ContextO : una plataforma de ingeniería lingüística orientada al filtrado semántico de textos

Resumen para demostración

Gustavo Crispino*, Jean-Pierre Desclés**, Slim Ben Hazez**, Jean-Luc Minel**, Ghassan Mourad**

*Universidad de la República
J. Herrera y Reissig 565
11300 Montevideo - Uruguay
crispino@fing.edu.uy

**CAMS/LaLIC, UMR du CNRS
96 Boulevard Raspail
75 006 Paris - France
{descles,benhazez,minel,mourad}@msh-paris.fr

1. Introducción

La construcción de esta plataforma se apoya en las experiencias obtenidas por el método de exploración contextual (Desclés 1996, 1997), el cual identifica los conocimientos lingüísticos ubicándolos en sus contextos y organizándolos en tareas especializadas. Este método presenta por un lado la ventaja de permitir que el trabajo del lingüista se realice de manera independiente de su implementación informática, y por otro lado, la de articular efectivamente en una misma arquitectura informática los dos tipos de trabajos.

En este enfoque, el trabajo previo del lingüista consiste en estudiar sistemáticamente un corpus de textos buscando regularidades léxicas y discursivas cuyo empleo es representativo de la categoría semántica considerada. De acuerdo a esta hipótesis que hasta el momento se ha revelado fecunda, las expresiones asociadas a esas categorías discursivas en el corpus son finitas. Por consiguiente, no es necesario ni la identificación de estructuras sintácticas específicas, ni la construcción de ontologías de dominio. Los lingüistas identifican entonces en el corpus los marcadores (*indicadores* e *índices*) gramaticales pertinentes para la resolución de un problema, y luego conciben y escriben las reglas de exploración del contexto para esos marcadores identificados en los textos.

Este proyecto recibió el apoyo del programa ECOS (Francia - Uruguay, Acción n° U97E01) en el marco de una cooperación entre el equipo LaLIC (UMR 8557 du CNRS, EHESS, Université Paris-Sorbonne) y el grupo de TALN del Instituto de Computación de la Facultad de Ingeniería (Universidad de la República - Uruguay).

2. Arquitectura general de ContextO

2.1. El motor de exploración contextual

Como respuesta a invocaciones determinadas en función de parámetros fijados por el usuario, el motor de exploración contextual dispara, para una o varias tareas especializadas, el proceso de reconocimiento de *indicadores* e *índices* presentes en un segmento textual. Este proceso es realizado por el sistema de gestión de conocimientos lingüísticos, el cual proporciona al motor de exploración contextual el conjunto de reglas potencialmente aplicables.

Hemos definido un lenguaje de descripción que permite al lingüista constituir su base de conocimientos especificando las *tareas*, los *indicadores*, los *índices* y las *reglas* de exploración contextual asociadas. Estas últimas se expresan en un lenguaje formal de tipo declarativo. Cada regla comprende una parte de *Declaración de un Espacio de Búsqueda*, una parte de *Condición* y una parte *Acción*, la cual es ejecutada solamente si se verifica la *Condición*. Como resultado de la aplicación de las reglas, se colocan etiquetas semánticas que "decoran" la jerarquía del texto a diversos niveles; por ejemplo, una regla puede atribuir una etiqueta semántica a una oración.

El motor de exploración contextual explota los conocimientos lingüísticos para una o varias tareas seleccionadas por un usuario. Está compuesto por dos sistemas que cooperan entre sí.

2.1.1. El analizador de texto

Este sistema tiene por objeto construir una primera representación que refleja la estructura del texto. Se apoya en un texto balizado por un segmentador construido a partir de un estudio

sistemático de las marcas de puntuación. La segmentación se basa en estas marcas y en un estudio de los contextos a derecha y a izquierda de las mismas. Aplica además reglas heurísticas para reconocer las secciones, con sus títulos, los párrafos y las oraciones.
Con esta información el analizador construye una estructura jerárquica, la cual es enriquecida por agentes especializados que incorporan estructuras cuyo objetivo es mejorar la cohesión y la coherencia de los extractos de textos.

2.1.2. El ejecutor

Este sistema dispara para todas las tareas seleccionadas por el usuario las reglas asociadas a ellas. Las reglas son consideradas de manera independiente, por lo que el orden en que son disparadas, para una tarea dada, es indiferente. Todas las deducciones efectuadas por las reglas afectan a los elementos que componen la jerarquía del texto produciendo una estructura jerárquica "decorada" por informaciones semánticas.

2.2. Los agentes especializados

Los agentes especializados tienen por objetivo explotar las "decoraciones semánticas" del texto en función de las necesidades definidas por el usuario. Hay entonces un agente que construye un resumen compuesto de oraciones del texto de entrada que corresponden a un perfil tipo y un agente que construye diferentes extractos del texto de entrada en función de perfiles seleccionados por el usuario. Estos agentes especializados permiten desarrollar tratamientos específicos para un usuario explotando el modelo genérico de tratamiento de conocimientos lingüísticos.

3. Conclusión

La plataforma ContextO está actualmente operativa con una base de conocimientos que contiene 11.000 marcadores (*indicadores* e *índices*) y 250 *reglas* de exploración contextual. Pensamos que su arquitectura, que privilegia el concepto de componentes de software y de agentes especializados, la hace apta para representar diferentes tipos de tratamiento lingüístico ya que es posible definir nuevas bases de conocimiento para nuevas tareas de etiquetado semántico.

Referencias bibliográficas

Desclés, Jean-Pierre. (1996). Systèmes d'exploration contextuelle. Actes du colloque sur le Calcul du sens et contexte. Université de Caen.

Desclés, Jean-Pierre, Cartier, Emmanuel; Jackiewicz, Agata; Minel, Jean-Luc (1997). Textual Processing and Contextual Exploration Method. In *CONTEXT'97*, Rio de Janeiro, Brasil.