Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN - InCo

2016

Introducción al Procesamiento de Lenguaje Natural

Introducción

- ¿Qué es el PLN?
  - Conjunto de métodos y técnicas eficientes desde un punto de vista computacional para la **comprensión** y **generación** de lenguaje natural.
  - Subdisciplina de la Inteligencia Artificial.

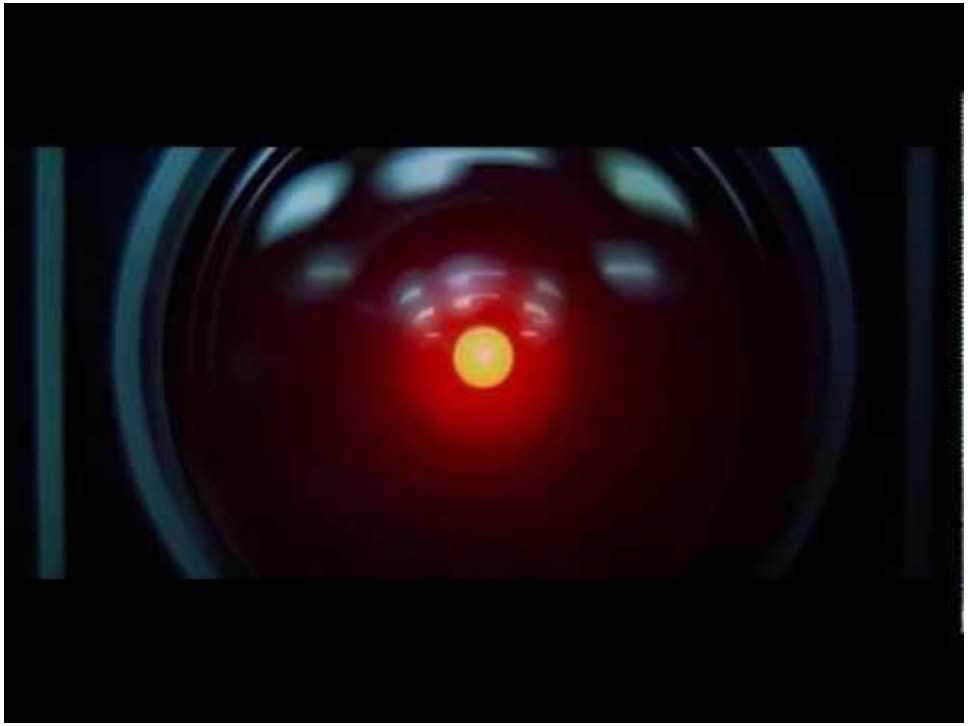
Introducción

- ¿PLN = Lingüística Computacional?

- Lingüística Computacional:
  - campo multidisciplinario de la **lingüística** y de la **computación**.
  - estudio científico del lenguaje con el fin de elaborar modelos de éste o de ciertos fenómenos específicos.
  - involucra a lingüistas, informáticos, lógicos, psicólogos cognitivos...

El PLN puede verse como la rama ingenieril de la LC

HAL 9000

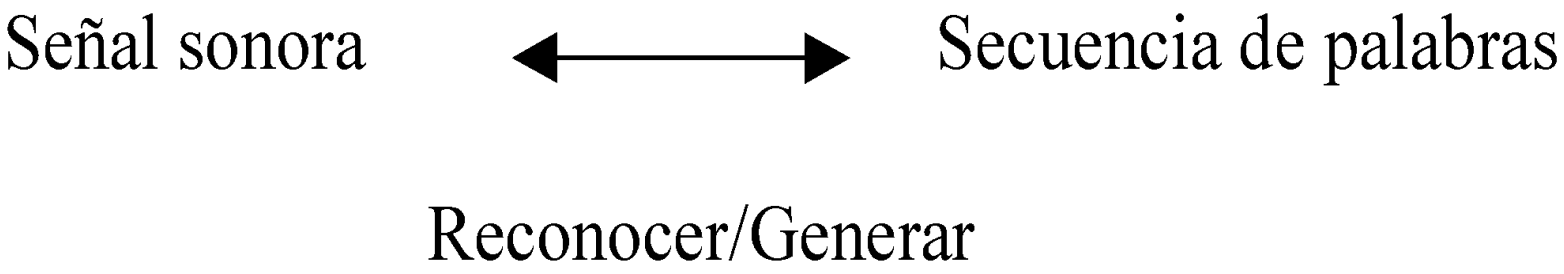


HAL 9000

Habilidades de HAL

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

HAL 9000



- 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

- 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. (*VC + ART + SUST + SP + SUST*)
  - 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

- Adicionalmente, HAL presenta una utilización educada del lenguaje: **Lo siento**, Dave. **Me temo** que **no puedo** hacerlo.
- **Significa**, en realidad: (1) no lo siente y (2) puede abrir las compuertas
- Conocimientos de:
  - **Pragmática**: estudio del modo en el que el contexto influye en la interpretación del significado. Cómo el lenguaje se utiliza para ciertos fines.
  - **Discurso**: estudio de las unidades mayores a la oración.

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.
- ***Pragmática***: estudio de cómo el lenguaje se utiliza para cumplir objetivos.
- ***Discurso***: estudio de las unidades mayores a la oración.

Un poco de historia...

Comienzos del PLN

Traducción Automática

- Interés hacia fines años 40' y años 50'
- En particular del Ruso al Inglés (Guerra Fría).

(Original) *"The spirit is willing, but the flesh is weak."* (El espíritu es fuerte pero la carne es débil)

(Doble traducción) *"The vodka is strong, but the meat is rotten."* (El vodka está bueno pero la carne está podrida)

¿Qué pasaría hoy con esta frase?

Nombres para recordar

- Alan Turing

Nombres para recordar

- Alan Turing
- Noam Chomsky

Nombres para recordar

- Alan Turing (1912-1954)
- Noam Chomsky (1928 -)
- Frederick Jelinek (1932 - 2010)
- Hinton, Bengio, LeCunn (the Canadian Mafia)
- Vladimir Vapnik (1936 -)

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.

(2010) The Italian league has not yet begun, but Inter Milan and Juventus, two of the most powerful clubs Calcium, and are playing a duel to stay with Diego Forlan, the Uruguayan forward who was chosen as the best player in the World Cup. The figure is far Inter manages the 36 million euros of the termination clause in the striker. But the club president Massimo Moratti will propose an improvement in the salary of the player, who the Italian daily receive about 4 million euros until 2013.

(2011) The Italian league has not yet begun, but Inter Milan and Juventus, two of the most powerful clubs of calcium, are already playing a match to stay with DiegoForlan, the Uruguayan striker who was voted the best player in the World Cup. The figure is far Inter manages the 36 million euro buyout clause in the striker.But the club president Massimo Moratti will propose an improvement in theplayer's salary, who the Italian daily receive about 4 million euros until 2013.

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 Calcium, and are playing a duel to stay with Diego Forlan, the Uruguayan striker was voted World Player of Sudáfrica.La Inter manages figure 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.

(2014) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the EPL, and are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted the best player in the World Sudáfrica.La Inter manages figure is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, according to the Italian newspaper who will receive about 4 million euros until 2013.

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.

(2015) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted best player of the World Sudáfrica.La Inter figure handles is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, who the Italian daily receive about 4 million euros until 2013.

(2016) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who ***was chosen as the best player*** *in the World Sudáfrica.****La Inter manages figure*** *is far from the 36 million euros of the* ***termination clause scorer****. But the club president Massimo Moratti propose an improvement in the player's salary,* ***according to the Italian daily who will receive*** *about 4 million euros until 2013.*

¿Vale la pena? (*)

- Ejercicio: interprete el siguiente texto en chino mandarín simplificado:

在加纳村惨剧后，暂停对黎南空袭48小时的以色列军队在8月1日恢复空袭，以色列内阁也通过决议扩大以军在黎巴嫩南部的地面攻势。同时，以色列开始大规模征召预备役人员。这一切表明，黎巴嫩南部的战火和硝烟在短期内难以平息。

(Traducción de Google) Ghana tragedy in the village, 48-hour suspension of air strikes against Lina in the Israeli army resumed air strikes on August 1, the Israeli cabinet passed a resolution to expand Israeli ground offensive in southern Lebanon. At the same time, Israel began a large-scale recruitment of reservists. All this shows that the fighting in southern Lebanon and smoke in the short term it is difficult to quell.

(*) Este título queda por razones históricas

Resumen automático

- Idea central: "condensación del contenido de la información de un documento para el beneficio de un lector" (Mani 2001).

- Primeros trabajos de Luhn (1958) y Edmunson (1960):
  - Basados en métodos estadísticos.
  - Extraen las oraciones más importantes.
  - Frecuencia de términos. Peso de oraciones.

- Los trabajos en el área resurgen a fines de los años 90'

Extracción de Información

- Restaurante Español cerca de Manchester en Inglaterra, busca camareros o camareras de salad con conocimiento de cockteleria y barra, deben de saber flambear y tener un mínimo de tres años de experiencia con un manejo de Ingles a nivel medio, conocimientos de vinos Españoles y resto del mundo una ventaja. Salario minimo 1500 euros mes con propinas. Cinco dias por semanas de unas 50/55 horas.

Industria: Restauración.

Puesto: Camarero/a.

Lugar: Manchester, Inglaterra.

Compañía: ? / Restaurante Español

Salario: 1500 euros/mes.

Dedicación: 50/55 hs. Semanales.

Interfaces a BD

- **Usuario**: Necesito un tren nocturno de París a Viena que llegue alrededor de las 10 de la mañana.
- **Sistema**: ¿Qué día desea viajar?
- **Usuario**: Mañana.
- **Sistema**: Los trenes disponibles son…

- Análisis de la entrada y "traducción" a una consulta.
  - P.ej: $\exists x(\text{tren}(x) \wedge \text{nocturno}(x) \wedge \text{recorrido}(x, \text{París}, \text{Viena}) \wedge \exists y \exists z(\text{horario}(x, y, z) \wedge \text{alrededor}(z, 10)))$
- El enfoque funciona bien con léxico y sintaxis restringidos.

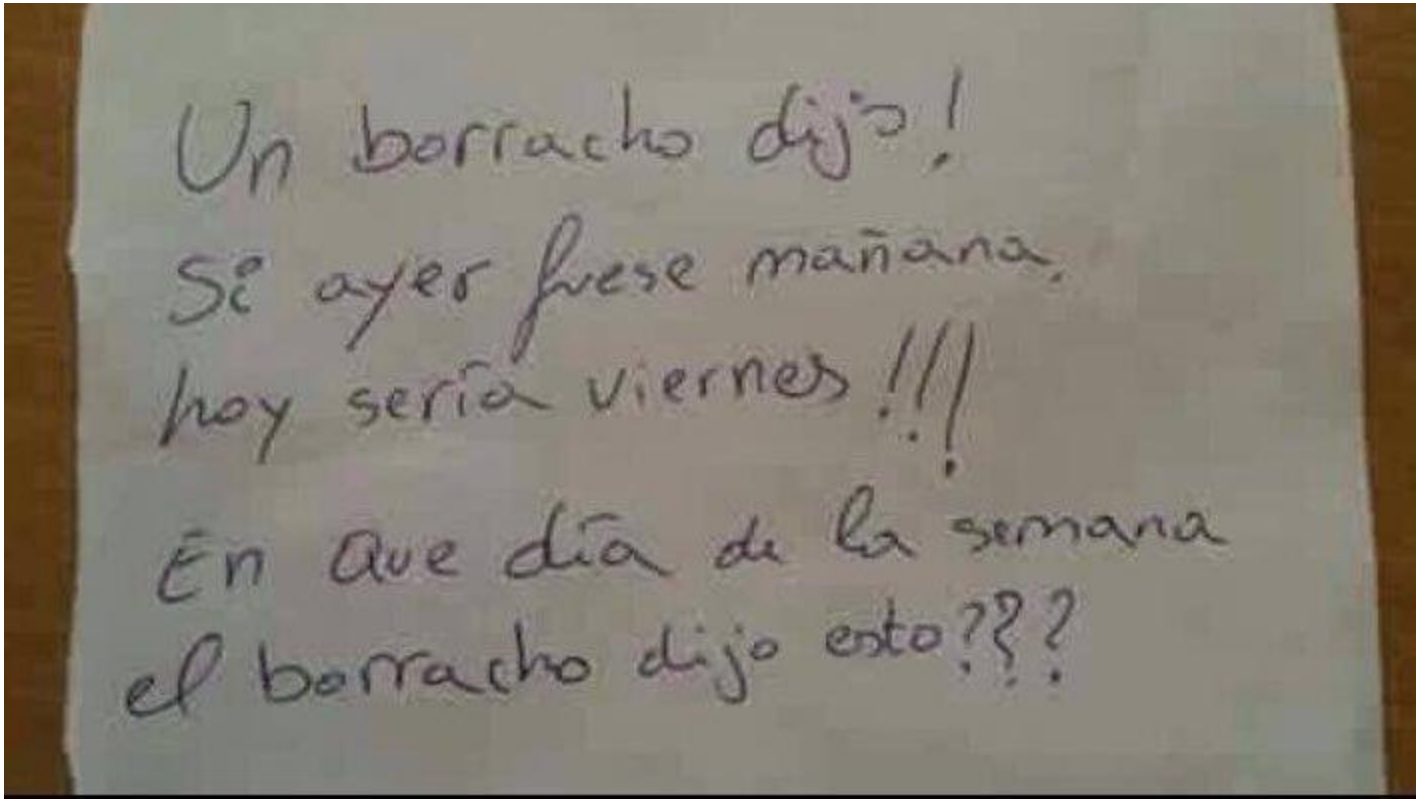
Más aplicaciones

- Recuperación de información.
- Verificadores de gramática y estilo.
- Categorización de documentos.
- Respuesta a preguntas.
- Proyectos Grupo PLN

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


Padre, he mentido
Te escucho, hijo
Dije que tenía 33 pal Envido.
Y tenía 24

Ambigüedad sintáctica


Un borracho dijo!
Si ayer fuese mañana,
hoy sería viernes!!!
En que día de la semana
el borracho dijo esto???

Ambigüedad

Fuentes de ambigüedad

- Ambiguo: que admite distintas interpretaciones.
- Homonimia: dos palabras con misma forma que tienen distintos significados.
  - Homografía: capital, banco,
  - Homofonía: Ola/Hola, As/Has, Cocer/Coser.
- Polisemia: una palabra con múltiples significados.
  - El hombre **desciende** del mono y el mono **desciende** del árbol.

Ambigüedad fonética

Ejemplos de calambures:

- **¡Qué beya plebella! (Les Luthiers)**
- 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)
- "*Now is the winter of our discontent made glorious summer by this* ***son*** *of York*“ (Shakespeare – Richard III)

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

La perra de mi vecina me ladró.

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 a nivel pragmático

- Llego a las ocho. Esperame.

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

- Nunca llegás 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.

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 y Algoritmos

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 **Redes Neuronales**
- **Representation Learning**

Algoritmos

- Búsquedas en espacios de estados:
  - buscar en un espacio de posibles secuencias fonológicas la correcta para una entrada dada
  - buscar en un espacio de árboles de análisis sintáctico el correcto para una entrada dada
- Programación dinámica:
  - convertir un autómata finito en una expresión regular equivalente
- Aprendizaje automático