

AnCoraPipe: Una herramienta para la anotación multinivel

Manuel Bertran, Oriol Borrega, Marta Recasens, Bàrbara Soriano

**CLiC - Centre de Llenguatge i Computació
Universitat de Barcelona
Gran Via Corts Catalanes,585
08007 Barcelona**

mbertran@lsi.upc.edu
oriol.borrega@thera-clic.com
mrecasens@ub.edu
bsoriano@ub.edu

1. Introducción

La anotación de corpus es una tarea que exige mucho tiempo de dedicación, y el desarrollo de *AnCora* hasta su estado actual ha supuesto mucho esfuerzo por parte de nuestro grupo de investigación. Durante este proceso, se han utilizado diversas herramientas y formatos, siempre con el riesgo de perder los datos al realizar la traslación de un formato a otro o unir datos que habían sido etiquetados con herramientas diferentes. Con el objetivo de resolver estos problemas, presentamos *AnCoraPipe*, que se basa en un formato de datos XML sencillo. Este formato de datos permite la anotación a distintos niveles de lenguas diferentes. Se realizó un esfuerzo para conseguir que la herramienta fuera ampliable en cuanto al volumen de datos y a las funciones desarrolladas.

En el proceso de construcción y prueba de *AncoraPipe* han participado diversos lingüistas con experiencia en la anotación de corpus. La interacción ha permitido crear una interfaz amigable y de uso sencillo para las operaciones más habituales. La nueva herramienta hace que el tiempo de anotación se reduzca en un 40% en el etiquetaje de roles semánticos, un 60% en la anotación de *named entities* y un 25% en la de correferencia.

2. Formato de los datos

Con el fin de ayudar a la anotación simultánea de diferentes niveles, la interfaz puede asociar los corpus de la máquina local con un servidor, de manera que los usuarios pueden conocer los cambios hechos en el servidor y sincronizarlos antes de realizar sus propios cambios en los archivos locales. Los cambios realizados en los archivos locales también se cargan en el servidor, para otros usuarios que añadan más anotaciones.

Las entradas e almacenan en formato XML codificado en UTF-8. El lenguaje XML permite la portabilidad y tiene la ventaja de la cantidad de herramientas y librerías para diferentes plataformas y lenguajes de programación. Además, la codificación UTF-8 permite que el formato sea multilingüe. El propio XML tiene estructura de árbol, de manera que representa fácilmente la estructura de los constituyentes sintácticos.

Nuestro XML se basa en los siguientes principios:

- Facilidad de lectura: la estructura es intuitiva.

- Facilidad de mantenimiento: la coherencia interna puede mantenerse con poco esfuerzo.
- Robustez: los pequeños cambios no afectan a la coherencia global. La estructura general se mantiene incluso cuando se produce un error.

Estos objetivos se reflejan en una serie de principios de diseño:

- Conjunto reducido de nombres de nodo: sólo se permiten 15 nombres de nodo. Por lo tanto, los nodos son únicamente genéricos, y la especificidad se consigue con los atributos.
- Los atributos son atómicos: cada atributo etiqueta sólo un rasgo del nodo. Ello reduce el número de valores posibles y hace que los niveles de anotación sean independientes.
- Los atributos describen sólo a su nodo. De esta manera, la creación, la supresión y el desplazamiento de los nodos se convierten en tareas muy simples, y, por lo tanto, se garantiza la coherencia.
- No hay datos redundantes.
- Se pueden añadir fácilmente nuevos niveles de anotación: sólo se requiere el diseño de un nuevo atributo y sus posibles valores.

3. Interfaz

Esta sección describe de manera resumida el editor *AncoraPipe*. Para una explicación más detallada, visiten la página <http://clic.ub.edu/mbertran/tbfeditor/help>.

3.1 Descripción

La interfaz se organiza en diferentes paneles donde se muestran los datos. Para realizar operaciones sobre los corpus, se pueden utilizar los botones y menús disponibles.

La IGU (Interfaz Gráfica de Usuario) destaca en amarillo las entradas en las cuales el codificador debe poner atención, es decir, sugiere los nodos del árbol que deberían anotarse o las oraciones que contienen estos nodos, dependiendo del nivel de anotación.

Los paneles disponibles son:

- Árbol de directorios de corpus: muestra la estructura de los directorios y permite que el usuario seleccione un archivo.
- Lista de oraciones: muestra las oraciones de cada archivo.
- Árbol de oraciones: contiene la estructura de la oración seleccionada. El usuario también puede ver las formas y lemas junto con los datos del nivel de anotación correspondiente.
- Panel de anotación: Este panel se utiliza para realizar operaciones sobre el árbol y anotar sus nodos. La visualización del árbol cambia según el nivel de anotación, lo que facilita la tarea.

Los niveles de anotación actual incluyen morfología, sintaxis (cambios en la estructura del árbol, agrupamiento y división de nodos, etc.), funciones, argumentos y papeles temáticos, *named* entities, *synsets* de *WordNet* y correferencia. La interfaz también proporciona algunas herramientas externas para niveles específicos, como son:

- Anotador de correferencia: la correferencia se puede anotar de una manera amigable para el usuario, que ve los archivos en formato de texto plano.

- *Synsets* de *WordNet*: sobre una base lema a lema, la herramienta externa busca todas las ocurrencias del mismo lema en el corpus, de manera que se anotan todas en una fila. Esta ayuda favorece la consistencia de la anotación.

La interfaz se puede ampliar creando herramientas adicionales para más niveles de anotación. Esto puede hacerse escribiendo dos nuevas clases Java, tras haber especificado el nuevo atributo y los posibles valores.

3.2 Funcionalidad

En el desarrollo de *AnCoraPipe* han participado muchos lingüistas. Ello ha dado como resultado una herramienta muy orientada al usuario, centrada en la facilidad de uso y la simplicidad de las operaciones. Con este objetivo, la utilización del ratón para realizar las operaciones se ha minimizado, y sólo se destacan los nodos relevantes de cada nivel de anotación, de manera que se evite el riesgo de descuido. De esta forma, se ha reducido el tiempo de anotación más de un 60%.

3.3 Instalación

Los requisitos para instalar *AncoraPipe* son: Java 1.5 y la librería gráfica de Java SWT. Nuestro paquete incluye la librería SWT para WindowsXP. Para otras plataformas, esta librería se encuentra en el paquete Eclipse, o bien puede obtenerse de la página <http://www.eclipse.org/swt/>.

4 Futuros trabajos

Los planes para ampliar la aplicación actual incluyen: hacer que la aplicación sea accesible a través de la web, proporcionar métodos para realizar búsquedas en el corpus a partir de la interfaz, proporcionar métodos para la elaboración de descripciones estadísticas del corpus, proporcionar herramientas para tratar los lexicones nominales y verbales, y añadir métodos semiautomáticos y funcionalidades de aprendizaje automático para un etiquetaje semiautomático.

Agradecimientos

Este artículo ha recibido la ayuda de Lang2World (TIN2006-15265-C06-06) - subproyecto de TEXTMESS - y de la beca FPU-200608 del Ministerio de Educación y Ciencia de España.