

Análisis Semántico

Procesamiento del Lenguaje Natural

José María Gómez Hidalgo
<http://www.esp.uem.es/jmgomez/>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Índice

- 1. **Ámbito de la semántica***
- 2. **El lenguaje de representación del significado***
- 3. **Marcos de casos***
- 4. **Restricciones selectivas y redes semánticas***
- 5. **Gramáticas semánticas***
- 6. **Semántica composicional***
- 7. **Lógica y semántica procedimental***
- 8. **Resolución de la ambigüedad léxica***

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

1. Ámbito de la semántica

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- La tarea de determinar el significado de una oración en LN se puede descomponer en dos fases
 - Calcular una expresión del significado independiente del contexto (típicamente una fórmula lógica) => semántica
 - Interpretar la expresión anterior en su contexto para obtener una representación final del significado => pragmática

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- La semántica cubre, por ejemplo
 - Eliminación de algunos significados de palabras explotando restricciones estructurales
 - Identificación de los papeles semánticos que cada palabra y sintagma juega en la representación (lógica)
 - Identificación de la restricciones de correferencia derivadas de la estructura de la oración

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- La semántica no cubre, por ejemplo
 - Determinación de las entidades referidas por medio de sintagmas nominales y otros sintagmas
 - Selección de una única representación del significado de entre las posibles
 - Determinación de la intención del uso de cada expresión en LN

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- Dos elementos fundamentales
 - El lenguaje elegido para representar el significado (Meaning Representation Language, MRL)
 - Algoritmo de análisis semántico
- Semántica
 - Proceso de traducción de una expresión en LN o una representación sintáctica suya al MRL

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- Integración con el análisis sintáctico
 - Modularización = acoplamiento débil
 - La entrada del análisis semántico es un árbol de análisis sintáctico
 - Modular
 - Acoplamiento fuerte
 - La entrada del análisis semántico es la expresión en lenguaje natural
 - Se construye la representación del significado al tiempo que se realiza el análisis sintáctico
 - Eficiente

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

1. Ámbito de la semántica

- **Semántica como apoyo al análisis sintáctico**
 - Uso de restricciones selectivas para filtrar análisis sintácticos incorrectos
 - Aparición de categorías sintácticas en la gramática => gramáticas semánticas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

2. El lenguaje de representación del significado

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

2. El lenguaje de representación del significado

- Hipótesis del MRL
 - Encontrar un MRL óptimo (que permita una perfecta comprensión del LN) es equivalente a comprender al ser humano y/o lograr una perfecta inteligencia artificial
 - Polémica: existe un tal MRL?
 - Enfoque práctico: buscar/usar un MRL que permita desarrollar aplicaciones prácticas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

2. El lenguaje de representación del significado

- Características del MRL
 - Mientras que el LN puede ser **ambiguo**, el MRL no debe serlo
 - Mientras que es complejo especificar reglas para determinar que una afirmación en LN es **cierta** o **falsa**, el MRL debe ir acompañado de un conjunto de reglas que permitan especificar que condiciones deben cumplirse en el mundo para que una expresión en el MRL sea cierta

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

2. El lenguaje de representación del significado

- Características del MRL
 - Mientras que es difícil establecer que conclusiones se pueden extraer de una afirmación en LN, el MRL debe ir acompañado de **reglas de inferencia** que permitan derivar otras expresiones en el MRL a partir de una expresión concreta (manteniéndose las condiciones de veracidad)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

2. El lenguaje de representación del significado

- Clasificación de MRLs propuestos
 - Basados en la lógica
 - Lógica de primer orden con ampliaciones (lógicas modales, etc.)
 - Basados en estructuras con soporte a métodos de inferencia específicos
 - Redes semánticas
 - Marcos (de casos)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

3. Marcos de casos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

- Se basa en las gramáticas de casos de Fillmore
- ¿Qué son los casos?
 - Tradicionalmente
 - Clasificación de los nombres con respecto a la función sintáctica que desempeñan en la oración
 - Típicamente determinado por el sufijo
 - Propios de lenguajes como el latín, ruso o finés
 - "Casos **superficiales** o sintácticos"

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

- ¿Qué son los casos?
 - Alternativamente
 - Clasificación de los sintagmas nominales de acuerdo a la función conceptual que desempeñan en la acción representada por una oración
 - Representación de una oración = acción (verbo) + características (sintagmas nominales)
 - "Casos **profundos** o semánticos"

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

- Diversos sistemas de casos semánticos
 - Sistema de Fillmore
 - Agente (agent)
 - Contra-agente (counter-agent)
 - Objeto (object)
 - Resultado (result)
 - Instrumento (instrument)
 - Fuente (source)
 - Objetivo (goal)
 - Paciente (experience)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

Agente - Instigador de la acción

Contra-agente - Fuerza o resistencia contra la que se realiza la acción

Objeto - Entidad acerca de cuya existencia, movimiento o cambio se refiere la acción

Resultado - Entidad que comienza a existir como consecuencia de la acción

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

Instrumento - Estímulo o causa física inmediata de la acción

Fuente - Lugar desde el que algo se mueve

Objetivo - Lugar hacia el que algo se mueve

Paciente - Entidad que recibe, acepta, experimenta o sufre los efectos de una acción

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

- Los casos se almacenan en un marco (frame)
 - Conjunto de pares atributo-valor
 - Atributos = casos
 - Valores = sintagmas nominales

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

3. Marcos de casos

- Ejemplo
 - Oración
 - "In Elm Street, John broke a window with a hammer"
 - Marco
 - [action = break
 - agent = john
 - object = window
 - instrument = hammer
 - source = elm street ...]

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

4. Restricciones selectivas y redes semánticas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Restricciones selectivas
 - Propuestas por Wilks
 - Indican los posibles tipos de los elementos de una oración (por ejemplo, los tipos de los casos)
 - Sirven para filtrar análisis sintácticos incorrectos semánticamente => eficiencia

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Ejemplo
 - Oración
 - "El perro come la salchicha"
 - El agente de la acción de comer debe ser una entidad **animada**
 - El objeto de la acción de comer debe ser una entidad **comestible**

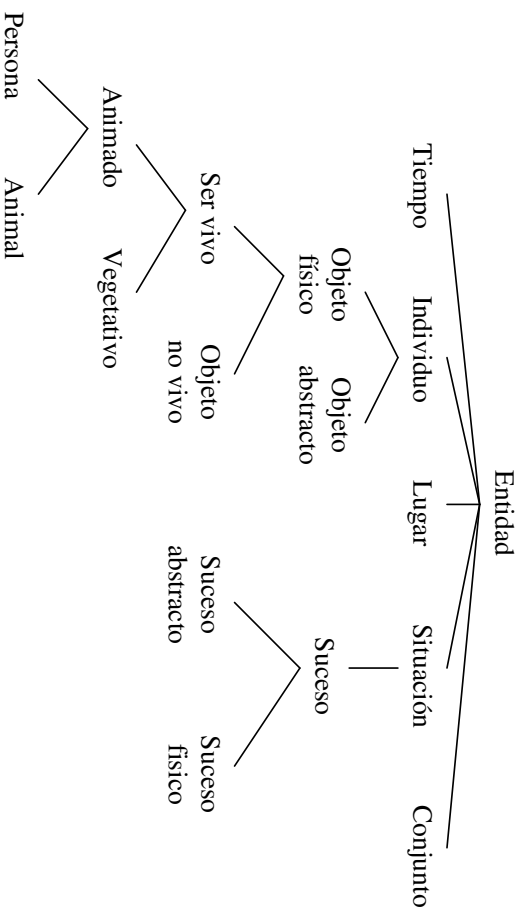
Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Las restricciones selectivas se suelen basar en **jerarquías de tipos**
- Relación "es un" entre tipos
 - Una entidad viva debe ser una entidad física
- Se obtiene una jerarquía o cuasi jerarquía

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Se heredan propiedades entre los tipos
- Se buscan las propiedades desde el tipo más específico al más general
- Las jerarquías facilitan el desarrollo del léxico
 - No es preciso describir todas las propiedades de un objeto

4. Restricciones selectivas y redes semánticas

- Las jerarquías de tipos se generalizan a redes semánticas
- Una **red semántica** es estructuralmente un grafo con nodos y arcos anotados
 - Los nodos suelen representar **conceptos**
 - Los arcos suelen representar **relaciones** entre conceptos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Se suelen incluir más relaciones entre conceptos además de la relación "es un"
- "es parte de"
 - Los objetos se relacionan con sus partes
 - Ejemplo: La cabeza es una parte del hombre

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Se pasa del concepto de jerarquía de tipos al de base de conocimiento
- **CYC**
 - Iniciativa para desarrollar una base de conocimiento que incluya todo lo que sabe un adulto medio o una enciclopedia de sobremesa
 - Microelectronics & Computer Technology Corporation
 - Muy ambicioso y poco fructífero

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- **WordNet**
 - Iniciativa para organizar un diccionario desde el punto de vista conceptual
 - Miller (Cognitive Science Laboratory) en Princeton
 - Base de conocimiento con información léxica extraída semiautomáticamente de diccionarios
 - Para el idioma inglés
 - Información sobre nombres, adjetivos, verbos y adverbios

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- **Concepto = conjunto de sinónimos (synset)**
 - "board" tiene dos significados (entre otros)
{board, gameboard} - tablero de juego
{dining table, board} - mesa donde se sirven comidas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- **Relaciones entre conceptos**
 - **Hiponimia**
 - Relación "es un" entre conceptos
 - {ambulance} es una clase especial de {car, auto, automobile, machine, motorcar}
 - **Meronimia**
 - Relación "es parte de" entre conceptos
 - {air bag} es una parte de {car, auto, automobile, machine, motorcar}

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- **Relaciones entre conceptos**
 - Antonimia
 - Relación "es lo contrario de" entre conceptos
 - {bad} es lo contrario de {good}

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- **Consulta interactiva**
 - "wn disguise"

The noun disguise has 3 senses (first 1 from tagged texts)

1. disguise, camouflage -- (an outward semblance that misrepresents the true nature of something; "the theatrical notion of disguise is always associated with catastrophe in his stories")
 2. disguise -- (any attire that modifies the appearance in order to conceal the wearer's identity)
- ...

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

...

3. disguise, camouflage -- (the act of concealing the identity of something by modifying its appearance; "he is a master of disguise")

The verb disguise has 1 sense (first 1 from tagged texts)

1. disguise -- (make unrecognizable; "The herb disguises the garlic taste"; "We disguised our faces before robbing the bank")

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Refinamiento (hipónimos de "disguise")

- "wn disguise -hynsn"

- 1 of 3 senses of disguise

- Sense 2

- disguise -- (any attire that modifies the appearance in order to conceal the wearer's identity)

- => fancy dress, masquerade, masquerade costume -- (a costume worn as a disguise at a masquerade party)

- ...

- => hairpiece, false hair, postiche -- (a covering or bunch of human or artificial hair used for disguise or adornment)

- => mask -- (a covering to disguise or conceal the face)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- Se puede integrar con un sistema de LN
 - Consultas en Java ó C
 - BD en Prolog
- Tamaño de WordNet 1.6
 - 32 Mb
 - más de 90000 palabras y expresiones
 - más de 70000 conceptos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

4. Restricciones selectivas y redes semánticas

- EuroWordNet
 - Proyecto europeo liderado por la Universidad de Amsterdam
 - Cubre los idiomas
 - Holandés (44015 synsets), Inglés (16361 synsets), Español (30485 synsets), Italiano, Alemán, Francés, Checo y Estonio
 - Versión reciente (Agosto 99)
 - Uso para recuperación de información multilingüe (Novell Linguistic Development)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

5. Gramáticas semánticas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Surgieron con el fin de caracterizar un subconjunto del lenguaje natural de manera suficiente para permitir la interacción de usuarios casuales
 - Se han aplicado sobre todo a interfaces a BD y sistemas de respuesta a preguntas
 - Son gramáticas en las que las categorías se refieren a conceptos semánticos y no sólo sintácticos
 - => el formalismo subyacente es independiente (CF-PSGs, DCGs, ATNs, etc.)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Ejemplo: SOPHIE, 1976 = Respuesta a preguntas sobre circuitos electrónicos
 - De reglas tipo
NP -> Det N Prep NP
 - Se pasa a reglas como
Measurement -> Det Measurable-Quantity Prep Part
 - Que permite sintagmas nominales como
"The voltage across R9"
"The current through the voltage reference capacitor"
"The power dissipation of the current-limiting transistor"

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Ventajas
 - Eficiencia
 - Se reduce el número de análisis posibles
 - No se precisa componente semántica
Query -> Query-Intro Measurement
 - Habitabilidad
 - Los usuarios pueden expresarse libremente en un subconjunto del lenguaje natural sin perderse en las limitaciones del subconjunto aceptado
 - Se permiten pequeñas variaciones como "Is something wrong?" = "Is there anything wrong?"

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Más ventajas
 - Fenómenos del discurso
 - Elipsis y referencia pronominal fácilmente resolubles
 - Ejemplo
 - "What is the population of Los Angeles?"
 - "What about San Diego?"
 - Por su posición, San Diego es de la categoría Ciudad

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Más ventajas
 - Entradas erróneas
 - Se reconocen fragmentos y se devuelve al usuario la regla para que sepa aplicarla
 - Ejemplo
 - Ante una expresión como "voltage across R9"
 - Devolver la regla
- Measurement -> Det Measurable-Quantity Prep Part

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Más ventajas
 - Auto explicación
 - Ante solicitudes de ayuda como "What is the voltage <help>"
 - El sistema puede devolver
Inputs that would complete the Measurement rule are
across Part
between Node and Node
at Node

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

5. Gramáticas semánticas

- Desventajas
 - Dificultad para cubrir fenómenos lingüísticos complejos
"Which ships does the admiral think the fourth fleet can spare?"
 - Dependencia del dominio
 - Debe recodificarse la gramática para cada nueva aplicación
 - Tamaño
 - Ciertas regularidades sintácticas como la coordinación o la pasiva deben recodificarse para cada regla semántica

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

6. Semántica composicional

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

6. Semántica composicional

- ¿Cómo organizar el proceso de análisis semántico?
- Principio de composicionalidad (Frege, s. XIX)

El significado del todo es una función del significado de las partes

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

6. Semántica composicional

- **Conclusión**
 - El significado de una oración se pone en función del significado de sus sintagmas
 - El significado de los sintagmas se pone en función del significado de los subsintagmas y palabras
 - ...
 - Se llega al significado de las palabras o incluso de los morfemas (lexemas)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

6. Semántica composicional

- **Organización = hipótesis regla-a-regla**
 - Para cada regla sintáctica que descompone un elemento en sus partes, se incluye una regla semántica que construye el significado del elemento en términos de los significados de las partes
 - Este enfoque no fuerza ningún grado de acoplamiento sintáctico-semántico

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

6. Semántica composicional

- Se deben resolver los puntos siguientes
 - ¿Cuáles son las partes adecuadas a considerar para obtener el significado de una oración?
 - Se supone que el análisis sintáctico produce estructuras semánticamente adecuadas
 - ¿Cómo depende el significado de una estructura del significado de sus subestructuras?
 - Construcción de las reglas semánticas
 - ¿Cuál es el significado de los elementos más básicos (palabras, lexemas)?

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

Análisis Semántico

7. Lógica y semántica procedimental

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Semántica procedimental
- Introducción a la lógica
- Lógica de primer orden (LPO)
- Representación del LN en términos de la LPO
- Interpretación de fórmulas lógicas
- Principios de diseño de un sistema de respuesta a preguntas
- Ejemplo

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Semántica procedimental
- En general
 - El significado de una orden es un procedimiento para realizar la acción requerida
 - El significado de una pregunta es un procedimiento para averiguar la respuesta
 - El significado de una afirmación es un procedimiento para agregar la nueva información al modelo del mundo o dominio

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Semántica procedimental
- Sistema de respuesta a preguntas
 - Base de conocimiento = base de datos
 - Objetivo: interpretar una forma lógica como una consulta a la base de datos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo
 - Consulta
 - "Does every flight to Chicago serve breakfast?"
 - Representación semántica (fórmula lógica)
 - $\forall F1 ((flight(F1) \ \& \ (dest(F1, name(chicago)))) \rightarrow (serve-breakfast(F1)))$
 - $todo(F1, implies(and(flight(F1), dest(F1, name(chicago))), ser-ve-breakfast(F1)))$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo
 - Interpretación
 1. Localizar todos los vuelos de la BD con destino Chicago
 2. Para cada vuelo encontrado, verificar si sirve desayuno.
Si así es, devolver "Yes". Devolver "No" en caso contrario.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Introducción a la lógica
- Lógica = intento de formalizar el razonamiento humano
- Numerosos sistemas prácticos de PLN con componente semántica basada en la lógica

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- **Idea central de la lógica**
 - Dada una situación descrita por un conjunto de afirmaciones ciertas (o asumidas como tales), determinar que otras afirmaciones son ciertas en la situación
 - Las afirmaciones iniciales son "premisas", las nuevas "conclusiones" y el proceso de obtenerlas "inferencia"

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- **Sintaxis de la lógica**
 - Conjunto de reglas que establecen que una afirmación está correctamente formada
- **Semántica de la lógica**
 - Conjunto de reglas que permiten deducir el valor de verdad de una fórmula

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Inferencia
 - ¿De un conjunto de premisas P, se puede deducir una conclusión Q? ($P \Rightarrow Q$)
 - Dos opciones
 - Se construye la fórmula $P \rightarrow Q$ y se demuestra que para todos los valores de verdad de las fórmulas de P, es cierta (es decir, es una tautología)
 - Se construye la fórmula $P \wedge \neg Q$ y se demuestra que para todos los valores de verdad de las fórmulas de P y Q, es falsa (es decir, es una contradicción - reducción al absurdo)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Típicamente se desarrolla un conjunto de reglas que permiten operar sintácticamente sobre las fórmulas para efectuar deducciones
 - Cálculo lógico ($P \rightarrow Q$)
 - Debe ser correcto y completo
 - Correcto
 - Si P es cierto y $P \rightarrow Q$, entonces Q es cierto
 - Completo
 - Si siempre que P es cierto se cumple que Q es cierto (es decir $P \Rightarrow Q$), entonces $P \rightarrow Q$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Sintaxis
 - Una fórmula es una constante (P, Q, R, \dots); o bien, si P y Q son fórmulas, entonces $\neg P$, $(P \wedge Q)$, $(P \vee Q)$, $(P \rightarrow Q)$, $(P \leftrightarrow Q)$ son fórmulas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Semántica
 - Una fórmula lleva asociado un valor de verdad (T, F) $\Rightarrow I(P)$ es T o F
 - Interpretación de las conectivas
 - $I(\neg P) = T$ si y sólo si $I(P) = F$
 - $I(P \wedge Q) = T$ si y sólo si $I(P) = T$ y $I(Q) = T$
 - $I(P \rightarrow Q) = T$ si y sólo si $I(P) = F$ ó $I(Q) = T$ ó se dan ambas cosas
 - etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Inferencia (basada en la semántica)
 - De P se deduce Q ($P \Rightarrow Q$) si siempre que $I(P) = T$ entonces $I(Q) = T$
 - Fácil con tablas de verdad

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Cálculo lógico
 - Es un conjunto de reglas que permiten obtener a partir de un conjunto de fórmulas otra fórmula por medio de operaciones puramente sintácticas
 - Consta de 3 axiomas y 1 regla

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Axiomas
 - $(P \rightarrow (Q \rightarrow P))$ (confirmación del consecuente)
 - $((P \rightarrow (Q \rightarrow R)) \rightarrow ((P \rightarrow Q) \rightarrow (P \rightarrow R)))$ (distributividad)
 - $((\neg P \rightarrow \neg Q) \rightarrow (Q \rightarrow P))$ (contraposición)
 - Regla
 - De P y $(P \rightarrow Q)$, se infiere Q (modus ponens)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo: Lógica proposicional
 - Este cálculo es correcto y completo
 - Correcto
 - Si de P se deduce Q, entonces $P \Rightarrow Q$
 - Es decir, si se puede operar P hasta llegar a Q, entonces Q se deduce semánticamente de P
 - » Si $I(P) = T$ entonces $I(Q) = T$
 - Completo
 - Si de P se deduce Q semánticamente, se puede operar con P hasta llegar a Q
 - Si siempre que $I(P) = T$ entonces $I(Q) = T$, entonces hay una forma de obtener Q a partir de P sintácticamente

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Numerosos sistemas lógicos
 - Lógica proposicional
 - Lógica de primer orden o lógica de predicados
 - Lógica modal - conceptos de necesidad y posibilidad
 - Lógica epistémica - concepto de conocimiento
 - Lógica doxástica - concepto de creencia
 - Lógica deóntica - conceptos morales como obligación y permiso
 - etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Enfoque práctico
 - Restringirse a la lógica proposicional o a la de predicados (primer orden) con ampliaciones dependientes del problema
 - Inferencia
 - Cálculo lógico completo y correcto (cálculo de predicados de primer orden)
 - Indecidible ...

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- El cálculo de predicados de primer orden es indecidible
 - No es posible, dado un conjunto de premisas y una potencial conclusión, construir un programa que
 - Responda "sí" si la conclusión se deduce de las premisas
 - Responda "no" si la conclusión no se deduce de las premisas
 - Y siempre termine devolviendo "sí" o "no" según lo anterior

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Lógica de primer orden (LPO)
 - Un término es una variable, o una constante, o un símbolo de función aplicado a otros términos
 - Ejemplos
 - X, a
 - juan
 - madre-de(pedro)
 - hijo-de(pedro,Hijo)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Una fórmula es un símbolo de predicado aplicado a una serie de términos; o si P y Q son fórmulas, entonces es de la forma $\neg P$, $(P \wedge Q)$, $(P \& Q)$, $(P \vee Q)$, $(P \rightarrow Q)$, $(P \leftrightarrow Q)$; o si V es una variable, y $P(V^*)$ es una fórmula con cero o más apariciones de V, entonces es de la forma $\forall V [P(V^*)]$ y $\exists V [P(V^*)]$
- Ejemplos
 - $a(X, Y)$
 - $en(torre-eiffel, paris), \neg sobrio(yeltsin)$
 - $(capital(españa, madrid) \wedge ciudad(X))$
 - $\exists X [humano(X) \rightarrow mortal(X)]$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Representación de LN en términos de LPO
 - Representación de nombres, adjetivos, verbos y determinantes
 - Ambigüedad en los cuantificadores
 - Primitivas versus representación léxica
 - Representación de la intención

7. Lógica y semántica procedimental

- Representación de nombres, adjetivos, verbos y determinantes
 - Nombres propios
 - Como símbolos de la base de conocimiento
Chicago = chi ó chicago
 - Con un predicado nombre y una variable
Chicago = nombre(X,'Chicago')
 - Nombres comunes y adjetivos
 - Con un predicado y una variable
perro = perro(X), alto = alto(X)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Verbos
 - Con un predicado de aridad posiblemente mayor que 1
come = comer(X,Y)
- Determinantes
 - Con cuantificadores aplicados sobre variables y fórmulas
todo = $\forall X [P(X) \rightarrow Q(X)] = \text{todo}(X,P(X),Q(X))$
un = $\exists X [P(X) \wedge Q(X)] = \text{existe}(X,P(X),Q(X))$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Necesitamos cuantificadores extendidos
 - el, muchos, pocos, algunos
 - most (la mayoría)
 - No sirven ni el cuantificador universal ni el existencial
 - Supongamos que existe un cunatificador M definido de modo que $M X [P(X)]$ sea cierto si más de la mitad de los objetos del dominio cumplen $P(X)$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- ¿Cómo representamos "Most dogs bark"?
- $M X [\text{dog}(X) \wedge \text{bark}(X)]$
- No sirve porque exige que la mayoría de los objetos del dominio sean perros
 - $M X [\text{dog}(X) \rightarrow \text{bark}(X)]$
- No sirve porque la implicación es cierta si la mayoría de objetos ladra aunque no sean perros
 - $\Rightarrow M X [\text{dog}(X), \text{bark}(X)] = m(X, \text{dog}(X), \text{bark}(X))$
- Análogamente
 - $eI = eI X [P(X), Q(X)] = eI(X, P(X), Q(X))$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Los cuantificadores presentan al menos dos tipos de ambigüedad
 - **Ámbito del cuantificador**
 - "Todo chico ama a un perro"
 - $\forall X$ (perro(X) & $\exists Y$ (chico(Y) \rightarrow ama(X, Y)))
 - $\exists Y$ (chico(Y) $\rightarrow \forall X$ (perro(X) \rightarrow ama(X, Y)))
 - **Lectura colectiva vs. distributiva**
 - "Dos hombres compraron un estéreo"
 - Distributiva = Cada hombre compró su propio estéreo
 - Colectiva = Los dos hombres compraron el mismo estéreo

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- **Primitivas vs. representación léxica**
 - Existen dos opciones
 - Para cada significado de una palabra, incluir un predicado (amar¹, amar², ...)
 - Poner las palabras en términos de un conjunto de primitivas del dominio (comer, tragar, devorar = ingerir (primitivo))
 - Llevado al extremo, se podrían diseñar primitivas para todos los conceptos de la mente humana (Conceptual Dependencies, Shank)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Representación de la intención
 - Tras cada interacción en LN hay una intención
 - Se pueden usar predicados para representarlás
 - assert/afirmar = la fórmula es una afirmación sobre el mundo
 - yn-query/pregunta-sino = se está preguntando por el valor de certeza de la fórmula
 - command/orden = la fórmula es una acción que el sistema debe realizar
 - wh-query/pregunta-cq = la fórmula describe un objeto que se quiere identificar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo
 - "El hombre come un melocotón"
 - Sin intención
 - el X [hombre(X), (\exists Y [melocoton(Y) & come(X, Y)]]]
 - Con intención
 - afirmar(el X [hombre(X), (\exists Y [melocoton(Y) & come(X, Y)]])]

7. Lógica y semántica procedimental

- ¿Qué falta por representar?
 - operadores modales para verbos como "creer" o "querer"
 - tiempo y aspecto de los verbos
 - número de los nombres
 - ...

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Interpretación de las fórmulas lógicas
 - Interpretación genérica de cuantificadores
 - Los cuantificadores se pueden interpretar en términos de conjuntos
 - algunos X [$P(X)$, $Q(X)$] = existe un subconjunto de los objetos que cumplen $P(X)$, que también cumplen $Q(X)$
 - mayoría X [$P(X)$, $Q(X)$] = existe un subconjunto de los objetos que cumplen $P(X)$ cuyo cardinal es mayor que la mitad de los elementos que cumplen $P(X)$, y que cumplen $Q(X)$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Semántica procedimental para un sistema de respuesta a preguntas
 - Problema de ejemplo
 - Interfaz a una base de datos sobre horarios de vuelos con la siguiente información
 - vuelo(ibe1).
 - aeropuerto(mad).
 - aeropuerto(bar).
 - salida(ibe1, mad, 1600).
 - llegada(ibe1, bar, 1730).

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Fases de la interpretación de la fórmula lógica
 - 1. Traducción de la fórmula lógica a un procedimiento de acceso de la BD
 - 2. Ejecución del procedimiento obtenido
- Interpretación de la fórmula lógica
 - Función I: MRL -> Procedimientos
 - Definida recursivamente sobre la estructura de las fórmulas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Interpretación de nombre(X, 'Madrid')
 - Se dispone de una "tabla de símbolos" similar a la de un compilador
 - Se recupera la constante de la base de datos (mad) y se actualiza la tabla de modo que toda referencia posterior a X sea interpretada como esa constante => I(nombre(X, 'Madrid')) = mad

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Interpretación de P (predicado)
 - Dos posibilidades
 - (a) El predicado P corresponde a un concepto de la BD => se devuelve el predicado directamente
 $I(P) = P$
 - (b) El predicado P no está en la BD (luego no hemos expresado los significados en términos de la BD sino en función de algo más general, como los significados de un diccionario) =>
 $I(\text{destino}(X, \text{nombre}(Y, 'Madrid')) = \text{llegada}(X, \text{mad}, Z)$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Interpretación de conectivas
 - Interpretación de $y(P, Q)$
 - $I(y(P, Q)) = \text{verificar-ambos}(I(P), I(Q))$
 - Ejecuta $I(P)$ e $I(Q)$, y si ambas son ciertas, devuelve cierto
 - Interpretación de $o(P, Q)$
 - $I(o(P, Q)) = \text{verificar-alguno}(I(P), I(Q))$
 - Ejecuta $I(P)$ e $I(Q)$, y si alguna es cierta, devuelve cierto
 - Interpretación de $no(P)$
 - $I(no(P)) = no(I(P))$
 - Ejecuta $I(P)$ y si es falso, devuelve cierto

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Interpretación de cuantificadores
 - Cuantificador "e!"
 - $I(e!(X(P(X), Q(X)))) = \text{encontrar-el}(X, I(P(X)), I(Q(X)))$
 - Se ejecuta $I(P(X))$ y se hallan el conjunto de todos los objetos que lo cumplen
 - Si el conjunto tiene un sólo elemento, se sustituye en $T(Q(X))$ y se devuelve si la cumple

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Si el conjunto es vacío, hay una "**violación de presuposición**"
 - Se supone que hay un elemento que cumple $I(P(X))$ y no lo hay
 - Se devuelve error
- Si el conjunto tiene 2 ó más elementos, hay una "**violación de presuposición**"
 - Se supone que sólo hay un elemento que cumple $I(PX)$ y resulta haber más
 - Se devuelve error o
 - Se devuelven aquellos que cumplen $I(Q(X))$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Cuantificador universal
 - $I(\text{todo}(X, P(X), Q(X))) = \text{iterar}(X, I(P(X)), I(Q(X)))$
 - Se buscan todos los elementos que cumplen $I(P(X))$, y si todos ellos cumplen $I(Q(X))$, se devuelve cierto
- Preguntas tipo "pregunta-cq"
 - $I(\text{pregunta-cq}(X, P(X), Q(X))) = \text{mostrar}(X, I(P(X), I(Q(X))))$
 - Muestra (devuelve) todos los objetos que cumplen $I(P(X))$ e $I(Q(X))$

7. Lógica y semántica procedimental

- Ejemplo
 - Oración

"¿Que vuelo a Madrid sale a las 1600?"
 - Fórmula (MRL)

pregunta-cq(X,y(vuelo(X), destino(X,nombre(Y,'Madrid'))), salida(X,Z,1600))
 - Interpretación (procedimiento)

mostrar(X,verificar-ambos(vuelo(X), llegada(X,mad,K)),salida(X,Z,1600))

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Principios de diseño de un sistema de respuesta a preguntas
 - Se pretende desarrollar un sistema de respuesta a preguntas
 - basado en Prolog y DCGs
 - con semántica procedimental
 - con conceptos próximos a los de la base de datos o conocimiento
 - fácil de interpretar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Pasos de desarrollo
 - Diseño de la base de datos o conocimiento en Prolog
 - Determinación del subconjunto del LN a procesar S
 - Construcción de una DCG que represente S
 - Determinación del MRL
 - Agregación de las reglas semánticas a la DCG
 - Construcción de un intérprete de las fórmulas lógicas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

7. Lógica y semántica procedimental

- Ejemplo
 - Base de datos (db41.pl)
 - Analizador sintáctico (dcg41.pl)
 - Analizador con representación del significado (dcg41.pl)
 - Intérprete de fórmulas (sem41.pl)
 - (Interfaz (interfaz41.pl))

Análisis Semántico

8. Resolución de la ambigüedad léxica (Word Sense Disambiguation, WSD)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD

- Definición
- Aplicaciones
- Evaluación
- Esquema general
- Métodos basados en aprendizaje
- Métodos basados en diccionarios
- Integración de técnicas y recursos

8. WSD – Definición

- Desambiguación del significado
 - Resolución de la ambigüedad léxica
 - *Word Sense Disambiguation* (WSD)
- Definición
 - Múltiples palabras con varios significados (polisemia)
 - Pero en un contexto de uso, sólo se activa uno
 - Esta es una hipótesis para la aproximación computacional

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

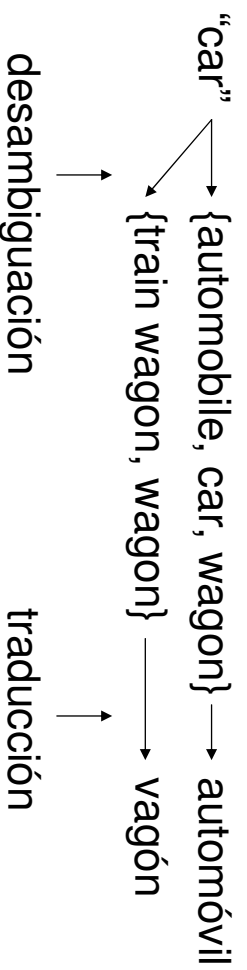
8. WSD – Definición

- Referencia
 - Significados en un diccionario
 - Clases de un thesaurus
 - Entradas en diccionario de transferencia para traducción
- Objetivo = identificar el significado *activado* en un contexto
- Problema difícil (Turing o IA – completo)
 - Comparativamente con otras tareas (etiquetado sintáctico = POS-Tagging)
 - Influyó notablemente en el informe ALPAC (60's)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Aplicaciones

- Tarea de clasificación *intermedia* – base para otros procesos o aplicaciones finales
- Aplicación estrella = Traducción automática



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Aplicaciones

- Clasificación de documentos
 - De gran importancia actual = R. Información
 - Polisemia
 - Recuperar documentos con "jaguar" como *animal* o como *automóvil*
 - Sinonimia
 - Recuperar documentos con "cosmonauta" si la consulta es "astronauta"
- Análogamente, categorización, agrupamiento, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Aplicaciones

- Especialmente, RI cross-lingüe
 - El usuario lee varios idiomas pero escribe en uno
 - Plantea consultas en un idioma y recibe documentos en varios
 - Técnica
 - Si los significados están unidos en distintos idiomas (e.g. EuroWordNet) = “car” – car/auto ~ coche/auto – “coche”
 - Y los documentos y consultas desambiguados
 - Se obtiene alta precisión en las consultas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Aplicaciones

- Otras tareas
 - Etiquetado sintáctico (POS-tagging)
 - Desambiguación referencial de preposiciones (PP-attachment)
 - Disminución de análisis sintácticos (restricciones selectivas)
 - Restauración de acentos (libro vs. libró)
 - Conversión a minúsculas (HE READ THE TIMES)
 - Reconocimiento del habla, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Aplicaciones

- No está claro si es ÚTIL
- SENSEVAL 3 (<http://senseval.org>)
 - Panel en aplicaciones de WSD
 - Personal de Google, Sharp, UPC, Microsoft, etc.
 - Debate centrado en aplicaciones que demuestren que la WSD explícita es útil para otras tareas
 - Considerables dudas y escasas evidencias prácticas
 - Mucho interés porque está bien definida y es fácil (?) de evaluar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- (Como siempre) centrada en la efectividad
- Otros factores
 - Eficiencia
 - Coste de producción
 - Implementación
 - Disponibilidad de recursos
- Nos centramos en la efectividad

8. WSD – Evaluación

- Evaluación directa vs. indirecta [Ide98]
 - Directa (*in vitro*) – Medición del grado de éxito en la tarea aislada (caja de cristal)
 - Imprescindible para apreciar las dificultades de la tarea
 - Indirecta (*in vivo*) – Medición del impacto de su éxito en la tarea superior (caja negra)
 - E.g. ¿Mejora la recuperación de información?
 - Imprescindible porque aun con eficacia limitada, posibles mejoras en la tarea superior

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación directa (*in vitro*)
- Índice de éxito = número de aciertos / número de intentos (*accuracy, error*)
 - Sólo palabras ambiguas
 - Si no es capaz de desambiguar, *recall-precision*
 - Relativo a
 - La frecuencia de los significados (95/5 vs. 50/50)
 - La dificultad para los humanos (consistencia entre anotadores en torno al 70-75%)
 - La granularidad (abstracción) de los significados

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación directa (*in vitro*)
- Recursos = textos con palabras etiquetadas con significados de un diccionario = colección de evaluación
- Referente = SemCor
 - Subconjunto del Brown Corpus
 - Significados de la BDL WordNet

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación directa (*in vitro*)
- Línea base = resultados mínimos
 - Un sistema que no los supera, debe descartarse
 - La más usual = el significado más frecuente dada la etiqueta sintáctica
 - Los etiquetadores sintácticos aciertan > 95%
 - Los diccionarios, etc. listan los significados por frecuencia (en corpus de referencia)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación directa (*in vitro*)
- Objeto de SENSEVAL
 - “The purpose of Senseval is to evaluate the strengths and weaknesses of such (WSD) programs with respect to different words, different varieties of language, and different languages.”
 - SENSEVAL 1 (1998) – Inglés, francés, italiano
 - SENSEVAL 2 (2001) – Euskera, chino, checo, danés, holandés, inglés, estonio, italiano, japonés, coreano, español, sueco

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- SENSEVAL 3 (2004) – 14 tareas
 - WSD masiva – inglés, italiano
 - WSD muestra – euskera, catalán, chino, inglés, italiano, rumano, español
 - Adquisición automática de sub-categorización
 - WSD muestra multilingüe
 - WSD de descripciones de significados en WordNet
 - Roles semánticos
 - Formas lógicas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- SENSEVAL 3 – WSD masiva – inglés
 - 26 sistemas de 16 equipos
 - 5000 palabras – WSJ y Brown Corpus
 - Línea base (más frecuente) ~ .61
 - Consistencia entre anotadores ~ .725
 - Mejor sistema (Bélgica) = .652
 - Mejor sistema España (UAI,UJa,UPoV) = .626 (4º)
 - 14 sistemas sobre la línea base
 - *OBS: muestra léxica - más fácil*

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación indirecta (*in vivo*)
 - E.g. [Gonzalo98]
 - Integración de WordNet en RI
 - Consultas y documentos =(WSD)=> significados de WordNet => vectores de conceptos => MEV
 - Resultados empíricos sobre colección artificial

Indexing by synsets	62.0
Indexing by word senses	53.2
Indexing by words (basic SMART)	48.0
Indexing by synsets with a 5% errors ratio	62.0
Id. with 10% errors ratio	60.8
Id. with 20% errors ratio	56.1
Id. with 30% errors ratio	54.4
Indexing with all possible synsets (no disambiguation)	52.6
Id. with 60% errors ratio	49.1

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Evaluación

- Evaluación indirecta (*in vivo*)
 - E.g. [Ureña01]
 - Integración de WordNet en categorización
 - Nombres de categorías =(WSD)=> significados de WordNet => expansión con sinónimos => enriquecimiento del vocabulario => aprendizaje
 - Resultados empíricos

F1	NoWN	NoWSD	AutoWSD	ManWSD
M	0.464	0.538	0.571	0.576
m	0.661	0.664	0.674	0.678

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Esquema general

- Clasificación de técnicas
 - Atendiendo al recurso empleado
 - Colección de datos etiquetados = colección de entrenamiento (e.g. SemCor) => *basado en aprendizaje, supervisado*
 - Recurso léxico (diccionario, Base de Datos Léxica = WordNet) => *basado en diccionarios, no supervisado*
 - Tendencia a la integración

8. WSD – Esquema general

- Ventajas y desventajas
 - Basado en aprendizaje
 - Dependiente del dominio
 - Datos objetivo similares a los de entrenamiento
 - Problemas usuales de aprendizaje
 - Carencia de datos, distribuciones desequilibradas, etc.
 - Más efectivo si lo anterior no ocurre
 - Coste de creación de recursos (?) – colecciones de entrenamiento

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en aprendizaje

- Idea general
 - Comparar los contextos de aparición de cada significado con el contexto actual
(Basado en WordNet)
CAR - {automobile, car, wagon}
“The car is running out of fuel”
“I sold my car and purchased a brand new BMW”
“My car’s rearviewmirror is being fixed”
CAR - {train wagon, wagon}
“I left the dog in the baggage car of the train”
“I boarded the wrong car, so I changed at Victoria Station”
¿“*VW and BMW are improving their **car technologies**”?*

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en aprendizaje

- **Proceso simplificado**
 1. Para cada significado, se recolectan contextos de aparición y las palabras en ellos
 2. Para la palabra objetivo, se compara el solapamiento / similitud entre su contexto y los de cada significado
 - Se puede usar el MEV (representación como vectores de pesos, similitud del coseno)
 - Mayoría de relaciones léxicas = entornos de 5 palabras

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en aprendizaje

- **Detalles**
 - Tamaño del contexto (evidencias lejanas)
 - Falta de datos de entrenamiento
 - Empatate entre varios significados
 - Técnicas de aprendizaje
 - Definición de atributos (lingüísticos, léxicos)
 - Selección de atributos
 - Algoritmos de aprendizaje

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en aprendizaje

- GAMBL (Top 1, SENSEVAL 3 English-AW)
 - Contexto = 7 palabras, centro en objetivo
 - Atributos (e.g.) = palabras, POS, grupos + relaciones (nominal/verbal/preposicional + sujeto/objeto) (<= analizador superficial)
 - Selección de atributos = algoritmos genéticos
 - Algoritmos = KNN (TIMBL)
 - Proceso en cascada sobre varias colecciones de datos (SemCor, SENSEVALs previos, etc.)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en aprendizaje

- Enriquecimiento sucesivo [Yarowsky95]
 - Esquema algorítmico
 1. Seleccionar algunas apariciones y etiquetarlas
 2. Entrenar en las apariciones etiquetadas
 3. Aplicar sobre las no etiquetadas y aceptar las más probables
 4. Volver a 2 hasta que el algoritmo converja
 - Aprendizaje automático = *bootstrapping*
 - Además: “*One Sense Per Collocation*”, “*One Sense Per Discourse*”

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en diccionarios

- Idea general
 - Comparar el contexto de uso con las definiciones de un diccionario, asignando la más similar (WordNet 2.0)
 1. Car - 4-wheeled motor vehicle; usually propelled by an internal combustion engine; “he needs a car to get to work”
 2. Car - a wheeled vehicle adapted to the rails of railroad; “three cars had jumped the rails”
- ¿“*The train derailed and some **cars** were very damaged*”?

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en diccionarios

- Proceso simplificado
 1. Para la palabra objetivo, se compara el solapamiento / similitud entre su contexto y las definiciones
 - Se puede usar el MEV (representación como vectores de pesos, similitud del coseno)

8. WSD – Métodos basados en diccionarios

- Algoritmo clásico de Lesk [Lesk86]
 - Se comparan las definiciones potenciales con las definiciones posibles de cada palabra del contexto
 - Se asigna la combinación de mayor solapamiento
 - Pine* (1) Kinds of evergreen tree with needle-shaped leaves
 - (2) Waste away through sorrow or illness
 - Cone* (1) Solid body which narrows to a point
 - (2) Something of this shape whether solid or hollow
 - (3) Fruit of certain evergreen tree
- Si $|\text{contexto}| > 2$, numerosas comparaciones
 - ¿“... pine cone ...”?*

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Métodos basados en diccionarios

- Densidad Conceptual, WordNet [Agirre96]
 - Conjuntos de nombres contiguos => seleccionar los significados que maximizan la DC
 - DC tiene en cuenta
 - La longitud del camino más corto entre conceptos
 - La profundidad de la jerarquía (+profunda => >relación)
 - La densidad de conceptos en la jerarquía (+densa => >relación)
- Independencia del número de conceptos medidos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Integración

- **Idea general**
 - Tomar lo mejor de ambos métodos, integrando a
 - Nivel de recurso
 - E.g. Extraer sinónimos, hipérrnimos, etc. para las palabras en los contextos de entrenamiento
 - Nivel de clasificación
 - E.g. Dos clasificadores, votos por confianza sobre un conjunto de validación
 - “Cuanto más informado esté un sistema, más efectivo será”

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Integración

- **R2D2 (Top 4, SENSEVAL 3 English-AW)**
 - **Sistemas**
 - Maximum Entropy (apr.)
 - UPV-SHMM-AW (apr., Modelos Ocultos de Markov)
 - Relevant Domains (dic., WordNet Domains)
 - LVQ-JAEN-ELS (apr., Learning Vector Quantization)
 - CIAOSENSO (dic., densidad conceptual, WordNet)
 - **Combinación**
 - Voto
 - En cascada

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea de Madrid

8. WSD – Resumen

- WSD – Problema importante en LN
- Aplicaciones importantes, pero con dudas
- Evaluación basada en efectividad
- Métodos clasificados por recursos
 - Aprendizaje, diccionarios, integrados
- Estado de la tarea
 - Efectividad limitada, lejos de otras tareas
 - Aplicabilidad potencial en clasificación de texto