Procesamiento del Lenguaje Natural, num. 28 (2002), pp.45-53 recibido 20-3-2002; aceptado 2-5-2002

Desambiguación del sentido y del dominio de las palabras con modelos de probabilidad de Máxima Entropía*

Armando Suárez y Manuel Palomar
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apartado de correos, 99
E-03080 Alicante, Spain
armando.suarez@ua.es

Resumen: En este artículo se presenta un sistema de aprendizaje supervisado para la desambiguación del sentido de las palabras. Dicho sistema se basa en los modelos de probabilidad condicional de Máxima Entropía. El conocimiento lingüístico se adquiere a partir de un corpus anotado y se representa en forma de atributos (*features*). Se han estudiado varios tipos de atributos para un conjunto limitado de palabras del corpus DSO. También se ha estudiado la sustitución de los sentidos de WordNet por etiquetas de dominio. En la actualidad, la implementación del sistema no soporta ninguna técnica de suavizado o preproceso complejo, pero sus resultados son buenos si son comparados, por ejemplo, con los de los sistemas presentados en el SENSEVAL-2.

Palabras clave: desambiguación del sentido de las palabras, desambiguación de los dominios de las palabras, basado en corpus, aprendizaje supervisado, máxima entropía

Abstract: In this paper, a supervised learning system of word sense disambiguation is presented. It is based on *maximum entropy conditional probability models*. This system acquires the linguistic knowledge from an annotated corpus and this knowledge is represented in the form of features. Several types of features has been analyzed for a few words selected from the DSO corpus. Moreover, substituting WordNet senses by domain labels have been studied too. Currently, the system implementation does not support any smoothing technique or complex pre-processing but its accuracy is good when it is compared with, for example, the systems at SENSEVAL-2.

Keywords: word sense disambiguation, word domain disambiguation, corpus-based, supervised learning, maximum entropy

1 *Introducción*

La asignación del sentido a las palabras (WSD, *word sense disambiguation*) es un área de investigación dentro del procesamiento del lenguaje natural (PLN) que aún requiere muchos esfuerzos. La tarea consiste, básicamente, en asignar el significado correcto a las palabras mediante la utilización de un diccionario electrónico como fuente de sus definiciones.

En la actualidad, se puede decir que coexisten, de forma general, dos enfoques principales al problema: métodos "basados en el conocimiento" y métodos "basados en corpus". El primer enfoque necesita la adquisición previa del conocimiento lingüístico, mientras que el segundo utiliza técnicas estadísticas y de aprendizaje automático para construir modelos del lenguaje a partir de grandes cantidades de ejemplos textuales (Pedersen, 2001). Estos últimos pueden realizar aprendizaje supervisado o no supervisado. Con el aprendizaje supervisado, la clase a la que pertenece cada ejemplo (en nuestro caso, su significado) es conocida de antemano, mientras que en el aprendizaje no supervisado no se conoce tal información (Manning y Schütze, 1999).

La asignación del significado correcto a las palabras puede ser abordado como un problema de "clasificación" de contextos

* Este artículo ha sido financiado parcialmente por el Gobierno Español (CICYT) dentro del proyecto número TIC2000-0664-C02-02.

ISSN 1135-5948

lingüísticos; dado un contexto, dentro del cual encontramos una palabra ambigua, las clases posibles son todos los significados definidos en algún diccionario para esa palabra. Por ejemplo, WordNet (WN) (Miller et al., 1993) es una base de datos léxica que asigna a cada palabra uno o varios números de sentido, dependiendo de la cantidad de conceptos (*synsets*) que se pueden expresar con ella. La tarea consistiría, pues, en asignar el número de significado correcto a esa palabra en ese contexto.

En este artículo se presenta un sistema que implementa un método de WSD basado en corpus. El método que se utiliza para realizar el aprendizaje a partir de un conjunto de ejemplos previamente clasificados es el de modelos de probabilidad de Máxima Entropía (ME). La información lingüística se representa mediante vectores de atributos, los cuales caracterizan los contextos lingüísticos (a partir de ahora, simplemente "contextos") informando de la aparición o no de ciertos atributos previamente definidos[1]. Los atributos o características[2] pueden ser de distinta naturaleza: palabras en posiciones concretas, categorías gramaticales, palabras clave, información del tópico o de género, relaciones gramaticales y, en general, todo dato que se considere relevante.

En el apartado siguiente, se citan algunos trabajos recientes en el área. A continuación, se introduce el concepto de desambiguación del dominio de las palabras. En el apartado 4 se muestran los fundamentos de ME para, seguidamente, describir el sistema que implementa este método. Posteriormente, se realiza una descripción más detallada de los atributos a utilizar en el aprendizaje y se presentan los resultados obtenidos a partir de este aprendizaje. El artículo finaliza con la exposición de las conclusiones acerca del trabajo presentado y los desarrollos que se está llevando a cabo actualmente.

2 Trabajos relacionados

Las últimas contribuciones a la resolución de este problema se pueden ver en SENSEVAL-2. Refiriéndonos a la tarea en inglés de clasificación de ejemplos (*English lexical sample*) algunos de los sistemas que compitieron se mencionan a continuación (SENSEVAL-2, 2001). El sistema de la *John Hopkins University* combina varios subsistemas de WSD basados en diferentes métodos: *decision lists* (Yarowsky, 2000); *transformation-based, error-driven learning* (Brill, 1995) (Florian y Ngai, 2001); *cosine-based vector models*; y dos sistemas basados en *Naive Bayes* (uno entrenado con palabras y otro con lemas). Los atributos utilizados incluyen propiedades sintácticas y relaciones gramaticales. La *Southern Methodist University* presentó un sistema cuyo método se basa en aprendizaje basado en ejemplos (*exemplar-based*) pero que también utiliza patrones de relaciones entre pares de palabras obtenidos a partir de WN en su versión 1.7 (Mihalcea y Moldovan, 2000). El sistema de la *Korea University* utiliza un modelo de información de clasificación basado en lo que ellos denominan *Discrimination Score* aplicado a los atributos (palabras cercanas en el contexto, tópicos y bigramas) (Seo, Lee, y Rim, 2001). Otro sistema presente en la misma competición se basa en *Boosting* (Escudero, Màrquez, y Rigau, 2000), una variante del algoritmo *AdaBoost.MH*. Es, también, un método basado en corpus y con aprendizaje supervisado.

En (Pedersen, 2002) se propone una metodología de referencia para WSD que se basa en el aprendizaje de árboles de decisión y clasificadores *Naive Bayes*. El trabajo presenta varios sistemas combinando los métodos mencionados, pero cada uno entrenado con conjuntos de atributos diferentes. También compitió en el SENSEVAL-2, tanto en la tarea *lexical sample* en inglés como en español.

En (García-Varea et al., 2001) se puede ver una aplicación de ME a WSD pero utiliza otro procedimiento estadístico entrenado para definir el conjunto de clases para cada palabra. Además, hay información publicada en Internet sobre WSD utilizando ME pero, a día de hoy, desconocemos si han sido presentados o incluidos en alguna publicación científica.

3 Desambiguación de los dominios de las palabras

Una variante de WSD es la desambiguación de los dominios de las palabras (WDD, *Word Domain Disambiguation*). La diferencia se encuentra en el conjunto de clases a tratar:

[1] Por contexto entendemos el texto que acompaña al dato ambiguo y que es relevante para el proceso de desambiguación

[2] "atributo" y "característica" se van a utilizar indistintamente: nos estamos refiriendo al concepto, en inglés, *feature*

las palabras están marcadas con una etiqueta de dominio en lugar de un sentido de WN. Este conjunto de etiquetas se puede obtener de un enriquecimiento de WN, *WordNet Domains* (Magnini y Strapparava, 2000), que utiliza códigos de clasificación por temas (*subject field codes*): *medicina*, *deportes*, *biología*, etc.

Por una parte, este conjunto de etiquetas produce una reducción de la polisemia puesto que agrupa sentidos que, a un nivel más abstracto que las propias definiciones de WN, pertenecen al mismo tópico. Se espera, entonces, que la tasa de acierto de WDD sea mayor que la de WSD. Por otra parte, la intención de este enriquecimiento es dar respuesta a la crítica de varios investigadores a la excesiva granularidad de WN: muchos significados parecen innecesarios o se diferencian mínimamente entre sí (o incluso son, en realidad, el mismo). Estos mismos investigadores defienden que ciertas tareas del PLN, como la recuperación de información (*information retrieval*) o los sistemas de pregunta-respuesta (*question answering*)podrían ser mejorados con información semántica de dominios y no de sentidos.

Otra propuesta de enriquecimiento de WN con información de mayor nivel de abstracción se puede ver en (Montoyo, Palomar, y Rigau, 2001), donde se utilizan las categorías IPTC[3].

4 Máxima Entropía: fundamentos

Los modelos de ME proporcionan un entorno para la integración de información útil para la clasificación obtenida de fuentes de información heterogéneas (Manning y Schütze, 1999). Los modelos de probabilidad de ME han sido aplicados con éxito a varias tareas del PLN tales como etiquetado gramatical (*part-of-speech tagging*) o detección de los límites de la frase (Ratnaparkhi, 1998).

El método de WSD aplicado en este trabajo se basa en los modelos de probabilidad condicional de ME. Es un método que realiza un aprendizaje supervisado que consiste en la definición de funciones de clasificación de palabras en significados. Una función de clasificación obtenida de esta forma incluye un conjunto de coeficientes o parámetros estimados por un procedimiento de optimización, cada uno asociado a un atributo lingüístico concreto de forma que el coeficiente determina el peso de la característica dentro de la función de clasificación. El fin último es obtener la distribución de probabilidad que maximiza la entropía, o lo que es lo mismo, no se asume nada que no esté en los propios datos de entrenamiento. Algunas ventajas de los modelos de probabilidad de ME son los buenos resultados mediante la utilización de información relativamente simple y que permite, virtualmente sin ninguna restricción, la representación del conocimiento especifico de un determinado problema en forma de atributos (Ratnaparkhi, 1998).

Sea X el conjunto de contextos y C el conjunto de clases. La función $cl : X \rightarrow C$ elige la clase c con la mayor probabilidad condicional en el contexto x: $cl(x) = \arg\max_c p(c|x)$. Cada atributo se calcula mediante una función que está asociada a una clase específica c', y tiene la forma de la ecuación (1), donde $cp(x)$ es alguna característica observable en el contexto[4]. La probabilidad condicional $p(c|x)$ se define en la ecuación (2), donde α_i es el coeficiente o peso del atributo i, K es el número de atributos definido, y $Z(x)$ una constante que asegura que la suma de todas las probabilidades condicionales para este contexto es igual a 1.

$$f(x,c) = \begin{cases} 1 & \text{if } c' = c \text{ and } cp(x) = true \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$p(c|x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} \quad (2)$$

El siguiente apartado muestra los detalles de implementación de los modelos de probabilidad de ME para este trabajo.

5 La propuesta del sistema de WSD con ME

La implementación del sistema de WSD que se describe a continuación se realizó en C++. El sistema incluye varios módulos: aprendizaje, clasificación, evaluación y conversor de

[3]El *IPTC Subject Reference System* se desarrolló para suministrar a los servicios de información un sistema de codificación universal independiente del lenguaje para la clasificación de noticias por temas. http://www.iptc.org

[4]Los modelos de probabilidad de ME no están limitados a funciones binarias pero el procedimiento de optimización, *Generalized Iterative Scaling (GIS)*, utiliza este tipo de funciones.

formato de corpus, siendo los dos primeros los principales.

El módulo de aprendizaje es el encargado de definir las funciones de clasificación para cada palabra utilizando un corpus que está etiquetado sintáctica y semánticamente. Dos submódulos pueden diferenciarse dentro de él. El primero de ellos procesa el corpus de aprendizaje para poder definir los atributos a observar en los contextos y, posteriormente, rellenar los vectores de atributos (uno por cada contexto). El segundo submódulo realiza la estimación de los coeficientes y almacena, finalmente, las funciones de clasificación. El sistema utiliza distintos parámetros para caracterizar los contextos con un conjunto determinado de atributos, que puede ser distinto para cada palabra.

El módulo de clasificación se encarga de la desambiguación de los nuevos contextos utilizando las funciones de clasificación previamente almacenadas. Cuando ME no tiene suficiente información acerca de un determinado contexto, varios sentidos pueden alcanzar el mismo valor máximo de probabilidad y, entonces, el contexto se mantiene ambiguo. En estos casos se elige el sentido más frecuente en el corpus. No obstante, esta heurística es necesaria sólo para un mínimo número de casos o cuanto el conjunto de atributos utilizados en el aprendizaje es excesivamente pequeño.

El módulo de evaluación se ha desarrollado para poder medir la exactitud del método, y compara el sentido elegido por ME con el anotado en el corpus. Los resultados que se muestran en la siguiente sección han sido obtenidos por este módulo.

Finalmente, el módulo conversor de formato de corpus es una herramienta auxiliar diseñada para aplicar un formato predefinido a los corpus utilizados. Con esto se facilita la evaluación de corpus de distinta procedencia. Si es necesario, cada corpus se procesa con un analizador para añadirle la información gramatical deseada; actualmente se está utilizando MINIPAR (Lin, 1998).

6 *Implementación de las funciones de atributo*

Un aspecto importante de la implementación de los modelos de probabilidad de ME es la forma de las funciones que calculan cada atributo. Estas funciones son las que se utilizan para rellenar un vector de características para cada contexto, en el que cada componente nos dice si el atributo al que representa se encuentra o no en el contexto, si su función asociada ha devuelto cierto o falso. Estos vectores son los que se utilizarán en el aprendizaje propiamente dicho.

Para cada palabra que se quiere aprender se examinan todos sus ejemplos para definir las funciones, una por atributo. Una definición usual (pero no la única) de atributos es la que substituye $CP(x)$ en la ecuación (1) con una expresión del tipo $info(x,i) = a$, en la que $info(x,i)$ nos informa de una propiedad que se puede encontrar en la posición i dentro del contexto x, y a es un valor predefinido. Por ejemplo, si consideramos 0 como la posición de la palabra a aprender y que i es una posición relativa a dicha palabra, entonces podríamos utilizar algo así como *palabra(x,-1) = "best" y c'=interest#1*: *¿es "best" la palabra anterior a "interest" con sentido 1?*. Los valores de a pueden ser predefinidos manualmente o, como es nuestro caso, con los valores que se encuentran en el propio corpus de aprendizaje. Generalizando, para cada posible valor de $(sentido, a)$, y para cada posición deseada i, se genera una función.

Por ejemplo, sean los contextos de aprendizaje para la palabra "interest" que se muestran a continuación, y para los que se dispone de toda la información sintáctica necesaria:

... considering the widespread **interest#1** *in the election ...*
... to the best **interest#5** *of both governments ...*
... anonymous persons expressing **interest#1** *in the trial ...*

Si los atributos a utilizar son la palabra anterior y la categoría gramatical de esa misma palabra, dados los contextos anteriores, se definirán seis atributos:

atrib1: *¿es "widespread" anterior a interest#1?*
atrib2: *¿es "best" anterior a interest#5?*
atrib3: *¿es "expressing" anterior a interest#1?*
atrib4: *¿es "ADJ" anterior a interest#1?*
atrib5: *¿es "ADJ" anterior a interest#5?*
atrib6: *¿es "VERB" anterior a interest#1?*

Estas funciones se evalúan para todos los ejemplos, y se rellena un vector de características para cada uno de ellos con los valores obtenidos. En general, para un tipo concreto de atributos (*palabra en la posición i, lema en la posición i, categoría gramatical en*

la posición i, etc.) se implementa una función por cada posible $(sentido, valor_i)$. A este tipo de funciones las denominaremos "no relajadas".

Se puede utilizar otro tipo de funciones, como se muestra en el siguiente ejemplo:

atrib1: *¿es "widespread" o "expressing" anterior a interest#1?*
atrib2: *¿es "best" anterior a interest#5?*
atrib3: *¿es "ADJ" o "VERB" anterior a interest#1?*
atrib4: *¿es "ADJ" anterior a interest#5?*

Estas funciones, que llamaremos "relajadas", utilizan conjuntos de valores presentes en el corpus de entrenamiento y relacionados con ciertos atributos (en nuestro ejemplo, *palabras en la posición anterior de interest#1*). Esta implementación reduce el número de atributos (funciones) a uno por cada sentido y posición i.

En el apartado 7 se muestra que el rendimiento del sistema no es penalizado excesivamente por el uso de estas funciones "relajadas". Debido a la naturaleza de la tarea de desambiguación, la cantidad de veces que se activa un atributo generado por el primer tipo de función es muy pequeña, y los vectores de características consisten, generalmente, en una larga lista de ceros con algún uno intercalado entre ellos. El nuevo tipo de función reduce drásticamente la cantidad de atributos con una degradación mínima de los resultados de evaluación. Al mismo tiempo, nuevas características pueden ser incorporadas al aprendizaje con un mínimo impacto en la eficiencia del método.

6.1 Descripción de tipos de atributos

El conjunto de atributos definido para el entrenamiento del sistema se describe en la Figura 1. Los atributos se definen automáticamente y dependen de los ejemplos en el corpus de entrenamiento. Estos atributos se basan, principalmente, en el conocimiento lingüístico del contexto cercano a la palabra ambigua: palabras y composiciones de palabras que la acompañan, categorías gramaticales, rol gramatical, dependencias, etc.[5]

Cada grupo es, en realidad, un conjunto de atributos. Por ejemplo, en el grupo de atributos s están todas las funciones para cada posible valor, en el corpus de aprendizaje, de (a, c, i); a es cualquier palabra que se encuentra en algún ejemplo clasificado como sentido c en una posición $i \in \{-3, -2, -1, +1, +2, +3\}$. Volviendo al ejemplo del apartado anterior, dentro del conjunto de atributos s se encontrarían los tres primeros "no relajados" (atributos del tipo s_{-1}), y los quince que se definirían para el resto (los otros s_i).

Una descripción un poco más extensa de varios de estos grupos (con diferencias mínimas en su denominación) se puede encontrar en (Suárez y Palomar, 2002). En el siguiente apartado se amplía esa descripción.

Grupo Km

Para este conjunto de atributos se seleccionan nombres dependiendo de su frecuencia de aparición junto a un sentido concreto. Por ejemplo, en un conjunto de 100 ejemplos del significado 4 del nombre "interés", si el nombre "banco" aparece 10 veces o más ($m = 10\%$), entonces se define una característica para cada posible sentido de "interés" con la palabra "banco", utilizando funciones del tipo de la mostrada en la ecuación (3).

$$W = \{w \mid \exists c(\ c \in C, freq(w, c) > m)\} \quad (3)$$

$$f_{(c',w)}(x, c) = \begin{cases} 1 & \text{if } c' = c \text{ and } w \in W \\ 0 & \text{otherwise} \end{cases}$$

Grupo r

Si se dispone de un analizador apropiado, podemos obtener el rol gramatical de la palabra ambigua, esto es, si es sujeto, objeto, complemento, etc. Cada etiqueta anotada por el analizador es utilizada para definir una función para cada significado posible. Estas características no se han redefinido como "relajadas" debido al relativamente bajo número de etiquetas.

Grupos D y d

El analizador puede proporcionar un árbol sintáctico en el que las palabras están enlazadas entre si por una relación de dependencia o inclusión. Estas características se definen mediante las dependencias de la palabra ambigua anotadas en el corpus de entrenamiento.

[5] La elección de características se ha desarrollado partiendo de los trabajos de (Ng y Lee, 1996) y (Escudero, Màrquez, y Rigau, 2000).

Figura 1: Lista de grupos de atributos

- *No relajados*
  - ***0***: la palabra ambigua
  - *s*: palabras en posiciones ±1, ±2, ±3
  - *p*: categorías gramaticales de palabras en ±1, ±2, ±3
  - *km*: lemas de nombres que aparecen en al menos el *m*% de contextos de un sentido
  - *r*: rol gramatical de la palabra ambigua
  - *d*: la palabra de la que depende la ambigua
  - *m*: palabra compuesta a la que pertenece la ambigua
- *Relajados*
  - *L*: lemas (de palabras llenas) en ±1, ±2, ±3
  - *W*: palabras llenas en ±1, ±2, ±3
  - *S*: palabras en ±1, ±2, ±3
  - *B*: lemas de pares de palabras en $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
  - *C*: pares de palabras en $(-2,-1)$, $(-1,+1)$, $(+1,+2)$
  - *P*: categorías gramaticales en ±1, ±2, ±3
  - *D*: la palabra de la que depende la ambigua
  - *M*: palabra compuesta a la que pertenece la ambigua

Grupos *M* y *m*

Si el analizador puede identificar palabras compuestas en las que se encuentra la palabra ambigua ("dark ages", "interest rate", etc.), el sistema puede definir funciones de atributo con estas composiciones. Por ejemplo, MINIPAR proporciona este tipo de información.

7 Evaluación

En esta sección se presentan los resultados de la evaluación del sistema. Se seleccionaron algunos nombres y verbos polisémicos de los incluidos en el *DSO sense-tagged English corpus* (Ng y Lee, 1996). Este corpus está estructurado en ficheros, cada uno con ejemplos de una palabra (nombre o verbo) etiquetados con el significado correcto. Las etiquetas de sentido se corresponden con las de WordNet 1.5 y las fuentes de dichos ejemplos son el *Brown corpus* y el *Wall Street Journal*. Para el entrenamiento y evaluación se han utilizado todos los ejemplos de cada fichero de las palabras seleccionadas.

Los resultados que se muestran a continuación corresponden a tres pruebas diferentes: WSD para un subconjunto de nombres y verbos del citado corpus DSO, WSD-WDD para todos los nombres del mismo corpus, y WSD para los datos del *Spanish lexical sample task* del último SENSEVAL-2.

7.1 WSD con DSO

La Tabla 1 muestra los mejores resultados utilizando *10-fold cross-validation* como método de evaluación[6]. Se probaron varias combinaciones de atributos con el objetivo de hallar la mejor de ellas para cada palabra. Se pretendía detectar la información más útil en el corpus para cada palabra en lugar de aplicar las mismas características a todas ellas.

Antes de realizar las diez pruebas para cada palabra, el corpus se procesó con el analizador MINIPAR (Lin, 1998) y se distribuyeron uniformemente los ejemplos: cada archivo seleccionado del DSO se dividió en 10 partes conteniendo una décima parte del total de ejemplos de cada sentido. Los sentidos con menos de 10 ejemplos se eliminaron. Se buscaba mantener la misma proporción de sentidos que la del archivo original en cada prueba. Es por eso que, aunque los sentidos eliminados fueron muy pocos, la columna "Sentidos" informa de los significados realmente aprendidos, no de los presentes en el corpus originalmente.

La columna *Atributos*, mediante una cadena de caracteres para cada palabra, indica la combinación de grupos de atributos que se han utilizado para obtener el resultado de acierto. Cada grupo, dentro de la cadena, está identificado por su letra (excepto en el caso del grupo *Km*, que además indica el valor de porcentaje utilizado), correspondiendo las letras minúsculas a la implementación "no

[6]El corpus se divide en 10 partes y se realizan 10 pruebas utilizando, alternativamente, una parte como corpus de evaluación y las otras nueve partes como corpus de entrenamiento; la tasa de acierto es el promedio de las obtenidas en las diez pruebas.

Tabla 1: Mejores resultados con el corpus DSO

	Sentidos	Ejemplos	Atributos	Funciones	Acierto	SMF
age,N	3	491	0CsprDMk5	1587	73,8	62,4
art,N	4	393	0sprdm	1594	65,2	48,0
car,N	2	1363	s	3036	97,1	96,3
child,N	2	1057	sp	2731	90,5	81,8
church,N	3	367	0rDMCk3	228	67,9	62,0
cost,N	2	1456	0WrDM	62	90,0	89,6
head,N	7	844	sprdm	2911	80,8	41,6
interest,N	6	1479	0sprDM	4059	70,1	45,9
line,N	22	1320	0LSBCrdm	1542	54,7	22,7
work,N	6	1419	0sprdm	4784	53,2	32,8
fall,V	6	1341	LSBCrdm	503	84,9	70,1
know,V	6	1425	0rDMCk10	230	47,9	34,9
set,V	11	1246	BsprDMk5	4569	57,3	36,9
speak,V	5	510	0sp	1667	74,5	69,1
take,V	19	794	LWBCsrDMk10	3706	43,0	35,6
Promedios	**7**	**1034**		**2214**	**70,1**	**55,3**
Nombres	6	1019		2253	74,3	58,3
Verbos	9	1063		2135	61,5	49,3
Todas			**0sprdm**	**3411**	**68,8**	
Nombres				3013	73,5	
Verbos				4208	59,4	

Sentidos: cantidad de significados distintos en el corpus
Ejemplos: cantidad de ejemplos en el corpus
Atributos: grupos de atributos utilizados
Funciones: cantidad de funciones (atributos) derivadas de la selección *Atributos*
Acierto: número de contextos correctamente clasificados dividido por el número total de contextos
SMF: tasa de acierto obtenida al clasificar con el sentido más frecuente en el corpus

relajada" y las mayúsculas a la "relajada". Por ejemplo, la selección 0*sprdm* incluye seis grupos de atributos, cada uno de ellos con un número de funciones que depende del tipo de información suministrada y de los datos del corpus de aprendizaje.

Los datos de la Tabla 1 sugiere que todos los tipos de atributos, relajados y no relajados, son útiles en algún momento. Además, cada palabra obtiene el mejor resultado con una selección de atributos concreta; si tal estrategia de aprendizaje pudiera ser establecida, para estas quince palabras se clasificarían correctamente el 70.1% de los contextos. Este valor supone una ganancia del 15% respecto a la clasificación por el sentido más frecuente. También se puede ver que los nombres obtienen mejores resultados que los verbos.

Aplicando un conjunto de atributos fijo a todas las palabras, el mejor resultado se obtiene con la selección "0*sprdm*", un 68.8% de éxito (14.4% más que el SMF).

7.2 WSD-WDD con DSO

En este subapartado se van a mostrar los resultados obtenidos al evaluar los nombres del corpus DSO con etiquetas de dominio y con etiquetas de *synset*. Como ya se ha dicho anteriormente, las etiquetas de dominio se obtienen del recurso *WordNet Domains*.

Dado el preproceso del corpus comentado en la subsección anterior, algunos nombres se convierten en monosémicos, con lo que la clasificación obtiene un 100% de éxito. Se han mantenido estos datos puesto que se quiere mostrar la ganancia en aciertos al comparar WDD con WSD. También por este preproceso, el nombre "college" es monosémico tanto con *synsets* como con dominios. La Figura 2 es un resumen estadístico de los datos de los nombres del corpus después del preproceso.

Figura 2: Datos de los nombres del DSO

120 Nombres
872 Ejemplos(*) por nombre
Dominios: 11 nombres monosémicos
Synsets: 1 nombre monosémico
3,5 dominios(*) por nombre
4,8 synsets(*) por nombre

(*) valores promedio

La Tabla 2 muestra los resultados de evaluación de los 120 nombres del corpus DSO[7] cuando los conjuntos de clases están formados por etiquetas de dominio en vez de sentidos de WN. La primera consecuencia es la disminución de clases posibles para algunas

[7]En el momento de escribir este artículo, sólo se dispone del conjunto de etiquetas de dominio para los nombres en WN.

de las palabras y el aumento en la tasa de acierto del método. Obviamente, aquellas palabras que no reducen su conjunto de clases no contribuyen a este aumento.

Tabla 2: Resultados de WDD y WSD

Atributos	Dominios	Synsets	FuncsD	FuncsS
SMF	68,7	58,7		
LB	73,5	64,6	44,8	45,2
SP	74,8	66,6	54,4	59,6
0LB	75,4	67,1	56,9	63,6
0SP	75,7	67,8	66,4	78,2
sp	77,2	69,5	1191,6	2363,0
0sp	77,7	70,2	1198,8	2381,5
sprdm	78,1	70,6	1452,2	2730,9
0sprdm	78,4	71,0	1459,8	2749,5
0LSsBCprdm	78,6	71,0	1405,3	2836,4
0sprdmk10	78,7	71,4	1393,0	2803,8
0sBCprdmk10	78,7	71,5	1415,1	2790,0

Siguiendo el método *10-fold cross-validation* se han realizado varias pruebas con distintas selecciones de atributos. La columna "Dominios" muestra la tasa de acierto promedio para cada selección de atributos cuando las clases son *WN Domains*, y la columna lo mismo muestra "Synsets" pero utilizando las definiciones de WN. La primera fila de resultados, la etiquetada con "*SMF*", son los valores obtenidos al clasificar cada contexto con el significado más frecuente (en el corpus). Las columnas *FuncsD* y *FuncsS* son las cantidades promedio de funciones o atributos utilizados en el aprendizaje de todos los nombres.

El mejor resultado para WDD se obtiene con la selección de atributos más compleja (*0sprdmk10*) y obtiene, aproximadamente, un 7% más de acierto que la obtenida para WSD. Al añadir los grupos de atributos "relajados" de composición de palabras y lemas al aprendizaje, B y C, el incremento en el acierto es inapreciable. Básicamente, los resultados para dominios y para *synsets* se incrementan de igual manera al ir añadiendo grupos de atributos al aprendizaje.

7.3 WSD con SENSEVAL-2

ME se ha evaluado también utilizando los datos de entrenamiento y evaluación suministrados por la organización del SENSEVAL-2 para la tarea en español. El sistema obtiene un 65.03% de éxito con la selección de atributos $0LWsBCp$, y un 64.04% con $OLBk5$. En ambos casos, comparados estos valores con los resultados oficiales de los trece sistemas que allí compitieron, ME hubiera quedado en cuarto lugar. Para esta evaluación se utilizó el Conexor FDG Parser (Tapanainen y Järvinen, 1997).

8 Conclusiones

Se ha presentado un sistema de asignación del significado correcto a las palabras basado en la implementación de un método de aprendizaje supervisado: modelos de probabilidad condicional de Máxima Entropía. Se han utilizado diferentes conjuntos de atributos para caracterizar los contextos basados en la ocurrencia de palabras en determinadas posiciones relativas a la de la palabra ambigua, y propiedades gramaticales de esas palabras.

El sistema se ha evaluado utilizando WordNet como diccionario del que extraer los sentidos de las palabras, tanto con definiciones de alto grado de detalle (*synsets*) como con etiquetas de dominio (*WordNet Domains*).

Se han efectuado varias pruebas utilizando el corpus DSO y los datos suministrados por SENSEVAL-2 para la tarea *Spanish lexical sample*. Con el primer corpus, una prueba de WSD con algunos nombres y verbos en la que se buscaba conocer que combinaciones de atributos eran las mejores, y otra de comparación de WDD y WSD con todos los nombres en el corpus. Con el segundo, se ha comparado el resultado obtenido por nuestro sistema con los resultados oficiales de la competición.

WDD tiene como ventaja frente a WSD la menor polisemia y, por tanto, su mayor tasa de acierto, además de su posible mejor adecuación a tareas tales como la recuperación de información y los sistemas de pregunta-respuesta. Así, se obtiene una mejora de aproximadamente un 13% cuando se comparan los dos tipos de desambiguación.

Como trabajo en curso, se está estudiando la posibilidad de incorporar la información de dominio a la desambiguación de sentidos de WN, incorporándola como un atributo más en el aprendizaje.

También se está estudiando la cooperación de varios métodos, supervisados y no supervisados, en un único sistema que los integre a todos, como ya se ha hecho en SENSEVAL-2.

Bibliografía

ACL, editor. 2001. *Proceedings of NAACL Workshop WordNet and Other Lexical*

Resources: Applications, Extensions and Customizations, Pittsburgh, PA, USA.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Escudero, Gerard, Lluis Màrquez, y German Rigau. 2000. Boosting applied to word sense disambiguation. En *Proceedings of the 12th Conference on Machine Learning ECML2000*, Barcelona, Spain.

Florian, Radu y Grace Ngai. 2001. Multidimensional transformation-based learning. En Walter Daelemans y Rémi Zajac, editores, *Proceedings of CoNLL-2001*, páginas 1–8. Toulouse, France.

García-Varea, Ismael, Franz J. Och, Hermann Ney, y Francisco Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. En *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, páginas 204–211.

Gelbukh, A., editor. 2002. *Proceedings of 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Lecture Notes in Computer Science, Mexico City, February. Springer-Verlag.

Lin, Dekang. 1998. Dependency-based evaluation of minipar. En *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain.

Magnini, Bernardo y C. Strapparava. 2000. Experiments in Word Domain Disambiguation for Parallel Texts. En *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, Hong Kong, China.

Manning, Christopher D. y Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Mihalcea, Rada y Dan Moldovan. 2000. An iterative approach to word sens disambiguation. En *Proceedings of FLAIRS-2000*, páginas 219–223, Orlando, FL, May.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, y Katherine J. Miller. 1993. Five Papers on WordNet. *Special Issue of the International journal of lexicography*, 3(4).

Montoyo, Andrés, Manuel Palomar, y German Rigau. 2001. WordNet Enrichment with Classification Systems. En ACL (ACL, 2001).

Ng, Hwee Tou y Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. En Arivind Joshi y Martha Palmer, editores, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, San Francisco. Morgan Kaufmann Publishers.

Pedersen, Ted. 2001. A decision tree of bigrams is an accurate predictor of word sense. En ACL (ACL, 2001).

Pedersen, Ted. 2002. A baseline methodology for word sense disambiguation. En Gelbukh (Gelbukh, 2002), páginas 126–135.

Ratnaparkhi, Adwait. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. tesis, University of Pennsylvania.

SENSEVAL-2. 2001. Second international workshop on evaluating word sense disambiguation systems: system descriptions. http://www.sle.sharp.co.uk/senseval2/.

Seo, Hee-Cheol, Sang-Zoo Lee, y Hae-Chang Rim. 2001. Classification information model. http://nlp.korea.ac.kr/ hcseo/senseval2/cim.htm, June.

Suárez, Armando y Manuel Palomar. 2002. Feature selection analysis for maximum entropy-based wsd. En Gelbukh (Gelbukh, 2002), páginas 146–155.

Tapanainen, Pasi y Timo Järvinen. 1997. A non-projective dependency parser. En *Proceedings of the Fifth Conference on Applied Natural Language Processing*, páginas 64–71, April.

Yarowsky, David. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1-2).