

Sesión de Prácticas 4

Resolución de la Ambigüedad Semántica de las Palabras

1. Objetivos

- Conocer los problemas de la resolución automática de las palabras.
- Analizar WordNet, su estructura y funcionamiento, y su aplicación a WSD y sus problemas.
- Reflexionar sobre métodos para desarrollar sistemas de resolución automática.

2. Tareas

La tarea a desarrollar esta semana es la anotación manual de corpus español. Vamos a trabajar sobre el primer fichero del corpus de la sesión 2 de prácticas (en español).

Para ello, sigue los siguientes pasos:

1. Descarga el corpus y ábrelo con un editor de textos cualquiera.
2. Abre el Wei de EuroWordNet, que se puede consultar desde esta web

<http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl>

3. Empezando el corpus por el principio, busca cada nombre y selecciona un solo sentido.
4. Para marcar el sentido, escribe en el corpus tras la palabra la marca # y a continuación el número de sentido. Ejemplo, banco#06227059n
 - a. Para cada palabra que anotes, responde a la cuestión 3 del punto 4.
5. Al final, responde a las cuestiones del punto 4.

3. Sobre el WEI de EuroWordNet

- Los sentidos están ordenados por probabilidad: del más probable al menos.
- Cada sentido incluye un conjunto de sinónimos que aparecen junto al número
- La lista de palabras de la izquierda son los dominios y los conceptos ontológicos relacionados con el sentido
- Para ver las relaciones (hiperónimos, hipónimos, etc.), selecciona la opción correspondiente.
- Prueba todas las opciones a ver qué información ofrece EuroWordNet

4. Cuestiones

1. Dificultad de la tarea, ¿crees que es difícil o fácil para un humano desambiguar el sentido de las palabras? Y con ello, ¿para una máquina?
2. ¿Cuál crees que es el principal problema? ¿El contexto, los sentidos, ...?

3. ¿Hay muchas palabras ambiguas? ¿A qué crees que se debe?
4. ¿Qué información utilizas para especificar el sentido de la palabra?
 - a. Información del contexto:
 - i. Posición: palabras anteriores y posteriores, otros.
 - ii. Información categorial: categoría de las palabras de la oración, otros.
 - iii. Información sintáctica: nombres anteriores y posteriores, preposiciones y otras categorías gramaticales, el verbo principal de la oración, el núcleo de la oración, otros.
 - iv. Información semántica: palabras monosémicas del contexto.
 - v. Información textual: tema general del texto, dominio, otros.
 - vi. Otra
 - b. Información del recurso léxico:
 - i. Glosa o definición
 - ii. Relaciones semánticas: sinónimos, hiperónimos, hipónimos, etc.
 - iii. Información de probabilidad (el primero de la lista).
 - iv. La traducción al inglés
 - v. Otra
5. Con esta información, ¿podrías proponer un método de resolución automática?