Proyecto Tagparsing

J. Gabriel Amores
Grupo de Investigación Julietta
Universidad de Sevilla

1 Ficha del Proyecto

- Título del Proyecto: Integración de técnicas y herramientas de tagging, parsing y unificación.
- Entidad Financiera: Dirección General de Investigación Científica y Técnica. Ministerio de Educación, Cultura y Deporte. (Código PB98–1151). Duración: Diciembre 1999–Diciembre 2002
- Grupos Participantes: Universidad de Sevilla y Universitat Pompeu Fabra.
- Investigador Responsable: J. Gabriel Amores. Departamento de Lengua Inglesa. Universidad de Sevilla. Palos de la Frontera s/n. 41004 Sevilla. Tel. 95 455 1549. E-mail: jgabriel@cica.es y Toni Badia. IULA. Universitat Pompeu Fabra. tbadia@upf.es.

2 Resumen

La integración de herramientas computacionales para el tratamiento del lenguaje es una prioridad común a los programas actuales de investigación y desarrollo en el campo de la Ingeniería del Lenguaje.

Dentro del campo del Procesamiento del Lenguaje Natural, la mayoría de las aplicaciones requieren el uso simultáneo, o en alguna de sus fases, de módulos de tagging, parsing y unificación. Entre otras aplicaciones, merecen destacarse la Recuperación de Información (especialmente relevante en la búsqueda de información en Internet), sistemas de gestión de diálogo con entrada por voz, traducción automática, etc.

Durante los últimos años, los dos grupos participantes han concentrado su investigación en el desarrollo de técnicas y herramientas para tagging y lexicografía computacional (Universitat Pompeu Fabra) y técnicas de parsing y unificación en procesamiento del Lenguaje Natural (grupo de la Universidad de Sevilla). Las herramientas desarrolladas por ambos grupos poseen un grado de madurez y eficiencia demostrado en otros proyectos que permite abordar su integración para la obtención de un entorno común de trabajo. Dicha integración implica un trabajo de investigación dirigido al desarrollo de protocolos de comunicación entre los distintos módulos, así como una labor de desarrollo de un entorno único que integre las diferentes herramientas.

3 Objetivos

3.1 Investigación Básica

Desde el punto de vista de la investigación básica, se pretende resolver la integración de las técnicas de tagging (habitualmente basadas en la lingüística de corpus) con los modelos clásicos de parsing y unificación, propios de los sistemas de procesamiento del lenguaje natural en general. Este apartado se subdivide en los siguientes subobjetivos:

1. Adaptación del sistema de representación de la salida del tagger Conexor a los modelos requeridos por los módulos de parsing y unificación. Esta adaptación está guiada por los principios de eficiencia computacional, preservación de toda la información relevante y transparencia de cara al usuario. En concreto, se pretende estudiar si la información obtenida por el tagger es suficiente para el módulo de parsing, o en qué medida ha de ser retocada. En concreto, habrá que tomar decisiones sobre cómo ambos sistemas representan y tratan fenómenos lingüísticos tales como ambigüedad léxica y funcional.
2. Tradicionalmente los modelos de unificación van dirigidos por un módulo de parsing. La integración de módulos adicionales de tagging exige replantear el

problema de la unificación y, en concreto, nos planteamos como objetivo la investigación de modelos formales y computacionales de unificación no guiada por estructuras de constituyentes.

3. Dentro del espíritu de los proyectos internacionales de investigación en ingeniería del lenguaje, la eficiencia es uno de los criterios básicos que condicionan la posible aplicación comercial del resultado. En este sentido, nos proponemos investigar cómo mantener la eficiencia obtenida independientemente por cada uno de los módulos, una vez que se integran en un solo entorno.

3.2 Objetivos de Desarrollo

Los objetivos de este apartado se centran en la implementación de un entorno de acuerdo con la arquitectura propuesta, que se desglosa en los siguientes subobjetivos:

1. Integración de las herramientas de gestión de bases de conocimiento léxico Mph y Vtree con el tagger.
2. Especificación e implementación del interfaz entre el tagger y Episteme.
3. Completar, armonizar y ampliar las bases de datos léxicas para el castellano, inglés y el catalán
4. Completar, armonizar y ampliar las bases de datos gramaticales para el castellano, inglés y el catalán.

4 Situación Actual

Hasta el momento se han obtenido los siguientes resultados:

- Se ha completado el 50% de un léxico y gramática del inglés para Episteme.
- Se ha iniciado un léxico y gramática para el castellano para Episteme.
- Se ha modificado Episteme para admitir como entrada un conjunto de pares atributo–valor, sin necesidad de acudir al análisis léxico.
- Se ha desarrollado un interfaz con **Awk** que transforma la salida del etiquetador **EngCG–2** de Conexor (http://www.conexor.fi/) en una cadena de pares atributo–valor como entrada para Episteme. Con esta versión se pretende aumentar rápidamente la gramática del inglés sin tener que preocuparse de la información léxica. Una de las cuestiones más interesantes que ha habido que resolver en esta fase han sido la creación de varias entradas en caso de ambigüedad categorial o funcional, que pasan como disyunciones a Episteme. Además, ha planteado problemas la presencia de rasgos y subrasgos en la salida del etiquetador, ya que éstos pueden aparecer o no, y en cualquier orden.

5 Problemas y trabajo futuro

1. El principal problema que se plantea a continuación es que el etiquetador no da información de subcategorización asociada con cada entrada léxica. Esta información está en la entrada léxica del diccionario del parser. Para solucionar este problema se están estudiando varias posibilidades. Nuestro grupo pretende utilizar la categoría sintáctica y la forma base devuelta por el etiquetador para hacer una búsqueda en el diccionario y utilizar esa información durante la unificación. El grupo de la Universitat Pompeu Fabra está más interesado en inferir la información de subcategorización de la información obtenida por la **Functional Dependency Grammar**.
2. Otro problema consiste en que las categorías léxicas que utiliza el parser pueden no coincidir con las que utiliza el tagger. Esto plantea el problema de su reutilización en otros contextos, o la dificultad de cambiar de etiquetador.
3. Los nombres propios y las multipalabras también plantean problemas, ya que en este sentido quisiéramos dar prioridad a la información proveniente del diccionario del sistema, obre todo para tratar verbos con partícula, preposiciones complejas, etc.