5. Análisis semántico

Contenido

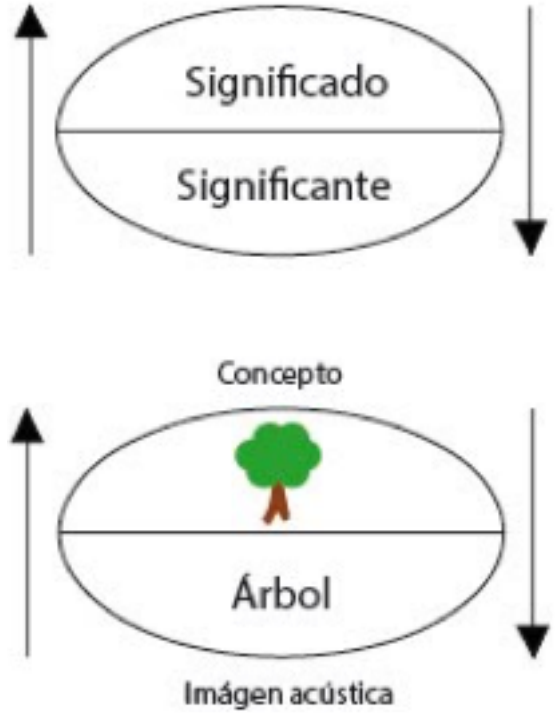
5.1 Análisis semántico

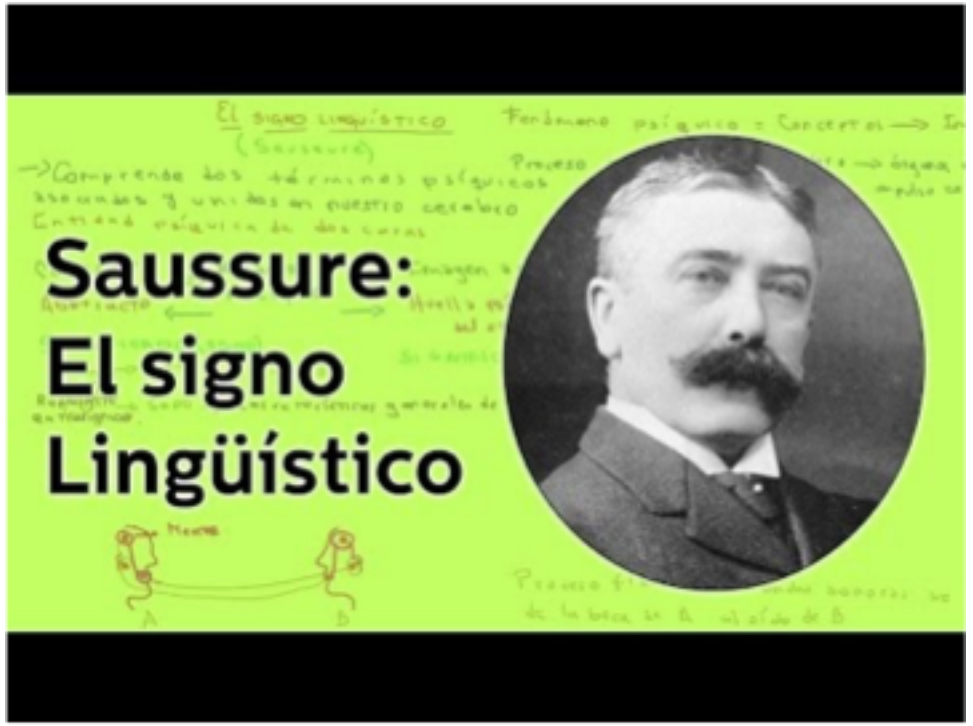
Hasta aquí hemos visto los niveles estructurales de la lengua, ahora vamos a entrar a un terreno un poco más abstracto, la semántica. Este nivel de la lengua puede ser uno de los que mayor dificultad de análisis representa para el PLN, pues el procesamiento de los significados no es fácil.

5.1 Análisis semántico

La semántica es la rama de la lingüística que se encarga del estudio del significado de los signos lingüísticos.

El signo lingüístico fue concebido por Ferdinand de Saussure como una unidad biplánica, la unión entre un concepto o idea (significado) y una imagen acústica (significante). Se trata de una entidad psíquica.





Cliquea en el video para ver una explicación más detallada.

5.1 Análisis semántico

La semántica se encarga tanto del significado de unidades léxicas (unidades simples), como de oraciones, sintagmas, etc. (expresiones complejas).

Las primeras son estudiadas por la semántica léxica, mientras que a las segundas las estudia la semántica composicional.

5.1 Análisis semántico

Semántica composicional

Hablemos primero, de la semántica composicional. Esta se encarga del estudio de las expresiones complejas, que pueden ser sintagmas u oraciones, aunque no se limita a estos ya que no restringe la longitud o grado de complejidad de los elementos que estudia.

Las expresiones complejas están formadas por unidades simples (esto como consecuencia de la recursividad de la que hablábamos la sesión 5) y deben cumplir con el principio de gramaticalidad, es decir, que estén apegadas a las reglas de una gramática.

5.1 Análisis semántico

En el caso de la sintaxis podemos decir que existe un número finito de palabras y un número finito de reglas de composición sintáctica, pero que estas reglas se pueden aplicar recursivamente.

Ejemplo:

O → SN SV

SN → (Det) N (SP)

SP → P SN

5.1 Análisis semántico

En semántica también podemos decir que existe un número finito de semas y un número finito de reglas de composición semántica, pero que estas reglas se pueden aplicar recursivamente.

[[O]] → [[SN]] ∈ [[SV]]

Gracias a esto es que se pueden analizar y comprender la infinidad de frases complejas que es posible producir en un lenguaje (visión algorítmica).

5.1 Análisis semántico

Principio de composicionalidad

Desarrollado por Gottlob Frege, el principio de composicionalidad establece que dependiendo de la estructura de la frase compleja y de sus componente es que determinará el significado.

Por ejemplo, no es lo mismo:

Carlos ama a Luisa

Que

Carlos odia a Luisa

O que

Luisa ama a Carlos

Este principio representa una generalización de cómo las unidades complejas construyen significado, además permite a la semántica composicional encontrar patrones de combinación semántica, los cuales sirven para diseñar muchos lenguajes formales.

5.2 Semántica léxica

La semántica léxica parte de la teoría de Chomsky (sesión 1), que establece que en nuestra mente existe una infinidad de palabras (lexicón), las cuales se combinan mediante reglas para generar relaciones de sentido. Estas reglas pueden tratarse desde la informática.

5.2 Significado léxico

Significado léxico

El significado léxico está compuesto por rasgos semánticos o sememas que son unidades mínimas de significación, estos a su vez se agrupan para formar el semema de un lexema .

Semema 'silla'

Determinar el semema (rasgos definitorios signiticativos)

Características		1	2	3	n	Σ	
• q1	Respaldo	+	+	+	+	+	S1
• q2	Terciopelo	+	-	-	+	+/-	S2
• q3	Sobre pie	+	+	+	+	+	S3
• q4	de madera	-	+	-	-	+/-	S4
• q5	para sentarse	+	+	+	+	+	S5
• q6	para una persona	+	+	+	+	+	S6

Rasgos distintivos (semas) comunes
La suma de todos ellos = semema (S): S1 + S3 + S5 + S6

Actividad formativa

Determina el semema de las palabras tiburón, bicicleta y mesa.

5.2 Análisis léxico

Este concepto también se relaciona con el de campo semántico, el cual se define como conjunto de palabras que comparten categoría gramatical y que están relacionadas por sus significado, es decir que comparten semas.

palabra	Sema compartido	semas distinguidores
pared	[+obstáculo][+vertical]	[+alto]
tapia	[+obstáculo][+vertical]	[+alto][+piedra][+delgado]
muro	[+obstáculo][+vertical]	[+alto][+piedra][+ grueso]
pretil	[+obstáculo][+vertical]	[+alto][+piedra]
cerca	[+obstáculo][+vertical]	[-alto][+rústico]
verja	[+obstáculo][+vertical]	[+metálico]
muralla	[+obstáculo][+vertical]	[+defensivo]

Actividad formativa

Desarrolla el campo semántico de transportes, que tengan como sema compartido ‘+ motorizado’.

5.2 Relaciones léxicas

Estas relaciones son importants cuando se trata la extracción de información.

Ahora bien, las palabras no son completamente independientes, sino que pueden guardar relación entre su significado léxico, algunas relaciones que podemos mencionar son las siguientes:

Sinonimia: hablamos de sinonimia cuando dos lexemas son utilizados en el mismo contexto y con el mismo significado, pueden ser sustituidos el uno por el otro sin cambiar el significado de la oración.

Por ejemplo: Computadora / ordenador

carro / automovil

chico/ pequeño

5.2 Relaciones léxicas

Antonimia: cuando el significado de dos lexemas se opone.

Por ejemplo: bueno/malo

alto/ bajo

bello/feo

5.2 Relaciones léxicas

Polisemia: cuando un mismo lexema tiene más de un significado (mismo origen etimológico, misma entrada en el diccionario).

Por ejemplo:	cometa	(astro) / (juguete)
	frente	(militar) / (cara)
	cresta	(de gallo) / (de la ola)

Homonimia: dos palabras que tienen la misma forma pero cuyos significados no se encuentran relacionados. Esta puede ser total (homógrafos: la misma escritura) o parcial (homófonos: misma pronunciación)

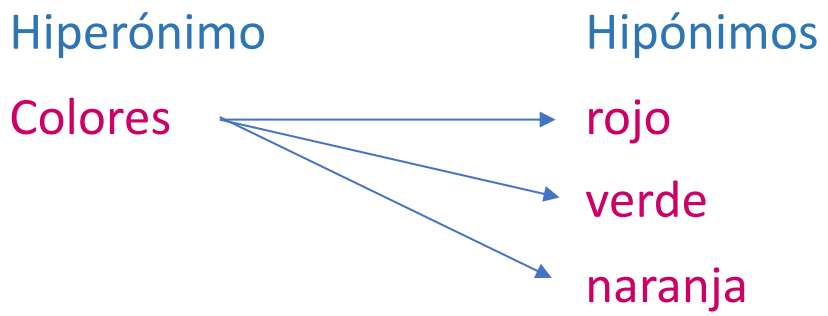
Por ejemplo:	homógrafos	cobre (metal) / cobre (Del verbo cobrar)
		haz (verbo hacer) / haz (Manojo, atado)
	homófonos	Sabia/ savia
		Votar/ botar

5.2 Relaciones léxicas

Hiperonimia: La hiperonimia es una relación que se establece entre una palabra de carácter más general y otra de carácter más específico. Los hiperónimos, entonces, son palabras cuyo significado incluye o engloba el de otra u otras palabras.

Hiponimia: es una relación que se establece entre una palabra de carácter más específico y otra de carácter más general. Los hipónimos son palabras cuyo significado es más restringido que el de otra palabra, que se interpreta como término genérico.

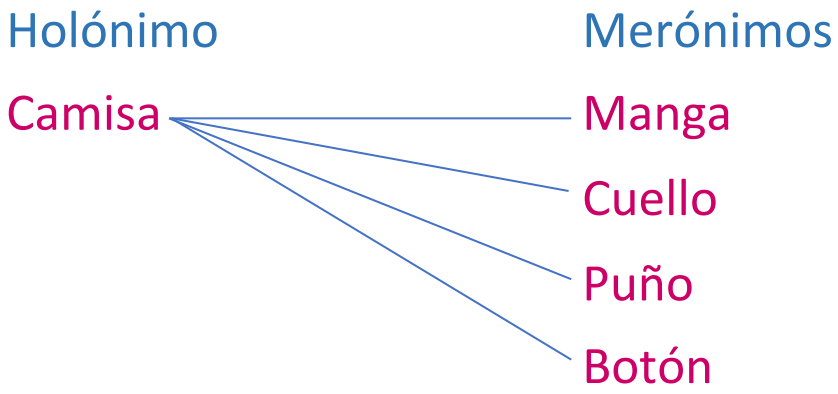
Por ejemplo:



5.2 Relaciones léxicas

Meronimia: establece una relación no simétrica entre los significados de dos palabras dentro del mismo campo semántico. Se denomina merónimo a la palabra cuyo significado constituye una parte del significado total de otra palabra, denominada holónimo.

Por ejemplo:



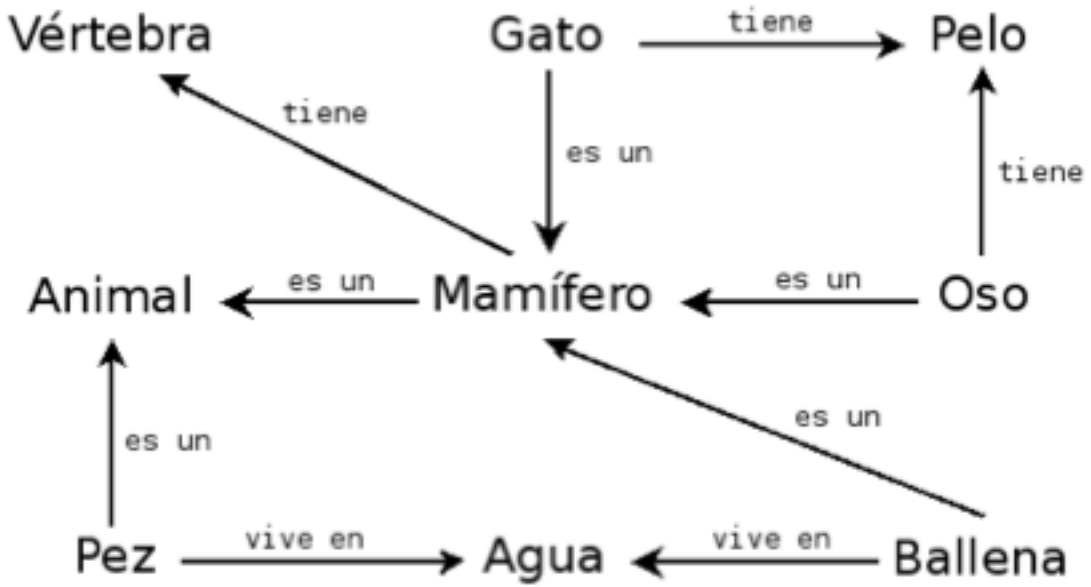
5.3 Redes semánticas

Podemos representar este conocimiento lingüístico mediante **redes semánticas**. Se trata de grafos en los que se representan conceptos y sus interrelaciones mediante nodos o vértices (elementos semánticos) y aristas (líneas de gráfos).

Dado un conjunto de términos $\{t_1, t_2,..., t_n\}$ se construye un grafo $G = (V,A)$

- Siendo V el conjunto de nodos formado por n elementos, que se corresponden al conjunto de términos $\{t_1, t_2,..., t_n\}$
- Y siendo A el conjunto de aristas. Dados t_i y t_j existirá una línea a_{ij} que una estos nodos si es que están relacionados.

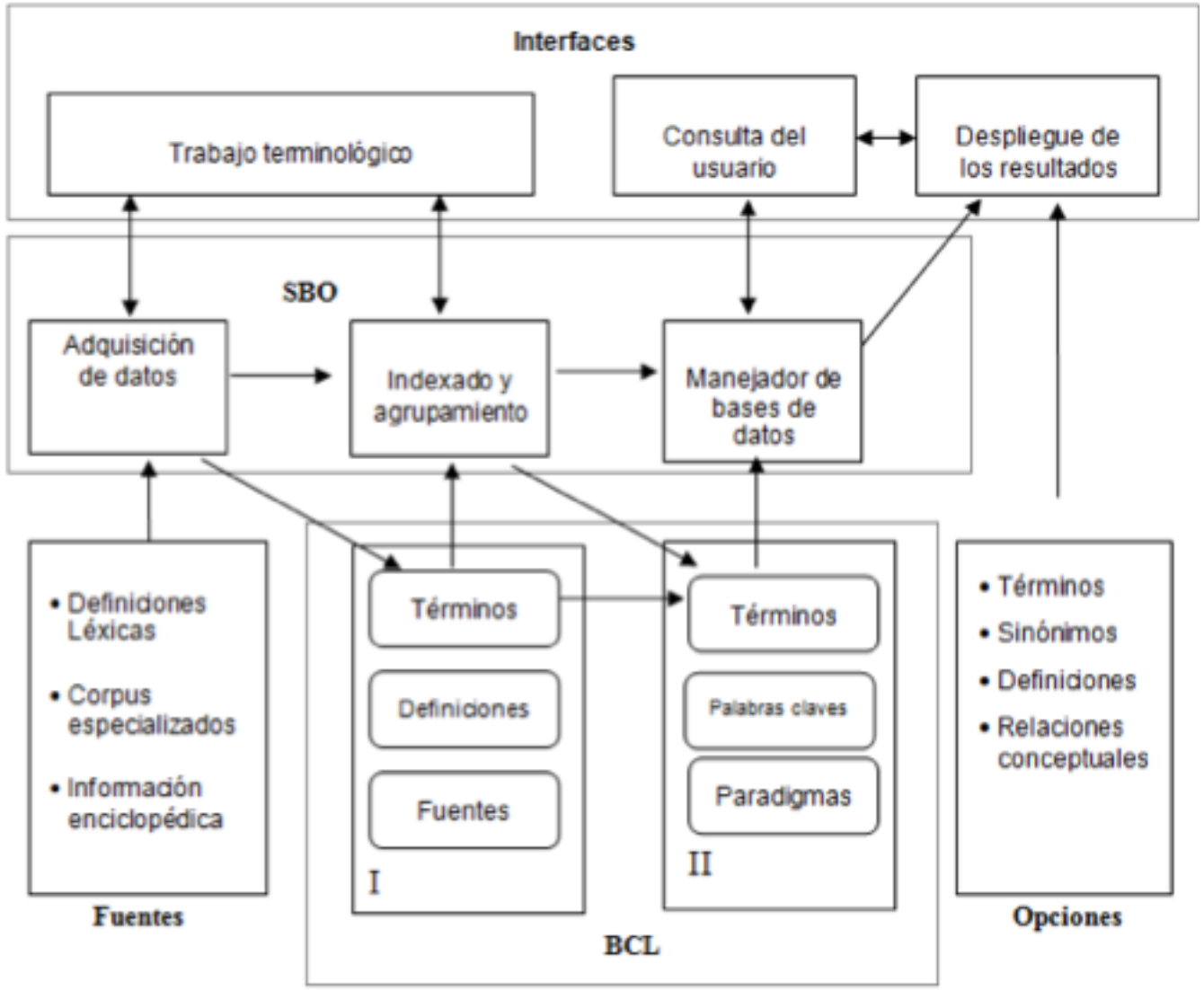
*si la relación no es simétrica, entonces se usan grafos dirigidos (con flechas) para representar la relación.



5.3 Redes basadas en relaciones léxicas

A partir de estas relaciones léxicas se han creado bases de conocimiento léxico (o BCLs), estas almacenan, administran y proporcionan conocimiento obtenido del lenguaje natural.

A la derecha vemos un esquema de cómo es que operan las BCLs.



5.3 WordNet

Una de las BCLs más conocidas e importantes es WordNet (ya en la sesión 4 hemos hablado un poco sobre ella), surgida a partir del interés de saber cómo el cerebro asocia y organiza conceptos, fue desarrollada en la Princeton University en 1985, y la investigación fue coordinada por George Armitage Miller.

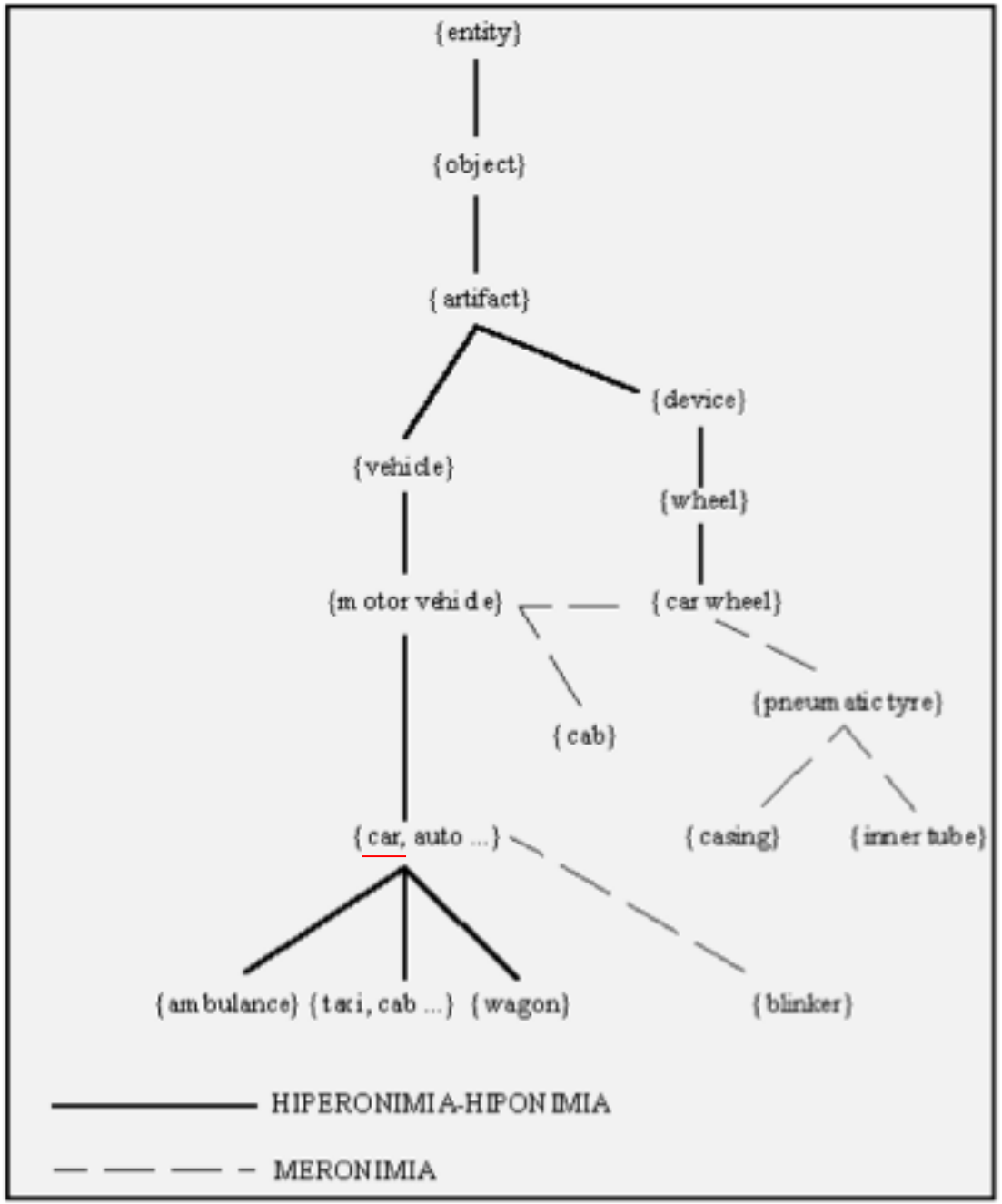
https://wordnet.princeton.edu/

5.3 Wordnet

WordNet es una base de datos léxico-conceptual del inglés estructurada en forma de red semántica, esto quiere decir que está compuesta de unidades léxicas y relaciones entre ellas.

El lexicón se divide en cinco categorías: nombres, verbos, adjetivos, adverbios y elementos funcionales. Con base en estas categorías, WordNet determina una serie de relaciones entre palabras a partir de su contenido semántico, el cual se denomina *synset*. Los *synset* equivalen a un descriptor sintáctico y semántico y permiten identificar y jerarquizar una palabra dentro de una red semántica.

Las relaciones léxicas qué considera WordNet son la sinonimia, antonimia, superordinación (hiperonimia), subordinación (hiponimia) y la meronimia.



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: car Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (frequency) {offset} <lexical filename > [lexical file number] (gloss) "an example sentence"

Display options for word: word#sense number (sense key)

Noun

- (71){02961779} <noun.artifact>[06] S: (n) **car#1 (car%1:06:00::)**, auto#1 (auto%1:06:00::), automobile#1 (automobile%1:06:00::), machine#6 (machine%1:06:01::), motorcar#1 (motorcar%1:06:00::) (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- (2){02963378} <noun.artifact>[06] S: (n) **car#2 (car%1:06:01::)**, railcar#1 (railcar%1:06:00::), railway car#1 (railway_car%1:06:00::), railroad car#1 (railroad_car%1:06:00::) (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- {02963937} <noun.artifact>[06] S: (n) **car#3 (car%1:06:03::)**, gondola#3 (gondola%1:06:03::) (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- {02963788} <noun.artifact>[06] S: (n) **car#4 (car%1:06:02::)**, elevator car#1 (elevator_car%1:06:00::) (where passengers ride up and down) *"the car was on the top floor"*
- {02937835} <noun.artifact>[06] S: (n) cable car#1 (cable_car%1:06:00::), **car#5 (car%1:06:04::)** (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*

http://wordnetweb.princeton.edu/perl/webwn

5.3 Similitud semántica

Este tipo de bases de datos se basan principalmente en la similitud semántica. En el campo del PLN se entiende similitud semántica como la medida de la interrelación existente entre dos palabras y esta se puede calcular mediante diversas métricas, veremos brevemente algunas de ellas.

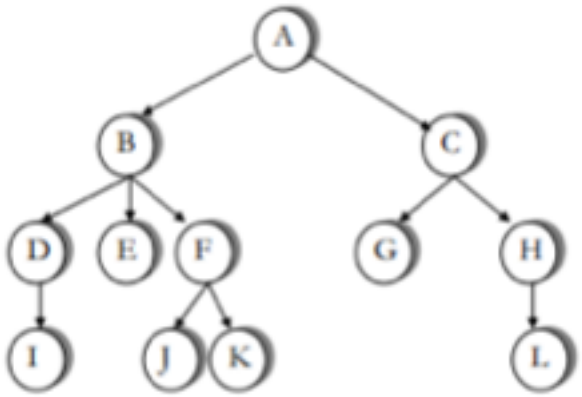
- Longitud de camino
- Algoritmo de Resnik
- Algoritmo de Dekang Ling

5.3 Similitud semántica

Longitud de camino

- En teoría de grafo el camino es una secuencia de nodos dentro de un grafo tal que exista una arista entre cada nodo. Dos nodos pueden estar conectados por varios caminos y el número de aristas dentro de estos es su longitud.

En un árbol la raíz tiene longitud de 1, sus descendientes directos longitud de camino 2 y así sucesivamente.



1. A es la raíz del árbol

2. B es hijo de A
   C es hijo de A
   D es hijo de B
   E es hijo de B
   H es hijo de L

3. A es padre de B
   B es padre de D
   D es padre de I
   C es padre de G
   L es padre de H

4. B y C son hermanos
   D, E y F son hermanos
   G y H son hermanos
   J y K son hermanos

5. I,E,J,K,G y L
   son nodos
   terminales
   u hojas.

6. B, D, F, C y H
   son nodo
   interiores.

7. El grados del nodo A es 2
   El grados del nodo B es 3
   El grados del nodo C es 2
   El grados del nodo D es 1
   El grados del nodo E es 0

8. El nivel de D es 1
   El nivel de B es 2
   El nivel de D es 3
   El nivel de C es 2
   El nivel de L es 4

9. La altura del
   del árbol es 4

5.3 Similitud semántica

Longitud de camino interno

La longitud de camino interno es la suma de las longitudes de camino de todos los nodos del árbol. Puede calcularse por medio de las siguientes fórmulas:

del

$$LCI = \sum_{i=1}^{h} n_i * i$$

i representa el nivel

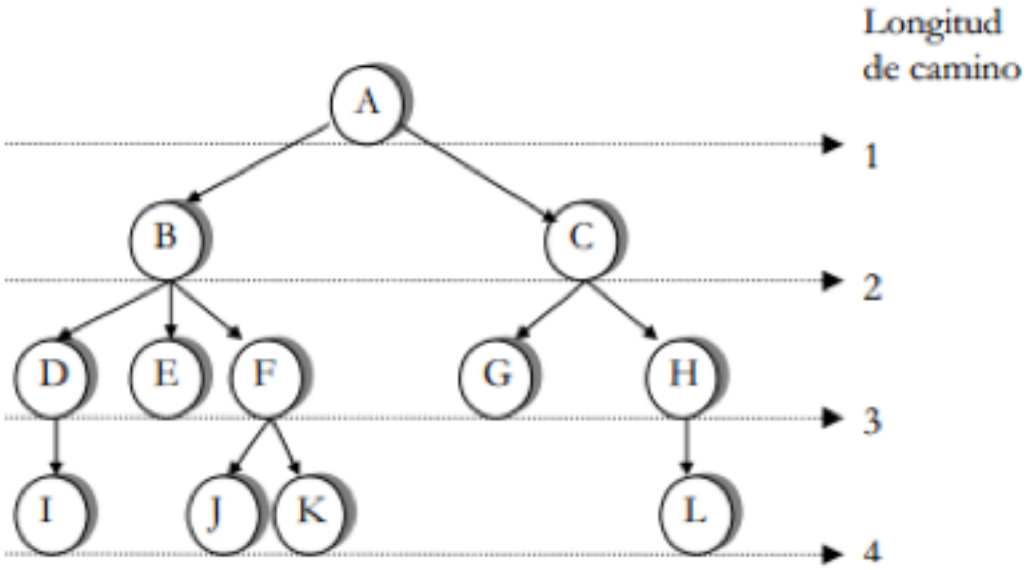
h su altura

n el número de

nodos especiales en el nivel *i*

5.3 Similitud semántica

Ejemplo:



LCI = 1 * 1 + 2 * 2 + 5 * 3 + 4 * 4 = 36

5.3 Similitud semántica

Longitud de camino externo:

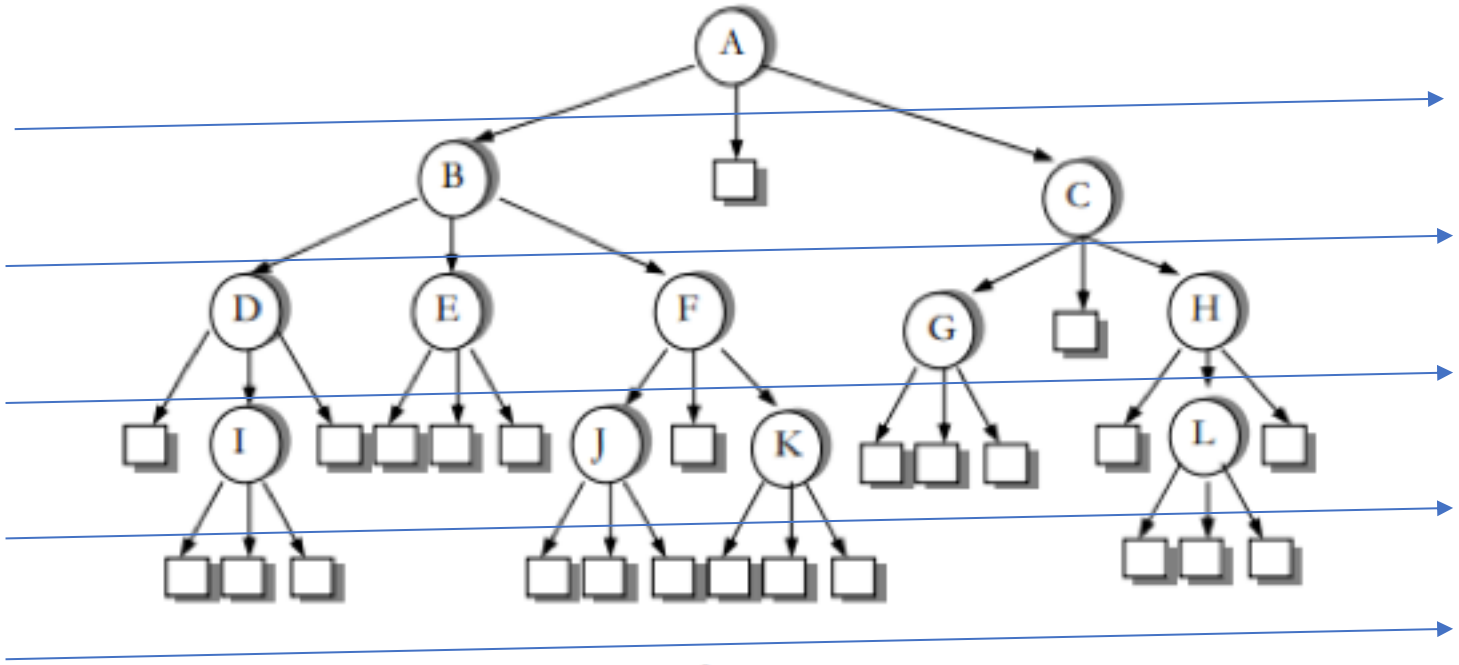
Un árbol extendido es aquel en el que el número de hijos de cada nodo es igual al grado del árbol, si esto no es así deben agregarse nodos especiales (reemplazan las ramas vacías o nulas y no tienen descendientes), tantos como sea necesario hasta cumplir la condición.

De este modo, la longitud de camino externo es la suma de las longitudes de todos los nodos especiales dentro del árbol y se mide con la siguiente fórmula:

$$LCE = \sum_{i=2}^{h+1} ne_i * i$$

5.3 Similitud semántica

Ejemplo:



LCE = 1 * 2 + 1 * 3 + 11 * 4 + 12 * 5 = 109

*Nota que el primer nodo especial aparece en el nivel dos.

5.3 Similitud semántica

La teoría de longitud de camino puede adaptarse para encontrar similitud semántica con los siguientes algoritmos:

$$wsim_{edge}(w_1, w_2) = 2 \times MAX - \min_{\substack{c_1, c_2 \\ c_1 \in S(w_1) \\ c_2 \in S(w_2)}} len(c_1, c_2)$$

Max es la probabilidad máxima de la taxonomía y se usa para transformar una medida de distancia en una de similitud.

S(w) se refiere a el conjunto de sentidos de la palabra *len(C1,C2)* es la longitud de camino.

Este método es muy sencillo de utilizar pero tiene sus limitaciones ya que no todos los caminos siguen de forma lineal.

5.3 Similitud semántica

Similitud de *Resnik*

Este algoritmo es uno de los más destacados en el cálculo de la similitud semántica. Propone que la similitud entre dos conceptos c1 y c2 de una estructura taxonómica, puede ser obtenida mediante la siguiente ecuación.

$$sim\ (c1, c2) = \max_{c \in S(c1,c2)} (-\log p(c))$$

Donde S (c1, c2) representa el conjunto de conceptos de los cuales tanto c1 como c2 descienden. Mientras que p(c) es la probabilidad del concepto c.

5.3 Similitud semántica

Similitud de Dekang Lin

Propone un algoritmo similar al de Resnik con la diferencia que este incluye no sólo coincidencias sino también discrepancias.

$$sim(x_1, x_2) = \frac{2 \times \log P(C_0)}{\log P(C_1) + \log P(C_2)}$$

5.4 Desambiguación

En semántica oracional, como en semántica léxica, también suelen darse casos de ambigüedad, pues, debido a la polisemia, una palabra puede ser interpretada de diferentes formas.

La ambigüedad semántica ocurre cuando una palabra o concepto tiene un significado de por sí difuso que se basa en el uso informal o generalizado. O cuando la organización sintáctica propicia diferentes interprestaciones de una oración.

Ejemplo: Juan le compró un coche a Pedro

¿Pedro le vendió el coche a juan?

¿Juan compró el coche para pedro?

5.4 Desambiguación

La Desambiguación del Sentido de las Palabras (WSD: Word Sense Disambiguation) busca eliminar o aminorar la ambigüedad semántica y favorecer la asignación automática de sentidos a las palabras de un texto.

La WSD emplea al menos dos tipos de métodos, uno basado en conocimientos y otro basado en corpus.

5.4 Desambiguación

Métodos basados en conocimiento

Estos métodos utilizan un conocimiento lingüístico previamente adquirido. Es decir, utilizan recursos externos, ya sea diccionarios, tesauros, textos y hasta recursos de la web. Los diccionarios utilizados por estos métodos se conocen como MRD (Machine Readable Dictionaries), entre los cuales encontramos:

Longman Dictionary of Contemporary English (LDOCE) (http://www.ldoceonline.com

Collins English Dictionary (CED) (http://www.collinslanguage.com/)

5.4 Desambiguación

Métodos basados en corpus

Estos usan técnicas estadísticas y de aprendizaje automático a partir de grandes ejemplos de corpus textuales (Corpus), estos métodos pueden ser tanto supervisados (corpus etiquetado) como no supervisados (corpus no etiquetado).

Aplicaciones del análisis semántico

El análisis semántico se emplea en:

- Traducción automática
- Robótica y agentes inteligentes
- Clasificación de documentos
- Motores de búsqueda
- Extracción de información
- Minería de textos
- Ontologías
- Detección de plagio
- Etc.

Bibliografía

Jurafsky, Daniel & James H. Martin. (2008). Speech and Language Processing: *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (2nd edition), Prentice Hall.

Antonín, M. A. M., Montraveta, A. F., & García, G. V. (Eds.). (2003). *Lexicografía computacional y semántica* (Vol. 64). Edicions Universitat Barcelona.