Temas de Investigación

Héctor Jiménez Salazar
Tecnologías de la Información, DCCD, UAM-C

1. Motivación
2. Procesamiento del lenguaje natural
3. Desambiguación
4. Relaciones semánticas

Papel: lujo no sustentable

1. Según UNESCO son necesarios 786 millones de árboles para generar el papel que consume el mundo en un año[a] (Un árbol viene a producir unas 80,500 hojas).
2. Cada habitante de la tierra consume 1,510 hojas al año (en EEUU 11,916 hojas por hab./año. Europa 7,280).
3. La mitad de todo ese volumen de papel se emplea en impresoras y fotocopiadoras de oficina.

[a]http://cibernauta.elcorreodigital.com, Peter Lyman, Hal Varian. U. de Berkeley.

Digitalización progresiva

Información	Estructurada	Bases de Datos
	No estructurada	Texto, imágenes...

Algunos datos sobre internet:

- 167 terabytes anuales de texto ($\approx$ 167 millones de libros)[a]
- 600 millones de personas tienen acceso (2003)[a].
- Crecimiento de acceso a internet en México: 14.5% (INEGI, 2009).
- Las TIC's crecen a un ritmo anual del 30% en el mundo (2010)[b].

[a] `cibernauta.elcorreodigital.com`, Peter Lyman et alter,
[b] tendencias21.net.

Conocimiento implícito en los textos

El conocimiento permite construir sistemas ambiciosos. La web no contiene conocimiento, sólo información. Por ello, el Procesamiento del Lenguaje Natural (PLN) se ha revitalizado.

Implicaturas. Convenciones del lenguaje: *Es gringo pero me cae bien.*→ [Poca simpatía por los gringos].

Presuposiciones. Hechos que se desprenden de una afirmación: *Rockdrigo (no) sobrevivió al maremoto.* → [Hubo un maremoto].

Conocimiento del mundo. Conocimiento adquirido por experiencia: *Salió sin cerrar la puerta.* → [La puerta quedó abierta].

La atención a la paradoja de la web **mucho texto poco acceso**, *requiere la adquisición de conocimiento a partir de la misma web y acudiendo al PLN.*

Mapa del PLN

Finalidad: *Desarrollar sistemas que utilicen el lenguaje natural*

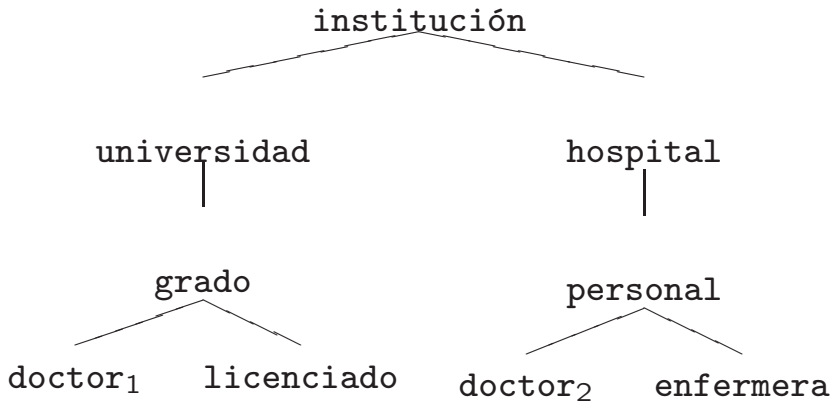
- Texto
  - Representación
    - Técnicas de selección de términos (1), Ponderación en la representación vectorial (2), *Parsing* ...
  - Generación de recursos lingüísticos
    - Bases de Datos Léxicas (BDL) (3), Terminología (4), Relaciones léxico-semánticas (RLS) (5), *corpora*, Extracción de Información (EI), ...
  - Clasificación
    - Categorización de textos (CT) (6), Agrupamiento de textos (AT) (8), Desambiguación del sentido (WSD) (9) ...
  - Búsqueda de información
    - Recuperación de Información (RI) (10), Búsqueda de Respuestas (QA) (11) ...
  - Acceso a la información
    - Recuperación de términos (12), Resumen Automático (13), Traducción Automática, Generación de texto ...
- Voz ...

El problema de la ambigüedad

La ambigüedad es una característica inherente a los lenguajes naturales; es común encontrar varias interpretaciones de una expresión. Este fenómeno ocurre en diversos niveles de los enunciados: léxico (*partido*), sintáctico (*El niño ve un gato con su telescopio*), o semántico (*Los estudiantes entregaron su trabajo*).
El ser humano usa el contexto, la situación y todo el conocimiento implícito; por ello (casi) no percibe la ambigüedad.
Word Sense Disambiguation (WSD) es *la elección de entre varios sentidos el que le corresponde a una palabra dentro de una oración*.

Ejemplo

Un enfoque muy común es usar una red de relaciones entre términos, como la BD-léxica *WordNet*. Considérese una parte de la jerarquía de relaciones de holonimia e hiperonimia:



La oración "El doctor instruyó a la enfermera" instancia "doctor" como personal de hospital, puesto que hay menor *distancia semántica* entre `doctor`$_2$ y `enfermera` que entre `doctor`$_1$ y `enfermera`.

Relaciones léxico semánticas

Para las tareas de PLN es necesaria la información; desde la menos elaborada (*corpora*), hasta la más sistematizada: relaciones. En lingüística se tienen clases de relaciones entre palabras. Por ejemplo:

- hipónimo-hiperónimo (`verde-color`)
- partitivo-holónimo (`viernes-semana`)
- sinónimo (`pavo-guajolote`)
- cohipónimo (`delantero-portero`)

También en cada dominio puede haber relaciones especiales. En `Juan pagó $10 por un helado`, se tendría: `Juan`$\rightarrow$*comprador*.

El problema es, a partir de un *corpus*, **identificar** si dos palabras están relacionadas, y **reconocer** el tipo de relación.

Ejemplo

- Para el texto: `En la corriente contemporánea encontramos autores como Agustín Yañez, Juan Rulfo, y Martín Luis Guzmán`
  se observa el patrón `X como Y (Y hip-de X)`, lo cual conduce a conjeturar que `Yañez, Rulfo` y `Guzmán` son hipónimos de `autor`.
- Los métodos se basan en el contexto y la representación de las palabras.
- El recurso más conocido son los *patrones léxico sintácticos*. Por ejemplo, para el reconocimiento de la hiponimia, estos son algunos patrones (15):

No.	Expresión	Peso
1	Ant1 word*, **pero** word* Ant2	5
2	**desde** word* Ant1 **hasta** word* Ant2	4
3	Ant1 word* [,|;] **sino** word* Ant2	5
4	Ant1 word{0,4}[**y**|**o**] word{0,4} Ant2	1

Resumen

1. Ante el desmesurado crecimiento de la información global es necesario el uso de herramientas que permitan manejar humanamente la web: **resultados precisos frente a una petición de información**.
2. La sistematización de la información (*web semántica*) es indispensable en la **generación de conocimiento útil para la construcción de sistemas de acceso a la información**.
3. El PLN contribuye con métodos, herramientas y aplicaciones para **resumir textos, clasificarlos, representarlos, y permitir operaciones en el contexto de la interacción humano computadora**.