Aproximación estadístico-algebraica a la anáfora discursiva

(Tesis doctoral leída por Celia Rico Pérez en la Universidad de Alicante)

El origen de esta tesis surge como un intento de dar respuesta a uno de los problemas más complejos de una lengua: la anáfora. En este sentido, nuestro objetivo es conseguir la automatización del proceso de identificación de relaciones anafóricas, para lo cual debemos contar no sólo con criterios gramaticales de concordancia de género y número, sino también con la información que se desprende de los datos sintácticos, semánticos y pragmáticos contenidos en el discurso.

Proponemos en esta tesis un modelo de identificación de la anáfora que se basa en la siguiente hipótesis: puede llevarse a cabo un tratamiento de los poblemas que plantea la identificación de los antecedentes anafóricos en el discurso si para ello se emplean modelos de inspiración vectorial. Esta estrategia considera que si las entidades discursivas y expresiones anafóricas pueden representarse en forma de vectores se definirá como relación anafórica aquella en la que los vectores de una entidad y una expresión están más próximos.

A partir de esta hipótesis, construimos un modelo de inspiración vectorial en el que se coordinan en un único patrón las fuentes de información morfológica, sintáctica, semántica y discursiva.

El plan de trabajo sobre el que construimos la tesis es el siguiente:

1) Formulación de una tipología de la anáfora discursiva con el fin de conocer en profundidad el fenómeno lingüístico que tratamos.

2) Estudio y análisis crítico de las aproximaciones computacionales más relevantes con respecto al problema de la anáfora. Con este fin, examinamos con una visión crítica los sistema y algoritmos desarrollados desde lo que podemos considerar los primeros tiempos de la lingüística computacional hasta las últimas propuestas de los años 90.

3) Análisis empírico de las relaciones anafóricas contenidas en un corpus de textos para conocer, con datos objetivos, el comportamiento real de la anáfora en el discurso así como los tipos de información lingüística que intervienen en la localización de antecedentes.

4) Creación de la estrategia de identificación de antecedentes empleando para ello un modelo de inspiración vectorial. La construcción de ésta se apoya, como hemos mencionado, en el empleo de notación vectorial para la codificación de las fuentes de información localizadas en el análisis de corpus.

5) Automatización de la estrategia propuesta mediante la creación de un programa informático que lleve a cabo todas las tareas.

6) Evaluación de la aproximación adoptada con el fin de determinar su validez.

Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española

(Tesis doctoral leída por Aurora Martín de Santa Olalla- Universidad Alfonso X El Sabio)

Esta tesis presenta una propuesta de codificación morfosintáctica para corpus de referencia en lengua española basada en los estándares de la Text Encoding Initiative (TEI), The Network of European Reference Corpora (NERC) y The Expert Advisory Group on Language Engineering Standards (EAGLES).

Nuestra propuesta de codificación morfosintáctica para corpus de referencia en lengua española consiste en la creación de un sistema taxonómico que toma como unidad de análisis la palabra ('conjunto de signos entre dos blancos') y describe todos aquellos rasgos que presentan una marca formal

explícita que supone, además, un comportamiento gramatical específico.

El formalismo se expresa mediante un sistema de pares atributo-valor. Las etiquetas o membretes tienen una estructura atómica o jerarquizada que, junto a la pertenencia de las palabras a clases ('categorías') y subclases ('tipos'), refleja rasgos recurrentes y específicos de las distintas formas.

Como resultado de la aplicación de estos estándares a la descripción morfosintáctica de nuestra lengua, nuestro conjunto de etiquetas consta de 660 etiquetas morfosintácticas o 'entidades de segundo orden' que son el resultado de la combinatoria de 117 pares atributo-valor o 'entidades de primer orden'.

Las principales clases coinciden básicamente con las 'partes de la oración' tradicionales a las que se añaden una clase 'residual' y una clase 'puntuación'. Los atributos codificados en cada clase son sus rasgos morfosintácticos específicos. Los valores son los correspondientes a cada clase más tres valores para subespecificación ('invariante', 'cualquiera' o 'no-aplicable') y dos valores booleanos Y y O.

Nuestra propuesta de codificación contiene además un 'manual del codificador' que caracteriza y describe cada una de las clases ('categorías gramaticales' o 'partes de la oración', en nuestro caso).

Finalmente, incluimos los formalismos necesarios para su utilización con un analizador SGML: definición de entidades de primer y segundo orden, Declaración de un Sistema de rasgos (o fichero FSD) con sus Declaración de Tipo de Documento (DTDs) correspondientes.

Aspectes del sintagma nominal en català des de la perspectiva de la traducció automàtica

Tesis presentada por Toni Badia en el marco del programa de doctorado *Formalització del lenguatge natural* del ICE de la Universitat Politècnica de Catalunya, para obtener el título de doctor, en 1992.

En esta tesis se discuten distintos aspectos de la sintaxis y semántica de los sintagmas nominales catalanes con la finalidad de obtener un análisis suficientemente fino como para poder ser utilizado en un sistema de traducción automática. Es decir, se pretende determinar la información que debe ser recogida en la representación abstracta de los sintagmas nominales catalanes en un sistema de traducción basado en la transferencia. Aunque en la tesis se presentan y discuten los distintos aspectos desde una perspectiva que pretende ser general, hay dos fenómenos que han sido tratados más a fondo y que marcan el desarrollo de la obra: las estructuras de complementación preposicional en el sintagma nominal y la organización de los determinantes (especialmente la que resulta de la combinación de distintos determinantes).

La tesis se estructura en tres partes diferenciadas: una de introductoria (formada únicamente por el primer capítulo), la central (con los capítulos segundo, tercero y cuarto), y una de final (constituida por los tres capítulos restantes). El plan de todos ellos es el que sigue.

El capítulo primero plantea algunas de las cuestiones previas (en relación a la traducción automática -y a la lingüística computacional en general- y al sintagma nominal). A partir de ellas se define el planteamiento lingüístico básico adoptado en la tesis.

Los tres capítulos siguientes constituyen el cuerpo central de la obra; en ellos se discuten sucesivamente los tres constituyentes básicos de los sintagmas nominales. El segundo capítulo está dedicado al nombre núcleo del sintagma. Fundamentalmente, se presentan las características básicas del nombre que influyen en la interpretación general de los sintagmas. El tercer capítulo examina con detenimiento los varios tipos de complementos que pueden tener los nombres. El aspecto fundamental tratado es la distinción entre los complementos

subcategorizados (argumentos) y los que no lo son (modificadores). Por ello, la distinción entre los nombres predicativos y los que no lo son (presentada en el segundo capítulo) toma un relieve importante. En la elaboración de la teoría sobre los argumentos de los nombres predicativos se toman en consideración los valores semánticos del predicado básico, entre los cuales destaca el del tipo de acción (o "Aktionsart"). El capítulo cuarto trata de los elementos que forman el sistema especificador de los sintagmas nominales catalanes. Se discute la aportación de los varios determinantes y se valora como interactúa con las características generales de los sintagmas en los que aparecen. Asimismo se presentan elementos que justifican una particular estructura de los determinantes complejos.

Todos los aspectos presentados y discutidos en estos tres capítulos centrales de la tesis tienen una incidencia tanto en relación con la estructura del sintagma nominal como con los cálculos semántico que pueden tener lugar en su interior. Estos dos puntos constituyen el contenido de los dos capítulos siguientes: en el quinto se presenta una propuesta de estructura (basada en formulada en términos de estructura sintagmática) y en el sexto una seria de cálculos semánticos que se pueden realizar sea en el interior del sintagma nominal, sea en su entorno inmediato. Finalmente el séptimo capítulo, muy breve, muestra algunos de las cuestiones pendientes en relación al tratamiento de los sintagmas nominales en los sistemas de traducción automática; naturalmente desde la perspectiva del momento en que fue escrito (a finales de 1991).

Aunque la obra adopta unos planteamientos relativamente tradicionales desde el punto de vista de la sintaxis (por ejemplo, no utiliza en absoluto la técnica de representación de la información lingüística basada en estructuras de rasgos), los planteamientos semánticos, su interacción con fenómenos sintácticos y las líneas básicas descriptivas de la estructura sintáctica nos parece que siguen en pie en la actualidad.

Finalmente, nos queda sólo consignar que la tesis ha sido publicada por la editorial Publicacions de l'Abadia de Montserrat en 1994, con el mismo título que la tesis.

DISEÑO Y CONSTRUCCION DE UN SISTEMA INTELIGENTE DE AYUDA DICCIONARIAL

Adquisición y representación del conocimiento diccionarial, implantación de los mecanismos deductivos y especificación de la funcionalidad básica.

Tesis Doctoral. Facultad de Informática de San Sebastián, Universidad del País Vasco. Mayo, 1993.
Autor: Xabier Artola Zubillaga

Resumen.

Se ha diseñado y construído un sistema inteligente de ayuda diccionarial. La representación del conocimiento diccionarial sobre el léxico de la lengua y los mecanismos de inferencia implantados sobre el mismo hacen posible la deducción de información que no era explícita en la fuente de conocimiento utilizada.

La fuente de conocimiento del sistema ha sido un pequeño diccionario convencional que se ha representado utilizando técnicas de inteligencia artificial. Se ha definido una metodología para la adquisición del conocimiento diccionarial, basada en procedimientos estadísticos y empíricos: el punto de partida de esta metodología es la caracterización del meta-lenguaje lexicográfico utilizado en las definiciones del diccionario con el objeto de obtener, por medio del análisis sintáctico-semántico de las mismas, una representación expresada en términos de las relaciones interconceptuales establecidas en estas definiciones.

Para la representación del conocimiento se ha elegido el modelo de las redes semánticas donde cada nodo -frame- es representación de un sentido o acepción, y los nodos están interconectados por medio de relaciones léxico-semánticas. La arquitectura específica diseñada agrupa tres sub-bases de conocimiento, representando de forma independiente por una parte el conocimiento sobre conceptos y sus interrelaciones, por otra el referido a las palabras y, por último, el meta-conocimiento sobre las estructuras utilizadas.

En lo que se refiere al sistema de ayuda inteligente se ha diseñado una arquitectura para el mismo y se ha hecho un bosquejo sobre el proceso de comunicación con el usuario. Se ha especificado e implementado un conjunto de funciones básicas tales como petición de definición, búsqueda de relaciones entre dos conceptos, búsqueda tesáurica de conceptos, etc. Este conjunto de funciones

está soportado por la infraestructura deductiva del sistema. La funcionalidad definida facilita al usuario diferentes vías de obtención de información que de otra manera sería inaccesible.