Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística

A. Ferrández; M. Palomar; L. Moreno; P. Martínez-Barco; J. Peral; R. Muñoz; M. Saiz-Noeda
{antonio, mpalomar, patricio, jperal, rafael, max}@dlsi.ua.es
lmoreno@dsic.upv.es

Dept. Lenguajes y Sistemas Informáticos
Univ. de Alicante - Apt. 99 – 03080 – Alicante

Dept. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

1. Introducción.

A continuación se presenta un sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística, que se caracteriza por su flexibilidad y modularidad, así como por su capacidad de operar sobre textos no restringidos.

En el campo del tratamiento de la correferencia lingüística, el sistema resuelve la anáfora pronominal, la *one-anaphora* y los sintagmas nominales definidos. Además, su posible incorporación a sistemas multilingües, tratamiento de diálogos y herramientas para extracción de información son algunas utilidades que justifican la aplicabilidad de este sistema.

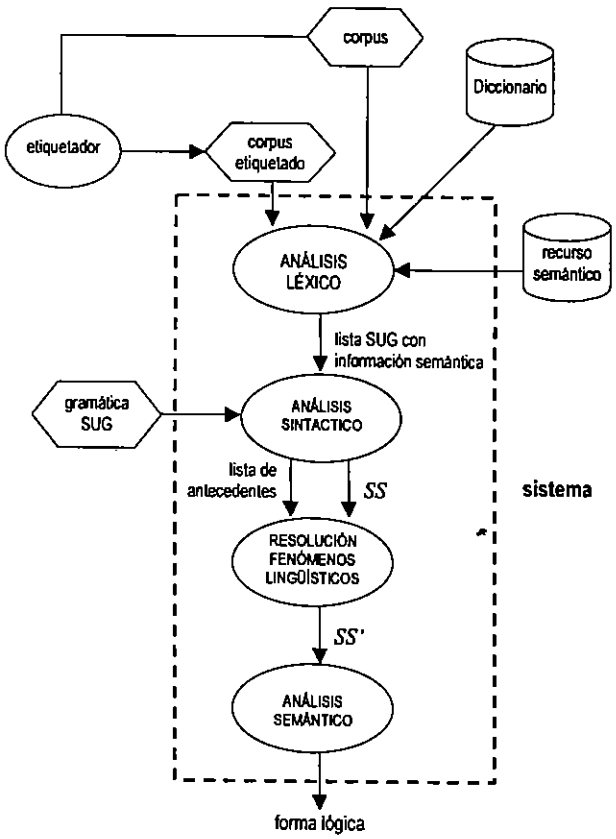

Figura 1: Arquitectura del sistema.

En este resumen se presentará en primer lugar, una descripción del sistema, donde se detallará cada uno de los módulos que lo componen. A continuación se profundizará en el tratamiento de la correferencia lingüística que realiza el sistema.

2. Descripción del sistema.

Es un sistema orientado fundamentalmente a la resolución de la correferencia lingüística[1] (Ferrández 1998) aunque también permite el tratamiento de otros problemas lingüísticos como la elipsis (Palomar 1995) y extraposición de elementos (Ferrández 1997). El sistema usa como formalismo gramatical las *Gramáticas de Unificación de Huecos* (*SUG, Slot Unification Grammar*) que permite almacenar el conocimiento sintáctico necesario para la resolución de diferentes problemas lingüísticos.

Su arquitectura, como se puede observar en la Figura 1, está formada por 4 módulos independientes que interactúan unos con otros: análisis léxico, análisis sintáctico, resolución de fenómenos lingüísticos y análisis semántico.

Éste utiliza como unidad de tratamiento básica la oración. Para ello se realiza un preprocesamiento que divide el texto a analizar en oraciones que serán la entrada del módulo de análisis léxico. Por lo tanto, el sistema sigue un proceso secuencial para cada oración. La comunicación entre módulos se realiza mediante estructuras de datos (listas y estructuras en Prolog[2]) que se utilizan como entrada/salida de cada uno de los módulos. Para

[1] Tal y como se define en Allen, la correferencia lingüística se define en el ámbito de la anáfora profunda en la que se establece una relación directa entre antecedente y expresión anafórica, a diferencia de la anáfora superficial en la que se introduce una nueva entidad del discurso basada en información obtenida de antecedentes anteriores.

[2] El sistema completo ha sido implementado en Prolog y ha sido probado en diferentes intérpretes tales como SICStus Prolog, Arity Prolog o LPA Win-Prolog.

la resolución de problemas lingüísticos intersentenciales el sistema hace uso de estructuras de datos que almacenan información del discurso. Por ejemplo, para la resolución de la anáfora se construye una lista de antecedentes de todas las oraciones previas.

3. Sistema de resolución de la anáfora.

El módulo de resolución de problemas lingüísticos, concretamente la parte que se centra en la resolución de la anáfora, está basado en la definición de restricciones y preferencias, donde las primeras descartan algunos candidatos, mientras que las segundas establecen criterios de preferencia sobre los candidatos restantes. Las restricciones se comprueban para cada candidato posible con el principal objetivo de descartar antecedentes incompatibles. Por ejemplo, la anáfora y el antecedente tienen que concordar en género y número, en caso contrario el candidato se descarta como posible antecedente. La resolución de cada tipo de anáfora puede requerir su propio conjunto de restricciones y preferencias así como su orden de aplicación.

El conjunto de restricciones definido es:

- Concordancia morfológica para la que se comprueba la concordancia de género, número y persona entre el antecedente y la expresión anafórica.
- Restricciones *c-dominio*: se aplican estas restricciones definidas en Reinhart (1983) sobre la información sintáctica de los constituyentes.
- Consistencia semántica: utilizado principalmente en textos de dominio restringido. Actualmente se está utilizando el método *IRSAS (Incorporar Restricciones Semánticas en el Análisis Sintáctico)*, Moreno (1992), para comprobar la consistencia semántica, tanto en la resolución de la anáfora, como en el análisis sintáctico.

Las fuentes de información que se aplican en las preferencias son:

- Información sintáctica: paralelismo sintáctico.
- Información léxica: información correspondiente al lema de cada palabra.
- Información estadística: proporciona información sobre el número de veces que ocurre una determinada circunstancia en el corpus de entrada.
- Estructura superficial de la oración.

4. Bibliografía.

Ferrández, A.; Peral, J.; Martínez-Barco, P.; Saiz, M.; Romero, R. *Resolución de la extraposición a izquierdas con las gramáticas de unificación de huecos*. Procesamiento del Lenguaje Natural, 21. 1997.

Ferrández, A.; Palomar, M.; Moreno, L. *Anaphor resolution in unrestricted texts with partial parsing*. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on computational linguistics. 1998.

Moreno, L. and Palomar, M. *Semantic Restrictions in a Syntactic Parser: Queries-Answering to Database*. Database and Expert Systems Applications, Springer-Verlag. 1992.

Moreno, L.; Palomar, M.; Molina, A.; Ferrández, A. *Introducción al Procesamiento del Lenguaje Natural*. Servicio de publicaciones de la Universidad de Alicante. 1999.

Palomar, M.; Ferrández, A.; Moreno, L. *Aportaciones a la resolución de la elipsis en la coordinación*. Procesamiento del Lenguaje Natural 17. 1995.

Reinhart, T. *Anaphora and Semantic Interpretation*. Croom Helm Backenham, Kent. 1983.