Colaboración entre información paradigmática y sintagmática en la Desambiguación Semántica Automática

Iulia Nica
D. Lingüística General
Universidad de Barcelona
Universidad de Iasi, Rumania
iulia@clic.fil.ub.es

Mª. Antònia Martí i Antonin
D. Lingüística General
Universidad de Barcelona
amarti@clic.fil.ub.es

Andrés Montoyo y Guijarro
DLSI
Universidad de Alicante
montoyo@dlsi.ua.es

Resumen. Proponemos un método alternativo para la desambiguación semántica automática, centrado en la interacción entre la información sintagmática y paradigmática. Se toma como unidad en el proceso de desambiguación una ocurrencia ambigua integrada en un patrón sintagmático. La estrategia no necesita corpus etiquetado al nivel de sentido, presupone tan sólo un análisis previo de tipo morfosintáctico y agrupación por *chunks*, no usa información estadística y su potencial desambiguador es amplio. Ilustramos las dos implementaciones propuestas con ejemplos concretos y estudiamos posibilidades de refinamiento del método.
Palabras clave: desambiguación semántica automática, etiquetación semántica

Abstract. We propose an alternative method for Word Sense Disambiguation, based on the interaction between syntagmatic and paradigmatic information. The unit of the disambiguation process is taken to be an ambiguous occurrence integrated into a syntagmatic pattern. The strategy needs not a semantically annotated corpus, it supposes only a morphological analysis and chunking, does not make use of statistical information and has en wide disambiguating potential. We illustrate the two implementations proposed with concrete examples and study ways for refinement.
Key words: Word Sense Disambiguation, semantic annotation

1. Introducción

En este trabajo[*] presentamos parte de una investigación más amplia sobre el uso intensivo de conocimiento lingüístico en el proceso de Desambiguación Semántica Automática (DSA)1. En concreto, para esta tarea nos centramos en cuestiones esenciales, como la caracterización de los sentidos de una palabra polisémica, el acercamiento entre el lexicón y el corpus y por último la identificación y el tratamiento del contexto relevante para la asignación del sentido de una ocurrencia ambigua.

Existen diferentes sistemas de DSA según la información de referencia que se utiliza para la identificación de los sentidos: los sistemas DSA basados en conocimiento (*Knowledge-Driven WSD*), que utilizan fuentes léxicas estructuradas (diccionarios accesibles por ordenador, ontologías, etc.), y los basados en corpus (*Corpus-Based WSD*), que utilizan ejemplos etiquetados a nivel de sentido.
En las dos competiciones SENSEVAL, los sistemas DSA basados en corpus obtuvieron mejores resultados debido al uso de información sintagmática. Estos sistemas plantean un problema de cobertura, ya que necesitan grandes cantidades de ejemplos etiquetados con sentidos para su correcto funcionamiento. La obtención de textos etiquetados semánticamente es difícil, aunque hay trabajos como (Mihalcea y Moldovan, 1999) que intentan adquirirlos de manera automática. Los sistemas de DSA basados en conocimiento usan la información, principalmente paradigmática, de fuentes

[*] La investigación ha sido posible gracias a una beca predoctoral MAE.
[1] Varios experimentos desarrollados últimamente confirman que la calidad de la DSA depende más de la información utilizada que de los algoritmos que la explotan: Pedersen, 2002, Yarowsky y Florian, 2002.

léxicas, por lo cual tienen limitaciones debido al "vacío" que hay entre el lexicón y el corpus (Kilgarriff, 1998). Para solucionar el problema, las propuestas actuales acercan las fuentes léxicas a los corpora mediante la incorporación de información sintagmática (Véronis, 2001). El enfoque supone un proceso costoso de extracción y de representación de tal conocimiento.

El contexto que se puede usar para la desambiguación de una ocurrencia ambigua se suele diferenciar en dos categorías básicas: contexto local y contexto tópico. El contexto local, que es el objeto de nuestro análisis, se ha definido como una ventana centrada en la ocurrencia por desambiguar, usualmente de dimensión predefinida.

Su explotación para la desambiguación semántica se ha realizado principalmente mediante un enfoque "bolsa de palabras": se toman en consideración sólo las palabras de contenido léxico y se ignoran las palabras funcionales. Sin embargo, cada vez son más frecuentes los sistemas que tienen en cuenta las palabras funcionales que vinculan la ocurrencia ambigua a las demás unidades de contenido léxico presentes en el contexto considerado. Los resultados demuestran la validez de la opción[2].

Así, se han usado las palabras funcionales para identificar, en un corpus etiquetado a nivel de sentido, bigramas o trigramas alrededor de la ocurrencia ambigua, que pueden aportar información sobre su sentido (Yarowksy, 1993, Pedersen, 2001).

En un nivel de análisis más profundo, los elementos funcionales permiten identificar las relaciones sintácticas con las demás unidades de contenido léxico de la oración. Los sistemas de DSA que usan estas relaciones lo suelen hacer en un enfoque basado en ejemplos, con lo cual necesitan un corpus etiquetado sintáctica y semánticamente. La información sintáctica usada para la DSA se ha limitado en general a las relaciones verbo-sujeto y verbo-objeto (Ng, 1996, Leacock *et al.,* 1998, Federici *et al*., 2000, Agirre y Martínez, 2001, etc.), con pocas excepciones (Lin, 1997, Stetina *et al*., 1998, etc.).

En el presente trabajo proponemos un enfoque alternativo para la DSA, centrado en la interacción entre la información sintagmática y paradigmática que caracteriza el lenguaje natural: la estrategia "P&S". Introducimos la noción de "patrón sintagmático" que engloba ambos tratamientos del contexto local, n-gramas y relaciones sintácticas, y tomamos como unidad en el proceso de desambiguación una ocurrencia integrada en un patrón sintagmático.

De esta manera, en nuestra aproximación tratamos de superar tanto el problema de la falta de datos de los sistemas basados en ejemplos como el vacío entre el lexicón y el corpus, aproximando éste último al léxico.

La propuesta presupone tan sólo un análisis previo de tipo morfosintáctico y una agrupación por *chunks*[3]. Inicialmente el método se ha aplicado para la desambiguación de nombres, pero igualmente se puede adaptar a la desambiguación de otras categorías sintácticas. En este trabajo se ha aplicado la estrategia sobre el castellano.

Después de esta introducción, en el apartado 2 se describe la propuesta con las modalidades de aplicación; en el apartado 3 se discuten posibles mejoras; finalmente, en el apartado 4 se definen las conclusiones y las líneas de investigación futura.

2. Propuesta

2.1. Aproximación

Nuestro enfoque a la desambiguación léxica parte de las consideraciones que presentamos a continuación sobre a) la caracterización de los sentidos, b) el contexto local, y c) la distancia entre el lexicón y el corpus.

a) Para la información de referencia sobre los sentidos, hemos utilizado el componente español de EuroWordNet (Vossen, 1998). Este componente se utiliza en su versión

[2] Yarowsky y Florian, 2002, Mihalcea, 2002, Hoste et al., 2002.

[3] Civit, 2003.

estándar para un tipo de implementación (M1, apartado 2.2.1.), y hemos desarrollado una adaptación para otro tipo de implementación (M2, apartado 2.2.2.).

En esta adaptación, cada uno de los sentidos de una palabra polisémica se caracteriza mediante el conjunto de *variants* de los *synsets* con los que está en relación, y que no comparte con ningún otro sentido de la palabra. Así, para cada sentido X_i de una palabra X, se extrae de EWN el conjunto de *synsets* con que se relaciona, y de éstos los *variants* que contienen. Con lo cual, a los sentidos X_i de X se les asocian respectivamente los conjuntos V_i de *variants*. Se eliminan los *variants* que aparecen en más de un conjunto V_i, de modo que las palabras de los conjuntos reducidos Di son específicas para cada uno de los sentidos correspondientes X_i y se convierten en discriminadores de sentido[4].

Por ejemplo, *órgano* tiene cinco sentidos en EWN[5]:

*órgano*_1: 'parte de una planta';
*órgano*_2: 'agencia gobernamental, instrumento';
*órgano*_3: 'parte funcional de un animal'
*órgano*_4: 'instrumento musical'
*órgano*_5: 'periódico'

Los conjuntos disjuntos que los caracterizan, extraídos de EWN, son respectivamente:

D1: {*órgano vegetal_1, lámina_3, raíz_2, tronco_4, troncho_1, tallo_1, pedúnculo_1, hoja_3, …*}
D2: {*oficina_2, agencia_2, unidad administrativa_1, organización_1, grupo social_1, colectivo_1, ...*}
D3: {*parte del cuerpo_1, trozo_8, porción_3, parte_9, lóbulo_2, patas_3, lengua_3, ojo_4, ...*}
D4: {*instrumento de viento_1, instrumento musical_1, mecanismo_3, aparato_3, teclado_1, ...*}
D5: {*publicación_2, periódico_4, medio de comúnicación_1, manera_4, obra_5, ...*}

b) El contexto local debe ser delimitado de manera distinta para cada ocurrencia, y según criterios lingüísticos. Nuestra hipótesis es que, excepto la información de tipo temático, el sentido de una palabra en una oración está determinado esencialmente por las relaciones sintácticas que ésta establece con las demás palabras de la oración.

Debido a que nos centramos en la desambiguación de nombres, consideramos una buena aproximación a las relaciones sintácticas el contexto local estricto: las palabras que se encuentran inmediatamente antes y después de la palabra en cuestión, hasta la siguiente palabra de contenido léxico. Identificamos el contexto local de una ocurrencia ambigua con la suma de los patrones sintagmáticos en que participa.

Un patrón sintagmático se define formalmente como una tripleta que corresponde a una relación sintáctica, formada por dos unidades de contenido léxico y un patrón léxico-sintáctico R que expresa la relación (de dependencia o de coordinación) que contraen las dos unidades léxicas:

L1 - R - L2.

Incluimos en este patrón general el caso en que R es nulo, como en la relación entre un nombre y un adjetivo[6]. Ejemplos: *grano*-n *de*-prep *azúcar*-n; *pasaje*-n *subterráneo*-adj.

Desde esta perspectiva, una palabra polisémica, y cada uno de sus sentidos, se podrán caracterizar mediante los patrones sintagmáticos en que participan.

En concordancia con la manera misma de delimitar el contexto local, nuestra aproximación toma en consideración las palabras funcionales, distanciándose del enfoque "bolsa de palabras".

c) Para reducir la distancia entre el lexicón y el corpus, evitando el esfuerzo que supone la incorporación de información sintagmática en las fuentes léxicas, hemos explorado la vía opuesta: el acercamiento del corpus al lexicón. La información implícita en los corpora, creemos, es explotable para la DSA mediante diferentes agrupaciones de palabras. Una manera de concretar esta segunda opción puede ser la explotación de la interacción semántica entre los ejes paradigmático y

[4] La extracción de la información para caracterizar los sentidos es completamente automática.

[5] Las pseudodefiniciones son nuestras.

[6] Ofrecemos una definición genérica y aproximativa de lo que entendemos por patrón sintagmático. Su delimitación no es trivial, y debe ser sujeta a restricciones sintácticas y semánticas. Es fuera del alcance de esta presentación, y será objeto de nuestro estudio futuro.

sintagmático del lenguaje[7]. Las condiciones sintagmáticas identicas delimitan conjuntos de tipo paradigmático de palabras. Inversamente, las palabras afines paradigmáticamente (por ejemplo, mediante las relaciones léxico-semánticas del EWN) se sustituyen recíprocamente en la cadena enunciativa.

2.2. Estrategia de DSA

A partir de estas consideraciones, proponemos una modalidad distinta de desarrollar el proceso de DSA: se toma como unidad de referencia en el proceso de desambiguación la ocurrencia ambigua integrada en un patrón sintagmático y no la ocurrencia ambigua aislada. Esta integración constituye el elemento clave de nuestra propuesta: sobre la base de los patrones sintagmáticos se realiza la transición entre los ejes sintagmático y paradigmático.
La estrategia se funda en las siguientes hipótesis:
H0: Las ocurrencias de una palabra ambigua en una determinada posición de un patrón sintagmático fijado tienen el mismo sentido (hipótesis que podemos denominar "*one sense per syntagmatic pattern*", y que proponemos como alternativa a la hipótesis "*one sense per collocation"* (Yarowsky, 1993).
H1: Las diferentes palabras que pueden aparecer en una determinada posición de un patrón sintagmático fijado tendrán sentidos relacionados, pertenecientes a una zona conceptual común.
H2: Dos palabras con sentidos relacionados son conmutables en un mismo patrón sintagmático.
La hipótesis H0 permite el paso de una ocurrencia aislada a una ocurrencia integrada en un patrón sintagmático como unidad de desambiguación; H1 y H2 proyectan los ejes sintagmático y paradigmático el uno en el otro y son la base de las implementaciones M1 y M2 respectivamente, que se explican a continuación.
- Modalidad M1: Se identifica en un corpus el conjunto de posibilidades para la posición de la ocurrencia ambigua en el patrón sintagmático, lo que define una clase de tipo paradigmático. Sobre esta clase se aplica un algoritmo de desambiguación que se basa en las relaciones paradigmáticas de EWN.
- Modalidad M2: Se sustituye la ocurrencia ambigua en el patrón sintagmático por cada una de las palabras de los conjuntos que caracterizan sus sentidos. Se verifica en el corpus la existencia de los patrones obtenidos para cada conjunto, como indicador para la identificación del sentido correcto.
Ambas implementaciones se basan en la colaboración entre fuentes de conocimento y corpora, sin necesidad de ejemplos etiquetados al nivel de sentido. Por lo tanto, se trata de desambiguación semántica no supervisada, basada en conocimiento.
Detallamos a continuación las dos posibilidades de aplicación de la estrategia.

2.2.1. La modalidad M1

En nuestros experimentos correspondientes a la modalidad M1 usamos como corpus el CREA (RAE)[8], como fuente léxica el EWN, y como heurística de DSA, la Marca de Específicidad Común, MEC (Montoyo y Palomar, 2000), que usa la información paradigmática de EuroWordNet en un enfoque "bolsa de palabras". La base intuitiva del algoritmo es: cuanta más información común comparten dos conceptos, más relacionados estarán. En EWN, la información común que comparten esos dos conceptos corresponde al concepto padre de

[7] "Syntagmatic sense relations [...] are an expression of coherence constraints. Paradigmatic sense relations, on the other hand, operate within the sets of choices. Each such set represents the way the language articulates, or divides up, some conceptual area, and each displays a greater or lesser degree of systematic structuring. Paradigmatic relations are an expression of such structuring. [...] Paradigmatic and syntagmatic relations function in tandem, syntagmatic relations delimiting the space within which paradigmatic relations operate." (Cruse, 2000: 149)

[8] http://www.rae.es/

ambos en la jerarquía (Marca de Específicidad, ME). La heurística toma como entrada las palabras no funcionales del contexto oracional de la ocurrencia ambigua, incluida la palabra en cuestión, y busca aquella ME en EuroWordNet que tenga mayor densidad de palabras de entrada debajo de su subárbol. Se elige como sentido de la ocurrencia ambigua el que se encuentra en el subárbol de la ME así identificada.

La modalidad M1 consiste en la siguiente secuencia de operaciones para la desambiguación de una ocurrencia del nombre polisémico X:

Paso 1º. Se establecen los patrones sintagmáticos S_k en que la ocurrencia ambigua de X aparece en la oración.

Paso 2º. Para cada patrón S_k identificado:

- Se buscan, en el corpus, los nombres que pueden aparecer en el patrón S_k como sustitutos de X. Se obtiene así un paradigma P_{Sk}.
- Se aplica la heurística de DSA (MEC) sobre el paradigma P_{Sk}.

Paso 3º. Se establece el sentido de la ocurrencia ambigua X, corroborando las propuestas para su sentido obtenidas en 2º.

Contrastamos (a) la aplicación estándar de la heurística MEC con (b) nuestra propuesta para la desambiguación de la ocurrencia resaltada del nombre *órgano* en la siguiente oración del corpus CREA (RAE):

Entre sus composiciones más célebres destacan las obras para ***órgano*** *"La Natividad del Señor" (1935), "Los cuerpos gloriosos" (1939), "Misa de Pentecostés" (1950), "Libro de órgano" (1952), obras corales como "Tres pequeñas liturgias de la presencia divina" (1944), y para piano, "Veinte miradas sobre el Niño Jesús" (1944) y "Cuatro estudios de ritmo" (1949).*

a) En la estrategia estándar, la heurística MEC toma como input el conjunto de nombres del contexto oracional, y asigna a *órgano* el sentido 5 ('periódico'), que es inadecuado.

b) En la estrategia por patrones sintagmáticos, se identifica el patrón en que participa la ocurrencia ambigua:

OBRA(S) PARA ÓRGANO.

Se buscan en el corpus los nombres que se pueden alternar con ÓRGANO en el patrón:

OBRA(S) PARA N(sg),

obteniendo el paradigma: {*órgano, piano, guitarra*}. El algoritmo MEC toma ahora como input las palabras de este paradigma, y desambigua correctamente *órgano*_4, y además *piano*_2, *guitarra*_1, como instrumentos musicales.

Aun más, a partir del patrón identificado se pueden operar generalizaciones de formas a categorías sintácticas. Una generalización sería *órgano* integrado en el patrón:

N PARA ÓRGANO.

La aplicación de la estrategia sobre esta variante se realiza en los siguientes pasos:

- Se determinan los nombres que aparecen en este patrón en la posición de N: N_i.
- Para cada Ni, se buscan en el corpus los nombres en la posición de N' (ÓRGANO) en el patrón:

$\mathbf{N}_i$ PARA **N'**.

Se obtiene el conjunto de nombres $\{N'_{ij}\}$.

- Para cada i, se aplica la heurística sobre los nombres N'_{ij}.

Así, los N_i hallados son: *concierto, obra, pieza*. Sustituyendo estos nombres en el patrón en la posición N y buscando en el corpus los nombres N', se obtienen los paradigmas correspondientes y, como resultado de la aplicación del algoritmo MEC, las asignaciones de sentido:

- Para *N1 = concierto:*

Patrón: CONCIERTO(S) PARA **N'**

Paradigma: {*piano, violín, guitarra, solista, órgano, clarinete, ...}*

Aplicación MEC: *piano_2, violín_2, guitarra_1, solista_1, órgano_4, clarinete_2, ...*

- Para *N2 = obra:*

Patrón: OBRA(S) PARA **N'**

Paradigma: *{piano, guitarra, órgano, ...}*

Aplicación MEC: *piano_2, guitarra_1, órgano_4, ...*

- Para *N3=pieza:*

Patrón: PIEZA(S) PARA **N'**

Paradigma: *{orquesta, piano, clarinete, órgano, ...}*

Aplicación MEC: *orquesta_2, piano_2, clarinete_2, órgano_4, ...*

Los sentidos asignados son correctos (instrumentos musicales), excepto en algunos pocos casos, para cuyo tratamiento se proponen soluciones en el apartado 3.

En el mismo proceso, se desambigua además cualquier combinación entre dos nombres en las posiciones N y N', en el marco del patrón sintagmático generalizado

N PARA **N'**,

donde N pertenece al conjunto unificado de los Ni (*concierto, obra, pieza*), y N', al conjunto unificado de los N'ij (*piano, violín, guitarra, solista, órgano, clarinete,…*).
Nos acercamos, en esta extensión, a propuestas como las de (Federici *et al*., 2000; Agirre y Martínez, 2001), en que se combinan variantes paradigmáticas para las dos posiciones léxicas. Sin embargo, en estos trabajos la combinación se realiza sólo para las relaciones verbo-(sujeto/objeto) sobre un corpus ya etiquetado sintáctica y semánticamente.

2.2.2. La modalidad M2

Utilizamos, como información de referencia sobre los sentidos los conjuntos disjuntos D_i extraídos de EWN (apartado 2.1.), y como corpus, LEXESP (Sebastián *et al*., 2000).
La "prueba de conmutabilidad", basada en la hipótesis H2, se define de la manera siguiente: partiendo de una ocurrencia del nombre por desambiguar, integrado en un patrón sintagmático, el algoritmo lo sustituye por cada una de los nombres presentes en los conjuntos D_i asociados a sus sentidos, y busca en el corpus ocurrencias del patrón sintagmático obtenido en cada sustitución. Si se encuentra en el corpus alguno de los patrones sintagmáticos obtenidos con los nombres de D_i, a la ocurrencia ambigua se le asigna el sentido i correspondiente al respectivo conjunto D_i.
Los pasos que se siguen, para una ocurrencia del nombre polisémico X, son los siguientes:
Paso 1º. Se identifican los patrones sintagmáticos S_k en que participa X.
Paso 2º. Para cada patrón S_k identificado y para cada sentido X_i de X:

- Se sustituye X en el patrón S_k por cada uno de los elementos d_{ij} (j variable) de D_i. Se obtendrán las variantes particulares S_{kij} del patrón S_k.
- Se buscan ocurrencias de S_{kij} en el corpus. Se cuentan las ocurrencias n_{ij} identificadas de cada variante S_{kij}.
- Se suman las ocurrencias identificadas n_{ij} para cada i (j variable): N_i.

En base del numero N_i de ocurrencias encontradas en el corpus para las variantes correspondientes a cada conjunto D_i, se establece el sentido que se obtiene usando el patrón S_k.
Paso 3º. Se establece el sentido de la ocurrencia ambigua, corroborando las propuestas de los diferentes patrones individuales S_k en el paso 2°.
Hemos realizado esta secuencia de operaciones para la desambiguación de *órgano* en el patrón:

TRÁFICO DE ÓRGANOS,

y hemos encontrado en el corpus un único patrón obtenido mediante la sustitución de *órgano* por las palabras de los conjuntos que caracterizan sus sentidos: *tráfico de ojos.* En base de este resultado, se asigna a *órgano* el sentido 3, y a *ojo* el sentido 4, ya que *ojo_4* pertenece al conjunto D_3 asociado a *órgano*.

3. Discusión

a) Hemos realizado un estudio de caso sobre el nombre *órgano*, en una variedad amplia de patrones sintagmáticos, con el objetivo de estudiar cuestiones fundamentales de nuestra propuesta: el impacto sobre la calidad de la desambiguación; los problemas de cada implementación; posibilidades de refinamiento; la comprobación de las hipótesis iniciales y eventuales revisiones.
Nos hemos centrado en el paso 2º, que es fundamental para la estrategia, en ambas modalidades. En este paso se considera un sólo patrón sintagmático a la vez. Respecto de esta reducción, el paso 3º se puede ver como una generalización.
Las principales observaciones relacionadas con la implementación de la estrategia son:
a1) Hay patrones sintagmáticos más "estrictos" (OBRA(S) PARA ÓRGANO) y otros más "débiles" (ÓRGANO DEL CUERPO). Los segundos no son suficientemente restrictivos, y el paradigma que se obtiene es heterogéneo, con lo cual la identificación del sentido no es satisfactoria.
Para delimitar subconjuntos homogéneos en este paradigma se pueden aplicar restricciones: que los elementos del paradigma compartan más de una palabra en la otra posición del patrón sintagmático, o que tengan la misma etiqueta de dominio.

Ejemplificamos el impacto del primer filtro sobre el paradigma obtenido en el apartado 2.2.1. para el nombre *obra* en el patrón OBRAS PARA ÓRGANO. Siendo el paradigma inicial:
{*arpa, atención, canto, cello, clarinete, clave, cobla, conjunto, coro, cuerda, febrero, flauta, gente, guitarra, noviembre, órgano, orquesta, percusión, piano, quinteto, teatro, tenor, violín, violonchelo, voz*},
se obtiene el conjunto homogéneo:
{*arpa, cello, clarinete, clave, flauta, guitarra, orquesta, órgano, piano, quinteto, violín, violonchelo, voz*},
adecuado al contexto oracional y sobre el cual el algoritmo MEC da óptimos resultados.
a2) Para los patrones sintagmáticos de coordinación, con dos nombres y una conjunción, hemos probado dos modalidades de aplicar el algoritmo MEC: vertical, sobre los nombres del paradigma, y horizontal, sobre los dos nombres del patrón sintagmático. Si la aplicación vertical se enfrenta con el mismo problema de heterogeneidad del paradigma, la aplicación horizontal lleva a resultados correctos, aunque la desambiguación puede ser parcial.
a3) La implementación M2, sobre todo, se ve afectada primero por la escasez de datos en la fase de búsqueda en el corpus. Además del uso de un corpus más grande o de la web, vemos como posible solución el relajamiento de la prueba de conmutabilidad: se buscan ocurrencias de los nombres de los conjuntos D_i integradas no solo en el patrón de partida sino también en variantes de este patrón, sustituyendo la otra palabra léxica por una palabra relacionada con ella en EWN.

b) El potencial desambiguador del patrón es amplio:
b1) la desambiguación de una ocurrencia ambigua integrada en un patrón sintagmático será supuestamente válida para todas las ocurrencias de la palabra respectiva en el mismo patrón en cualquier otra oración;
b2) en este proceso se desambiguan a la vez ocurrencias de las diferentes palabras alternativas a la palabra de partida en el patrón dado;
b3) si se desambiguan ambas palabras léxicas del patrón, se desambiguan implícitamente cualquier combinación de sus posibles sustitutos en el patrón inicial.
c) Una característica importante de nuestra propuesta es la no dependencia de un corpus etiquetado al nivel de sentidos y de funciones o relaciones sintácticas. La flexibilidad respecto del corpus utilizado permite que los resultados puedan mejorar de manera continua, en paralelo con la ampliación de los corpora, ya que se va reduciendo el problema de la escasez de datos.

4. Conclusiones e investigación futura

Hemos propuesto una estrategia para la DSA que se basa en la colaboración entre la información sintagmática, presente en los textos, y la información paradigmática de las fuentes léxicas. El método toma como unidad por desambiguar una ocurrencia ambigua integrada en un patrón sintagmático.
El potencial del proceso de desambiguación es amplio, ya que permite la desambiguación a la vez de varias ocurrencias de una misma palabra en cualquier texto, y de varias palabras relacionadas con la primera en base a un patrón sintagmático.
La estrategia no necesita un corpus etiquetado al nivel de sentido ni al nivel sintáctico, con lo cual es un método de desambiguación totalmente automático.
La prioridad absoluta de nuestra labor futura es la aplicación a gran escala, para una evaluación real de la propuesta. Se compararán las dos modalidades por separado, y luego se combinarán en un sistema único.
Interesan las posibles generalizaciones: por una parte, considerar los varios patrones en que participa una ocurrencia ambigua; por otra, pasar de variantes flexivas a lemas, y de lemas a categorías sintácticas.
Las aplicaciones inmediatas son la obtención de ejemplos etiquetados al nivel de sentido, y la ampliación del conocimiento asociado a los sentidos.

5. Agradecimientos

Esta investigación ha sido financiada por la Agencia Valenciana de Ciencia y Tecnología (OCyT) con el proyecto CTIDIB/2002/151.

6. Bibliografía

Agirre, E. y D. Martínez, 2001. Learning class-to-class selectional preferences, en *Proceedings of the ACL CONLL'2001 Workshop*, Tolouse

Civit, M., 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*, tesis doctoral, Universidad de Barcelona (en preparación)

Cruse, Alan, 2000. *Meaning in Language. An Introduction to Semántics and Pragmatics*, Oxford University Press

Federici, S., S.Montemagni y V.Pirelli, 2000. ROMANSEVAL: Results for Italian by SENSE, en *Computers and the Humanities. Special Issue: Evaluating WSD Programs*, **34 (1-2)**

Hoste, V., I.Hendrickx, W.Daelemans y A. van den Bosch, 2002. Parameter optimization for machine-learning of WSD, en *Natural Language Engineering, **8 (4)***

Kilgariff, A., 1998. Bridging the gap between lexicón and corpus: convergence of formalisms, en *LREC'1998*, Granada

Leacock, C., M.Chodorow y G.A.Miler, 1998. Using Corpus Statistics and WordNet Relations for Sense Identification, en *Computational Lingüístics. Special Issue on Word Sense Disambiguation*, **24 (1)**

Lin, D., 1997. Using Syntactic Depency as Local Context to Resolve Word Sense Ambiguity, en *Proceedings of ACL and EACL'97*, Morgan Kaufman Publishers, San Francisco

Mihalcea, R., 2002. WSD with pattern learning and feature selection, en *Natural Language Engineering, **8(4),*** Cambridge University Press

Mihalcea, R. y D. Moldovan, 1999. An Automatic Method for Generating Sense Tagged Corpora, en *Proceedings of AAAI '99*, Orlando

Montoyo, A. y Palomar M., 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. *Proc. 11th International Workshop on DEXA*, Greenwich, London

Ng, H.T. y H.B. Lee, 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach, en *Proceedings of the 34th Annual Meeting of the ACL*

Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense, en *Proceedings of NAACL 2001*, Pittsburg

Pedersen, T., 2002. *A Baseline Methodology for Word Sense Disambiguation*: http://www.d.umn.edu/~tpedersen

Sebastián, N., M.A. Martí, M. F. Carreiras, F. Cuetos Gómez, 2000. *Lexesp, léxico informatizado del español*, Edicions de la Universitat de Barcelona

Stetina, J., S. Kurohashi, M. Nagao, 1998. General WSD Method Based on a Full Sentential Context, en *Proceedings of COLING-ACL Workshop*, Montreal

Véronis, J., 2001. Sense tagging: does it make sense? Trabajo presentado en *The Corpus Lingüístics'2001 Conference*, Lancaster

Vossen, P., 1998 (ed.). *EUROWORDNET. A Multilingual Database with Lexical Semántic Networks*, Kluwer Academic Publishers, Dordrecht

Yarowsky, D., 1993. One Sense per Collocation, en *DARPA Workshop on Human Language Technology*, Princeton

Yarowsky, D. y R. Florian, 2002. Evaluating sense disambiguation across diverse parameter spaces, en *Natural Language Engineering, **8(4),*** Cambridge University Press.