Procesamiento de Lenguaje Natural

Logros

Desafíos

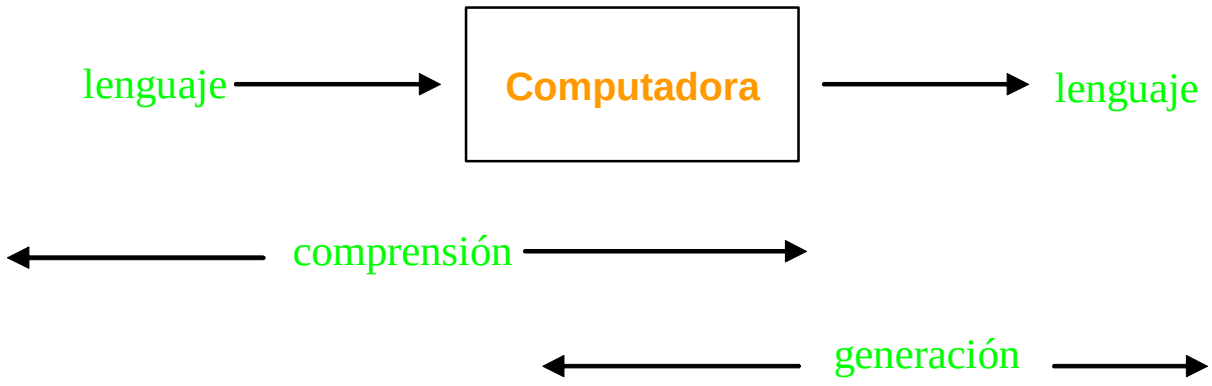
Impacto

Temario

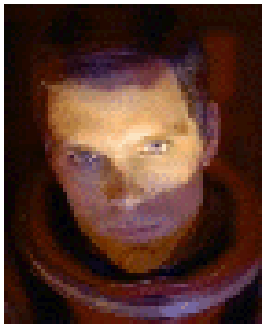
- ¿Qué es el PLN?
- 6 niveles de procesamiento.
- Un poco de historia, éxitos y desafíos.
- Proyectos del grupo PLN del InCo.

Temario

- ¿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 IA.



HAL - 2001, Odisea del Espacio 1967



- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry Dave. I'm affraid I can't do that.*

- Dave: Abre las compuertas, HAL.
- HAL: Lo siento, Dave. Me temo que no puedo hacerlo.

HAL - 2001, Odisea del Espacio

Habilidades de HAL (1967)

- comprensión de humanos vía:
  - reconocimiento del habla
  - comprensión de lenguaje natural
- comunicación con humanos vía:
  - generación de lenguaje natural
  - síntesis del habla
- pero también:
  - capacidades gráficas
  - juega al ajedrez
  - percepción visual

Habilidades de HAL

Señal sonora  Secuencia de palabras

Reconocer/Generar

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

Habilidades de HAL

- 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.

Habilidades de HAL

- 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.

Habilidades de HAL

- La sintaxis no es suficiente:
  - Abre las compuertas, HAL. (*Estructura: VC + ART + SUST + SP + SUST*)
  - Baja las persianas, HAL.
  - Saca los dados, HAL.
  - Suelta los perros, 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**.

Habilidades de HAL

- 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
- HAL podría haber respondido:
  - No.
  - De ninguna manera.
- 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.

6 niveles de procesamiento

- ***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.

Ambigüedad: el mayor problema en PLN

Fuentes de ambigüedad

- Ambiguo: que admite distintas interpretaciones.

- Homonimia: dos palabras con misma forma que tienen distintos significados

  (distinta etimología, distintas entradas en el diccionario).
  - Homografía: *vino (bebida) / vino (llegó*)
  - Homofonía: o*la / hola, as / has / haz, cocer / coser*.

- Polisemia: una palabra con múltiples significados

  (una entrada en el diccionario con distintos significados).
  - *El hombre **desciende** del mono y el mono **desciende** del árbol.*
  - *banco, capital*

Ambigüedad en los niveles de análisis

Ambigüedad a nivel fonético

Homofonía

- ola / hola
- as / has / haz

Segmentación

- Ató dos palos. / A todos, palos.
- Entre el clavel y la rosa, su majestad escoja. (Quevedo)

Ambigüedad en los niveles de análisis

Ambigüedad a nivel morfológico

Nosotros plantamos papas.

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

Ambigüedad en los niveles de análisis

Ambigüedad a nivel sintáctico

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 en los niveles de análisis

Ambigüedad a nivel semántico

Todos los hombres aman a una mujer.

Todos los estudiantes leyeron un libro.

a) Es la misma *mujer/libro* para todos

b) Para cada *hombre/estudiante* existe una *mujer/libro*

Ambigüedad en los niveles de análisis

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 en los niveles de análisis

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.

El procesamiento de lenguaje es difícil porque:

- Alta ambigüedad en todos los niveles.
- Complejo y sutil.
- Involucra razonar acerca del mundo.
- Se debe considerar la inserción en un sistema social de gente que interactúa:
  - exponiendo, convenciendo, ordenando, insultando, …
  - cambiando a lo largo del tiempo

Un poco de historia…

Breve historia : 50s, 60s

Primeras aplicaciones en computadoras menos poderosas que una calculadora

- Traducción Automática del Ruso al Inglés (Guerra Fría).
  - Famosa leyenda urbana:
    - (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 es muy mala)

- Trabajo fundacional en Autómatas, Lenguajes Formales, Probabilidades y Teoría de la Información

Breve historia : 70s, 80s

- Primer sistema de comprensión completa en un dominio limitado (Winograd, SHRDLU, 1971)
  - ¿La pirámide verde está sobre el cubo rojo?
- Separación de procesamiento (parsers) y descripción del conocimiento lingüístico
- Explicitación de nivel de representación semántica
- Se percibe necesidad de utilizar conocimiento sobre el mundo (proyecto CYC, Lenat)
- Traducción automática en dominios limitados (meteorología)

Breve historia : 90s

- Métodos de estado finito: gran eficiencia
  - Karttunen, Kaplan & Kay, FST
- La disponibilidad de grandes cantidades de texto (Web) reorienta el área
- Primeros resultados robustos con métodos probabilísticos
- Utilización de aprendizaje automático

Breve historia : 2000s

- Enfásis en semántica y representación del conocimiento
- Enfásis en discurso y diálogo
- Integración de técnicas simbólicas y probabilísticas
- Mayor integración de componentes LN en otros sistemas

- Pero también : proliferación de aplicaciones “guiadas por patrones”, sin análisis profundo

Algunas aplicaciones

Traducción Automática

- Actualmente
  - Original: *el día que las vacas vuelen*
  - Doble Traducción (español-> inglés -> español) con Google
    - *el día que las vacas* ***lo*** *vuel****a****n* (2008?)
    - ***las vacas día volar*** (2009)
    - *el día que las vacas vuelen* (2012)

    (traducción intermedia: *the day the cows come home* -> frase hecha)
- Tasa de error entre 20% y 30%

Traducción Automática

- Cuestionamiento: con tasas de error tan elevadas, ¿es realmente útil la traducción automática?
- 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.

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

Texto Original

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

Ficha

Industria: Restauración.

Puesto: Camarero/a.

Lugar: Manchester, Inglaterra.

Compañía: Restaurante Español

Salario: 1500 euros/mes.

Dedicación: 50/55 hs. Semanales.

Extracción de Información

- Objetivo: mapear una colección de documentos a una base de datos estructurados.

- Motivaciones:

  – Permitir búsquedas complejas: quiero trabajos en restauración en Manchester que paguen por lo menos 1200 euros al mes.
  – Permitir consultas estadísticas: ¿el número de trabajos en restauración creció en los últimos cinco años?

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.
- ...

Grupo PLN – InCo - UDELAR

Algunos proyectos

- Análisis sintáctico
  - Segmentación de oraciones en proposiciones
  - Desambiguación de comas
- Reconocimiento de eventos
  - ¿Cuáles son los eventos a los que se hace referencia en un texto?
  - ¿Ocurrieron efectivamente?
- Análisis temporal de textos

  Ubicación temporal y ordenamiento de los eventos mencionados.
- Opiniones

  ¿Quién opinó sobre el tema X? ¿Qué dijo? ¿Opinó a favor o en contra?
- BIO-NLP (Proyecto Microbio)

Algunas herramientas y recursos

- FreeLing (etiquetador morfo-sintáctico, distribución libre Universitat Politècnica de Catalunya)
- Clatex (segmentador en proposiciones, PLN-InCo)
- Editor de reglas contextuales (PLN-InCo)
- Lavinia (ambiente web para procesamiento de textos, PLN-InCo)
- Anotadores de textos (Clark, Knowtator-Protégé, MMAX2)
- NLTK (kit de herramientas de PLN para Python)
- Spanish WordNet (Universitat Politècnica de Catalunya, licencia gratuita para usos académicos)
- Corpus (Corin (Lingüística), CREA, “Corpus del Español”, Temantex (PLN-InCo), Opiniones (PLN-InCo))