Desambiguación del Sentido de las Palabras

(Word Sense Disambiguation)

Introducción

- ¿qué es ambigüedad?
- ¿qué significa desambiguar?
- ¿qué entendemos por sentido de las palabras?
- ¿en qué consiste la tarea de desambiguación del sentido de las palabras?

Ambigüedad en el Lenguaje

- Léxica
  - *habla*, *aviso* – ¿verbo o sustantivo?
- Sintáctica
  - *Veo un gato con el telescopio* (*con cola larga*).
- De referencia
  - *Juan tomó la torta de la mesa y la comió* (*y la limpió*).
- Y otras más…

Ambigüedad Semántica de la Palabra

- Dados estos ejemplos de uso de la palabra ***gato***:
  - *“Saca el gato a pasear”*
  - *“Saca el gato de la cajuela”*
  - *“Saca el gato tridimensional del closed”*

- ¿en los cuatro casos hablamos del mismo gato?
- ¿cuál es el sentido correcto en cada oración?
- ¿por qué o cómo lo pudimos determinar?

Polisemia y Ambigüedad Semántica

- *Polisemia* – Una palabra puede poseer varios significados posibles.
- Por ejemplo, *gato* posee 3 sentidos:
  1. Animal felino
  2. Herramienta para levantar cosas pesadas
  3. Juego de 3 en línea o Tic-Tac-Toe.

Desambiguar el Sentido de las Palabras

- Es el proceso de decidir (seleccionar) el sentido de una palabra en su contexto.
- Decimos “seleccionar” porque cada palabra tiene un conjunto determinado de sentidos posibles.
- Este problema surge debido a que las palabras pueden asumir diferentes significados dependiendo del contexto en el que se usan.
- Nos referiremos a esta tarea como WSD

Aplicaciones prácticas

- Traducción automática
  - Traducir "*bill*" del inglés al español:
    - ¿El pico de un pájaro o cuenta bancaria?
- Recuperación de información
  - Encontrar páginas web sobre "*Java*"
    - ¿el lenguaje de programación, el tipo de café, ó la isla en el archipiélago de Indonesia?
- Búsqueda de respuestas
  - *¿Dónde está la estación Hidalgo?*
    - *¿me preguntan por un lugar o por una persona?*

El problema de WSD

1. Elegir un repositorio de sentidos
   - Diccionario o tesauros de referencia, donde se indiquen los distintos sentidos de las palabras.
2. Diseñar un procedimiento de desambiguación.

   (Este es el punto que trataremos en detalle)
3. Evaluar el rendimiento del procedimiento
   - Corpus etiquetado manualmente
   - Usar el sentido más frecuente (*baseline*)

¿cómo hacerlo automáticamente?

Enfoques de WSD

1. Métodos basados en conocimiento
   - Hacen uso de recursos léxicos tales como *diccionarios o tesauros*.
2. Métodos supervisados
   - Requieren de un *corpus etiquetado*.
   - Incluyen métodos mínimamente supervisados.
3. Métodos no supervisados
   - Discriminación del sentido de las palabras

Métodos Basados en Conocimiento

Podemos distinguir 3 tipos básicos:

- Basados en diccionarios
  - Usan las distintas definiciones de un término
- Basados en Tesauros/Ontologías
  - Consideran categorías semánticas de las palabras.
  - Aprovechan relaciones léxicas entre las palabras.
- Basados en traducciones en otro lenguaje
  - Usar “modelo de lenguaje” de otro lenguaje para decidir mejor traducción, y con ello, sentido correcto.

Basados en Diccionarios

- Las palabras usadas en las definiciones de una palabra suelen ser indicadores adecuados de sus sentidos.
- Algoritmo de LESK (1986):
  - Dada una palabra *w*, en un contexto *c*, y con sentidos $s_1,\ldots,s_k$.
  - Extraer del diccionario la bolsa de palabras correspondiente a cada sentido s_k.
  - Comparar cada bolsa de palabras con las palabras del contexto. Seleccionar el sentido con el mayor traslape.

Basados en Tesauros

- Las categorías semánticas de las palabras del contexto determinan la categoría semántica de todo el contexto, y a su vez el sentido de las palabras que lo conforman.
- Algoritmo de WALTER (1987):
  - Cada palabra tiene uno o varios categorías semánticas que corresponden a sus diferentes sentidos.
  - Para cada categoría semántica se cuentan las palabras que la disparan.
  - Se selecciona la categoría semántica con mayor conteo. Esta establece el sentido de todas las palabras.

La Ontología “WordNet”

- Una ontología de relaciones léxicas
- Se organiza en conjuntos de sinónimos:
  - Una palabra que tiene varios sentidos pertenece a varios SYNSETS (conjuntos de sinónimos)
- Considera varias relaciones entre SYNSETS
  - Hiponimia, hiperonimia, meronimia, etc.
- Incluye información sobre sustantivos, verbos y adjetivos+adverbios
- También incluye un conjunto de glosas (definiciones) por sysnset (¡sirven para método de Lesk!)

Basados en Ontologías

- La densidad de las palabras del contexto es mayor en el vecindario del sentido correcto.
- Algoritmo Aguirre-Rigau (1996):
  - Se identifican todos los nodos que corresponden a los distintos sentidos de la palabra a desambiguar *W* y de las palabras del contexto
  - Se mide la densidad conceptual alrededor de cada sentido de *W*
  - Se elige el sentido con la mayor densidad

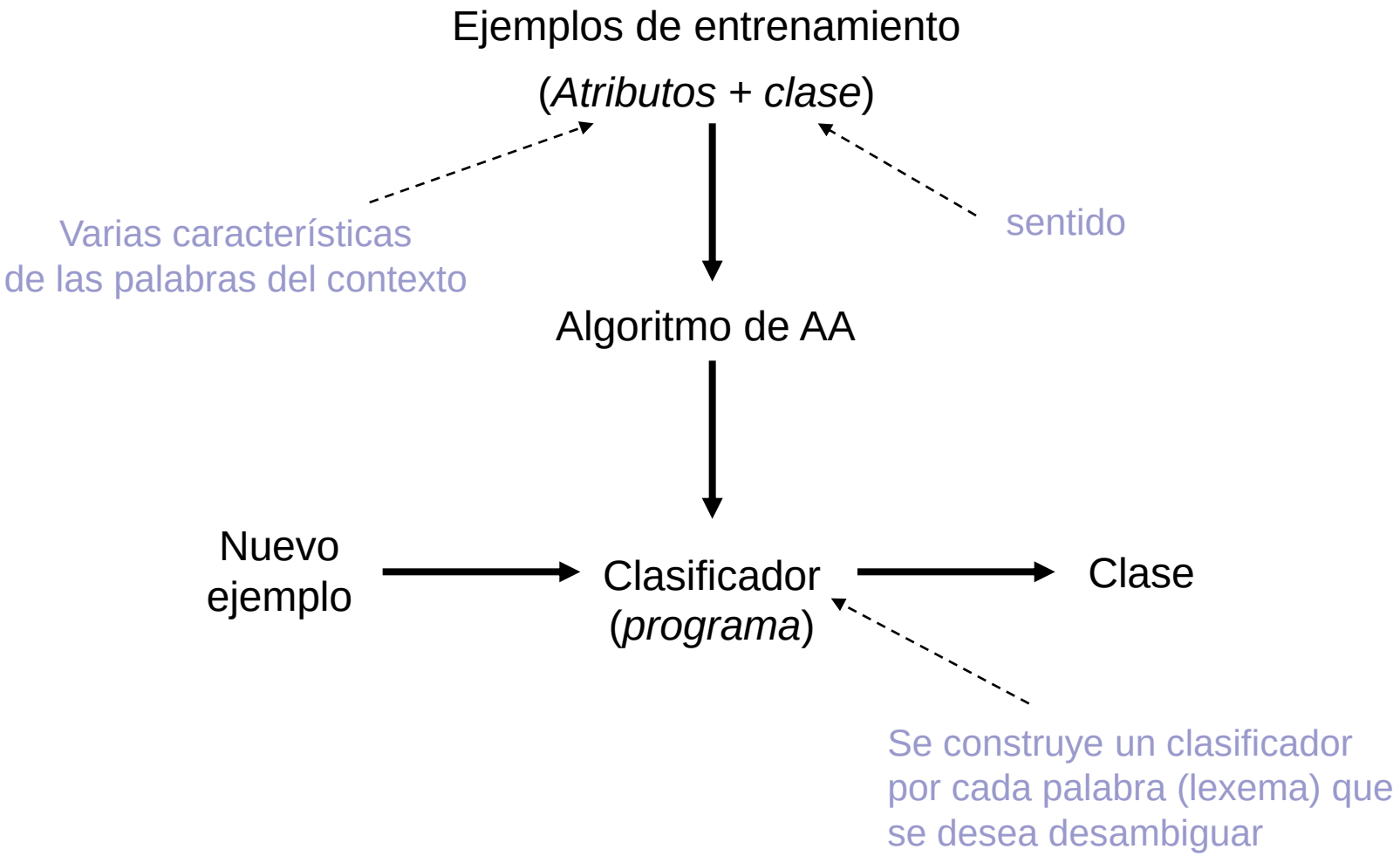
Basado en traducciones

- La desambiguación de las palabras puede realizarse considerando sus traducciones en otros lenguajes.
- Algoritmo de Dagan-Itai (1991):
  - Identificar en un corpus en un segundo lenguaje todas las traducciones de la palabra que se desea desambiguar.
  - Contar las veces que cada traducción ocurre junto a las traducciones de las palabras del contexto.
  - Seleccionar el sentido con la mayor cuenta.

Problemas/novedades del enfoque

- Gran problema: *pocas ontologías* disponibles.
  - Poco cobertura de conocimiento de dominio restringido.
  - Disponibles en unos cuantos lenguajes.
- Trabajo reciente en modificar métodos clásicos
  - Por ejemplo, en *Lesk* se miden similitudes de segundo orden, es decir, se usan las definiciones de los términos de la definición.
- Mucho trabajo en *enriquecimiento automático* de ontologías.
  - Idea es obtener de texto plano varias glosas por concepto, asi como extraer varias instancias por concepto.

Métodos supervisados



Principales características

- Requieren de corpus etiquetado
  - Por ejemplo *SemCor* o *Senseval*
- Aplican algoritmos de AA
  - Bayes y SVM son los más usados actualmente, aunque también se usan ensambles.
- Los atributos más utilizados son:
  - las palabras, sus lemas, POS, y posiciones.
- Más precisos que los basados en conocimiento, pero su aplicación se limita a pocas palabras.

Datos problemáticos

Sense	n-secmic Nouns	Average number of examples
1	9082	13.51
2	1368	4.61
3	544	3.68
4	228	3.55
5	117	3.24
6	59	2.74
7	43	3.52
8	22	3.13
9	8	3.17
10	4	2.33
>10	11	1.75

- *Alto grado de polisemia*, especialmente en verbos aunque también en sustantivos.
- Ejemplos por clase muy *desbalanceados*. La mayor parte se concentran en el primer sentido.

Métodos mínimamente supervisados

- La idea es tener el *menor número de ejemplos* etiquetados manualmente por sentido.
  - El etiquetado es muy costoso; además no puede hacerlo cualquiera.
- Se han usado algoritmos de *bootstrapping*
  - Co-training
  - Self-Training
- También se ha usado la *Web* para extraer automáticamente más ejemplos por sentido

Receta de Bootstrapping

- Ingredientes
  - Unos POCOS ejemplos etiquetados por sentido
  - MUCHOS ejemplos sin etiquetar
  - UNO o VARIOS (al gusto) clasificadores base
- Resultado
  - Un clasificador que MEJORA el rendimiento de los clasificadores base.

Procedimiento general

- Un conjunto *L* de ejemplos etiquetados
- Un conjunto *U* de ejemplos no-etiquetados
- Un conjunto de clasificadores {C_i}

1. Seleccionar un subconjunto de ejemplos *U'*
   - Se seleccionan aleatoriamente ***P*** ejemplos de *U*
2. Se hacen ***I*** iteraciones
   - Se entrena C_i con *L* y se etiqueta *U'*
   - Se seleccionan los mejores ***G*** ejemplos, y se agregan a *L*
   - Se vuelve a llenar *U'* con ejemplos de *U*

La Web como corpus

- Construir corpus anotado usando oraciones obtenidas de la Web. La idea es usar frases *monosémicas* para buscar estos ejemplos.
- Algoritmo Mihalcea (1999)
  - Extraer, usando algunas heurísticas, un conjunto de frases de búsqueda a partir de un diccionario.
  - Buscar en la Web usando las frases de búsqueda
  - Remplazar, en los ejemplos bajados, las frases de búsqueda por el sentido de la palabra correspondiente.
  - Almacenar los nuevos ejemplos.

Métodos NO supervisados

- Tarea conocida como *Discriminación del sentido de las palabras* (Pedersen and Bruce, 1997).
- El objetivo es agrupar palabras considerando la similitud de sus contextos. Se basa en las siguientes hipótesis:
  - Las palabras con significados similares tienden a ocurrir en contextos similares
  - Uno puede conocer el significado de una palabra por las palabras que la acompañan.

Métodos NO supervisados (2)

- Entonces, agrupan las palabras basándose en la similitud de su contexto.
- Cada grupo representa un sentido distinto, aunque no están etiquetados.
- El único recurso necesario es un conjunto de datos NO etiquetado, es decir, muchos ejemplos de uso de la palabra.
  - ¡y un buen *algoritmo de agrupamiento*!

Métodos NO supervisados (2)

- Ventaja: su independencia hacia los datos etiquetados.
- Desventajas principales:
  - Los grupos NO representan los verdaderos sentidos.
  - La cantidad de sentidos resultantes pueden variar.
  - Difícil medir similitud de (pequeños) contextos
    - Muchas veces no presentan palabras en común
    - Para mejorar esta medición se han usado representaciones de segundo orden (Purandare and Pedersen, 2004).

Evaluación

- Conjuntos de prueba estándar: SENSEVAL-3.
  - Tareas: etiquetar todas las palabras o sólo un pequeño conjunto predeterminado.
  - Para el segundo caso proporciona ejemplos de entrenamiento y prueba.
- Los puntos de referencia son:
  - La evaluación humana (etiquetas de corpus)
  - Sentido más frecuente

Métricas de Evaluación

- Precisión
  - ejemplos clasificados correctamente / ejemplos clasificados
- Recuerdo
  - ejemplos clasificados correctamente / total de ejemplos

SENSEVAL-3: Resultados

- 47 Sistemas participantes:
  - 38 en supervisados
  - 9 en basados en conocimiento (y algunos híbridos)

Categoría		Precisión	Recuerdo
MFS		**55.2%**	**55.2%**
Supervisados	**Mejor**	72.9%	72.9%
	Peor	78.2%	31.0%
	Promedio	67.5%	65.2%
Basados en conocimiento	**Mejor**	66.1%	65.7%
	Peor	19.7%	11.7%
	Promedio	44.0%	41.9%