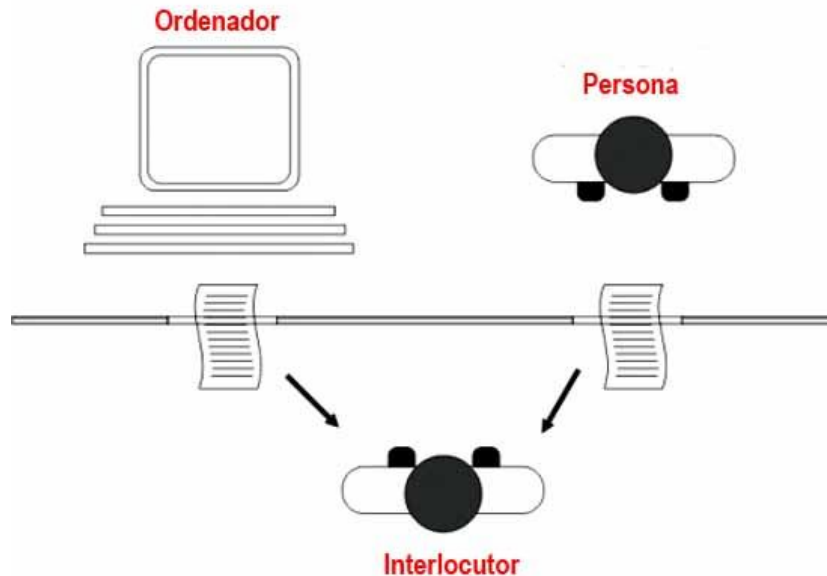
- Traducción Automática
 - En particular del Ruso al Inglés (60 oraciones)
 - Guerra Fría
 - Experimento *Georgetown* (1954)
 - tenía 6 reglas gramaticales y
 - 250 palabras
 - En tres años la traducción estará resuelta....
-

Década del '50

Alan Turing: "Computing Machinery and Intelligence"

(I propose to consider the question, "Can machines think?")

Test de Turing

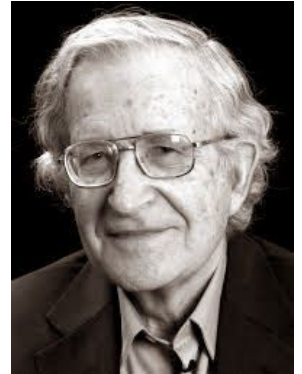


Década del '50

Noam Chomsky: "Syntactic Structures" (1957)

Colorless green ideas sleep furiously - 1955

(Las ideas verdes incoloras duermen furiosamente)



"probabilistic models give no insight into the basic problems of syntactic structure"

Gramática Universal → *todos los idiomas que usamos los seres humanos tienen unas características o principios comunes en su propia estructura*

Década del '60

- **Cocke - Kasami - Younger (1965)**
(*Parser*)



- **Earley (1968)**
(*Otro Parser*)



- **Key (1967)**
(*Otro Parser*)



El foco está puesto en el Análisis Sintáctico

Década del '70

Karen Spärck Jones



- Recuperación de información: obtener documentos más relevantes dada una consulta
 - Inverse Document Frequency (1972)
-

Década del '70

Richard Montague



“English as a Formal Language”

- Fue pionero en el enfoque lógico de la semántica del lenguaje natural
 - La gramática de Montague está basada en lógica formal:
 - alto orden
 - lambda cálculo
 - semántica intensional, mundos posiblesEj: "*la estrella matutina*" y "*la estrella vespertina*"
-

Década del '70

Alain Colmerauer: PROLOG



- Lenguaje de Programación basado en lógica
 - Pensado originalmente para procesamiento de Lenguaje Natural
-

Década del '80

- Se construyen sistemas de Laboratorio, con reglas hechas a mano
- Sistemas de interrogación de BDs relacionales
- Uno de los grandes problemas es la portabilidad

*Surgen los sistemas basados en
aprendizaje automático*

Década del '90

Frederik Jelinek



- Modelos IBM de traducción estadística y reconocimiento de VOZ
 - EL PLN se mueve hacia métodos basados en datos
-

Comienzo de este siglo...

Vladimir Vapnik

- Support Vector Machines
- AT&T, NEC, Facebook IA Research



2010s en adelante...

Geoffrey Hinton, Yoshua Bengio, Yann LeCun

(The Canadian Mafia)

- Deep Learning
- Word Embeddings



La investigación en PLN hoy

- Casi completamente data-driven.
 - Las compañías de Internet se llevan a los grandes investigadores (casi todos).
 - Se publica principalmente en conferencias (con frecuencia primero en Arxiv), se publica poco en journals (avance rápido del área vs procesos lentos de publicación).
 - Las publicaciones principales del área son Open Source ([ACL Anthology](#)).
 - Muchos recursos, herramientas, datasets disponibles.
 - Las redes neuronales son las estrellas.
 - Hoy: Large Language Models (LLM)
-

Aplicaciones



Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2013) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted World Player of South Africa. The amount handled by Inter is far from the 36 million euros for the striker's release clause. But the club president Massimo Moratti propose an improvement in the player's salary, who according to the Italian daily receive about 4 million euros until 2013.

(2018) The Italian championship has not yet begun but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan forward who was chosen as the best player in the World Cup in South Africa. The amount handled by Inter is far from the 36 million euros of the rescission clause of the scorer. But the club chaired Massimo Moratti propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.

(2022) The Italian championship has not yet started but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. Inter's figure is very far from the 36 million euros of the striker's termination clause. But the club chaired by Massimo Moratti will propose an improvement in the player's salary, who according to the Italian newspaper will receive around 4 million euros until 2013.

Generación de texto

Prompt: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Generated: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic." (...)

Generación de texto

Prompt: En un hallazgo impactante, un científico descubrió una manada de unicornios que vivían en un valle remoto, previamente inexplorado, en las montañas de los Andes. Aún más sorprendente para los investigadores fue el hecho de que los unicornios hablaran un inglés perfecto.

Generado: El científico nombró a la población, por su cuerno distintivo, Unicornio de Ovidio. Estos unicornios de cuatro cuernos, de color blanco plateado, eran previamente desconocidos para la ciencia.

Ahora, después de casi dos siglos, el misterio de qué provocó este extraño fenómeno finalmente se resuelve.

El Dr. Jorge Pérez, biólogo evolutivo de la Universidad de La Paz, y varios compañeros, exploraban la Cordillera de los Andes cuando encontraron un pequeño valle, sin otros animales ni humanos. Pérez notó que el valle tenía lo que parecía ser una fuente natural, rodeada por dos picos de roca y nieve plateada.

Pérez y los demás luego se aventuraron más adentro del valle. “Cuando llegamos a la cima de un pico, el agua se veía azul, con algunos cristales en la parte superior”, dijo Pérez.

Pérez y sus amigos quedaron atónitos al ver la manada de unicornios. Estas criaturas se podían ver desde el aire sin tener que moverse demasiado para verlas, estaban tan cerca que podían tocar sus cuernos.

Mientras examinaban a estas extrañas criaturas, los científicos descubrieron que las criaturas también hablaban un inglés bastante regular. Pérez afirmó: “Podemos ver, por ejemplo, que tienen un ‘lenguaje’ común, algo así como un dialecto o dialéctica”.

Generación de imágenes

Prompt: Godzilla eating Facultad de Ingeniería, Uruguay.



¿Qué tiene el lenguaje natural que no tienen los lenguajes formales?

Ambigüedad

Ambigüedad



Fuentes de ambigüedad

- Ambiguo: que admite distintas interpretaciones
-

Ambigüedad fonética

Ejemplos de calambures:

- *Ató dos palos. / A todos palos.*
 - *Yo loco, loco, y ella loquita. / Yo lo coloco y ella lo quita.*
 - *El dulce lamentar de los pastores. / El dulce lamen tarde los pastores. (Garcilaso de la Vega)*
 - *Entre el clavel y la rosa, su majestad escoja. (Quevedo)*
-

Ambigüedad a nivel morfológico

Nosotros *plantamos* papas.

¿El verbo plantar está conjugado en pasado o en presente?

Ambigüedad sintáctica

Pedro vio a Juan con el telescopio.

- a) Pedro vio [a Juan] con el telescopio.
- b) Pedro vio [a Juan con el telescopio].

Los hombres y las mujeres que hayan cumplido 60 años pueden solicitar una pensión.

- a) [Los hombres y las mujeres que hayan cumplido 60 años] pueden solicitar una pensión.
 - b) [Los hombres] y [las mujeres que hayan cumplido 60 años] pueden solicitar una pensión.
-

Ambigüedad semántica

Homonimia: dos palabras con misma forma que tienen distintos significados (normalmente, distinta etimología).

- Homografía: *capital, banco, vino*
- Homofonía: *Ola/Hola, As/Has, Cocer/Coser*

Polisemia: una palabra con múltiples significados, que de alguna manera están relacionados, normalmente con un origen común.

*El hombre **desciende** del mono y el mono **desciende** del árbol.
Plantar un **árbol** // Recorrida DFS de un **árbol***

Ambigüedad semántica

La perra de mi vecina me ladró.

- a) Mi vecina realmente tiene una perra.
 - b) No tengo un buen trato con mi vecina.
-

Ambigüedad semántica

Cuantificadores

Todos los hombres aman a una mujer.

Todos los estudiantes leyeron un libro.

- a) Nos referimos a una misma mujer/un mismo libro.
 - b) Para cada hombre/estudiante nos referimos a una mujer/un libro diferente.
-

Ambigüedad a nivel pragmático

Llego a las ocho. Esperame.

- ¿A qué hora llegarás?
 - Llego a las ocho. Esperame. (**Previsión**)

 - Nunca llegarás en hora.
 - Llego a las ocho. Esperame (**Promesa**)

 - Eso me lo vas a tener que decir cara a cara.
 - Llego a las ocho. Esperame. (**Amenaza**)
-

Ambigüedad a nivel de discurso

Tomé el alfajor del escritorio y lo comí.

- a) Tomé el alfajor que estaba en el escritorio y comí el alfajor.
 - b) Tomé el alfajor que estaba en el escritorio y comí el escritorio.
-

¿Se puede resolver la ambigüedad?

Juan mató al carpincho con la escopeta.

- No puede ser el carpincho quien lleve la escopeta.

Puse la camisa en la lavadora y la lavé.

- Las lavadoras lavan. La ropa se lava.

Se requiere conocimiento del mundo

Modelos

- **Máquinas de estado finito:** autómatas finitos, transductores, autómatas con peso...
 - **Sistemas de reglas:** gramáticas regulares, expresiones regulares, gramáticas libres de contexto, gramáticas con atributos...
 - **Lógica:** cálculo de predicados.
 - **Teoría Probabilística.**
 - **Modelos basados en Aprendizaje Automático, en particular, Redes Neuronales.**
-

Algoritmos

- Métodos simbólicos
 - Programación dinámica
 - Aprendizaje automático
 - Redes Neuronales
 - Grandes Modelos de Lenguaje (LLM)
-

Referencias

- D.Jurafsky & J.Martin. Speech and Language Processing. Second edition. Stanford. 2008.
-