Aplicación de Técnicas de Extracción de Información a Bibliotecas Digitales

Alejandro Bia
abia@dlsi.ua.es
Bib.Virtual Miguel de Cervantes
Tel: 34-96-5903400 #9567

Rafael Muñoz
rafael@dlsi.ua.es
Dpto. Lenguajes y Sistemas Informáticos
Tel: 34-96-5903653 Fax: 34-96-5909326

Universidad de Alicante, apartado de correos 99, E-03080, España

Resumen Con frecuencia, las Bibliotecas Digitales tienen la necesidad de extraer información a partir de documentos pobremente marcados para almacenarla en bases de datos o crear nuevos documentos hipertexto con un marcado altamente estructurado. En este trabajo, abordaremos el problema de extraer información bibliográfica a partir de informes literarios en formato HTML para alimentar una base de datos de publicaciones gallegas de una Biblioteca Digital para ser consultada a través de Internet. Para este fin se ha utilizado con éxito una metodología que aprovecha la información contenida en las marcas HTML y que a su vez aplica técnicas de Procesamiento del Lenguaje Natural (PLN).

1 Introducción

Tal y como Sperberg y Burnard [9] afirman: "Un sistema de marcado descriptivo utiliza marcas o etiquetas que simplemente proporciona nombres para categorizar, estructurar parte del documento". Son claras las ventajas del marcado estructural o descriptivo, el cual define los componentes estructurales de un documento, comparado con el marcado procedural, que únicamente define características de salida o de impresión (fuentes, tamaños, caracteres enfatizados -itálica, negrita-) tal y como Abaitua cita en [1]. Los lenguajes de marcados descriptivos, como XML, añaden algunos valores semánticos al texto que pueden ser útiles para posteriores procesamientos, como la extracción de información para el relleno de bases de datos o para la generación de diferentes formatos de documentos, resúmenes o listados. Pero normalmente, los textos utilizados como punto de partida vienen con pequeñas o ninguna marca, como las salidas de un OCR[1], o documentos electrónicos en formato como RTF, PDF, PS o HTML.

2 El problema

En este trabajo se presenta una aproximación al problema del la extracción de información bibliográfica a partir de informes literarios en formato HTML para su almacenamiento en una base de datos de referencias bibliográficas que será usada para responder a preguntas realizadas desde una página de búsqueda de una Biblioteca Digital. Estos informes literarios están escritos en lengua gallega. El formato de estos informes literarios consisten en texto libre en el que se hace una descripción de varias obras. Cada obra se encuentra agrupada en un grupo de frases (párrafos) en las que se describe todos los detalles bibliográficos (nombre del autor, título de la obra, lugar de publicación, editores, artista gráfico, colección, numero, volumen, año de publicación, numero de páginas, ISBN y un pequeño resumen de la obra). Algunos de estos detalles están siempre presentes en todas las obras, pero otros son opcionales, como artista gráfico, colección, numero, etc. Además puede aparecer otro tipo de información sobre la obra que no es significativa para la base de datos de las referencias bibliográficas.

Estos párrafos pueden contener cabeceras, comentarios literarios, etc. El formato de dichos párrafos suele ser consistente (en lo relativo a convenciones sobre títulos y tamaños de las fuentes que se mantienen a lo largo de todo el texto), pero hay diferencias en el formato y delimitadores de campo entre estas descripciones de libros que hacen que el uso de un analizador (parser) convencional sea poco práctico.

---

[1] Los mejores programas de OCR de hoy en día pueden reproducir el formato procedural del texto origen, pero nada mas.

Por un lado, HTML, aún siguiendo el estándar SGML no es tan rico como XML con su capacidad de marcado estructural-descriptivo. HTML define algunos elementos estructurales básicos como título (⟨TITLE⟩ ... ⟨/TITLE⟩), cuerpo del texto (⟨BODY⟩ ... ⟨/BODY⟩), cabeceras (⟨H#⟩, ⟨/H#⟩, dónde # representa un número de cabecera), párrafos (⟨P⟩, ⟨/P⟩), etc, pero su fuerza radica en el aspecto de marcado procedural. Por lo tanto nuestro punto de partida es un texto bien estructurado pero pobremente marcado para nuestro propósito (sin etiquetas léxicas ni sintácticas).

Por otro lado, dentro de las múltiples aplicaciones para las que puedes ser aplicada la *extracción de información (EI)*, quizás la más interesante y a su vez la más desarrollada, es el almacenamiento de la información extraída en una base de datos. Es decir, la trasformación de información no estructurada en información estructurada (en función de unas plantillas que se corresponderán con la estructura de las diferentes tablas de la base de datos). Tal y como citan algunos autores, [3, 7], el objetivo de la extracción de información es el relleno de una plantillas (definidas previamente) a partir del texto a manejar para su inserción en una base de datos. Mientras que el objetivo de la *recuperación de información (RI)* es el proporcionar los textos en los que se encuentra la información deseada y rechazando aquellos textos o documentos en los que no se encuentra dicha información. es por ello que muchos autores consideran a la recuperación de información como una etapa previa a la extracción de información.

Nuestro problema está a mitad de camino entre los documentos no estructurados completamente y los documentos altamente estructurado. Nuestros documentos son documentos HTML pobremente estructurados con un formato consistente. Este formato consistente puede ser usado para la selección de los fragmentos de textos en los que se encuentra la información a extraer. La aplicación de algunos procesos de extracción de información con la ayuda de algunos recursos de NLP como diccionarios nos permitirá rellenar las plantillas usadas para el almacenamiento en la base de datos.

3 Extracción de información: Tareas a realizar

La EI no es una nueva idea, podemos encontrar sus raíces a mediados de los años 60. Pero es en los 80 cuando la extracción de información empieza a crecer rápidamente. Esto es debido a la intervención de DARPA[2], la cual fomentó la competición entre diferentes grupos de investigación para que desarrollasen sistemas de extracción de información. Estas reuniones dieron origen a las hoy conocidas como Messages Understanding Conference (MUC). El objetivo de estas conferencias ha sido establecer un régimen de evaluación cuantitativo para los sistemas de extracción de información. Estableciendo el conjunto de textos sobre los que todos los sistemas eran evaluados. En las MUC-3 (mayo-1991) se adoptaron como medidas de evaluación las métricas utilizadas en *recuperación de información.* Estas métricas son la *cobertura* y la *precisión.* Se entiende por cobertura (recall) el número de extracciones correctas realizadas con respecto al número de extracciones posibles existentes en el texto. Se entiende por precisión (precision) el número de extracciones correctas con respecto del total extraídas.

$$Precision = \frac{n^o\ extracciones\ correctas}{n^o\ extracciones\ posibles}$$

$$Cobertura = \frac{n^o\ extracciones\ correctas}{n^o\ extracciones\ realizadas}$$

En los MUC-6 (noviembre-1995) se definieron cuatro tareas fundamentales que siguiendo las definiciones realizadas por Cunningham [4], consisten en:

- **Reconocimiento de entidades (NE).** El objetivo de esta tarea es encontrar y clasificar las entidades. Se entiende por entidades los nombres de personas, organizaciones, lugares, expresiones temporales y expresiones numéricas relacionadas con monedas.
- **Resolución de correferencias (CO).** El objetivo es identificar las expresiones en el texto que hagan referencia al mismo objeto. Las relaciones de correferencias son sólo marcadas entre ciertas clases de expresiones sintácticas (sintagmas nominales definidos, pronombres).

---

[2] DARPA es la agencia de defensa de los Estados Unidos

- **Producción de plantillas de elementos (TE).**

  El objetivo de esta tarea es añadir información descriptiva a los resultados de NE.

- **Extracción de plantillas de escenarios (ST).**

  El objetivo de esta tarea es reunir los resultados de TE en el escenario específico.

En la MUC-7 (abril-1998) se definió una nueva tarea que se denominó **relación de plantillas (TR)**. El objetivo de esta tarea es identificar las relaciones entre las diferentes plantillas de elementos. Existen tres tipo de relaciones: *empleado de, localizado en* y *producto de*.

Actualmente existe una gran cantidad de sistemas que se basan en la estructura propuesta por Grishman [5], realizando pequeñas variaciones. La Figura 1 muestra esta arquitectura, en la que se observa que tiene una estructura modular, o mejor dicho tubular ya que consiste en una serie de módulos independientes que realizan las tareas básicas de los sistemas de extracción de información y en la que la salida de un módulo sirve como entrada del módulo siguiente.

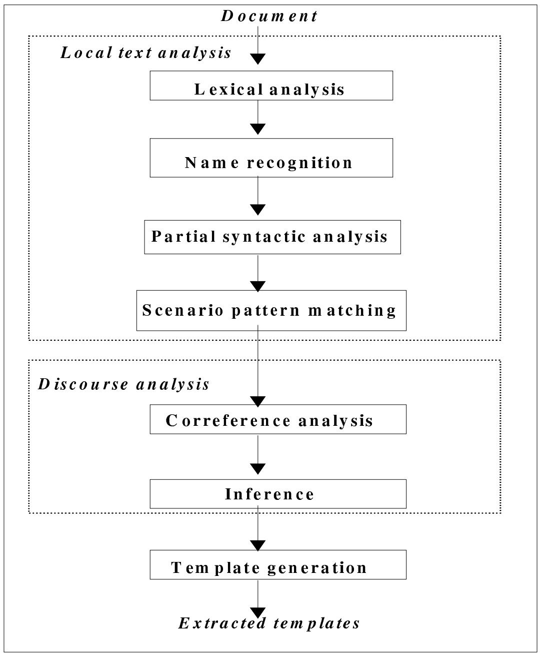


Figura 1: *Arquitectura de un sistema de extracción de información*

Los dos sistemas más representativos de la arquitectura propuesta por Grishman son: LaSIE-II [6] y Proteus [10].

4 El método

El método que presentamos a continuación combina tanto técnicas de marcado (HTML markup techniques) como técnicas de Procesamiento de Lenguaje Natural. Como se puede ver en la figura 2 en nuestro método se diferencian dos etapas o fases:

4.1 Aplicación de técnicas de marcado HTML.

En esta etapa un segmentador procesa un texto en formato HTML, y extrae todos los agrupamientos de información relevante ("clusters") basándose en el marcado HTML (la figura 3 muestra un "cluster"). De este modo, buena parte del texto irrelevante es pasado por alto. Para ello, debemos observar primero las características de formato de los segmentos en los que estamos interesados y tomar nota de los cambios que pueden indicar dónde esta la información que estamos buscando. Esto es lo que hace un humano cuando hace lectura rápida: los cambios en tamaño de letra, texto destacado, títulos, justificación, etc., nos ayudan a identificar las partes del texto donde puede estar la información que estamos buscando (en este caso descripciones bibliográficas), e ignorar el resto del texto. Es por este motivo que creemos que la información de formato no debe ser ignorada en aplicaciones prácticas para la extracción de información.

Esta observación previa del material fuente se convierte en instrucciones de filtrado para el programa segmentador. Cuando el texto tiene un formato homogéneo y regular, esto ayuda a extraer los potenciales segmentos de información interesante con alta precisión (100% en nuestro caso, en que los textos presentan un formato regular y se han usado reglas adecuadas de segmentación). Esto reduce el problema de EI a procesar sólo estos segmentos y no todo el texto fuente. Por otro lado sabemos, por observación de los textos fuente, que hay una correspondencia biunívoca entre las descripciones bibliográficas y los segmentos. Haciendo una analogía con la Ingeniería Electrónica podemos decir que como resultado, este proceso de segmentación separa exitosamente la señal útil del ruido.

En la siguiente etapa, por cada uno de los segmentos extraídos se creará una plantilla con la información relevante usando técnicas de PLN.

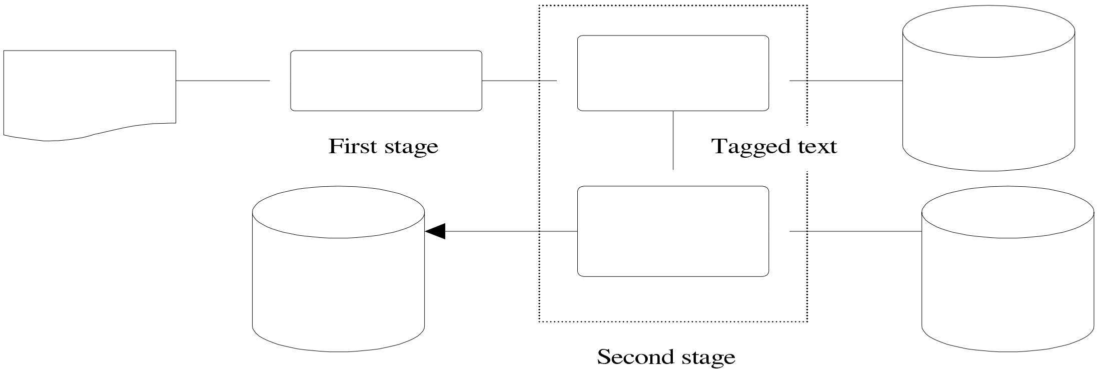
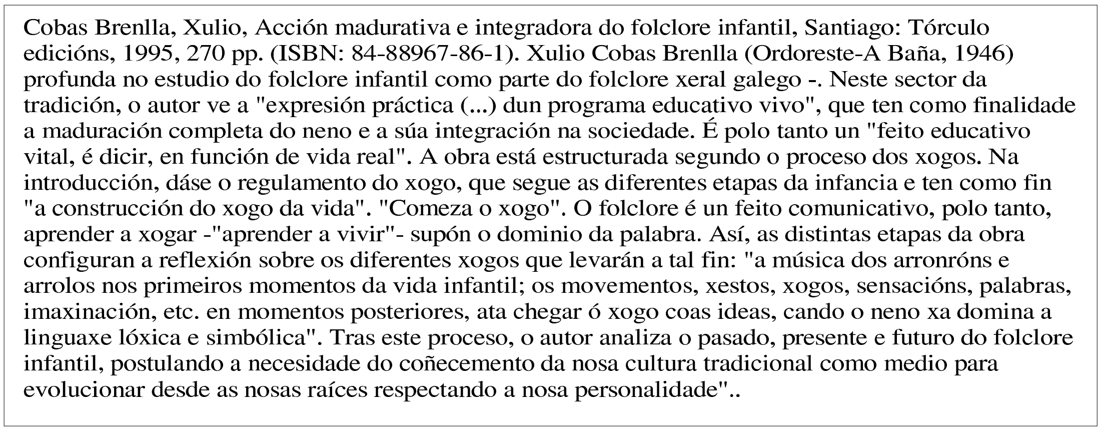
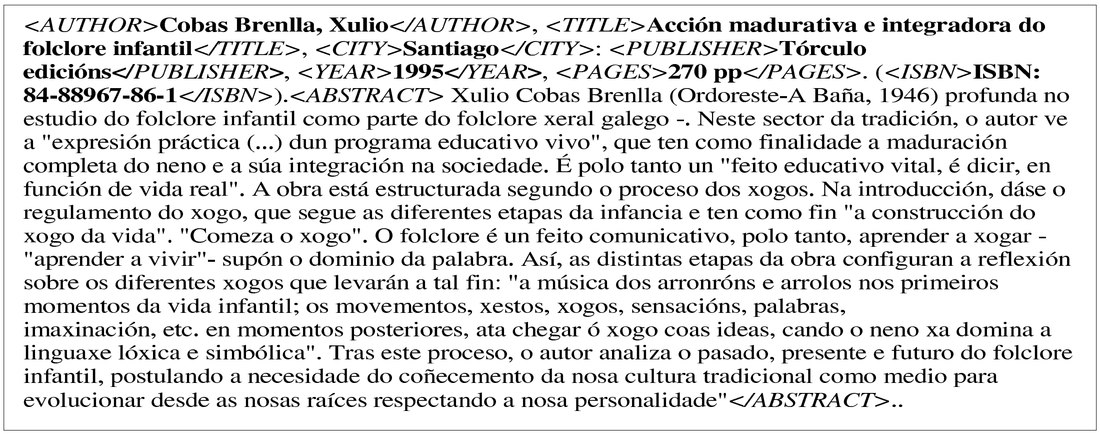


Figura 2: *Diagrama de flujo de datos de proceso de extracción de información*



Cobas Brenlla, Xulio, Acción madurativa e integradora do folclore infantil, Santiago: Tórculo edicións, 1995, 270 pp. (ISBN: 84-88967-86-1). Xulio Cobas Brenlla (Ordoreste-A Baña, 1946) profunda no estudio do folclore infantil como parte do folclore xeral galego -. Neste sector da tradición, o autor ve a "expresión práctica (...) dun programa educativo vivo", que ten como finalidade a maduración completa do neno e a súa integración na sociedade. É polo tanto un "feito educativo vital, é dicir, en función de vida real". A obra está estructurada segundo o proceso dos xogos. Na introducción, dáse o regulamento do xogo, que segue as diferentes etapas da infancia e ten como fin "a construcción do xogo da vida". "Comeza o xogo". O folclore é un feito comunicativo, polo tanto, aprender a xogar -"aprender a vivir"- supón o dominio da palabra. Así, as distintas etapas da obra configuran a reflexión sobre os diferentes xogos que levarán a tal fin: "a música dos arronróns e arrolos nos primeiros momentos da vida infantil; os movementos, xestos, xogos, sensacións, palabras, imaxinación, etc. en momentos posteriores, ata chegar ó xogo coas ideas, cando o neno xa domina a linguaxe lóxica e simbólica". Tras este proceso, o autor analiza o pasado, presente e futuro do folclore infantil, postulando a necesidade do coñecemento da nosa cultura tradicional como medio para evolucionar desde as nosas raíces respectando a nosa personalidade"..

Figura 3: *Cluster*

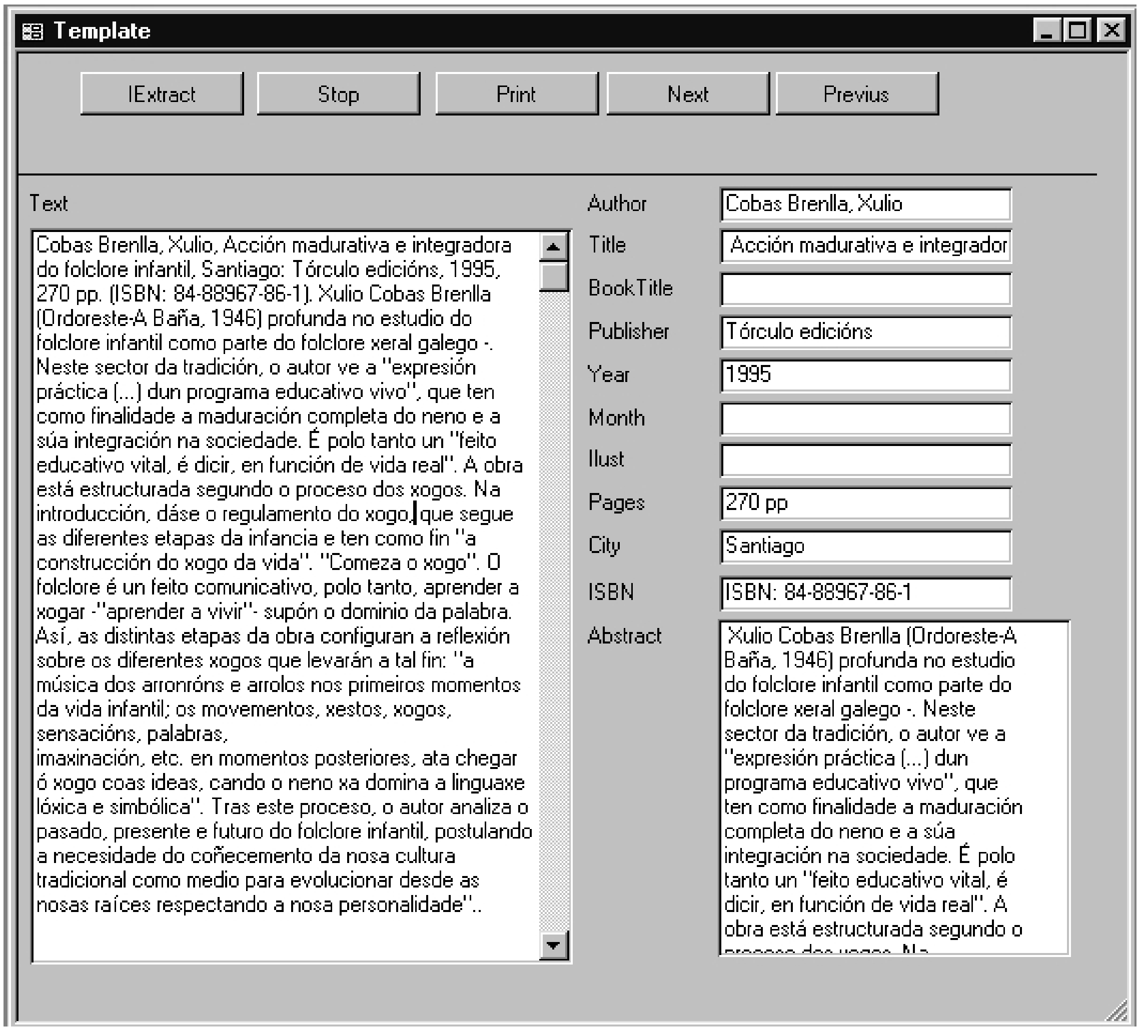


*<AUTHOR>***Cobas Brenlla, Xulio***</AUTHOR>*, *<TITLE>***Acción madurativa e integradora do folclore infantil***</TITLE>*, *<CITY>***Santiago***</CITY>*: *<PUBLISHER>***Tórculo edicións***</PUBLISHER>*, *<YEAR>***1995***</YEAR>*, *<PAGES>***270 pp***</PAGES>*. (*<ISBN>***ISBN: 84-88967-86-1***</ISBN>*).*<ABSTRACT>* Xulio Cobas Brenlla (Ordoreste-A Baña, 1946) profunda no estudio do folclore infantil como parte do folclore xeral galego -. Neste sector da tradición, o autor ve a "expresión práctica (...) dun programa educativo vivo", que ten como finalidade a maduración completa do neno e a súa integración na sociedade. É polo tanto un "feito educativo vital, é dicir, en función de vida real". A obra está estructurada segundo o proceso dos xogos. Na introducción, dáse o regulamento do xogo, que segue as diferentes etapas da infancia e ten como fin "a construcción do xogo da vida". "Comeza o xogo". O folclore é un feito comunicativo, polo tanto, aprender a xogar -"aprender a vivir"- supón o dominio da palabra. Así, as distintas etapas da obra configuran a reflexión sobre os diferentes xogos que levarán a tal fin: "a música dos arronróns e arrolos nos primeiros momentos da vida infantil; os movementos, xestos, xogos, sensacións, palabras, imaxinación, etc. en momentos posteriores, ata chegar ó xogo coas ideas, cando o neno xa domina a linguaxe lóxica e simbólica". Tras este proceso, o autor analiza o pasado, presente e futuro do folclore infantil, postulando a necesidade do coñecemento da nosa cultura tradicional como medio para evolucionar desde as nosas raíces respectando a nosa personalidade"*</ABSTRACT>*..

Figura 4: *Segmento marcado*

REFERENCE: *Cobas95*,
author ...... = Xulio Cobas Brenllas
title .... = Acción madurativa e integradora do folclore infantil
booktitle .. =
publisher.. = Tórculo edicións
month ..... =
year .... = 1995
ilust .... =
pages ...... = 270
city .... = Santiago
ISBN ...... = ISBN: 84-88967-86-1

Figura 5: *Plantilla rellena*



Text

Cobas Brenlla, Xulio, Acción madurativa e integradora do folclore infantil, Santiago: Tórculo edicións, 1995, 270 pp. (ISBN: 84-88967-86-1). Xulio Cobas Brenlla (Ordoreste-A Baña, 1946) profunda no estudio do folclore infantil como parte do folclore xeral galego -. Neste sector da tradición, o autor ve a "expresión práctica (...) dun programa educativo vivo", que ten como finalidade a maduración completa do neno e a súa integración na sociedade. É polo tanto un "feito educativo vital, é dicir, en función de vida real". A obra está estructurada segundo o proceso dos xogos. Na introducción, dáse o regulamento do xogo, que segue as diferentes etapas da infancia e ten como fin "a construcción do xogo da vida". "Comeza o xogo". O folclore é un feito comunicativo, polo tanto, aprender a xogar -"aprender a vivir"- supón o dominio da palabra. Así, as distintas etapas da obra configuran a reflexión sobre os diferentes xogos que levarán a tal fin: "a música dos arronróns e arrolos nos primeiros momentos da vida infantil; os movementos, xestos, xogos, sensacións, palabras,
imaxinación, etc. en momentos posteriores, ata chegar ó xogo coas ideas, cando o neno xa domina a linguaxe lóxica e simbólica". Tras este proceso, o autor analiza o pasado, presente e futuro do folclore infantil, postulando a necesidade do coñecemento da nosa cultura tradicional como medio para evolucionar desde as nosas raíces respectando a nosa personalidade"..

Author: Cobas Brenlla, Xulio
Title: Acción madurativa e integrador
BookTitle:
Publisher: Tórculo edicións
Year: 1995
Month:
Ilust:
Pages: 270 pp
City: Santiago
ISBN: ISBN: 84-88967-86-1
Abstract: Xulio Cobas Brenlla (Ordoreste-A Baña, 1946) profunda no estudio do folclore infantil como parte do folclore xeral galego -. Neste sector da tradición, o autor ve a "expresión práctica (...) dun programa educativo vivo", que ten como finalidade a maduración completa do neno e a súa integración na sociedade. É polo tanto un "feito educativo vital, é dicir, en función de vida real". A obra está estructurada segundo o

Figura 6: *Pantalla de la aplicación*

4.2 Aplicación de técnicas de PLN.

4.2.1 Reconocimiento de entidades.

Tal y como se definieron en las MUC entendemos por entidades a las unidades de información relevante a extraer a partir de los textos. A partir del estudio de una parte de los textos a procesar (considerado como corpus de entrenamiento) se han extraído un conjunto de reglas heurísticas que cumplen esas entidades. Entre otras, debemos extraer información del autor, del titulo, del editor, etc. Todas las entidades tienen un conjunto de características comunes que quedan plasmadas en las heurísticas que posteriormente propondremos. Diferenciamos dos tipos de heurísticas, por un lado tenemos una serie de reglas heurísticas que nos da información acerca de donde aparece cada tipo de información (autor, título, etc.). A estas reglas denominaremos patrones generales. Y por otro lado, tenemos patrones específicos para reconocer cada tipo de entidad (autor, título, editor, etc.).

Estas reglas heurísticas específicas se basan en diversas características útiles para la detección de entidades: 1) los nombres son palabras que empiezan con mayúsculas 2) algunas entidades incluyen palabras especiales o abreviaturas que actúan cono disparadores. (por ejemplo palabras como "Ilust."(ilustrador), "eds."(editores) o ISBN). La dificultad intrínseca de esta etapa es identificar el comienzo y fin de cada entidad. Uno de los problemas que encontramos es la aparición de preposiciones, artículos y conjunciones, dentro de la entidad que queremos reconocer (por ejemplo "del"en el nombre "Fernández del Riego", o "A"en el la ciudad "A Coruña").

A continuación mostramos algunas de las heurísticas usadas (los componentes entre "$\ll$"y "$\gg$"son opcionales):

- Patrones generales
  - *Author, Title, City: Editors, Year, Pp. "("Identification ")".*
  - *Author, Title. Comments, City: Editors, year, Pp. "("Identification ")".*
  - *Author, Title, City: Editors, Collection, Year, Pp. "("Identification ")".*
  - *Author, Title. Comments, City: Editors, Collection, Year, Pp. "("Identification ")".*
- Patrones específicos
  - $Author \rightarrow Surname1 \ll Surname2 \gg, FirstName1 \ll FirstName2 \gg, \ll "e" Author \gg$
  - *Title* → {*Words*} until looking for a comma(",") and a city name
  - *Year* →$\ll$ *Monthname* $\gg$ *Number.*
  - *Pp* → *number + "pp.".*
  - *Pp* → *"pp."+ number + "+ number.*
  - *Identification* → *"ISBN."+ number.*
  - *Identification* → *"D.L."+ number.*
  - *Identification* → *"ISSN."+ number.*

4.2.2 Manipulación de plantillas

1. Relleno de plantillas.

   Se añade una etiqueta a cada entidad reconocida para marcar cada elemento de la plantilla a rellenar, como se ve en la figura 4. Al reconocer todas las entidades de un segmento lo que se obtiene es un texto con una serie de etiquetas. Posteriormente, un proceso completa los huecos de las plantillas relacionando cada tipo de etiqueta con los huecos correspondientes. De esta forma se obtiene una plantilla por cada segmento, como puede verse en la figura 5 que muestra una plantilla llena.

2. Almacenamiento en la base de datos.

   En la base de datos se almacenan todas aquellas plantillas que tienen rellenos todos los atributos obligatorios, aunque algunos de los atributos opcionales estén vacíos. Con ello conseguimos una base de datos con las referencias bibliográficas de todas las obras literarias que aparecen en los segmentos detectados.

   Las plantillas en que no se pueden completar los elementos obligatorios se almacenan en un fichero para posterior corrección manual. Con ello se consigue que la BD sea correcta, segura, consistente y completa.

File	Clust	NE Recog	Filled Templ	Prec
G005	25	75	20	80.0%
G006	23	84	20	86.9%
G007	31	106	26	83.9%
Total	**79**	**265**	**66**	**83.5%**

Tabla 1: Resultados obtenidos

5 Evaluación

Este sistema ha sido evaluado sobre diferentes fragmentos de texto HTML (diferentes de aquellos usados para deducir las heurísticas). Por un lado, logra una precisión [3] de 83.5% en rellenado completo de plantillas. Lo que es más, éste enfoque logra una precisión del 92% en la tarea de reconocimiento de autor, título y lugar geográfico.

La tabla 1 muestra los resultados obtenidos luego de procesar 3 ficheros HTML. Estos ficheros están compuestos de 79 segmentos diferentes. Cada segmento corresponde a una plantilla. Nuestro método reconoce 66 plantillas exitosamente. En la tarea de reconocimiento de entidades con nombre nuestro método logra una precisión del 92% (243 de 265 entidades).

La figura 6 muestra un ejemplo procesado por nuestra aplicación. Podemos ver un segmento sin marcas HTML extraído del fichero HTML dentro de una ventana de texto a la derecha (etapa 1). En el lado izquierdo se observan todos los huecos de la plantilla (rellenos o no).

6 Conclusiones y trabajos futuros

Hemos desarrollado una herramienta para la estructuración de textos y su posterior almacenamiento en una base de datos bibliográfica a partir de textos HTML utilizando técnicas de extracción de información. Hemos usado tanto la segmentación del texto basada en las marcas procedurales del HTML así como las propiedades del lenguaje natural para detectar y extraer la información puntual. El texto fuente estaba bien estructurado pero pobremente marcado (en el sentido de que las marcas era puramente de formato y no indicaban componentes estructurales interesantes para la aplicación). Las técnicas de procesamiento de lenguaje natural utilizadas, basadas en reglas heurísticas, ayudan a detectar entidades para la creación o actualización de bases de datos. El conocimiento estructural adquirido a través de este proceso de extracción de información puede también ser usado para generar nuevos documentos con un marcado descriptivo más rico [2, 1]. El único recurso de PLN necesario fue un diccionario de nombres y lugares para lograr una precisión del 83.5%.

Referencias

[1] J. Abaitua. Material de referencia para un curso de introducción a sgml. http://orion.deusto.es /~abaitua /konzeptu /sgml /sgml0.htm, Visited 3-2-1999.

[2] J.H. Coombs, A.H. Renear, and S.J. DeRose. Markup systems and the future of scholarly text processing. *Communications ACM*, 30/11:933–947, 1987. Cf. CACM 31/7 (July 1988) 810-811.

[3] M. Crawdford. Information extraction. http://www.dcs.shef.ac.uk /research /groups /extraction, Visitada el 10-12-1997.

[4] H. Cunningham. Information Extraction a User Guide. Technical report, Research memo CS-97-02. Institute for Language, Speech and Hearing (ILASH), and Department of Computer Science. University of Sheffield. UK, 1997.

[5] R. Grishman. Information extraction: Techniques and challanges. *Lecture Notes in Computer Science*, 1299:10–27, 1997.

[6] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, and B. Mitchell. University of Sheffield: Description of the LaSIE-II System as used for MUC-7. In Publishers [8].

[7] University of Massachussetts. Information extraction. http://www.dcs.shef.ac.uk /research /groups /extraction, Visitada el 10-12-1997.

[8] Morgan Kaufman Publishers, editor. *Proceedings of Seventh Message Understandig Conference*, http://www.muc.saic.com /proceedings/, Spring 1998.

---

[3] Precisión es el cociente entre el número de plantillas correctamente resueltas y el número de segmentos procesados.

[9] C. M. Sperberg-McQueen and Lou Burnard, editors. *A Gentle Introduction to SGML*, chapter 2, page 23. TEI P3 Text Encoding Initiative Chicago, Oxford, May 1994.

[10] R. Yangarber and R. Grishman. NYU: Description of the Proteus/PET system used for MUC-7. In Publishers [8].