Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos

Tesis Doctoral
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, julio de 1998

Tesis Doctoral presentada por Antonio Ferrández Rodríguez
Dirigida por Lidia Moreno Boronat y Manuel Palomar Sanz

En esta Tesis hemos tratado la resolución de la anáfora, uno de los problemas más difíciles pendientes de solución en el campo de estudio del procesamiento del lenguaje natural. Esta resolución de la anáfora la hemos integrado dentro de un sistema modular denominado SUP (Slot Unification Parser) desarrollado en Prolog que utiliza como mecanismo de almacenamiento de la información sintáctica el formalismo gramatical SUG o gramática de unificación de huecos. Dentro de este sistema SUP hemos añadido nuevas reglas a la gramática que permiten un análisis parcial del texto, de tal modo, que el mismo algoritmo de resolución de la anáfora que aquí proponemos se pueda aplicar tanto a textos restringidos como no restringidos, tanto mediante un análisis sintáctico completo, como mediante un análisis parcial.
A continuación, se exponen las contribuciones más destacadas realizadas en esta Tesis, y para ello las clasificamos en los siguientes apartados:

? ?Hemos propuesto un nuevo formalismo gramatical: gramática de unificación de huecos (SUG, Slot Unification Grammar) como una extensión de las gramáticas de cláusulas definidas (DCG, Definite Clause Grammar). La principal extensión de este formalismo gramatical es el uso de constituyentes opcionales dentro de cada regla gramatical, y prever la solución de determinados problemas lingüísticos como son la coordinación y yuxtaposición de constituyentes.

? ?Hemos desarrollado en Prolog un sistema modular denominado SUP (Slot Unification Parser). Este sistema lo componen una serie de módulos independientes entre sí: módulo de almacenamiento de la información sintáctica mediante la gramática SUG, módulo del diccionario, módulo de análisis sintáctico, módulo de resolución de problemas lingüísticos y módulo de análisis semántico.

SUP es un sistema completo de procesamiento del lenguaje natural, cuyas características más importantes se pueden resumir en los siguientes puntos:

?? Facilita la resolución modular de problemas lingüísticos (elipsis, extraposición, anáfora, etc.).

?? Relaja la relación entre cada uno de los componentes del sistema: gramática, diccionario, analizador sintáctico y semántico, y módulo de resolución de problemas lingüísticos.

?? Simplifica el trabajo con diferentes diccionarios, y la aplicación de análisis sintácticos parciales o relajados sobre textos no restringidos en los que no tenemos la seguridad de obtener módulos de análisis sintáctico lo suficientemente robustos.

?? Desarrolla un analizador eficiente. Este objetivo se consigue al reducir el número de reglas gramaticales y el número de símbolos no terminales. También se alcanzará este objetivo ya que será capaz de recordar lo que ha sido analizado en

una frase y lo que no, disminuyendo el número de intentos de análisis de cada constituyente de la gramática.

? ?Hemos llevado a cabo un estudio completo del fenómeno lingüístico correspondiente a la anáfora. Inicialmente partimos de la relación de la anáfora con otros temas de estudio del procesamiento del lenguaje natural como por ejemplo la elipsis. Mediante estas relaciones con otros campos se ha reseñado la importancia que tiene en la actualidad un correcto tratamiento de la anáfora. Y a continuación se ha realizado una exhaustiva clasificación desde diferentes aspectos.

? ?Debido a esta exhaustiva clasificación hemos podido centrar el objetivo principal de este trabajo: resolver la anáfora que sucede en un contexto lingüístico, considerando para su resolución no sólo la oración en la que se encuentra la expresión anafórica, sino todo el discurso (anáfora discursiva). Nos hemos centrado en la anáfora morfosintáctica, o sea, la de mayor accesibilidad de su antecedente, tratando tanto referencias superficiales como profundas, tanto relaciones de correferencia como de no correferencia, y hemos resuelto la anáfora pronominal y la de tipo adjetivo, es decir, expresiones anafóricas de tipo pronombre y sintagmas nominales cuyo núcleo sea un adjetivo. Ésta es precisamente una de las principales aportaciones que realizamos a la resolución de la anáfora, ya que la aproximación que proponemos nos permite el tratamiento de otros tipos de anáfora aparte de los habitualmente tratados de tipo pronominal.

? ?También hemos llevado a cabo un profundo estudio de las aproximaciones al tratamiento de la anáfora pronominal y de tipo adjetivo existentes actualmente. Nuestra propuesta de resolución de estos tipos de anáfora se encuadra entre los sistemas integrados democráticos con un enfoque basado en restricciones (eliminan antecedentes factibles de una determinada anáfora) y preferencias (seleccionan uno entre varios candidatos posibles). Como sistema integrado, se basa en el conocimiento, es decir, maneja una serie de fuentes de información que se consideran necesarias para el correcto tratamiento de la anáfora: morfológica, léxica, sintáctica y semántica. Y dentro de los sistemas integrados utiliza un enfoque democrático para coordinar estas fuentes de información, es decir, que aquellas entidades que pueden ser susceptibles de convertirse en antecedente pueden surgir por igual de aportaciones de la información morfológica, léxica, sintáctica o semántica.

? ?Hemos aplicado sobre textos restringidos y no restringidos el algoritmo de resolución de la anáfora propuesto. Para afrontar el reto de los textos no restringidos hemos disminuido la cantidad de información con que cuenta el sistema: léxica, morfológica e información sintáctica reducida. Al reducir la cantidad de información que se utiliza para resolver la anáfora, pasamos a incluir nuestro sistema dentro de un grupo denominado *sistemas pobres en conocimiento*. Una aportación importante de nuestro sistema es que precisamente realiza un análisis sintáctico parcial del texto, extrayendo de modo automático la información indispensable para la resolución de la anáfora, y proponemos un modo de empleo de las restricciones c-dominio para que puedan aplicarse sobre información sintáctica incompleta. Aún reduciendo esta información, hemos conseguido un porcentaje de éxito o precisión del 83% en la resolución de la anáfora pronominal junto con una cobertura del 100%.