Una herramienta para la edición y manipulación de corpus

Article *in* Procesamiento de Lenguaje Natural · September 2006

Source: OAI

3 authors, including:

F. Javier Ortega
Universidad de Sevilla
49 PUBLICATIONS **426** CITATIONS

Una herramienta para la edición y manipulación de corpus*

Fco. Javier Ortega, Víctor J. Díaz y Luisa María Romero
Departamento de Lenguajes y Sistemas Informáticos
E. T. S. Ingeniería Informática - Universidad de Sevilla
Avda. Reina Mercedes s/n 41012-Sevilla (Spain)
vjdiaz@lsi.us.es

Resumen: La anotación de corpus es una tarea muy laboriosa aunque esencial a la hora de desarrollar algoritmos estadísticos para el procesamiento del lenguaje. Presentamos la primera versión de una herramienta, encargada de aliviar esta tarea. La herramienta incluye una serie de características que facilitan la edición de textos anotados mediante distintos formatos y la anotación de textos planos de forma manual o mediante la ejecución de etiquetadores externos.
Palabras clave: PLN basada en corpus, anotación lingüística

Abstract: Corpus annotation is a very laborious task which is essential to guide the development of statistical language-processing algorithms. To help this task we have designed the first version of a system which provides a number of features which make it an easy tool to use for editing annotated text with several formats and for tagging plain text manually or by triggering external tools.
Keywords: Corpus-based NLP, linguistic annotation

1. Introducción

El procesamiento del lenguaje natural estadístico exige la existencia de corpus previamente anotados lingüísticamente a partir de los que construir los modelos. Sin embargo, la anotación es una tarea muy laboriosa y de gran dificultad. Por una parte la anotación varía dependiendo del conocimiento que se desea anotar, por otra, y más grave, no existen criterios únicos de anotación (Civit, 2003). En esta demostración presentamos una herramienta de visualización, edición y gestión de corpus anotados general que ha sido desarrollada fundamentalmente como apoyo para una de las tareas del proyecto NERO (Name Entity Recognition using Ontologies). A continuación mostraremos los objetivos que nos marcamos a la hora de abordar su desarrollo:

- Realizar una aplicación portable y extensible, que gestione los corpus de la forma más eficiente posible.
- Proporcionar una interfaz gráfica que facilite el uso de la aplicación, visualizando los corpus de manera intuitiva.
- Permitir gestionar distintas anotaciones sobre corpus de textos, mediante la definición de proyectos de etiquetado.
- Definición y modificación (crear, editar y eliminar etiquetas) de etiquetarios.
- Permitir definir, modificar y eliminar formatos de etiquetado de corpus, así como leer y escribir corpus anotados en dichos formatos.
- Interacción con herramientas de etiquetado externas a la aplicación.
- Realizar consultas sobre los corpus, acerca del etiquetado del mismo.

2. Aspectos tecnológicos del sistema

Caben destacar ciertas decisiones tomadas, sobre todo en cuanto a aspectos tecnológicos se refiere. Así, para cubrir el requisito de portabilidad de la aplicación a diversos sistemas operativos, se optó por una implementación en lenguaje Java. Así, con un diseño adecuado, también se favorece la posible adición de nuevas funcionalidades a la aplicación.

En el aspecto estático del sistema, se eligió una implementación apoyada en en lenguaje de representación XML. La primera razón, y más importante de todas, es la capacidad de aplicación inmediata de este lenguaje de marcas para la etiquetación de textos. Además, hemos podido definir de una

* Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia (TIN 2004-07246-C03-03)

manera sencilla un formato de etiquetado muy flexible, extensible, y fácil de utilizar.

En segundo lugar, también se ha optado por el XML para almacenar datos relativos a configuraciones de los diversos aspectos de la aplicación, así como datos necesarios para facilitar su uso, como por ejemplo: archivos de definiciones de tipos de etiquetados por columnas, definición de proyectos, definición de etiquetarios, órdenes de uso de los etiquetadores automáticos, etc.

Y como complemento casi indispensable al XML hemos utilizado archivos XMLSchema para validarlos, de manera que nuestra aplicación siempre trabaje con archivos válidos, en cuanto a estructura y tipos de datos se refiere.

Otro motivo para el uso de esta tecnología es la utilidad inmediata de traducir los archivos XML con texto etiquetado, a archivos HTML fácilmente visibles gracias a las utilidades de la Swing de Java. El único problema en este apartado ha sido la eficiencia de la implementación de esta característica en la API de Java, lo que nos ha llevado a fragmentar virtualmente los archivos de texto, para visualizarlos de una manera más rápida.

Con lo comentado anteriormente, la aplicación desarrollada hasta el momento cumple los requisitos expuestos, pudiendo etiquetar textos, mostrar textos etiquetados, abrir corpus de textos predefinidos, etc.

3. Descripción básica del sistema

El sistema se basa en un entorno gráfico organizado en torno a tres elementos básicos: un conjunto de menús desplegables donde se pueden seleccionar todas las acciones disponibles actualmente en la aplicación, un conjunto de ventanas donde fundamentalmente se visualizan los textos y la estructura del corpus, y una serie de barras donde se incluye información contextual.

Cada corpus está asociado con un proyecto en el que se incluyen todos los archivos en los que está dividido. La ventana principal se subdivide en dos partes: la parte izquierda contiene la estructura y archivos del proyecto (corpus) actual, y en la derecha se visualizarán aquellos archivos del proyecto que el usuario desee ver su contenido. En la barra inferior, se incluye información relevante sobre la palabra seleccionada en el archivo activo.

Los ficheros constituyentes del corpus se pueden visualizar de dos formas. Por defecto, las palabras aparecen en distintos colores, según la etiqueta que se les haya dado. El color negro se reserva para las palabras que no tienen etiqueta. Aquellas que estén en negrita son las que tienen varias posibilidades de etiquetación. El usuario puede cambiar la etiqueta de cualquier palabra. Basta con que la seleccione y haga clic en el botón derecho del ratón. Esta operación desplegará un menú contextual que le ofrecerá las distintas etiquetas.

La otra manera de visualizar los corpus es tal y como aparecen en los archivos originales. En esta visualización no es posible modificar la etiquetación del corpus.

4. Trabajo futuro

Respecto a la visualización estamos especialmente interesados en ampliar la capacidad del sistema con objeto de mostrar gráficamente anotaciones anidadas. Esta ampliación es especialmente relevante en el caso de corpus anotados sintácticamente ya que la visualización de la anotación realizada es más intuitiva si se muestra jerárquicamente usando árboles.

Debido a lo laborioso del proceso de anotación, es frecuente la participación de equipos. Esta estrategia implica costes adicionales de ingeniería relacionados con el mantenimiento de la coherencia en el proceso de etiquetación y la gestión y monitorización de versiones de corpus. Como línea futura desearíamos que la herramienta incorporara funcionalidades que facilitaran este tipo de procesos.

Bibliografía

Civit, M. 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*, volumen 3 de *Colección de monografías de la Sociedad Española del Procesamiento del Lenguaje Natural*. CEE Limencop S.L., Alicante.