Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 365-366 recibido 1-05-2009; aceptado 5-06-2009

Bitextor, un cosechador automático de memorias de traducción a partir de sitios web multilingües

Bitextor, an automatic bitext harvester from multilingual web sites

Miquel Esplà
Grupo Transducens, DLSI, Universitat d'Alacant
E-03071 Alacant
miquel.espla@ua.es

Resumen: Bitextor es una aplicación que tiene como objetivo la generación de memorias de traducción utilizando sitios web multilingües como fuente de corpus. La aplicación descarga y preprocesa todos los ficheros HTML de un sitio web. Después aplica una serie de heurísticas (principalmente basadas en la estructura de etiquetas HTML y la longitud de los bloques de texto) mediante las cuales es capaz de emparejar los *textos paralelos* y generar memorias de traducción en formato TMX mediante el uso de la librería LibTagAligner.
Palabras clave: recolección de bitextos, alineamiento de textos, generación de corpus multilingüe, alineamiento de ficheros HTML.

Abstract: Bitextor is an application whose objective is to generate translation memories using multilingual websites as a corpus source. It downloads and preprocesses all the HTML files in a website. Later applies a set of heuristics (based mainly on HTML tag structure and text block length) to make pairs of files which are candidates to contain the same text in different languages. From these candidates, translation memories are generated in TMX format using the library LibTagAligner.
Keywords: bitext harvesting, text alignment, multilingual corpus generation, HTML file alignment.

1. Introducción y motivación

En la actualidad, la fuente de corpus multilingüe más grande del mundo es Internet. Además, podemos encontrar webs en más de una lengua o traducciones a diversos idiomas de una misma web (textos paralelos) que podrían ser aprovechadas para la obtención de bitextos. Teniendo en cuenta que cada vez son más las aplicaciones del campo de la traducción automática que requieren de grandes cantidades de corpus bilingüe, se plantea el funcionamiento de Bitextor. El objetivo es el de obtener memorias de traducción con una precisión y cobertura razonables y de manera automática. De esta forma, pese a que la supervisión humana de los resultados siempre será necesaria, se agilizaría la recolección de bitextos para una tarea de traducción dada.

2- Etapas del proceso de obtención de bitextos

1. Descarga: La primera tarea es la de descargar todos los ficheros HTML del sitio web indicado por el usuario.[1]

2. Preprocesamiento: Seguidamente, la aplicación recorre recursivamente los subdirectorios convirtiendo los ficheros al estándar XHTML[2] para evitar las etiquetas sin cerrar y los errores en la estructura HTML. Además, la codificación de caracteres se normaliza a UTF-8.[3]

Después, se obtiene y almacena una serie de información de cada fichero: el idioma en que está escrito,[4] la extensión de fichero, su tamaño en bytes y la longitud total (en caracteres) del texto llano que contiene. Además, se obtiene información sobre su contenido la cual será utilizada en la comparación entre ficheros.[5]

3. Generación de las memorias de traducción: La información obtenida a partir de los ficheros es comparada para discernir cuales de ellos son

1 Mediante Httrack: `http://www.httrack.com`

2 Mediante LibTidy: `http://tidy.sourceforge.net`

3 Utilizando LibEnca, para detectar la codificación original: `http://sourceforge.net/projects/freshmeat_enca/`

4 Mediante la librería LibTextCat: `http://software.wise-guys.nl/libtextcat/`

5 Ver la sección *Representación del contenido de los ficheros*

ISSN: 1135-5948

candidatos a ser textos paralelos. Aquellos que pasan el filtro son alineados, frase por frase, utilizando la librería LibTagAligner.

3. Representación del contenido de los ficheros

Como se comenta en la sección anterior, tanto Bitextor como LibTagAligner utilizan la información contenida en los ficheros HTML para su comparación y posterior alineamiento. Esto se hace mediante un sistema de representación de la información de los ficheros utilizando cadenas de números enteros que actuarían como *huellas digitales* de los mismos. Básicamente, lo que se hace es filtrar su contenido dejando solo los nombres de las etiquetas (eliminando los comentarios y los atributos de la mismas) y el texto. Después se utilizan números negativos para representar cada una de las etiquetas y números positivos para la longitud, en caracteres, de los bloques de texto. Posteriormente, se utiliza una adaptación del algoritmo de distancia de edición entre estas cadenas para calcular las similitudes.

4. El proceso de comparación entre ficheros

La comparación se realiza utilizando una serie de heurísticas, a partir de la información obtenida sobre cada uno de los ficheros XHTML del sitio web. Hay que puntualizar que Bitextor y LibTagAligner son completamente configurables: prácticamente todos los umbrales y parámetros son ajustables mediante un fichero de configuración, permitiendo así ajustar la búsqueda a las características de cada sitio web.

En lo referente a las heurísticas, podemos considerar, por una parte, un grupo que tienen como objetivo descartar las parejas de ficheros con diferencias *evidentes a simple vista* para evitar la necesidad de realizar cálculos más complejos sobre parejas que, evidentemente, no son traducciones de un mismo texto. Estas serían:

- La comparación de las extensiones de fichero.
- La comparación del idioma.
- La comparación del tamaño (en bytes) de los ficheros.
- La comparación de la longitud del texto plano.

A parte de estas heurísticas superficiales, el método principal de comparación entre ficheros se basa en el sistema de representación de la información de los ficheros descrita en el punto anterior. Esta especie de *huellas digitales* permiten una comparación mucho más precisa (mediante el cálculo de la distancia de edición).

5. El papel de la librería LibTagAligner

Si bien Bitextor tiene como función comparar los ficheros de un sitio web y escoger aquellos que tengan más probabilidades de ser traducciones de un mismo texto, la generación de las memorias de traducción no sería posible sin la librería LibTagAligner, que fue creada paralelamente por el mismo equipo que ha desarrollado Bitextor.

LibTagAligner utiliza una clasificación de las etiquetas HTML creada por el propio usuario mediante un fichero de configuración. Así mismo, se aplica un sistema de pesos para las operaciones de comparación (mediante distancia de edición) de estas etiquetas. Con esto, la librería es capaz de alinear ficheros etiquetados con XHTML proporcionando una calidad de alineamiento considerable. Además, mediante un sistema de reglas de división de frases (que también es definido por el usuario), produce como resultado una memoria de traducción, en formato TMX, en la que empareja cada una de las frases de los ficheros.

6. Sobre la aplicación

Tanto Bitextor como LibTagAligner se encuentran disponibles bajo licencia GPL.[6]

Ambas aplicaciones han sido creadas para plataformas basadas en Unix y han sido probadas en la mayoría de las distribuciones GNU/Linux más utilizadas.

7. Agradecimientos y menciones

Enrique Sánchez Villamil fué el autor de la primera versión de TagAligner[7][8] y Miquel Simón perfeccionó la configurabilidad en la segunda versión. El proyecto original que permitió el desarrollo posterior de Bitextor[9] fue financiado por el Ministerio de Ciencia y Tecnología entre los años 2004 y 2006. El proyecto Bitextor ha sido financiado por la Universidad de Alicante.

6 `http://sf.net/projects/bitextor` y `http://sf.net/projects/tag-aligner`

7 Enrique Sánchez-Villamil, Susana Santos-Antón, Sergio Ortiz-Rojas, Mikel L. Forcada. 2006. Evaluation of alignment methods for HTML parallel text. En *Advances in Natural Language Processing*, volumen 4139 en *Lecture Notes in Computer Science*, páginas 280-290. Springuer.

8 También fue autor de una primera versión de Bitextor, pese a que las siguientes versiones de la aplicación no se basaron en ella, sino que heredaron tan solo la idea general.

9 *Traductores de estados finitos a partir de bitextos alineados recolectados en Internet* (TIC2003-08681-C02). `http://www.dlsi.ua.es/eines/projecte.cgi?id=eng&projecte=33`