L. ALFONSO UREÑA LÓPEZ
MANUEL GARCÍA VEGA
FERNANDO MARTÍNEZ SANTIAGO
Departamento de Informática. Universidad de Jaén
Jaén, España
E-Mail: {laurena, mgarcia, dofer}@ujaen.es

Explotando las Relaciones Léxicas y Semánticas de WORDNET en la Resolución de la Ambigüedad Léxica

Resumen

La etiquetación semántica es la asignación automática del significado apropiado de las palabras de un texto de acuerdo con los sentidos proporcionados por un lexicón. La mejora en la identificación del significado correcto de las palabras puede ser muy útil para variadas aplicaciones del procesamiento de lenguaje natural. En este trabajo se muestra que adaptando el modelo del espacio vectorial a la resolución de la ambigüedad léxica (WSD) con las relaciones léxicas y semánticas de WORDNET se obtienen resultados significativos, especialmente en cuanto a la precisión obtenida. Mostraremos también en qué medida se produce ese incremento según la relación utilizada de las disponibles en WORDNET.

Palabras clave: Word Sense Disambiguation (WSD), Text Categorization (TC), WORDNET, SEMCOR, Information Retrieval (IR), Ventana Contextual (VC), Modelo del Espacio Vectorial (MEV).

Introducción

La tarea de la resolución de la ambigüedad léxica del significado de las palabras consiste en identificar el significado de una palabra en un determinado contexto. Esta es una tarea que puede incrementar muy significativamente el rendimiento de algunas aplicaciones del procesamiento en lenguaje natural [Kilgarriff97], como Categorización de Texto (TC) [Buenaga97]; traducción automática [Brown91]; restauración de acentos [Yarowsky94]; encaminamiento y filtrado de textos; agrupamiento y segmentación de textos y, en general, en la recuperación de información [Salton83, Kilgarriff97, Sanderson96].

Un recurso lingüístico a escala real como es WORDNET puede considerarse muy importante para WSD, ya que proporciona palabras semánticamente relacionadas. Por ejemplo, *plant, flora, plant life* se pueden identificar como ocurrencias del mismo concepto *"a living organism lacking the power of locomotion",* frente a *plant, works, industrial plant* que se identifican como ocurrencias de otro concepto (significado) *"building for carrying on industrial labor".* Se puede utilizar la sinonimia en WORDNET para medir la distancia semántica entre ocurrencias de términos y así obtener maneras más sofisticadas de comparar los términos consultados con los posibles significados. Asimismo, se pueden utilizar otras relaciones, además de la propia sinonimia como la hiperonimia, hiponimia, meronimia, etc.

Hemos fundamentado nuestro algoritmo en el Modelo del Espacio Vectorial (MEV) [Salton83], utilizando el concepto de *Ventana Contextual* de tamaño variable [Ureña98a]. Los vectores de pesos son calculados para cada ventana contextual, empleando WORDNET. Hemos calculado el vector de pesos basándose en todas las relaciones léxicas y semánticas procedentes de WORDNET.

Hemos realizado un conjunto de experimentos, centrándonos en nombres y verbos, mostrando que un enfoque utilizando las relaciones léxicas y semánticas mejora la efectividad de la resolución de la ambigüedad léxica.

Este trabajo está organizado como sigue. En primer lugar, introducimos la tarea de resolución de la ambigüedad léxica y los recursos utilizados. Seguidamente, describimos el modelo en el que estos elementos son integrados y la combinación de las relaciones léxicas. Después de esto, presentamos nuestra evaluación y los resultados obtenidos, y finalmente, describimos nuestras conclusiones y trabajos futuros.

Utilización de las relaciones léxicas y semánticas conceptuales de WORDNET

WORDNET [Miller95, Landes98] proporciona un número importante de relaciones conceptuales, léxicas y semánticas, además de la sinonimia y antonimia que poseen muchos de los thesaurus existentes. Estas relaciones se pueden utilizar para ingeniar un cambio en la representación de la información textual, transformando los vectores de palabras en vectores de significados. Así por ejemplo, la relación de sinonimia se utiliza para clasificar las palabras con significados similares (o

usos similares). Y la relación de hiperonimia[1] (que corresponde a la relación "es un") se utiliza para generalizar los significados de los nombres en el nivel más alto de abstracción, como se ilustra en figura 1.

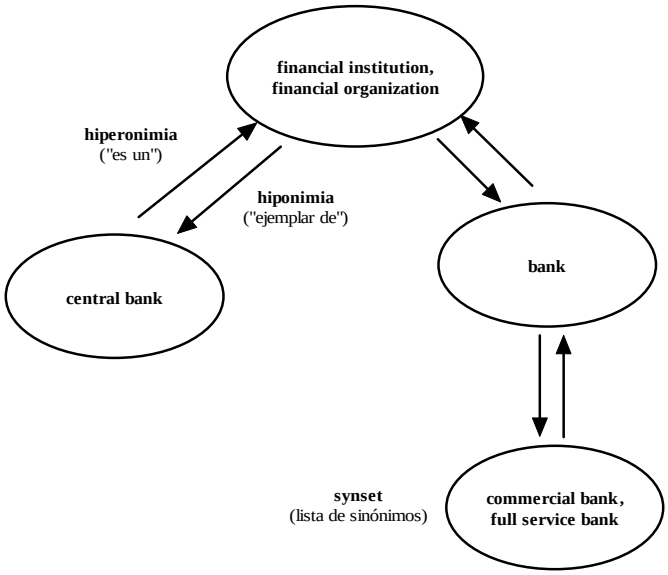

Figura 1. Fragmento de la relación jerárquica en WORDNET.

En otros trabajos de desambiguación se analiza la expansión de términos que denominamos “expansión de sentidos" con información meramente de sinonimia [Ureña98b, Agirre96] en el entrenamiento. Sin embargo, en este enfoque se estudia el aporte de las propias relaciones de WORDNET a WSD, no sólo mediante información de sinonimia, sino a través de diferentes relaciones léxicas y semánticas como la hiperonimia, meronimia, hiponimia, antonimia, holonimia, etc. Para explotar toda esta información hemos recurrido al concepto de Ventana Contextual [Ureña98a]. La VC de un término atendiendo a su significado está formada por todos los términos que se relacionan con tal término, según las relaciones léxicas y semánticas existentes en WORDNET. Existe por lo tanto, una ventana contextual para cada significado de un término y para cada una de las relaciones de que disponga el término dado. Gracias a que mantenemos en ventanas contextuales independientes cada una de las relaciones de que dispone el término, podemos estudiar con qué tipo de relaciones se obtiene mayor precisión.

La información que aporta WORDNET con sus relaciones léxicas y semánticas nos puede ayudar a predecir el significado de una palabra. El proceso de entrenamiento[2] se realiza como sigue. Se representa cada término de WORDNET con un vector, conforme a [Ureña98a], cuyas componentes son: el peso del término y los pesos de los términos que constituyen la ventana contextual para cada una de las relaciones mencionadas, si es que existen. Así, para cada uno de los términos del conjunto de entrenamiento, calcularemos su ventana contextual, construyendo tantas ventanas como palabras con diferentes sentidos existan en la colección, así como relaciones existan para ese término. La fase de prueba se realiza confrontando la consulta, por medio del cálculo de la similitud con los vectores creados en el entrenamiento[3] con WORDNET.

Los experimentos realizados no se limitan a estudiar por separado qué aporta cada una de las relaciones, también realizaremos una combinación de las distintas relaciones. Para llevar a cabo este proceso, proponemos un algoritmo desambiguador fundamentado en técnicas probabilísticas, concretamente un algoritmo Monte Carlo *p-correcto*, donde *p* es la precisión obtenida en los entrenamientos para cada una de las relaciones de WORDNET propuestas. Al tomar la precisión obtenida en el entrenamiento como probabilidad de acierto, la construcción del algoritmo es

[1]Relación de inclusión de unidades léxicas que va de lo más específico a lo más general
[2]La totalidad de los nombres que se encuentran en WORDNET ha constituido el conjunto de entrenamiento.
[3]Es decir, con las distintas ventanas contextuales.

inmediata, ya que basta con hacer las desambiguaciones para todas las relaciones y ponderar los sentidos obtenidos con la precisión correspondiente. De esta forma, tenemos en cuenta la aportación semántica de todas ellas, devolviendo el sentido más probable.

Experimentos y evaluación

En la tabla 1 se muestran los resultados para cada una de las relaciones proporcionadas por WORDNET. Podemos considerar como relaciones para nuestro enfoque las siguientes: sinonimia, hiponimia, hiperonimia, meronimia, holonimia, antonimia y coordinados, destacando la gran precisión para todas ellas.

Relaciones WORDNET	Microaveraging	Macroaveraging	Recall
Ants	93,33 %	89,14 %	0,12 %
Holon	80,91 %	80,29 %	0,30 %
Hholon	79,70 %	77,83 %	0,36 %
Hypo	65,05 %	66,04 %	2,64 %
Tree	61,31 %	62,01 %	4,30 %
Meron	64,61 %	60,82 %	0,48 %
Hmeron	60,20 %	60,65 %	0,82 %
Hype	43,06 %	43,17 %	7,52 %
Coorn	43,41 %	46,08 %	10,22 %
Syns	25,29 %	35,67 %	2,82 %
Todas	51,43 %	52,93 %	19,85 %
Combinación de relaciones	76,11 %	75,64 %	30,61 %

Tabla 1. Resultados obtenidos por las distintas relaciones léxicas y semánticas del enfoque basado en la evaluación del Brown1

Se ha empleado como colección de prueba para realizar la evaluación de este enfoque, el Brown1 de SEMCOR [Brown91] en su totalidad. Y como colección de entrenamiento se han utilizado todos los nombres y verbos de WORDNET 1.6, construyendo tablas de entrenamiento independientes relativas a cada relación.

En los resultados de este método sí aparece un valor de recall para cada relación empleada, esto es como consecuencia de que este enfoque no decide en la totalidad de los casos. Por ejemplo, podemos observar la gran precision obtenida por este enfoque con la relación de holonimia.

Nótese como este enfoque para cada una de las seis primeras relaciones de la tabla obtiene una precisión superior al 60 %. Un resultado quizás sorprendente es la precisión tan alta, del orden del 90 %, que obtienen ambos algoritmos empleando la relación de antonimia. Este valor tan alto sorprende y parece paradójico, pues cabría esperar mayor precisión con otras relaciones más que con la propia antonimia, sin embargo el recall es extremadamente bajo, sólo en el 0,1 % de los casos aproximadamente es capaz de decidir.

En virtud de los datos reflejados en la tabla 1, podemos afirmar que las relaciones más equilibradas en cuanto a la relación *precision-recall* son la sinonimia e hiperonimia. Además obsérvese que los mejores valores de *recall* son los obtenidos por las relaciones de hiperonimia (hype) y coordinados (coorn), lo que provoca una pequeña disminución de precision.

La fila marcada como “Todas” no es una relación, sino la unión de todas ellas. Como podemos apreciar en esta unión la precision supera el 50 %, y el recall aunque es superior comparado con cada una de las relaciones, es relativamente bajo (21 %).

Combinar los resultados mostrados en la tabla 1 puede llevarnos a mejorar el desambiguador. Podemos ver la precisión como una probabilidad de acierto en la desambiguación con las relaciones léxicas y semánticas. Si hacemos un mismo experimento para todas las relaciones, y ponderamos los resultados según dichas relaciones, obtendremos la probabilidad de cada uno de los sentidos posibles.

Por último, a la vista de los resultados, podemos afirmar que WORDNET es una herramienta de muy buena calidad para la desambiguación y discriminación de los sentidos de las palabras, aportando

información tremendamente relevante con el contexto que hemos definido. Sin embargo, no es perfecta, ya que no cubre la totalidad de los contextos, por lo que estamos limitados en recall, y por tanto, este enfoque de desambiguación combinando la relaciones no es capaz de decidir en el 70% de los casos aproximadamente.

Conclusiones y futuro trabajo

En este trabajo se ha presentado un enfoque automático WSD para nombres y verbos, basado en la utilización de WORDNET, demostrando la explotación y aporte de todas las relaciones de WORDNET a la resolución de la ambigüedad léxica, permitiendo cuantificar la relevancia de cada una de esas relaciones. Los resultados obtenidos son positivos, a pesar de la dificultad de la tarea y de la gran variedad de sentidos definidos en WORDNET por palabra.

La aportación que puede realizar un recurso lingüístico como WORDNET al problema de la resolución de la ambigüedad léxica es bastante más significativa en términos de precisión que de *recall*. Por lo tanto, nuestros próximos esfuerzos deben encaminarse a incrementar los valores de *recall*. Posiblemente, tal incremento pase por la integración de las relaciones de WORDNET con un corpus de entrenamiento de manera análoga al propuesto en [Ureña 98b]. Asimismo, se pretende incorporar nuevas fuentes de conocimiento al sistema de resolución de la ambigüedad léxica.

Referencias

[Agirre96] Agirre E., Rigau G. *Word sense disambiguation using conceptual density.* In Proceedings of COLING 1996.

[Buenaga97] Buenaga Rodríguez M., Gómez Hidalgo J. M., Díaz Agudo B. *Using WORDNET to Complement Training Information in Text Categorization.* Second International Conference on Recent Advances in Natural Language Processing, 1997

[Brown91] Brown P. B., Pietra S. A., Pietra V. J.*Word Sense Disambiguation Using Statiscal Methods*. In Proc. Of ACL, pp. 264-270, 1991.

[Kilgarriff97] Kilgarriff A *What is word sense disambiguation good for?* Proc. Natural Language Processing Pacific Rim Symposium. Phuket, Thailand. December 1997. pp 209-214.

[Landes98] Landes, S., Leacock, C., Tengi, R., Building semantic concordances. In WORDNET: An Electronic Lexical Database. MIT Press. 1998.

[Miller95] Miller G. *WORDNET: lexical database.* Communications of the ACM Vol 38, No. 11.

[Salton83] Salton G., McGill, M.J.*Introduction to modern information retrieval*. McGraw-Hill.1983.

[Sanderson96] Sanderson, M., *Word sense disambiguation and information retrieval*. Ph.D. Thesis, Department of Computing Science, University of University of Glasgow. 1996.

[Ureña98a] Ureña López, L. A., García Vega, M., Buenaga Rodríguez, M., Gómez Hidalgo, J. M. *Resolución automática de la ambigüedad léxica fundamentada en el modelo del espacio vectorial usando ventana contextual variable.* Asociación Española de Lingüística Aplicada. AESLA. 1998.

[Ureña98b] Ureña López, L. A., García Vega, M., Gómez Hidalgo, J. M., Díaz Esteban, A. *Integrando una Base de Datos Léxica y una Colección de Entrenamiento para la Desambiguación del Sentido de las Palabras.* Procesamiento del Lenguaje Natural, Revista nº 23, septiembre 1998.

[Yarowsky94] Yarowsky D. *Decision lists for lexical ambiguity resolution: Aplication to accent restoration in Spanish and French*. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL'94).1994.