

Tecnologías del Lenguaje aplicadas al Modelado de Sistemas de Información (TeLMoSIS)

Language Technologies applied to Information System Modeling

Antonio Molina, Lidia Moreno, Natividad Prieto, Juan Sánchez, David Buitrago
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
{amolina, lmoreno, nprieto, jsanchez}@dsic.upv.es
dabuiar@gmail.com

Isabel Díaz
Escuela de Economía
Departamento de Métodos Cuantitativos
Universidad Central
Caracas- Venezuela
idiaz@cantv.net

Resumen: El objetivo principal del proyecto es integrar los avances logrados hasta el momento en Tecnologías del Lenguaje Humano (TLH), principalmente por las aproximaciones de análisis semántico, y las estrategias de modelado de Sistemas de Información (SI). Esta integración se orientará al desarrollo de métodos, técnicas y herramientas que permitan derivar automáticamente modelos estructurales y dinámicos de software a partir de textos escritos en lenguaje natural.

Palabras clave: lenguaje natural, rol semántico, sistemas de información

Abstract: The main objective of this project is to combine advances in Human Language Technologies (HLT), particularly semantic analysis, with information system modeling. The work will be focus on developing methods, techniques and tools that allow us to automatically derive structural and dynamic software models from requirements written on natural language.

Keywords: natural language, semantic role, information systems

1 Datos del proyecto

El proyecto Telmosis (Tecnologías del Lenguaje Aplicadas al Modelado de Sistemas de Información) está financiado por el Vicerrectorado de Investigación, Desarrollo e Innovación de la Universidad Politécnica de Valencia (PAID-06-08-3294).

También forma parte del proyecto Enterprise 2.0: Técnicas de Comprensión de textos aplicadas a las necesidades de la Empresa 2.0 (CICYT TIN2009-13391-C04-03) financiado por el Ministerio de Ciencia e Innovación.

2 Descripción del proyecto

El objetivo del proyecto es integrar los avances logrados hasta el momento por las aproximaciones de análisis semántico y las estrategias de modelado, para el desarrollo de métodos, técnicas y herramientas que permitan derivar automáticamente modelos estructurales y dinámicos de software a partir de textos escritos en

lenguaje natural. Para ello se propusieron varias líneas de actuación:

- i. Estudio y desarrollo de un sistema de anotación de roles semánticos.
 - Estudio y análisis comparativo de los recursos y sistemas de anotación de roles desarrollados principalmente para el inglés y el castellano, tanto para dominios generales como específicos.
 - Experimentar con las herramientas de análisis y anotación disponibles. La eficiencia de un sistema de anotación de roles depende de la calidad de los procesos previos de análisis. Por ello, se hace necesaria la adaptación y mejora de herramientas desarrolladas previamente, como etiquetadores morfosintácticos, analizadores sintácticos, reconocedores de entidades nombradas o herramientas de desambiguación de sentidos de las palabras.
 - Desarrollo de herramientas propias para la anotación de roles semánticos mediante

técnicas de aprendizaje automático, principalmente para el castellano.

ii. Realizar un análisis comparativo de aproximaciones (métodos, técnicas y herramientas) que utilizan TLH para la deducción (semi)automática de modelos de sistemas de información a partir de textos de especificación de requisitos. Se trata de establecer una sólida base de conocimiento sobre el estado actual de las herramientas desarrolladas para deducir, de forma automática, modelos de sistemas de información. Evaluando aspectos tales como la estrategia conceptual y tecnológica de TLH utilizada por estas herramientas para obtener información de los textos de especificación de requisitos; la estrategia conceptual y tecnológica aplicada para obtener los modelos a partir de la información conseguida con los recursos de TLH. Este estudio facilitará:

- Establecer una caracterización de las herramientas en términos de sus fortalezas y debilidades.
- Establecer las características deseables de las herramientas de este tipo según se concibe actualmente el modelado de sistemas y de los avances de TLH en los últimos años. Esto permitirá elaborar la especificación de los requisitos (funcionales y no funcionales) que deben satisfacer estas herramientas, insumo fundamental para el posterior diseño e implementación de TeLMoSis.

iii. Construir un prototipo de la herramienta TeLMoSis que permita la deducción automática de modelos de sistemas a partir de textos escritos en lenguaje natural.

- Diseñar los patrones sintáctico-semánticos para la transformación de especificaciones de requisitos en modelos conceptuales.
- Desarrollar el prototipo de una herramienta de ayuda al modelado conceptual que permita obtener un modelo inicial a partir de una especificación de requisitos expresada en lenguaje natural. Esta herramienta integrará las herramientas de TLH desarrolladas y/o adaptadas y la implementación de los patrones de traducción diseñados. Esta herramienta deberá ser independiente del dominio y de la lengua. Para la construcción de este

prototipo se seguirán los procesos del ciclo de vida de desarrollo de software y se documentará utilizando, en lo posible, estándares de especificación (por ejemplo: el Unified Modeling Language –UML– y el Object Constraint Language –OCL–).

3 Estado actual del proyecto

Inicialmente se ha conseguido desarrollar un corpus de un caso de estudio constituido por un conjunto de Casos de Uso y sus correspondientes Diagramas de Interacción. De este corpus se han extraído una 500 frases en castellano y en inglés con el fin de anotarlas con roles semánticos siguiendo la aproximación de PropBank.

Se han realizado diversos experimentos de anotación de roles para la versión inglesa utilizando los etiquetadores de Dan Roth y de Paloma Moreda. Como resultado de ese trabajo se ha obtenido el corpus de frases referido anteriormente etiquetado con roles semánticos.

También se han realizado aproximaciones de etiquetado para el castellano con unos resultados muy preliminares.

Por otra parte, se ha conseguido acotar las características de la herramienta a desarrollar para deducir, de forma automática, modelos de sistemas de información. Se han diseñado los patrones sintáctico-semánticos para la transformación de especificaciones de requisitos en modelos conceptuales que en estos momentos se encuentra en fase de implementación.