Procesamiento del Lenguaje Natural, nº39 (2007), pp. 275-276 recibido 06-06-2007; aceptado 22-06-2007

Una herramienta para la manipulación de corpora bilingue usando distancia léxica*

Rafael Borrego Ropero y Víctor J. Díaz Madrigal
Departamento de Lenguajes y Sistemas Informáticos
E. T. S. Ingeniería Informática - Universidad de Sevilla
Avda. Reina Mercedes s/n 41012-Sevilla (Spain)
{rborrego, vjdiaz}@us.es

Resumen: En este artículo se presenta una herramienta que permite anotar corpora bilingue y realizar alineamiento entre textos usando heurísticas basadas en frecuencia, posición y cercanía léxica (con Edit Distance). La anotación de corpora bilingue es una tarea muy laboriosa pero esencial a la hora de desarrollar bases de conocimiento para la realización de traducciones automáticas entre distintos idiomas. Esta herramienta ayuda esta tarea, permitiendo anotar de forma rápida y sencilla. Incluye características que facilitan la edición de textos planos y de textos anotados.
Palabras clave: Alineamiento, Etiquetado de entidades, Edit Distance, Corpora Bilingüe

Abstract: In this article is presented a tool for labeling bilingual parallel corpora and aligning texts using heuristics based on word frequency, position and lexicographical similarity (using Edit Distance). Bilingual corpora annotation is a very laborious task but essential at the time of developing knowledge bases for the accomplishment of automatic translations between different languages. This tool helps to this task, allowing to annotate texts in a fast and simple way. It includes characteristics that help editing plain and annotated texts.
Keywords: Alignment, Name Entity Recognition, Bilingual corpora, Edit Distance

1. Introducción

El sistema que presentamos ha sido desarrollado como apoyo a una de las tareas del proyecto NERO (TIN 2004-07246-C03-03) y facilita el alineamiento de entidades con nombre en corpora paralelo basándose en varias heurísticas descriyas en (Borrego y Díaz, 2007). El alineamiento de textos consiste en identificar en un corpus biligüe qué partes (párrafos, frases, palabras) de uno de los corpus se corresponden con las del otro. Dado que la anotación es una tarea muy laboriosa y de gran dificultad, se ha desarrollado una herramienta de visualización y edición de corpus como apoyo a la anotación, que detecta alineamientos entre conjuntos de palabras. A continuación mostraremos los objetivos marcados a la hora de abordar su desarrollo:

- Realizar una aplicación portable y extensible, que permita anotar corpora paralelo de forma eficiente.
- Proporcionar una interfaz gráfica que facilite el uso de la aplicación, visualizando los corpus de manera intuitiva (sin que sea necesario tener conocimientos ni sobre las heuristicas usadas ni sobre XML).
- Permitir anotar corpora paralelo, relacionando un conjunto de palabras en un lenguaje con su equivalente en el otro.
- Aplicar heurísticas y un sistema de votación para obtener alineamientos entre conjuntos de palabras en un idioma con su equivalente en el otro
- Definición y modificación (crear, editar y eliminar etiquetas) de etiquetarios.
- Leer y escribir corpus anotados con distintos formatos de etiquetado, realizando la división de textos usando expresiones regulares o de forma automática.
- Realizar consultas sobre los corpus acerca de sus etiquetados, y ver sus propiedades.

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

ISSN: 1135-5948

- Generar automáticamente informes sobre el resultado de las anotaciones realizadas.

2. Aspectos tecnológicos del sistema

Caben destacar ciertas decisiones tomadas relativas a aspectos tecnológicos. Así, para cubrir el requisito de portabilidad de la aplicación a diversos sistemas operativos, se optó por una implementación en lenguaje Java.

En el aspecto relativo a los datos, se eligió una implementación apoyada en el lenguaje de etiquetado XML. La primera razón es la capacidad de aplicación inmediata de este lenguaje de marcas para la etiquetación de textos. Ésto ha permitido definir de una manera sencilla un formato de etiquetado muy flexible, extensible, y sencillo de utilizar, que es fácilmente tratable por aplicaciones externas. Además, es un formato de almacenamiento portable, que no requiere tener instalado ningún programa específico.

También se ha optado por 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: definición de proyectos, definición de expresiones regulares para dividir el texto por frases o por palabras, palabras huecas que se desea ignorar, etc.

Para facilitar al usuario su manejo la aplicación permite convertir de forma automática documentos en texto plano a XML, indicando la ruta de los ficheros y, de forma opcional, información sobre su contenido o autores. Con ello se puede empezar a manejar la aplicación sin tener que conocer XML ni tener que hacer conversiones entre formatos de codificación. Además, permite trabajar con un corpus sin alterar su contenido, ya que en ningún momento se modifica el contenido de los ficheros en texto plano.

Con lo comentado anteriormente, la aplicación desarrollada cumple los requisitos expuestos, pudiendo etiquetar textos, mostrar corpus etiquetados en distintos idiomas, etc.

3. Descripción básica del sistema

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

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. Las ventanas internas que muestran el contenido de cada fichero se encuentran divididas en dos zonas, una para cada idioma, mostrando con distinto tipo de letra aquellas palabras que se encuentran anotadas. Además, tras seleccionar un conjunto de palabras en una de las zonas, indica en la otra zona la frase equivalente.

Los ficheros constituyentes del corpus se pueden visualizar de dos formas. La primera forma es en las ventanas asociadas a los ficheros que nos muestra el contenido de cada fichero, teniendo un color distinto aquellos conjuntos de palabras que han sido anotados. La otra es en una ventana especial que permite ver el conjunto de palabras que contiene, indicando la posición origen y fin, así como el tipo de palabra.

En cualquier momento se puede anotar, para lo cual solo hay que seleccionar el texto deseado con el ratón, e indicar que se desea anotar la selección. También se puede hacer el proceso inverso, para eliminar una anotación hecha previamente.

4. Trabajo futuro

Respecto al reconocimiento de entidades sería interesante incluir más heurísticas para realizar el alineamiento. Además, debido a lo laborioso del proceso de anotación, es frecuente la participación de equipos. Esto implica dificultades relacionadas con el mantenimiento de la coherencia en el proceso de etiquetación y la gestión de versiones de corpus. En este aspecto, pretendemos enriquecer la herramienta para incorporar funcionalidades que faciliten este tipo de procesos.

Bibliografía

Borrego, R. y V. Díaz. 2007. Alineamiento de Entidades con Nombre usando Distancia Léxica. *Procesamiento del Lenguaje Natural*, 38(1):61–66.