Procesamiento del Lenguaje Natural, num. 28 (2002), pp.129-130

Tratamiento y resolución de las descripciones definidas y su aplicación en sistemas de extracción de información*

Rafael Muñoz
Universidad de Alicante
Campus de Sant Vicent del Raspeig
rafael@dlsi.ua.es

Resumen: Este trabajo presenta un sistema de resolución de las descripciones definidas en español. Este sistema está basado en la agrupación semántica de los sintagmas nominales que han aparecido previamente en el texto a la expresión anafórica (descripción definida) utilizando la ontología de EuroWordNet. Para la resolución de las referencias producidas se utiliza diversas fuentes de información como son la información léxica, sintáctica y semántica. Además, se presenta un estudio de la influencia de su resolución en los sistemas de extracción de información.
Palabras clave: fenómenos lingüísticos, resolución de la anáfora, descripciones definidas, extracción de información

Abstract: This work presents a definite description resolution system in Spanish. This system is based on the clustering of previous noun phrases to the anaphoric expression (definite description) using the EuroWordNet's ontology. To solve the references produced by definite description different sources are used: lexical, syntactic and semantic information. Moreover, we also present a study of the definite description resolution influence in information extraction systems.
Keywords: linguistic phenomena, anaphora resolution, definite description, information extraction

1 *Introducción*

Esta Tesis fue dirigida por Dr. Manuel Palomar Sanz, y presentada en la Universidad de Alicante con fecha 30 de mayo de 2001. El tribunal estuvo formado por: Dra. Dña. Lidia Moreno Boronat, Dr. D. Antonio Ferrández Rodríguez, Dra. Dña. Encarna Segarra Soriano, Dra. Dña. Arantxa Díaz de Ilarraza, y Dr. D. Ruslan Mitkov. La calificación obtenida fue de Sobresaliente *cum laude*.

2 *Resumen de la Tesis Doctoral*

En este trabajo se presenta un sistema para la resolución de las descripciones definidas (sintagmas nominales introducidos por un artículo definido o por un demostrativo) en español y su aplicación al sistema EXIT (Llopis et al., 1998).

Se ha presentado un sistema de resolución de las DDs desde dos enfoques. El primer enfoque realiza en un mismo proceso la identificación de las DDs no anafóricas y la resolución de las DDs anafóricas basado en un conjunto de heurísticas aplicadas en forma de filtro. Por otro lado, en el segundo enfoque, se plantean dos procesos, uno de ellos para la identificación de algunas de las DDs no anafóricas basado en la generación automática de una red semántica y otro para la resolución de las DDs anafóricas basado en un conjunto de heurísticas aplicadas en base a su factor de importancia (sistema de pesos). El marco global de aplicación de la resolución de las DDs ha sido un sistema de extracción de información, aunque consideramos que ambos enfoques por sus características son independientes del dominio y aplicables a textos no restringidos. Las principales fuentes de información utilizadas en ambos enfoques han sido las informaciones del tipo léxico-morfológico, sintáctico y semántico.

Las principales contribuciones son:

- Estudio y clasificación de los diferentes tipos de DDs que se encuentran en un texto escrito en español (Muñoz, Palomar, y Ferrández, 2000). Se desarrolló una clasificación de los diferentes tipos de DDs. Esta clasificación, realizada para el español, es la primera que existe desde un punto de vista de la información necesaria para poder establecer una relación anafórica entre la DD y su posible antecedente. En esta clasificación

* Esta investigación ha sido parcialmente financiada por la CICYT a través del proyecto núm. TIC2000-0664-C02-01/02

ISSN 1135-5948

se distinguen dos grandes grupos: las DDs *no anafóricas* y las DDs *anafóricas*. Las DDs anafóricas, se establece una nueva división en función del tipo de información necesaria para su resolución: *uso anafórico de nivel sintáctico-semántico-textual* y uso *anafórico a nivel pragmático*.

- Desarrollo de dos enfoques diferentes en la resolución de las DDs. El primer enfoque (Muñoz y Palomar, 2000; Muñoz, Palomar, y Ferrández, 2000) realiza la identificación de las DDs no anafóricas y la resolución de las DDs anafóricas en un único proceso mientras que el segundo enfoque (Muñoz y Palomar, 2001) realiza dos procesos. El primer enfoque a pesar de realizar la identificación de las no anafóricas y la resolución de las anafóricas en un único paso conlleva un mayor gasto computacional ya que aplica el algoritmo de resolución a todas las DDs, y se ha demostrado que mas del 50% de ellas son no anafóricas. Además, para buscar el antecedente se utilizan todos los SN previos que han aparecido en el texto, por lo que en textos grandes el sistema se vuelve lento. Para evitar estos problemas se desarrolló el segundo enfoque que se basa en la agrupación de los SN previos en función a su compatibilidad semántica. Si es la única que pertenece a una determinada clase semántica se identifica como no anafórica y si más SN que pertenezcan a la misma clase se utilizan éstos como lista de candidatos para encontrar el antecedente. Con este segundo enfoque se aplica el algoritmo de resolución sólo a las anafóricas y se reduce la lista de candidatos manteniendo la efectividad del sistema.

- La evaluación del sistema obtuvo los siguientes resultados. En la tarea de identificación de las DDs no anafóricas, el primer enfoque alcanza una cobertura del 88% y una precisión del 93,1%, mientras que el segundo enfoque obtiene una cobertura del 88,3% y una precisión del 93,5%. En la tarea de resolución de las referencias, el primer enfoque obtiene una cobertura del 71,5% y una precisión del 75,6%, mientras que el segundo enfoque alcanza una cobertura del 75,3% y una precisión del 79,5%.

- Aplicación del sistema de resolución de las DDs a un sistema de EI (Palomar y Muñoz, 2000). La aplicación de este módulo a sistemas de extracción de información incrementa la efectividad de dichos sistemas. Los textos utilizados por los sistemas de EI no suelen tener un formato fijo, aunque trabajen en un dominio restringido, y la información a extraer aparece de manera implícita (dispersa) en el texto. Es decir, que no es normal encontrar en los textos a tratar la información a extraer en el formato exacto. La resolución de las correferencias, en general, y la de las descripciones definidas, en particular, ayudan a establecer las relaciones necesarias entre todas las partes del texto, de forma que la información implícita se transforma en información explícita que facilita el relleno de las plantillas correspondientes.

Bibliografía

Llopis, F., R. Muñoz, A. Suárez, y A. Montoyo. 1998. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Novática*, 133:26–30.

Muñoz, R. y M. Palomar. 2000. Processing of Spanish Definite Descriptions with the Same Head. En Dimitris Christodoulakis, editor, *Proceeding of NLP2000*, volumen 1835 de *Lectures Notes in Artificial Intelligence*, páginas 212–220, Patras, Greece. Springer-Verlag.

Muñoz, R. y M. Palomar. 2001. Clustering technique based on semantic for definite description resolution. En *Proceedings of the International Conference Text Speech and Dialogue*, volumen 2166 de *Lecture Notes in Artificial Intelligence*, Czech Republic. Springer-Verlag.

Muñoz, R., M. Palomar, y A. Ferrández. 2000. Processing of Spanish Definite Descriptions. En *Proceeding of MICAI'2000*, volumen 1793 de *Lectures Notes in Artificial Intelligence*, páginas 526–537, Acapulco, Mexico. Springer-Verlag.

Palomar, M. y R. Muñoz. 2000. Definite Descriptions in Information Extraction System. En Carolina Monmard y Jaime Simao, editores, *Proceeding of Iberamia-SBIA'2000*, Lectures Notes in Artificial Intelligence, páginas 320–328, Atibaia, São Paulo, Brazil. Springer-Verlag.