

Adquisición automática de información léxica

Núria Bel
IULA – UPF
nuria.bel@upf.edu

N.Bel

Objetivo de la investigación

- ¿Qué?
 - Automatizar el proceso de desarrollo de léxicos computacionales
- ¿Por qué?
 - Porque es caro, cuesta tiempo, es difícil hacerlo bien y no se puede decir que ya está hecho para todo y para siempre.
- ¿Cómo?

N.Bel

Muchas preguntas por contestar ...

N.Bel

Esquema

- Antecedentes
- Un poco de información sobre léxicos computacionales y el problema de la cobertura
- Adquisición automática de información léxica
 - a partir de corpus
 - con la ayuda de léxicos ya codificados

N.Bel

Antecedentes

- Desarrollo de gramáticas computacionales de alto nivel para traducción automática e interfaces en lenguaje natural (Eurotra, Trade, Internat)
- Desarrollo de léxicos computacionales al tiempo que las gramáticas se volvían más léxicas: EAGLES, PAROLE, SIMPLE, ISLE, LS-gram, Melissa, etc.
- Multext (HMM) y Peking (clasificación automática con Winnow y SVM's)

N.Bel

Léxico computacional

- Un léxico computacional contiene información sobre cada ítem léxico (lema y lectura, normalmente) que es necesaria para hacer análisis/generación lingüística: codifica información bien que discrimina posibles interpretaciones, o que es obligatoria para expresar determinado contenido.

N.Bel

Léxico computacional

- La información sirve para codificar qué características de combinación y/o de interpretación tiene un elemento:

abierto_1 {cat=adj, prep_reg=a, modo=subj, control=csubj_subj, cop=estar}

“la puerta está abierta”

“estar abierto a que me pregunten”

“estar abierto a discutir”

abierto_2: {cat=adj, prep_reg=no, modo=no, control=no, cop=ser.}

“su carácter es abierto”

N.Bel

El costo de un léxico

- PAROLE (LE2-4017 IVFP): información morfológica y sintáctica de 20.000 entradas costó, en codificación, 27 p/m.: más de 2 personas durante un año a dedicación completa

(<http://www.ub.es/gilcub/SIMPLE/simple.html>)

- Paralelo DescAdjserApcA

- Es un léxico de uso general que no está completo ni en número de entradas ni en diferentes descripciones

N.Bel

Cobertura del léxico

- ¿Cómo de general puede/ha de ser un léxico computacional? ¿qué lecturas/usos no ha de recoger? ¿hemos de tener léxicos con enumeración de sentidos y usos posibles?
- ¿Hay alguna teoría que pueda predecir cualquier y todos los usos?
- Con cada nuevo dominio, hay que ajustar el léxico:
 - por razones de productividad léxica
 - por razones prácticas, para reducir las entradas y la ambigüedad

N.Bel

Cobertura del léxico: Un ejemplo práctico

- Codificación del adjetivo “paralelo” = DescAdjserApcA
- En la gramática:
 - Se asume que el complemento es opcional, pero habrá que decidir, para cada oración, si está o no está.
 - En el caso de que haya en la oración un SP con la preposición ‘a’ habrá que decidir si es el complemento del adjetivo o no. En muchos casos habrá dos posibles interpretaciones:

“una posición paralela a la mesa”

N.Bel *“indicar la posición de la salida paralela al usuario”*

Cobertura del léxico: Ejemplos de corpus

- Paralelo=DescAdjserApcA
- Corpus de medicina:
 - 70/199 (0,35%) junto a un SP_A
 - 22/199 junto a ‘ser’
- Corpus de informática:
 - 7/80 (0,0875%) junto a un SP_A
 - 0/80 junto a ‘ser’
- ¿son relevantes estos datos?

N.Bel

Adquisición automática

- Sería ideal poder generar automáticamente los léxicos computacionales: “plug and play”
- Hay antecedentes:

N.Bel

Trabajos en adquisición con métodos puramente estadísticos

- Patrones de subcategorización verbal:
 - Brent 1991, 1993
 - Ushioda et al. 1993
 - Briscoe & Carroll 1997
- información semántica:
 - Riloff & Shepherd, 1997, 1999
 - Church & Hanks 1990
 - Grefenstette & Hearst, 199X

N.Bel

Con métodos de aprendizaje automático (Machine Learning)

- Daelemans & Durieux, 2000: "Inductive Lexica", en F. Van Eynde & D. Gibbon. *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers.
- "Inductive lexica: to use an **available lexicon**, and a **corpus**, as a source to bootstrap lexical acquisition. Lexical predicates of newly encountered words are computed by reference to similar words previously encountered, for which the lexical information wanted is available."

N.Bel

Estudiando el problema

- ¿qué se puede sacar del texto?
- ¿qué podemos tener en el léxico?

N.Bel

¿Qué información hay en el texto? ¿es suficiente y necesaria?

- Podemos extraer información relacionada con la **coaparición** (interrelación) de elementos (propiedades distribucionales).
- A diferentes escalas, dependiendo del nivel de anotación del corpus, ¿qué nivel de anotación es suficiente y necesaria para **inducir** la información de un léxico?
- Es importante formular hipótesis que no obliguen a tener un alto nivel de contenido lingüístico en el corpus, por la dificultad práctica de codificarla.

N.Bel

Codificación de los datos del corpus: Vector de distribución

- Podemos trabajar con la información de distribución convirtiéndola en un vector de distribución:
 - Datos obtenidos a partir de Expresiones Regulares sobre el corpus: número de veces que se da un determinado contexto
- <3,67,22,1,130,69,78,121,0,17,6,2,14,70>
<4,30, 0,0, 58,22,25, 55,0, 4,7,2, 4, 7>
(Pre,post,ser,estar,pl,sing,f,m,Q,con,en,para,de,a)
 - En este ejemplo son números absolutos, pero hay que estudiar qué y cómo lo representamos.

N.Bel

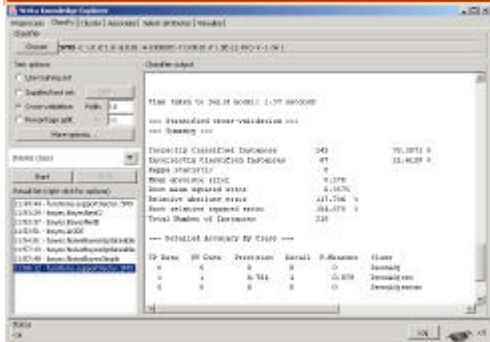
El léxico básico

- Teniendo un léxico básico ya codificado, podemos preparar los ejemplos para enseñar al sistema. Tomamos datos de clasificación de Parole, y de distribución del corpus:

<4,30,0,0,58,22,25,55,0,4,7,2,4,7, DescAdjsr>
<0,0,1,0,1,0,0,1,0,0,0,0,0, DescAdjsr>
<1,7,1,0,14,4,2,16,0,0,0,0,0,1, DescAdjsr>

N.Bel

Ejemplo de experimento con el paquete Weka



Diseño de los experimentos

- ¿Qué observaciones/datos distribucionales necesitaremos para cada rasgo?
- ¿Qué es ruido y qué información?
- ¿Se puede aprender más de un rasgo a la vez?

N.Bel

Diseño de los experimentos

- ¿Qué implicación puede tener para el modelo del léxico que estamos utilizando?
Si el corpus no está anotado con diferentes lecturas, estamos tomando datos de la pieza léxica sin más distinciones, tenemos una "meta-entrada", es decir no distinguimos entre 'estar blando' y 'ser blando', pero, ¿podemos incluirlo en el modelo?
- Podemos utilizar aproximaciones como la de Zipf que dice que el número de lecturas es proporcional a la raíz cuadrada de su frecuencia relativa.
- Pero, ¿se puede aplicar también a la terminología?
En contradicción con algunos trabajos para texto especializado.

N.Bel

Confianza en las observaciones/medidas

- ¿Qué grado de confianza puedo tener en las observaciones en un corpus ocasional?
- ¿y en la asignación de clase correspondiente?
 - Aplicar técnicas sobre la incertidumbre o indeterminación de las medidas.
Probabilidad bayesiana

N.Bel

Un posible modelo sobre confianza: Probabilidad bayesiana

- Es una manera de calcular la incertidumbre.
- La probabilidad bayesiana es como hacer una regla de 3 con probabilidades

N.Bel

Incertidumbre e inferencia bayesiana

- **Nos permite calcular la probabilidad de una hipótesis dados unos datos relevantes.**
- ¿qué probabilidad hay de que si veo un adjetivo n veces como atributo de 'ser', la hipótesis de que sólo sea *DescAdjs* sea correcta, al menos para ese corpus/dominio?
 - d : datos: veces que lo he visto
 - h : hipótesis: es *DescAdjs*
 - $p(h|d)$ [probabilidad de que la hipótesis sea verdadera dados unos datos d]

N.Bel

Incertidumbre e inferencia bayesiana

- $p(h|d)$ sólo se puede calcular si:
 - sabemos la probabilidad a priori de la hipótesis
 - $p(\text{DescAdjser}) = \text{¿cómo se puede calcular?}$
 - qué probabilidad hay de que siendo cierta la hipótesis hubiéramos visto esos datos
 - $p(d|h) = \text{¿cómo se puede calcular?}$

- Pero, si los tenemos
$$p(h|d) = \frac{P(h) \cdot P(d|h)}{P(d)}$$

N.Bel

Conseguir la probabilidad a priori $p(\text{DescAdjser})$

- Habría que hacer las mediciones en un léxico inicial, sería el 'léxico básico'
- podemos medir la probabilidad a priori de la hipótesis: ¿cuántos adjetivos de los codificados en Parole son *DescAdjser*?

N.Bel

Calculando $p(d|h)$

- Tenemos también los datos que hemos sacado del corpus y los adjetivos codificados, pero hemos de buscar una manera de computar la probabilidad a partir de los datos de distribución.
- Si lo conseguimos podremos modelar la incertidumbre.

N.Bel

Extensión de la inferencia bayesiana: Redes bayesianas

- Hay información que es fácilmente derivable de observaciones en el corpus (con mayor o menor error), pero hay otros rasgos que no pueden inducirse directamente.
- ¿Hemos de renunciar a intentar inducirlos o podemos inducir información sobre un rasgo A a partir de la información que tenemos sobre otros rasgos: B,C,D, etc.?
- Podemos explotar la idea de hacer una regla de tres con probabilidades si tenemos la probabilidad a priori de todos los elementos de nuestro léxico.

N.Bel

Red bayesiana

- La probabilidad conjunta de todo el léxico es difícil de obtener debido a los cálculos.
- Una red bayesiana es un Grafo Dirigido Acíclico (DAG) que especifica las probabilidades condicionadas de diferentes eventos dadas sus relaciones entre ellos. Los nodos representan las variables del problema que se desea resolver. El conocimiento del problema se representa mediante la instanciación de aquellos nodos cuyo valor es conocido, propagándose tal conocimiento a través de la red mediante ciertas reglas probabilísticas.

N.Bel

Ejemplo de red



N.Bel

