Procesamiento del Lenguaje Natural, num. 28 (2002), pp.121-124

Desambiguación léxica mediante Marcas de Especificidad*

Andrés Montoyo
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
montoyo@dlsi.ua.es

Resumen: Esta Tesis presenta un método para resolver la ambigüedad léxica pura (semántica) en textos de dominios no restringidos en cualquier lengua que tenga un repositorio de sentidos organizado como una base de conocimiento léxica. A este método de resolución de la ambigüedad léxica pura propuesto se denomina Método de Marcas de Especificidad y se basa en el uso de conocimiento lingüístico (información léxica y morfológica) y de conocimiento a partir de las relaciones léxicas y semánticas de la taxonomía de nombres de la base de conocimiento léxica WordNet. Además, se presenta la aplicación del método de Marcas de Especificidad con el objetivo de enriquecer semánticamente WordNet con etiquetas de dominio o categorías de otros sistemas de clasificación.
Palabras clave: Ambigüedad léxica, Word Sense Disambiguation, WordNet, Enriquecimiento de WordNet semánticamente, sistemas de clasificación.

Abstract: This Thesis presents a method for the automatic disambiguating of nouns, using the notion of Specification Marks and employing the noun taxonomy of the WordNet lexical knowledge base. The method resolves the lexical ambiguity of nouns in any sort of text, and although it relies on the semantic relations (Hypernymy and Hyponymy) and the hierarchic organization of WordNet, it does not, however, require any sort of training process, no hand-coding of lexical entries, nor the hand-tagging of texts. Besides, this Thesis presents a new method to enrich semantically WordNet with categories from general domain classification systems. The method is performed in two consecutive steps. First, a lexical knowledge word sense disambiguation process. Second, a set of rules to select the main concepts as representatives for each category. The method has been applied to label automatically WordNet synsets with Subject Codes from a standard news agencies classification system.
Keywords: lexical ambiguity, Word sense disambiguation, WordNet, Classification Systems, WordNet Enrichment.

Esta Tesis Doctoral presentada por Andrés Montoyo Guijarro para la obtención del título de Doctor Ingeniero en Informática ha sido dirigida conjuntamente por Dr. Manuel Palomar Sanz de la Universidad de Alicante y por Dr. German Rigau Claramunt de la Universitat Politècnica de Catalunya. El tribunal de Tesis, compuesto por los doctores Lluís Padró Cirera de la Universitat Politècnica de Catalunya, Julio Gonzalo Arroyo de la Universidad Nacional de Educación a Distancia, Eneko Agirre Bengoa de la Universidad del País Vasco, Ferrán Pla Santamaría de la Universidad Politécnica de Valencia y Patricio Martínez Barco de la Universidad de Alicante, le concedió por unanimidad la calificación de Sobresaliente *Cum Laude*.

1 *Introducción*

Todos los sistemas de Procesamiento del Lenguaje Natural (PLN) tienen asociado un problema común difícil de resolver, sus diferentes ambigüedades. Por este motivo, cuando se diseña un sistema de PLN, uno de los objetivos fundamentales es resolver sus múltiples ambigüedades (estructural, léxica, ámbito de cuantificación, función contextual y referencial) mediante la definición de procedimientos específicos para cada una de estas.

En concreto en esta Tesis nos centraremos en la resolución de la ambigüedad léxica, la cual se presenta cuando, al asociar a cada una de las palabras del texto la información

* Esta investigación ha sido parcialmente financiada por el Ministerio de Ciencia y Tecnología a través del proyecto núm. TIC2000-0664-C02-01/02

ISSN 1135-5948

léxico-morfológica, hay palabras que tienen más de un sentido o significado. Al campo que se encarga del estudio y resolución del problema de la ambigüedad léxica pura se conoce como *Desambiguación del sentido de las palabras* (WSD, en inglés *Word Sense Disambiguation*).

Para WSD existen diferentes métodos de trabajo como puede verse en el trabajo de (Ide y Véronis, 1998), sin embargo, la presente Tesis se centra en el método que se basa en el emparejamiento del contexto de la palabra a ser desambiguada con cualquier información de un recurso de conocimiento léxico externo, conociéndose como desambiguación del sentido de las palabras basada en el conocimiento (WSD knowledge-driven). Particularmente, para este trabajo se ha utilizado WordNet como base de conocimiento léxica.

Muchas investigaciones, sobre WSD basado en el conocimiento, han sido realizadas durante los últimos años. El trabajo de (Lesk, 1986) propone un método para descifrar el sentido de una palabra en un contexto, contando el número de coincidencias que aparecen entre el contexto y la definición del diccionario. (Cowie, Guthrie, y Guthrie, 1992) describen un método para resolver la ambigüedad léxica de textos usando la definición dada en Longman´s Dictionary of Contemporary English (LDOCE) obteniendo unos resultados del 47% en cuanto a distinguir los sentidos y un 72% para homógrafos. (Yarowsky, 1992) deriva clases de palabras a partir de palabras en categorías comunes del Roget´s International Thesaurus. (Wilks et al., 1990) utilizan co-ocurrencia de datos, extraídos del LDOCE, para construir vectores de contexto y de sentidos asociados a las palabras. (Voorhees, 1993) define la construcción denominada hood utilizando los hipónimos para nombres en WordNet. (Sussna, 1993) define una métrica basada en la distancia semántica entre los términos de un texto, la cual consistía en asignar pesos a los enlaces de WordNet según los tipos de relación (sinónimos, hiperónimos, etc) y en contar el número de arcos del mismo tipo que salen del nodo y la profundidad del arco en total. (Resnik, 1995) define una métrica basada en la similaridad semántica para las palabras en la jerarquía WordNet. (Agirre y Rigau, 1996) describe un algoritmo no supervisado usando la Distancia Conceptual para desambiguar nombres en Semcor. (Rigau, Agirre, y Atserias, 1997) combina un conjunto de algoritmos no supervisados para desambiguar el sentido de las palabras en un corpus no etiquetado. (Stetina, Kurohashi, y Nagao, 1998) introduce un método para WSD, basado en un corpus de entrenamiento etiquetado sintácticamente y semánticamente. Este método explota la información del contexto de la oración y sus relaciones semánticas. (Mihalcea y Moldovan, 1999) exponen un método para desambiguar nombres, verbos, adverbios y adjetivos de un texto, referenciando el sentido proporcionado por WordNet.

En esta Tesis se presenta un método, que resuelve la ambigüedad léxica de nombres, basándose en el conocimiento que nos proporciona la taxonomía semántica de nombres de WordNet.

2 *Aportaciones de la Tesis*

Las principales aportaciones de esta Tesis son:

- La definición de la noción de Marca de Especificidad, la cual formaliza la relación entre sentidos que se basan en taxonomías. La Marca de Especificidad se aprovecha de la información suministrada por las relaciones de sinonimia, hiperonimia e hiponimia. Aunque en esta Tesis se ha utilizado para nombres, se puede adecuar para trabajar con verbos. Con la utilización de las Marcas de Especificidad se aporta una noción como base para la desambiguación del sentido de las palabras, los cuales están unidos a través de conceptos en las taxonomías. Además, la aplicación de esta noción no requiere desambiguación manual previa para aplicarse.
- La definición de un método basado en las Marcas de Especificidad, el cual utiliza el paradigma de los métodos basados en conocimiento, y en concreto el conocimiento que suministra la base de conocimiento léxica WordNet. Debido a las características de las Marcas de Especificidad, se ha diseñado un método que desambigua nombres de un texto libre según estén organizados los sentidos de las palabras en la taxonomía. Por tal motivo, este método puede aplicarse a textos de diferentes idiomas sin realizar adaptaciones del mismo, siempre que

tenga un WordNet.

- La definición de un conjunto de heurísticas para mejorar la propuesto del método inicial, que denominamos método Marcas de Especificidad. Estas heurísticas utilizan el conocimiento suministrado por WordNet, como son el *synset* y la glosa en su idioma, por lo tanto también pueden aplicarse a textos de diferentes idiomas sin realizar adaptaciones.

- El diseño y desarrollo de una interfaz para implementar el método Marcas de Especificidad. Esta interfaz puede ser ejecutada desde cualquier ordenador que tenga acceso a Internet y que acceda a la dirección `http://gplsi.dlsi.ua.es/wsd`.

- El entrenamiento y ajuste del método Marcas de Especificidad implementado. Para ello el método se ha validado y ajustado con el objetivo de estudiar su efectividad mediante un trabajo experimental dividido en tres experimentos. El experimento 1 del capítulo *Experimentación* consiste en comprobar el funcionamiento del método cuando se aplica solamente la noción de Marca de Especificidad. El experimento 2 consiste en comprobar que al complementar el método con un conjunto de heurísticas, estas aportan mejores porcentajes de desambiguación y por lo tanto mejoran el método. Y el experimento 3 consiste en analizar y definir la ventana óptima de contexto para obtener la mejor desambiguación. Con estos tres experimentos se consigue que el método desambigue los sentidos de las palabras sobre WordNet con una "precisión" del 67,1 %, una "cobertura " del 66,2 % y una "cobertura absoluta" del 98,5 %.

- La comparación del método de Marcas de Especificidad con otros métodos de WSD y su evaluación final. Se ha realizado una comparación *directa* e *indirecta* del método mediante dos experimentos. En el experimento 4 del capítulo *Experimentación* se realiza una comparación con métodos basados en el conocimiento, es decir métodos WSD que pertenecen a la misma clasificación que el propuesto en esta Tesis. Y en el experimento 5 se realiza una comparación entre el método propuesto y uno basado en corpus, concretamente en un modelo probabilístico que utiliza el principio de Máxima Entropía. A partir de esta comparación y comprobar los resultados obtenidos, podemos probar que el método de Marcas de Especificidad es útil para desambiguar el sentido de las palabras, ya que se han obtenido mejores resultados que otros métodos basados en conocimiento, y en concreto el conocimiento que suministra la base de conocimiento léxica WordNet.

  Para la evaluación final del método Marcas de Especificidad se han realizado dos experimentos. El experimento 6 consiste en participar con el método en la competición de sistemas de WSD denominada SENSEVAL-2 para los nombres seleccionados por el comité en la tarea de "lexical sample" tanto para inglés como para español. Y el experimento 7 consiste en evaluar al método cuando se aplica con las heurísticas activadas en cascada o secuencialmente y cuando se aplican independientemente unas de otras, y en ambos casos sobre todos los documentos del corpus *SemCor*.

- La definición de un método para enriquecer semánticamente el recurso léxico WordNet con categorías o clases de otros sistemas de clasificación, mediante la aplicación del método de Marcas de Especificidad. El sistema de clasificación utilizado para etiquetar automáticamente los *synsets* de WordNet ha sido IPTC Subject Reference System.

  El recurso léxico WordNet, también muy utilizado en PLN, presenta una división de los sentidos de las palabras con demasiado detalle (en inglés conocido como fine-grained). Por eso, las categorías, como *Agriculture, Health, etc*, aportan una forma más adecuada para diferenciar los sentidos de las palabras. Por lo tanto para tratar y resolver este problema asociado a WordNet, se define y describe un método automático para enriquecer semánticamente WordNet versión 1.6. con las categorías utilizadas en el sistema de clasificación IPTC.

- El diseño y desarrollo de la interfaz construida para extender y mejorar la base

de datos léxica WordNet con categorías del sistema de clasificación. Gracias al desarrollo de esta interfaz se han podido realizar los experimentos para evaluar la efectividad del método propuesto para enriquecer WordNet. Como resultado se obtuvo una "cobertura absoluta" del 93.7 %, una "precisión" del 95.7 % y una "cobertura" del 89.8 %.

Como conclusión final indicaremos que en esta Tesis hemos estudiado dos métodos, uno aplicado a la desambiguación del sentido de las palabras para textos no restringidos en cualquier idioma que tenga un WordNet particular y otro aplicado al enriquecimiento de WordNet con categorías de sistemas de clasificación. En ambos métodos los resultados obtenidos son satisfactorios y consideramos que el método de Marcas de Especificidad es útil para WSD y el de enriquecimiento de WordNet es útil para mejorar y extender la base de datos léxica WordNet con categorías de los sistema de clasificación.

Bibliografía

Agirre, Eneko y German Rigau. 1996. Word Sense Disambiguation using Conceptual Density. En *Proceedings of the 16th International Conference on Computational Linguistic (COLING´96)*, Copenhagen, Denmark.

Cowie, Jim, Joe Guthrie, y Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. En *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, páginas 359–365, Nantes, France.

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

Lesk, Michael. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. En *Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery*, páginas 24–26, Toronto, Canada.

Mihalcea, Rada y Dan Moldovan. 1999. A Method for word sense disambiguation of unrestricted text. En *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, páginas 152–158, Maryland, Usa.

Resnik, Philip. 1995. Disambiguating noun groupings with respect to WordNet senses. En *Proceedings of the Third Workshop on Very Large Corpora*, páginas 54–68, Cambridge, MA.

Rigau, German, Eneko Agirre, y Jordi Atserias. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. En *Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97*, Madrid, Spain.

Stetina, J., S. Kurohashi, y M.Ñagao. 1998. General word sense disambiguation method based on full sentencial context. En *Proceedings of Usage of WordNet in Natural Language Processing. COLING-ACL Workshop*, Montreal, Canada.

Sussna, Michael. 1993. Word sense disamiguation for free-text indexing using a massive semantic network. . En *Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM´93*, páginas 67–74, Arlington, VA.

Voorhees, Ellen. 1993. Using WordNet to disambiguate word senses for text retrieval. En *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 171–180, Pittsburgh, PA, June/July.

Wilks, Yorick, Dan Fass, Cheng-Ming Guo, James E. MacDonald, Tony Plate, y Brian Slator. 1990. Providing machine tractable dictionary tools. En James Pustejovsky, editor, *Semantics and the Lexicon*. MIT Press, Cambridge, MA.

Yarowsky, David. 1992. Word sense disambiguation using statistical models of Roget´s categories trained on large corpora. En *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, páginas 454–460, Nantes, France.