XTRA-Bi: Extracción automática de entidades bitextuales para software de traducción asistida

Inés Jacob, Joseba Abaitua, Josuka Díaz, Josu Gómez, Koldo Ocina
Universidad de Deusto
ines@eside.deusto.es

Thomas Diedrich
Stella die KommunikationsFabrik

Organismo financiador: Departamento de Industria del Gobierno Vasco. Stella die KommunikationsFabrik, Plan Vasco de Ciencia y Tecnología (OD-00UD05)

Resumen: El principal inconveniente de los sistemas de memorias de traducción es que para que lleguen a ser productivos requieren un costoso proceso previo de alimentación manual. XTRA-Bi desarrolla métodos de extracción y alimentación automática de segmentos bilingües a partir de corpora paralelos. La clave del método radica en la utilización del formato TMX para la importación de corpus previamente segmentados y alineados.

1 Introducción

Desde hace una década ha ido adquiriendo auge un tipo de herramienta de traducción asistida cuyo rendimiento depende crucialmente del aprovechamiento de traducciones previamente realizadas por métodos manuales. Son los denominados gestores de memorias de traducción, sistemas que permiten el almacenamiento y reutilización masiva de bloques paralelos de texto, en su versión original y traducida. La ventaja de estas herramientas es que funcionan sobre la base de versiones validadas en la lengua meta, con lo que se garantiza la homogeneidad terminológica y de estilo en la traducción. El inconveniente es que para que lleguen a ser productivas requieren un costoso proceso previo de alimentación que habitualmente es manual.

XTRA-Bi automatiza este proceso en dos fases: Primero extrae automáticamente segmentos bilingües, utilizando algunas de las técnicas descritas en Abaitua y otros (1998), Martínez (1999) y Casillas (2000). El resultado es un corpus alineado en formato XML/TEI. Posteriormente el formato TEI se convierte a TMX, de forma que los segmentos alineados puedan ser importados por los sistemas de memorias de traducción comerciales (DéjàVu, Transit o TradosTW).

El ámbito de aplicación son textos administrativos bilingües en euskera y castellano.

2 Situación actual

El proyecto comenzó en octubre de 2000 y finalizará en diciembre de 2001. En la actualidad se están desarrollando simultáneamente las siguientes tareas:

2.1 Compilación de corpus

Las memorias de traducción deben basarse en colecciones amplias y representativas de textos para ser productivas. Por ello se necesitan métodos que permitan incrementar el tamaño del corpus de manera fácil y efectiva. Desde que la mayor parte de las instituciones publican en Internet las versiones bilingües de sus boletines oficiales, es ahora mucho más fácil compilar y actualizar corpora paralelos, como es el caso de nuestro corpus LEGE-Bi. En la actualidad se ha automatizado el proceso de bajada de boletines de las diputaciones de Bizkaia (BOB), Gipuzkoa (BOG), Álava (BOA) y Navarra (BON), así como del Gobierno Vasco (BOPV). Para completar y enriquecer el corpus con otros modelos textuales estamos bajando de Internet copias de prensa diaria en euskera y castellano, que sin ser paralelas, son al menos comparables en gran parte del contenido.

2.2 Segmentación en TEI/XML

Los textos que se obtienen de Internet se almacenan en su versión y distribución original.

Existe una gran disparidad de formatos y criterios de almacenamiento (Word, PDF, HTML, TXT). La reconversión y unificación de los formatos presenta un serio problema. Para resolverlo se ha desarrollado un conjunto de filtros y segmentadores. Algunos son *ad hoc*, adaptados a las propiedades concretas de los archivos dependiendo de su procedencia; otros son de propósito general. Una vez procesados, los archivos se almacenan en formato XML, según las directrices de TEI-P4. Esta tarea se ha resuelto para los boletines del BOG y del BOA.

2. 3 Alineación

Dada la naturaleza plenamente paralela de los boletines oficiales, existe un elevado índice de correspondencia (99%) entre las principales unidades textuales: títulos, epígrafes y párrafos. Es posible afinar en la granularidad de la alineación hacia segmentos menores, como la oración o los nombres propios (Martínez, 1999), aunque con un grado de fiabilidad menor.

2. 4 Conversión a TMX

Para la conversión a TMX, dada la naturaleza particular del corpus, el nivel de segmentación elegido es el de párrafo (por la uniformidad de correspondencias). Una vez en TMX, el corpus es fácilmente importable desde sistemas comerciales como (DéjàVu, Transit o TradosTW).

Referencias

Joseba Abaitua, Arantza Casillas, Raquel Martínez. 1998. Value Added Tagging for Multilingual Resource Management. *First International Conference on Language Resources and Evaluation*: 1003-1008. ELRA, Granada

Arantza Casillas. 2000. *Explotación de corpus alineados para el desarrollo de entornos de composición de documentos estructurados bilingües.* Tesis doctoral. Universidad de Deusto.

Raquel Martínez. 1999. *Alineación automática de corpus paralelos: una propuesta metodológica y su aplicación a un dominio de especialidad.* Tesis doctoral. Universidad de Deusto.