WWW como Fuente de Recursos Lingüísticos para su Uso en PLN

Fernando Martínez Santiago
L. Alfonso Ureña Lopez
Manuel Garcia Vega
Departamento de Informatica.Universidad de Jaen. Spain
{dofer, laurena, mgarcia}@ujaen.es

Resumen Crear un corpus extraído a partir de la Web está lejos de ser una tarea trivial. El elevado grado de heterogeneidad que es usual encontrar en el formato HTML, la gran cantidad de información irrelevante tanto en el sitio Web como dentro de una misma página y otros problemas de diversa índole, dificultan la obtención de un conjunto de documentos de aspecto homogéneo, estructurado y libre de ruido. Es presentada aquí una herramienta que pretende no sólo recuperar y almacenar selectivamente determinados sitios Web, sino dotar a los documentos obtenidos de un formato conveniente y homogéneo para su procesamiento automático, con independencia del origen de cada documento.

1 Introducción

La Web como fuente de recursos lingüísticos es realmente interesante [1]. Así por ejemplo, podría pensarse en la Web como una fuente inagotable de documentos durante el periodo de entrenamiento de un sistema de categorización. Cuantos más y mejores documentos tengamos, mejor será el sistema categorizado [2]. Si bien la idea es atractiva, encontramos un serio escollo en la heterogeneidad extrema de los documentos HTML: ni comparten un formato determinado, ni posiblemente nos interese todo su contenido. Una primera aproximación completamente automática podría basarse en la suposición de que existen ciertas etiquetas HTML en la mayoría de los documentos que recuperemos, y que tales etiquetas marcan de alguna manera la información que realmente nos interesa. Así en [6], John M. Pierre desarrolla una metodología de categorización de sitios Web, buscando, entre otras cosas, ciertas etiquetas HTML como son los metadescriptores META o el título del documento (TITLE). Otros enfoques, como el descrito en [7], prentenden extraer información de la Web partiendo de la hipótesis de que los documentos HTML presentan un cierto grado de estructuración (textos semiestructurados). Ambos métodos se ven limitados justamente por su generalidad: presuponen ciertas características en todas las páginas visitadas. Por el contrario, el enfoque que presentamos permite una descripción individualizada de la estructura HTML para cada sitio Web. Con tal descripción es posible recuperar sólo los fragmentos interesantes, y además dotarlos de un determinado formato con independencia del estilo HTML original. Si bien el sistema no es automático (debido a esa descripción que debemos suministrar para cada sitio Web) el resultado obtenido es muy satisfactorio: documentos estructurados con independencia de su origen, y libres de información irrelevante. Actualmente se está recolectado un corpus comparable, que ya cuenta con unos 20.000 documentos a un ritmo de unos 60 nuevos documentos añadidos diariamente. Las fuentes son muy heterogéneas: las secciones nacional e internacional de los diarios “El País”, “ABC” y “El Mundo”, así como la sección internacional de “Washington Post”, “The Guardian Observer” y “CNN News”. La finalidad es obtener un corpus bilingüe comparable utilizable en tareas de recuperación de información multilingüe (CLIR).

2 Historia de los problemas

Para la consecución de un corpus como el descrito a partir de los sitios en línea de un conjunto de diarios, es necesario solventar ciertos inconvenientes que dificultan sensiblemente la tarea:

- La práctica totalidad de los documentos disponibles están escritos en HTML, el cual liga indisolublemente la presentación y la semántica. Tal es así que como ya apunta Tim Berners-Lee [9], la Web está en un formato legible por el ordenador, pero la compresión del documento leído es prácticamente nula.
- Los navegadores admiten documentos Web mal formados, además de soportar etiquetas (tags) no admitidos por el estándar W3C [10].
- Incluso si nos restringimos a los sitios que están codificados con HTML estricto, la amplia variedad de estilos, efectos, plantillas, etc. hace imposible suponer nada sobre la codificación de un sitio. Por ejemplo, no siempre el título de un artículo se corresponde con la cláusula HTML “<TITLE>”.
- Usualmente, no nos interesará todos los documentos existentes en un sitio, tan sólo algunos de ellos.

En definitiva, la Web es demasiado abrupta. Es necesario suavizarla, limarla. Es necesario extraer la información que realmente es relevante y transformar su presentación hasta darle un formato adecuado a nuestras necesidades. Una herramienta que pretenda afrontar tal tarea debe:

- Tratar sólo aquellas páginas acordes con la colección de documentos o corpus que deseamos crear. Siguiendo con el anterior ejemplo, artículos de internacional, no artículos de otras secciones, ni cualquier enlace accesible desde internacional.
- Tratar aquella parte de la página relevante para nuestros intereses. En una página que contiene un artículo de internacional, nos interesará el texto del artículo, no publicidad, enlaces a otros sitios y demás.
- Conferir un aspecto homogéneo a todos los documentos generados. Una vez obtenidos los artículos, sería deseable que cada uno de ellos se etiquetara con un titulo, fecha, diario y cuerpo del artículo, transformando o eliminando las marcas HTML encontradas en la página original.

3 Obtención de recursos lingüísticos en la Web

Una herramienta ideal que hiciera una tarea como la descrita, debería poseer un lenguaje de representación del conocimiento del sistema sobre el dominio del problema con una capacidad expresiva tal que, muy posiblemente, resultaría intratable computacionalmente [11]. Sin embargo, si estamos dispuestos a perder cierta capacidad expresiva, todo el proceso es susceptible de ser automatizado con un grado de éxito bastante aceptable:

- Dado un sitio Web, todos los documentos relevantes deben ser accesibles desde lo que denominamos páginas índice. Una página índice es una o más URLs relativas al sitio que estamos explorando, tales que contienen los enlaces necesarios para acceder a los documentos relevantes. Así por ejemplo, para procesar los artículos de la sección internacional de un diario, bastaría con conocer la URL de tal sección y especificar los enlaces de esa página que nos lleven a artículos de internacional. Para conseguir mayor flexibilidad, se permite usar patrones tanto en la especificación de las URLs de las páginas índices como en la de los enlaces, de tal manera que se procesen todas aquellas páginas cuya URL cumpla con el patrón asociado bien a una página índice o bien a los enlaces que constituyen tal página.
- Una vez decidido que una página es de interés, debe presentar algún indicio en su formato que nos permita asignar secciones o zonas de tal página a secciones del documento equivalente ya normalizado. Supongamos que queremos incluir una sección “.TITULO” en nuestros documentos normalizados. Procesando una página de internacional, podríamos descubrir que la sección “.TITULO” del documento normalizado se corresponde con el texto que encontramos etiquetado en la página HTML como "<TITLE>", o quizás esto no se cumpla, pero encontrásemos que el tipo de letra del título es más grande y marcado que el usado en el resto de la página HTML, usando entonces esa peculiaridad como indicio de que se trata del título.
- El documento una vez normalizado, debe

presentar un formato sencillo, consistente básicamente en pares (nombre de sección, contenido) o bien (<nombre de sección, contenido, </nombre de sección>)

Sin duda alguna, la restricción más fuerte es la primera, pues presupone que dado un sitio Web, éste estructura su contenido de tal manera que todos aquellos documentos relevantes para nuestros intereses, penden de ciertas páginas que actúan como índice. Esta restricción podría relajarse sensiblemente si aplicásemos técnicas de encaminamiento de documentos [12] a la totalidad del sitio Web, eliminado así la necesidad de la existencia de la página índice. Sin embargo, nos veríamos limitados por la precisión y la cobertura [12] de la técnica de encaminamiento usada, por lo que muy probablemente se tratarían algunos documentos no relevantes, además de pasar desapercibidos documentos que sí lo son.

4 Descripción del proceso de recuperación y transformación de documentos: el archivo de configuración

Una vez asumidas estas restricciones, podemos representar el conocimiento que tendrá la herramienta sobre el dominio del problema mediante un documento escrito en XML. Las razones que nos han animado a usar XML son básicamente dos:

- La solución diseñada sigue un procedimiento de naturaleza jerárquica y XML trata estructuras jerárquicas de una forma natural.
- Tal como define Data Conversion Laboratory, en el ciclo de vida que proponen para un proyecto de conversión de documentos [13], una de las piezas claves en la etapa de diseño es el documento de especificación de la conversión, cuya finalidad es definir formalmente qué se desea convertir exactamente y cómo. Pues bien, ya que XML es fácilmente legible por un lector humano [10, 14], un documento XML puede actuar como tal documento de especificación de la conversión. Pero además, ya que se trata de un texto formal, sin ambigüedades y por ello fácilmente procesable computacionalmente, es válido también como la base de conocimiento que posee el sistema. Por lo tanto, el documento XML que a continuación se describe, tiene una doble vertiente como documento de especificación (para el humano) y base de conocimiento (para la máquina).

El aspecto de uno de estos documentos XML es el siguiente:

```
<!DOCTYPE webreader SYSTEM
"/home1/dofer/projects/webreader/webreader.dtd">
<webreader>
    <site name="Diario “ABC”"
    URL="http://www.”ABC”.es/”ABC”/fijas/internac/index.a
    sp" process_links="true"   process_content="false">
        <link name="Sección–internacional">
            <URL>pa00</URL>
            <format>
                <append>.SECCION Internacional
                </append>
            </format>
        </link>
            <format>
                <translate>
                    <tag_in>titulo</tag_in>
                    <tag_out>.TITULO</tag_out>
                </translate>
                <translate>
                    <tag_in>entradilla</tag_in>
                    <tag_out>.ENTRADILLA</tag_out>
                </translate>
                <translate>
                    <tag_in>firma</tag_in>
                    <tag_out>.AUTOR</tag_out>
                </translate>
                <translate>
                    <tag_in>texto</tag_in>
                    <tag_out>.CUERPO</tag_out>
                </translate>
            <append>
                .PERIODICO “ABC”
                </append>
            </format>
    </site>
    <site name="Diario “El Mundo”"
    URL="http://www.elmundo.es/diario/internacional/index.
    html" process_links="true" process_content="false">
        <link name="Seccion Internacional">
            <URL>$DD$N0</URL>
            <format>
            <translate>
                <tag_in>TITLE</tag_in>
                <tag_out>.TITULO</tag_out>
            </translate>
            <translate>
                <tag_in>DIV</tag_in>
                <tag_out>.CUERPO</tag_out>
            </translate>
            <append>
                .SECCION Internacional
            </append>
            </format>
        </link>
        <format>
            <append>.PERIODICO “EL MUNDO”
            </append>
        </format>
    </site>
    <site name="Diario El Pais" URL=”
    http://www.elpais.es/p/d/$YYYY$$MM$$DD$/internac/ “
    process_links="true" process_content="false">
        <link name="Seccion Iternacional">
            <URL>internac</URL>
```

```
        <format>
            <translate>
                <tag_in closed="false"
                attr_name="NAME"
                attr_value="TITULO"
                attr_get="CONTENT">META</tag_in>
                <tag_out>.TITULO</tag_out>
            </translate>
            <translate>
                <tag_in>BODY</tag_in>
                <tag_out>.CUERPO</tag_out>
            </translate>
            <append>.SECCION Internacional
            </append>
        </format>
    </link>
        <format>
            <append>.PERIODICO EL PAIS
            </append>
            </format>
    </site>
        <format>
            <append>
                .FECHA $MM$-$DD$-$YYYY$
            </append>
            <append>
                .LINK $URL$
            </append>
            <ignore>P</ignore>
            <ignore>BR</ignore>
            <ignore>STRONG</ignore>
        </format>
        <target append="true">
            <path>/home1/dofer/projects/webreader</path>
            <file>spanish_news.data</file>
        </target>
</webreader>
```

Fig. 1: Un documento de especificación

En concreto, aquí se describe el procedimiento mediante el cual se recuperan y procesan los artículos de la sección internacional de los diarios “ABC”, “El País” y “El Mundo”. Cada artículo recuperado será transformado hasta presentar un aspecto semejante al mostrado en las fig.2, 3 y 4.

El documento XML está estructurado en cuatro niveles:

- el primer nivel (etiqueta webreader, ver fig. 1) representa la totalidad de los sitios. En nuestro ejemplo, se podría leer como "artículos de la sección de internacional de algunos diarios españoles".
- Un segundo nivel (etiqueta site) representa cada uno de esos sitios: esto es, “El País”, el “ABC” y “El Mundo”-
- El tercer nivel (etiqueta link) representa las páginas índice. Podemos especificar tantas páginas índice como se desee, si bien no suele ser necesario especificar demasiadas gracias al uso de patrones. En nuestro ejemplo es suficiente con una página índice (la sección de internacional).
- Por último, cada sección link consta de uno o más patrones (etiquetas URL) que representan los enlaces a las páginas relevantes que finalmente serán procesadas. Es el caso del patrón "internac" especificado para “El País”, pues los enlaces que encontramos en la página índice a artículos son del tipo "internac01.html", "internac02.html",etc.

Con esta estructura jerárquica hemos descrito una manera de acceder a las páginas relevantes. El siguiente paso es procesar cada una de estas páginas y transformarlas convenientemente. Estas transformaciones se consiguen de una manera sencilla a través de las reglas *translate*, *append* e *ignore*.

- Las reglas *translate* podríamos leerlas como "busca la etiqueta *tag_in* en la página en proceso, y sustitúyela por *tag_out* en el documento equivalente, ya normalizado”. Así por ejemplo, para los artículos de “El País”, el cuerpo del artículo lo encontramos delimitado por la etiqueta "<BODY>". En “El Mundo”, sin embargo, esa misma información aparece enmarcada por "<DIV>". Por último, en el “ABC”, usan una única etiqueta que denominan "<texto>". En los tres casos, las etiquetas originales son sustituidas en su documento equivalente ya normalizado por la cabecera ".CUERPO", seguido por el contenido de tal sección. Además es posible afinar en la búsqueda de etiquetas, referenciando en la regla no sólo el nombre de la etiqueta si no también el nombre y valor de algún atributo. Tal es el caso de la primera regla descrita para los artículos de “El País” (tabla I). Éste es el principal mecanismo establecido para normalizar los documentos. Por otra parte, supongamos que procesando un artículo del “ABC”, entre las marcas <texto> y </texto> encontrásemos otras marcas cualesquiera (B, FONT...), entonces éstas serían eliminadas, pero se mantendría el texto delimitado por ellas si lo hubiera, pues se entiende que ese texto es parte de la sección más amplia "<texto>...</texto>". En el caso de que estas etiquetas HTML o textos se encuentren al margen de las etiquetas anotadas en la secciones translate,

son eliminados tanto las etiquetas como el texto que delimitan. Este es el mecanismo establecido para eliminar aquellas partes irrelevantes de la páginas.(ver tabla I).

Regla aplicable	***Documento HTML***	***Documento normalizado***
<translate> <tag_in>texto</tag_in> <tag_out>.CUERPO<tag_out> </translate>	<texto>Este texto <B>se mantiene</B> </texto>pero éste otro <B>se desecha</B><texto>y éste de aquí se añade</texto>	.CUERPO Este texto se mantiene y éste de aquí se añade
<translate> <tag_in closed="false" attr_name="NAME" attr_value="TITULO" attr_get="CONTENT"> META </tag_in> <tag_out>.TITULO</tag_out> </translate>	<META name=”TITULO” content=”Finaliza la cumbre europea”> <META name=”FECHA” content=” 12/11/2000”>	.TITULO Finaliza la cumbre europea

Tabla 1. En el segundo ejemplo, se usa una variante extendida de las reglas *translate* que permite procesar los posibles atributos de las etiquetas HTML

- Las reglas *append* permiten añadir información al documento resultante, no extraerla de la página HTML original. Tal es el caso de la sección ".FECHA" que representa la fecha en la que tal documento fue procesado, o añadir una sección ".DIARIO", que indica la procedencia del texto. Ambas secciones forman parte de los documentos ya normalizados, sin que exista equivalencia alguna entre estas secciones y las que podamos encontrar en la página HTML original.
- Por último, las reglas *ignore* son útiles si deseamos que ciertas marcas HTML se mantengan inalteradas cuando sean encontradas dentro de alguna sección de nuestro interés. Esto nos permite mantener cierto formato básico del documento original. Por ejemplo, puede ser útil mantener las marcas "<P>", pues representan retornos de carro en el lenguaje HTML.

Una de las características más interesantes para conseguir un aspecto homogéneo en todos los documentos que se generen, es el uso de la herencia. Las tres reglas arriba descritas pueden especificarse a cualquier de los cuatro niveles disponibles, heredándose desde los niveles superiores hacia los inferiores. En caso de que haya reglas incompatibles (por ejemplo, definir dos reglas translate a dos niveles distintos, pero que trabajan sobre la misma etiqueta), se opta por aplicar la regla especificada en el nivel más local. De esta manera podemos conseguir que todos los documentos recuperados cuenten con una sección .FECHA, o que todos los documentos del el diario “El Mundo”, tengan una sección con el aspecto:

.DIARIO
El Mundo

y además, únicamente para aquellos que pertenecen específicamente a la sección de internacional, se anote en el documento el texto:

.SECCION
Internacional

Ejemplos de documentos obtenidos con la configuración de la fig. 1, son los mostrados en la fig.2, 3 y 4

```
.FECHA
12-08-2000

.LINK
http://www.elpais.es/p/d/20001208/espana/retirada.htm

.PERIODICO
EL PAIS

.SECCION
Internacional

.TITULO
El PSOE busca apoyos en el Congreso para exigir al Reino Unido la retirada del 'Tireless'
```

```
.CUERPO
<BR><BR>

El PSOE busca apoyos en el Congreso para exigir al Reino Unido la retirada del submarino
<P> Zapatero afirma que la actuación del Gobierno "es un monumento a la incompetencia"
<P>A.DÍEZ/P.EGURBIDE,
Madrid/Niza
<BR> La revelación del presidente del Gobierno, José María Aznar, … tiene ahora que obtener resultados". <BR><BR>
```

Fig. 2. Extracto de un artículo ya normalizado procedente de “El País”

.FECHA
12-09-2000

.LINK
http://www.”ABC”.es/”ABC”/fijas/nacional/006pa00.asp

.PERIODICO
ABC

.SECCION
Nacional

.TITULO
PP y PSOE se comprometen a acabar con ETA y cualquier intento de darle «legitimación política»

.AUTOR

.ENTRADILLA
El pacto «por las libertades… y llama a todos los demócratas a «compartir estos principios».

.CUERPO
MADRID. Gonzalo López Alba
El acuerdo entre el PP y el PSOE… Rodríguez Zapatero.
El documento… todos los demócratas».

Fig. 3. Extracto de un artículo ya normalizado procedente de la sección de nacional del “ABC”

.FECHA
12-09-2000

.LINK
http://www.elmundo.es/diario/espana/09N0001.html

.PERIODICO
EL MUNDO

.SECCION
Nacional

.TITULO
ESPAÑA | PP y PSOE cierran el acuerdo contra el terrorismo de ETA. Exigen al PNV la «ruptura formal» con Estella y sus organismos

.CUERPO
ESPAÑA. PP y PSOE exigen al PNV la «ruptura formal» … contra el terrorismo de ETA.

Fig. 4. Extracto de un artículo ya normalizado procedente de la sección nacional de “El Mundo”

5 Conclusiones y trabajo futuro

Se ha mostrado aquí una manera sencilla y bastante eficaz de explotar la Web como recurso linguístico. Mediante la especificación de reglas y alguna heurística, es posible recuperar páginas de muy variados origenes y formatos, para posteriormente obtener a partir de ellas un conjunto de documentos con un formato homogéneo, pero respetando el contenido original, permitiendo de esta manera su posterior procesamiento automático con fines experimentales en PLN o, más concretamente en nuetro caso, recuperación de información. Nuestro método se distancia de otros sistemas de conversión de documentos tal como el descrito en [14], pues si bien estos sistemas parten de un conjunto de reglas aplicables al documento origen, no facilitan en manera alguna la consecución de un conjunto de documentos con un formato homogéneo. Esto es, no habilitan mecanismos de herencia para las reglas, el uso de patrones en la especificación de las URLs de los documentos a tratar, ni su agrupación jerárquica, según sea su formato original.

Los resultados que hemos obtenido en distintos ambientes (periódicos, boletines oficiales de ámbito nacional, boletines de la CEE) han sido muy satisfactorios, eliminando la necesidad de supervisión humana en la tarea de recuperación y normalización de documentos. Sin embargo, para que esto sea posible, es necesaria la existencia de lo que venimos denominando páginas índice, que son puertas a los documentos relevantes, y que deben ser conocidas de antemano. El concepto de “página índice” es muy flexible, pues puede corresponderse con una o más partes del sitio, y además se permite el uso de patrones tanto en la especificación de las URLs de las páginas índices, como de los enlaces contenidos en éstas. Aún así, hay casos en los que no es posible la descripción de la página índice. Entonces la respuesta a la anterior pregunta pasa con toda probabilidad por el uso de técnicas de encaminamiento de documentos, quedando esta mejora pendiente para futuras versiones. Otras mejoras que consideramos interesantes es la de conseguir mayor flexibilidad en las reglas translate, mediante la posibilidad del uso de hojas de estilo XSL.

6 Referencias

[1] Gregory Grefenstette. "The WWW as a Resource for Example-Based MT Tasks". conference. In ASLIB'99 Translating and the Computer 21, London, UK, Nov 10-11, 1999.

[2] D. Lewis. Text Representation for Intelligent Text Retrieval: A Classification-Oriented View. En P. Jacobs, editor, Text-Based Intelligent Systems, Capítulo 9. Lawrence Erlbaum, 1992.

[3] Yahoo!, http://www.yahoo.com/

[4] Altavista, http://www.altavista.com/

[5] Open Directory Project, http://www.dmoz.org/

[6] John M. Pierre. On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science. Vol. 6(2001): nr 0.

[7] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts. University of Library and Information Science.

[8] Eugenio Picchi and Carol Peters. Cross-Language Information Retrieval: A system for comparable corpus querying. En G. Grefenstette, editor, Cross-Language Information Retrieval, capítulo 7, Kluwer Academic Publishers, Boston, 1998.

[9] T.Berners-Lee. Semantic Web Road Map.http://www.w3.org/DesignIssues/Semantic.html, 1998.

[10] Charles F. Goldbarg and Paul Prescod. Why XML?. In The XML Handbook, chapter 1, Prentice Hall PTR.

[11] Hector J, Levesque and Ronald J. Brachman. A Fundamental Tradeoff in Knowledge Representation and Reasoning. Readings in Knowledge Representation, p41-70, 1984.

[12] Gerald Kowalski. Information Retrieval Systems. Theory and Implementation. Kluwer Academic Publishers, 1997.

[13] Mark Gross and John Lynch. Planing for document conversion. In The XML Handbook, second edition, chapter 39, Prentice Hall PTR.

[14] Charles F. Goldbarg and Paul Prescod. DynaTag visual conversion environment. In The XML Handbook, chapter 24, Prentice Hall PTR.

[15] Web Consortium. www.w3c.org/XML.