Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes

Sonia Vázquez y Andrés Montoyo
Grupo de Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.
Teléfono: 965903772 Ext. 2725 Fax: 965909326
E-mail: {svazquez, montoyo}@dlsi.ua.es

German Rigau
TALP Research Center
Universitat Politècnica de Catalunya
Pau Gargalló 5
08028 Barcelona, Spain
g.rigau@lsi.upc.es

Resumen

En este artículo presentamos un nuevo método WSD basado en la idea propuesta por Magnini y Strapparava, cuyo método utiliza el recurso léxico WordNet Domains. Sin embargo, desde nuestro punto de vista no se ha explotado la información que suministran las glosas de WordNet Domains asociadas a cada sentido y dominio. Así, presentamos la obtención de un nuevo recurso léxico a partir de las glosas de WordNet Domains denominado Dominios Relevantes. También se presenta un nuevo método WSD basado en este nuevo recurso léxico (Dominios Relevantes). Este método ha sido evaluado con los textos de la tarea *"English all-Words"* de SENSEVAL-2, obteniendo unos resultados satisfactorios.

1. Introducción

En el presente artículo nos centramos en la resolución de la ambigüedad léxica, la cual aparece cuando las palabras presentan diferentes significados. A esta tarea se le conoce como Desambiguación del sentido de las palabras (Word Sense Disambiguation, WSD). Esta es una "tarea intermedia" (Wilks Y. And Stevenson M., 1996), que sirve de ayuda cuando necesitamos conocer el sentido de las palabras en algunas aplicaciones del PLN, como en Traducción Automática (TA), Recuperación de la Información (RI), Clasificación de Textos, Análisis del Discurso, Extracción de Información (EI), etc.

De forma más genérica, la WSD consiste en la asociación de una palabra dada en un texto con una definición o significado, el cual la distingue de otros significados atribuibles a esa palabra. La asociación de palabras a los sentidos, se cumple dependiendo de dos recursos de información (contexto[1] y recursos de conocimiento externos[2]).

Para WSD existen diferentes métodos de trabajo como puede verse en (Ide N. and Véronis J., 1998), sin embargo, el presente artículo se centra en el método que se basa en el emparejamiento del contexto de la palabra a ser desambiguada con la información suministrada por el recurso de conocimiento léxico WordNet, conocido como desambiguación del sentido de las

[1] El **contexto** de la palabra a ser desambiguada es considerado como un conjunto de palabras que acompañan a la palabra a desambiguar, junto con las relaciones sintácticas, categorías semánticas, etc. En este artículo el contexto se compone de oración en oración.

[2] Los **recursos de conocimiento externos** son los recursos léxicos, enciclopédicos, recursos de conocimiento léxico (WordNet) desarrollados manualmente que proporcionan datos valiosos para asociar palabras con sentidos, etc.

palabras basada en conocimiento (*WSD knowledge-driven*).

WordNet no es un recurso perfecto para desambiguar el sentido de las palabras, ya que tiene el problema de la granularidad fina para la distinción de los sentidos (Ide N. and Véronis J., 1998). Esto crea muchas dificultades a la hora de desambiguar automáticamente el sentido de las palabras, debido a que hay que hacer elecciones en cuanto al significado, que a veces es difícil hasta manualmente. Varios autores como (Wilks and Stevenson, 1998, Gonzalo et al. 1998, Kilgarriff and Yallop, 2000), afirman que la granularidad fina para la distinción de los sentidos suministrada por WordNet hace más difícil la desambiguación del sentido de las palabras y no es necesaria para muchas tareas del PLN.

(Magnini and Strapparava, 2000) proponen resolver este problema presentando una variante de WSD denominada Word Domain Disambiguation (WDD), como una alternativa práctica para aplicaciones que no requieren una granularidad fina para la distinción de sus sentidos. Esta variante consiste en etiquetar las palabras de los textos con la etiqueta de un dominio en vez de la etiqueta de un sentido. Se entiende por dominio un conjunto de palabras que tienen una fuerte relación semántica. Por lo tanto, la idea fundamental de aplicar los dominios a WSD es aportar información relevante para establecer relaciones semánticas entre los sentidos de las palabras. Por ejemplo, la palabra "bank" tiene 10 sentidos en WordNet 1.6 pero tres de ellos "bank#1, bank#3 y bank#6" pueden agruparse en el dominio "ECONOMY", mientras que "bank#2 y bank#7" pertenecen a los dominios "GEOGRAPHY and GEOLOGY". Para aplicar el método WDD es necesario un recurso léxico donde los sentidos de las palabras estén asociados con los dominios. Así, el recurso que utilizaron en este trabajo fue una extensión de WordNet, denominada WordNet Domains (Magnini and Cavaglia, 2000), que tiene todos los synsets etiquetados con uno o más dominios. En el apartado 2 se explicará con más detalle este recurso.

El método de desambiguación que proponen Magnini y Strapparava utiliza el recurso léxico WordNet Domains. Sin embargo, desde nuestro punto de vista no se ha explotado la información que suministran las glosas de WordNet Domains asociadas a cada sentido y dominio. Así en este artículo se presenta la obtención de un nuevo recurso léxico a partir de las glosas de WordNet Domains denominado Dominios Relevantes. También se presenta un nuevo método WSD basado en este nuevo recurso léxico (Dominios Relevantes). Este método ha sido evaluado con los textos de la tarea *"English all-Words"* de SENSEVAL-2, obteniendo unos resultados satisfactorios.

Después de esta introducción, en la sección 2 se describe el recurso WordNet Domains. En la sección 3 se describe con detalle la obtención y la estructura del nuevo recurso léxico Dominios Relevantes. En la sección 4, se presenta el método WSD basado en el nuevo recurso léxico anterior. En la sección 5 se presenta la evaluación y discusión del método WSD propuesto. Y finalmente se definen las conclusiones y los trabajos futuros.

2. WordNet Domains

WordNet Domains es una extensión de WordNet 1.6, donde cada synset tiene asociado una o varios dominios (categorías semánticas). Estos dominios, están clasificados en una jerarquía con distintos niveles de especialización, cuanto más profundo es el nivel sobre el que nos movemos, mayor es el grado de especialización.

Como hemos comentado, cada synset tiene asociado uno o varios dominios, los mismos dominios pueden estar asociados a synsets de diferentes categorías sintácticas. Esta información añadida a WordNet 1.6. permite relacionar palabras pertenecientes a distintas subjerarquías y englobar dentro de un mismo dominio, varios sentidos de una misma palabra.

Por ejemplo, veamos cómo quedaría la clasificación de la palabra "*music*" en *"WordNet Domains":*

Synset	Dominio	Nombre	Glosa
05266809	Music	Music#1	an artistic form of auditory …
04417946	Acoustics	Music#2	any agreeable (pleasing…
00351993	Music Free_time	Music#3	a musical diversion; his music…
05105195	Music	Music#4	a musical composition in…
04418122	Music	Music#5	the sounds produced by singers..
00755322	Law	Music#6	punishment for one's actions;…

Tabla 1: Dominios asociados a music

Como se puede observar, la palabra *"music"* tiene seis sentidos diferentes, cuatro de los cuales, se han agrupado dentro del dominio "Music", debido a que las diferencias semánticas entre ellos son casi inapreciables, con lo que se disminuye la polisemia de la palabra.

Esta clasificación, va a ser el punto de partida para la extracción y aprovechamiento del conocimiento proporcionado por WordNet Domains.

3. Nuevo recurso: Dominios Relevantes

En este trabajo, WordNet Domains se usará como una recopilación de ejemplos para los diferentes dominios asociados a los significados semánticos de las palabras, gracias a las definiciones y ejemplos proporcionados para cada uno de los sentidos de las palabras. Por lo tanto, se utilizarán las glosas de WordNet Domains asociadas a cada sentido con el objetivo de reunir aquellas palabras más representativas para cada uno de los dominios. Así, el nuevo recurso léxico, denominado Dominios Relevantes, estará formado por todas las palabras que aparecen en las glosas de WordNet Domains etiquetadas con los dominios a los que pueden pertenecer, ordenados de mayor a menor según la importancia de estas palabras respecto al dominio.

En nuestra propuesta, para reunir las palabras más representativas para un dominio particular, se ha decidido utilizar la formula de "Información Mutua" (Church & Hanks, 1990) siguiente:

$$IM(w, D) = \log_2 \frac{\Pr(w \mid D)}{\Pr(w)} \quad (1)$$

Donde:

W: Palabra.
D: Dominio

Intuitivamente, una palabra representativa es aquella que aparece considerablemente más a menudo en el contexto de un dominio.

Pero lo que se va buscando es la importancia de las palabras asociadas a un dominio, entendiendo por importancia de las palabras aquellas que deben ser representativas y frecuentes a un dominio. Esta importancia se va a medir por medio de la formula denominada "Ratio de Asociación"[1]:

$$RA(w, D) = \Pr(w \mid D) \log_2 \frac{\Pr(w \mid D)}{\Pr(w)} \quad (2)$$

Donde:

W: Palabra.
D: Dominio.

El proceso de aplicación de esta fórmula se realiza primero para todas las palabras de las glosas de WordNet Domains cuya categoría

[1] El "Ratio de Asociación" de una palabra con un dominio se puede definir como el producto de la "Información Mutua" y la probabilidad de que una palabra pertenezca a ese dominio.

gramatical sea nombre, para posteriormente aplicarlo a los verbos, adjetivos y adverbios. Estos procesos se detallan en las subsecciones siguientes.

3.1. Ratio de Asociación para nombres.

Para obtener el "Ratio de Asociación" de los nombres que forman las glosas de WordNet Domains, es necesario la utilización de un analizador sintáctico que extraiga todos los nombres presentes en cada una de las glosas. Para ello, utilizaremos el analizador sintáctico *"Tree Tagger"* (Schmid, 1994).

Por ejemplo, para el sentido 1 de *"music"*, la glosa correspondiente es la siguiente:

"An artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner"

Con el *"Tree Tagger"* obtenemos los nombres de la glosa de WordNet Domains para la palabra *"music"*, y les asociamos el dominio *"Music",* que pertenece al sentido 1:

Dominio	Nombre
Music	Form
Music	Communication
Music	Tone
Music	Manner

Tabla 2: Asociación de dominios a los nombres de la glosa del sentido 1 de *"music"*.

Este proceso se realiza sobre todo Wordnet Domains, y una vez obtenidos todos los nombres junto con sus dominios asociados, se procede a calcular el "Ratio de Asociación".

Los resultados obtenidos, se almacenan en un fichero (Dominios Relevantes) que contendrá todos los nombres de WordNet 1.6. ordenados alfabéticamente, y junto a cada nombre, una lista de todos sus dominios asociados, ordenados de mayor a menor según los valores obtenidos mediante la fórmula "Ratio de Asociación". Con este formato, se establece que aquellos dominios asociados a un nombre, que aparecen en los primeros lugares de la lista son los más representativos y frecuentes.

Los resultados obtenidos tras el cálculo del "Ratio de Asociación" para el nombre *"music"* se muestran en la Tabla 3.

Nombre	Dominio	A.R.
Music	Music	0.240062
Music	Free_time	0.093726
Music	Acoustics	0.072362
Music	Dance	0.065254
Music	University	0.046024
Music	Radio	0.042735
Music	Art	0.020298
Music	Telecommunication	0.006069
…	…	…

Tabla 3: Ratio de Asociación de *"music"*

Como se aprecia en la Tabla 3, los tres dominios más representativos y frecuentes para la palabra *"music"* que aporta todo WordNet Domains son: Music, free-time y acoustics.

3.2. Ratio de Asociación para verbos.

Una vez obtenidos los valores del "Ratio de Asociación" para los nombres, se realiza un proceso similar para obtener los valores asociados a los verbos de WordNet.

Dado que únicamente tenemos etiquetados los nombres de WordNet 1.6. con sus dominios correspondientes en WordNet Domains, lo que se hace es asignar a los verbos que componen las glosas de cada uno de los sentidos el mismo dominio asignado al nombre al que corresponde la glosa. De esta forma, y siguiendo el mismo proceso de análisis sintáctico, extrayendo ahora los verbos de las glosas, se obtendrán los valores del "Ratio de Asociación" para los verbos.

Del mismo modo que para los verbos, se ha realizado el mismo proceso para obtener los valores del "Ratio de Asociación" para los

adjetivos y los adverbios que aparecen en las glosas de WordNet Domains.

4. Método WSD

Según se ha comentado anteriormente el nuevo método WSD que se va a describir tiene las características de los métodos basados en conocimiento. Así, el método WSD, que se propone, utiliza el nuevo recurso denominado Dominios Relevantes como fuente de información para desambiguar el sentido de las palabras que aparecen en un texto.

El método que se presenta consiste básicamente en desambiguar automáticamente el sentido de las palabras que aparecen dentro del contexto de una oración. Normalmente, las palabras que aparecen en un mismo contexto tienen sus sentidos muy relacionados entre sí. Por lo tanto se necesita de una estructura que contenga a aquellos dominios más representativos y frecuentes ordenados por el "Ratio de Asociación" para el contexto de entrada. Esta estructura se ha denominado Vector de Contexto.

Además, cada palabra polisémica contenida en el contexto tendrá diferentes sentidos, por lo tanto, para cada una de ellos también se necesita una estructura que contenga a aquellos dominios más representativos y frecuentes ordenados por el "Ratio de Asociación". Esta estructura se ha denominado Vector de sentidos.

Para obtener el sentido de las palabras que forman el contexto tendré que ver la coincidencia del vector de contexto con respecto a cada uno de los vectores de sentido. Esta coincidencia entre vectores se puede medir por el coseno formados entre ambos. Es decir, cuanto mayor es el coseno mayor coincidencia hay entre los dos vectores comparados.

En las subsecciones siguientes se describirán con mayor detalle cada una de las estructuras definidas anteriormente y su integración en el nuevo método de desambiguación.

4.1. Vector de contexto

El vector de contexto, agrupa en una única estructura los dominios más representativos y frecuentes relacionados con las palabras que forman el texto a desambiguar. Es decir, la información aportada por el conjunto de todas las palabras del texto de entrada (nombres, verbos, adjetivos y adverbios). Con esta información, lo que tratamos de averiguar es qué dominios son los más relevantes en el texto.

Para la obtención de este vector se utiliza la información aportada por el recurso Dominios Relevantes, obtenido a partir de *"WordNet Domains"*. De manera que, una vez extraídos los nombres, verbos, adjetivos y adverbios del texto a tratar, se obtendrían los dominios ordenados mediante los valores del "Ratio de Asociación".

Obtenidos todos los dominios posibles, y dado que muchos de ellos aparecerán repetidamente, se ponderan aquellos que aparecen con mayor frecuencia en el texto. Así obtenemos un vector ordenado, de forma que los dominios que aparecen en los primeros lugares son los más representativos y frecuentes en el texto de entrada.

La representación matemática correspondiente al vector de contexto corresponde a la mostrada en la fórmula (3):

$$VC = \sum_{w \in contexto} RA(W, D) \quad (3)$$

Por ejemplo, dado el siguiente texto de entrada:

"There are a number of ways in which the chromosome structure can change, which will detrimentally change the genotype and phenotype of the organism"

El vector de contexto obtenido tendría los datos mostrados en la Figura 1.

Dominio	**A.R.**

$$VC = \begin{pmatrix} Biology & 0.00102837 \\ Ecology & 0.00402855 \\ Botany & 3.20408e-06 \\ Zoology & 1.77959e-05 \\ Anatomy & 1.29592e-05 \\ Physiology & 0.000226531 \\ Chemistry & 0.000179857 \\ Geology & 1.66327e-05 \\ Meteorology & 0.00371308 \\ ... & ... \end{pmatrix}$$

Figura 1: Vector de contexto

4.2. Vector de sentidos

El vector de sentidos agrupa en una única estructura los dominios más representativos y frecuentes a partir de la glosa asociada a cada uno de los sentidos de las palabras. Es decir, para obtenerlo se parte de la idea de aprovechar la información presente en las glosas asociadas a los sentidos de las palabras que aparecen en un texto. De esta forma, se analizarían sintácticamente las glosas asociadas a cada sentido y se obtendrían sus palabras etiquetadas con su categoría gramatical (nombres, verbos, adjetivos y adverbios), y a continuación se realizarían los mismos cálculos que en el caso del vector de contexto. Obteniendo así, un vector para cada sentido de las palabras del texto.

Por ejemplo, para el sentido número 1 de la palabra *"genotype"*, obtendríamos el vector de sentido mostrado en la Figura 2:

$$VS = \begin{pmatrix} \textbf{Dominio} & \textbf{A.R.} \\ Ecology & 0.084778 \\ Biology & 0.047627 \\ Bowling & 0.019687 \\ Archaeology & 0.016451 \\ Sociology & 0.014251 \\ Alimentation & 0.006510 \\ Linguistics & 0.005297 \\ ... & ... \end{pmatrix}$$

Figura 2: Vector de sentido asociado a *"genotype#1"*

4.3. Comparación de vectores

El nuevo método de desambiguación propuesto comienza con el análisis sintáctico del texto a desambiguar, utilizando para ello el analizador sintáctico *"Tree Tagger"*. A partir de estas palabras etiquetadas con su categoría gramatical se procede a realizar el cálculo del vector de contexto y de los vectores de sentidos. A partir de estos vectores, se necesita saber que vector de sentido es el más cercano al vector de contexto, ya que esto me indica la relación semántica entre ellos. Esta coincidencia entre vectores se mide con el coseno que forman ambos. Así, se seleccionará aquel sentido cuyo coseno con el vector de contexto esté más aproximado a 1, ya que tendrán una mayor coincidencia. El coseno es equivalente al coeficiente de correlación normalizado, el cual, se utiliza para este propósito:

$$\cos(VC, VS) = \sum_{i=1..n} \frac{VC * VS}{\sqrt{\sum_{i=1..n} VC^2} * \sqrt{\sum_{i=1..n} VS^2}} \quad (4)$$

Donde:
VC: Vector de contexto
VS: Vector de sentido

Para la elección del sentido se realizará una comparación de todos los vectores de sentidos con el vector de contexto, seleccionando aquel sentido cuyo vector esté más cercano al vector de contexto.

Por ejemplo, el coseno de cada uno de los vectores de sentido de *"genotype"* con el vector de contexto, sería el siguiente:

genotype#1 = 0.00804111
genotype#2 = 0.00340548

Por lo que el sentido seleccionado para *genotype* sería el #1, dado que su coseno se aproxima más a 1.

5. Evaluación y discusión

En esta sección se evalua el nuevo método WSD propuesto. Para dicha evaluación se han utilizado los textos de la tarea *"english all-words"* de SENSEVAL-2. En estos textos, aparecen etiquetados nombres, verbos, adjetivos y adverbios, de los cuales se conoce su sentido. Estas palabras, son las que se desambiguarán utilizando el nuevo método WSD, para posteriormente comparar los sentidos obtenidos con los sentidos anotados en SENSEVAL-2. Para medir la evaluación del método se utilizan la *precision, recall y coverage*. *Precision* es el número de sentidos correctamente desambiguados dividido por el número de sentidos contestados. *Recall* es el número de sentidos correctamente desambiguados dividido por el número de sentidos totales. *Coverage* es el número de sentidos contestados dividido por el número de sentidos totales

La evaluación del método se ha realizado adoptando diferentes criterios según el tamaño de la ventana que forman el contexto. En la primera evaluación, se ha adoptado el criterio de tomar como entrada al método de WSD, una ventana correspondiente a una oración. De esta forma, el proceso de desambiguación etiqueta el sentido asociado a la palabra ambigua basándose en las palabras que forman la oración.

En este caso, el contexto de la palabra a desambiguar no es muy extenso, debido a que el número de palabras de las que podemos extraer información relevante es muy reducido.

Los resultados obtenidos para esta evaluación y adoptando este criterio se muestran en la fila 1 de la Tabla 4.

En la segunda evaluación, se ha adoptado el criterio de tomar una ventana de 100 palabras alrededor de la palabra a desambiguar. Con esta medida pensamos que la palabra ambigua estará acompañada de más contexto y por lo tanto se asociará el dominio del sentido correcto al del contexto.

Los resultados de esta segunda evaluación se muestran en la fila 2 de la Tabla 4.

Una tercera evaluación se realizó con el objetivo de reducir la especialización de los dominios, es decir se agruparon dominios más especializados en un dominio más general. Por lo tanto se obtuvieron 43 dominios, que son los niveles más representativos, sobre los 165 iniciales. Esta reducción se realiza teniendo en cuenta la estructura jerárquica de los dominios basados para la realización de WordNet Domains. En definitiva, se agrupan dentro de un mismo dominio el conjunto de dominios que pertenecen a su misma jerarquía pero que están en los niveles inferiores.

Por ejemplo, en el caso de la jerarquía del dominio *"Medicine"*, tenemos que por debajo de él se encuentran los dominios: *Dentistry, Pharmacy, Psychiatry, Radiology y Surgery*. Estos dominios se engloban dentro del dominio *"Medicine"*, y así se reduce el espacio de búsqueda y el grado de especialización.

Los resultados de esta tercera evaluación se muestran en la fila 3 de la Tabla 4.

Por último, dado que Wordnet posee una granularidad muy fina, como comentábamos anteriormente, es muy difícil realizar distinciones entre los significados de algunos sentidos. Así pues, en la realización de la última evaluación, se intenta reducir esa granularidad agrupando aquellos sentidos etiquetados con el mismo dominio. Para esta evaluación se utilizaron los 165 dominios de toda la jerarquía. De esta forma, el resultado de la desambiguación para una palabra, no sería un único sentido, sino todos aquellos sentidos de la palabra que tengan asociado el mismo dominio que se obtenga tras el proceso de desambiguación.

Por ejemplo, supongamos que tras el proceso de desambiguación de la palabra *"bank"*, se obtiene el dominio "*Economy"*, entonces daríamos como resultado los tres sentidos

asociados a este dominio: “bank#1, bank#3 y bank#6”.

Los resultados de esta evaluación se muestran en la fila 4 de la Tabla 4.

Criterio	Precision	Recall	Coverage
Oración	0.44	0.32	0.71
Ventana 100 palabras	0.47	0.38	0.81
43 dominios	0.48	0.41	0.85
Desambiguación a nivel de dominio	0.54	0.43	0.85

Tabla 4: Resultados obtenidos en las evaluaciones del método WSD

6. Conclusiones y trabajos futuros

7. Bibliografía

Church K. and Hanks P., *Word association norms, mutual information, and leogicography*. Computational Linguistics, vol. 16, ns. 1, 22-29. 1990. Also in proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL’89). Pittsburg, Pennsylvannia, 1989.

Ide N. and Véronis J. (1998) *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art*. Computational Linguistics. 24 (1), 1-40.

Killgarriff A. and Yallop C. *What’s in a thesaurus?* In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June 2000.

Magnini B. and Cavagliá G., *Integrating Subject Field Codes into WordNet.* In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June 2000

Magnini B. and Strapparava C., *Experiments in Word Domain Disambiguation for Parallel Texts.*In Proc. Of SIGLEX Workshop on Word Senses and Multi-linguaty, Hong-Kong, October 2000.

Schmid Helmut (1994) *Probabilistic part-of-speech tagging using decision tre,* Proceedings International Conference on New Methods in Language Processing.Manchester, pp 44-49. UK

Wilks Y. And Stevenson M. (1996) *The grammar of sense: Is word sense tagging much more than part-of-speech tagging?* Technical Report CS-96-05, University of Sheffield, UK.

Wilks Y. and Stevenson M. Word Sense Disambiguation using optimised combination of knowledge sources. In Proc. Of COLING-ACL’98, 1998.