UNIVERSIDAD CARLOS III DE MADRID

JOTRI 2003

8-9 SEPTIEMBRE 2003

LEGANÉS, MADRID

II JORNADAS DE TRATAMIENTO Y RECUPERACIÓN DE LA INFORMACIÓN

Editado por:
Dpto. de Biblioteconomía y Documentación y
Dpto. de Informática
Universidad Carlos III de Madrid

Programa

Martes, 9 de septiembre de 2003

Mañana

9:30	Panel - Mesa redonda	
10:30	Café.	
11:00	Desarrollo de una aplicación para la gestión de relaciones en tesauros generados automáticamente.	Página 151
11:15	Ontologías, metadatos y agentes: recuperación "semántica" de la información.	Página 157
11:30	Visualización en topic maps: tendencias y propuestas.	Página 166
11:45	El nuevo tesauro Eurovoc como instrumento de recuperación de la información multilingüe.	Página 175
12:00	Automatic generation of a multilingual similarity thesaurus from the web.	Página 183
12:15	Análisis morfosintáctico para la extracción de información.	Página 191
12:30	Named entity recognition and classification for texts in Basque.	Página 198
12:45	Extracción de términos índice mediante cascadas de expresiones regulares.	Página 204
13:00	La extracción de palabras clave en OmniPaper: primeras experiencias y conclusiones.	Página 212
13:15 - 13:45	Debate.	

Tarde

16:00	Indicadores de rendimiento en bases de datos bibliográficas: la tasa de filtrado del campo autor. Una aplicación al caso de los nombres de autores españoles.	Página 220
16:15	Sistema para la indización semiautomática (SISA) de artículos de revista sobre biblioteconomía y documentación.	Página 228
16:30	Un algoritmo segmentador basado en frecuencias de letras sucesoras de palabras con el mismo significado.	Página 233
16:45	Criterios de evaluación de la interacción indización/recuperación de la información en las pasarelas temáticas.	Página 239
17:00	Indización en la recuperación de información.	Página 250
17:15	Análisis de la medida de distancia entre documentos y consultas en el modelo lógico de recuperación de información PLBR.	Página 257
17:30	Pausa.	
17:45	Confluencia de paradigmas en el resumen documental: de la misión del bibliotecario de Ortega a la Biblioteca de Babel de J. L. Borges.	Página 265
18:00	Un modelo para la organización y recuperación de la información léxica contenida en los diccionarios.	Página 273
18:15	Multilingual markup automation for better document production and retrieval.	Página 281
18:30	Asignación automática de palabras clave en tiempo real.	Página 289
18:45	Debate de Clausura	

Automatic generation of a multilingual similarity thesaurus from the Web

Martínez Santiago, Fernando, García Vega, Manuel, Martín Valdivia, Mª Teresa, Ureña López, L. Alfonso
Departamento de Informática. Universidad de Jaén
Av. Madrid, 37. 23071. Jaén. Spain
{dofer, mgarcia, maite, laurena}@ujaen.es

In this paper, we describe the construction of a multilingual similarity thesaurus by using a comparable corpus extracted from the Web. We present a tool called WebReader that is used to create such comparable corpora. Selected multilingual Web sites are given to WebReader, and it generates structured, homogeneous and low noise documents from the semi-structured, heterogeneous and noisy Web. Also, we describe a method to align the obtained multilingual documents by using clustering techniques. The aligned documents are used to create a multilingual similarity thesaurus, with English and Spanish documents from several online newspapers. Finally, we apply the multilingual similarity thesaurus to cross language information retrieval tasks. The quality of the generated corpus from the Web and the proposal alignment method is shown in the performance reached.

Palabras clave: Similarity thesaurus; Multilingual resources; Cross language information retrieval;

1. INTRODUCCIÓN

Today, a large number of linguistic resources compiled by different organisations are available, fundamentally in the form of textual corpora. However, the Web provides us with a new source of linguistic resources. Among its more outstanding characteristics are its heterogeneous character, its dynamism and availability, and the wide variety of languages used.

The shortage of linguistic resources drives us to generate our own. The Web is an immense source of documents in many languages that allow us to generate textual corpora.

Recently, some studies have been aimed at extracting collections of documents from the Web in order to use them in a wide variety of tasks. For example, the Web has been used as lexical resource, and as a source of text data for word sense disambiguation [1]. Other researchers use the Web to generate corpora for languages where electronic resources are scarce [7], [9]. Also, the Web has been used as source of parallel corpora [14]. A parallel corpus is one that contains translationally equivalent documents in several languages. Usually this consists of original documents in one language and a translation of those documents in one or more other languages[17]. Although considered a useful linguistic resource, parallel corpora are not easy to obtain, and therefore another option is the comparable corpus. A comparable corpus is formed by two or more monolingual corpora dealing with a common topic, but documents in different languages are not necessarily exact translations of others.

There are several approaches for obtaining text from Web sites. Pierre [12] develops a methodology for the categorization of Web sites. Other approaches [4] extract information from the Web. The hypothesis used is that HTML documents are semi-structured documents. Both approaches are very general and therefore limited.

Our approach allows us to make a description of the HTML structure of each Web site. Only the relevant fragments are extracted and given an appropriate format. In this work, we have used several heterogeneous sources from different international newspapers in order to generate a bilingual comparable corpus, useful by Cross-Language Information Retrieval (CLIR) tasks. A CLIR system is basically an IR system capable of operating over a cross-lingual documents collection. Thus, if a user consults a CLIR system, all relevant documents in the collection are retrieved, independently of the language used in the query and the documents. So the result of one of these systems will frequently be a heterogeneous list of documents written in English, Spanish, French, German, etc. and

ordered according to the ranking given to each document for a given query.

Once we have used the Web to create a comparable corpus, we should to apply it in order to test its quality. To this end, we have built a multilingual similarity thesaurus [13]. A multilingual similarity thesaurus provides a relation of terms in language L1 to similar terms in language L2 and allows a query formulated by the user to be transferred into the target language by substituting the query terms with some of their most similar counterparts.

The paper is organized as follows. The next section defines a similarity thesaurus. Section 3 describes the proposed tool (WebReader) and we explain the fully automatic generation of the comparable corpus. Then, we explain how to build the multilingual similarity thesaurus to test the quality of the comparable corpus generated. Finally, we present out the conclusion and further works.

2. SIMILARITY THESAURUS

A similarity thesaurus is an information structure derived from corpora and generated in automatically. The similarity thesaurus represents term similarities, which reflect domain knowledge of the collection over which the thesaurus is constructed. When we use the similarity thesaurus in a query, we expand each term in the query returning for every term a list of similar terms, ranked in decreasing order of similarity.

The similarity thesaurus was first used to expand the query to include certain knowledge in a monolingual IR system [13]. Briefly, to construct the similarity thesaurus, we must interchange the roles of documents and terms in the traditional view of an IR system. A similarity thesaurus consists of term-term similarities that are determined by how the terms of the collection are indexed by the documents. The documents serve as indexing features and the terms represent retrievable items. This process for calculating a thesaurus is fully automatic.

If we have a multilingual corpus with documents containing terms in more than one language, then we can generate a multilingual similarity thesaurus. A multilingual similarity thesaurus is a data structure that provides a list of terms in one language that are statistically similar to a head term in another language. Such a multilingual similarity thesaurus can be built using the comparable corpus generated from the Web with the WebReader tool. In addition, we must identify the pair of comparable English/Spanish documents dealing with the same topic. In other words, we need a comparable corpus aligned at document level. We have used clustering techniques to align the documents, the SLINK algorithm [15]. Then, we merge the aligned comparable documents into multilingual documents and we build a retrieval index over the multilingual document collection. The similarity thesaurus is then constructed in the usual way over the multilingual index [17].

3. CREATING A CORPUS FROM THE WEB

The Web is an immense store or resources for the extraction and acquisition of linguistic knowledge. There are millions of pages, giving us an impressive potential to create as many corpus as we need. But the Web is not exactly a corpus:

- HTML is a user-friendly language, but is not ideal for giving information about the content of the text. The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. Until the Semantic Web project becomes reality ([3]), the Web is designed for humans and the structure of the data is not evident to a robot or program browsing the Web [2].
- HTML pages are usually ill-formed. The most popular navigators will show nearly any text document. We must obtain structured text from semi-structured text.
- Even if a document is well-formed and well structured, the whole of the information contained in the document is not useful to us: elements such as images, links, announcements, site index and so on may be superfluous to our requirements.
- There are as many Web styles as Web designers. However, the text structure of the downloaded corpus must be homogeneous.

Briefly, we must extract relevant information from several sources; the Web sites we wish to explore, with heterogeneous or even ill-formed structures and non-relevant information, and generate a collection of documents with a common syntactical structure. However, the Web is so huge, varied and poorly

structured, that an ideal tool with enough ability to accomplish the task fully automatically must have so much expressive power that such a tool is computationally untractable [10]. Our proposed tool attempts to balance expressiveness and tractability in order to be useful in most cases.

3.1. WebReader: extracting information from the Web

WebReader [11] is a tool used to obtain structured text from HTML pages. This is a data conversion task that can be depicted with a Conversion Specification Document (CSD) [8]. Guidelines to obtain new documents from source documents (the Web) are specified in such a document. If the CSD is sufficiently formal, it can be processed automatically by a program. Thus, WebReader is configured with a CSD, and such a document must be written by an expert human. Because CSD is written by a human and read by a machine, an effective language to write such a document is XML [5]. Figure 1 shows a CSD used to download and convert international and national news from two Spanish online newspapers. Figure 2 shows examples of converted files, ready to be part of our corpus. Figure 3 shows a document extracted from a Web page, using the CSD of Figure 1

The WebReader configuration file or CSD has a hierarchical structure:

- WebReader level: the root node represents al the documents we wish to download.
- Site level: Every "child" node describes a Web site such as http://www.guardian.co.uk or http://www.elpais.es.
- Section level: Every "site" node has a child node for each section of the site we want process. For example, Figure 1 shows that we are interested in international and national news from www.abc.es and www.elpais.es sites. So, every international article from ww.abc.es must be accessible from www.abc.es/ABC/fijas/internac/index.asp. This type of page is called the index page, because all downloaded pages must be accessible from it.
- Link level: We do not normally need to download all the pages linked to the given index page. For example, the international index page of www.abc.es contains links to announcements, services (Web mail, archives, search engine…) and so on. An index node has several link nodes. Every link node is a regular expression. Only when a hyperlink included in an index page is matched to one of these regular expressions is the linked page downloaded and converted.

We have depicted a mechanism to download only selected documents. The downloaded documents are transformed into target documents by applying some straightforward rules (Figure 2). There are three sorts of rules:

- Append rules: add certain static text and WebReader variables to target document, such as download date, name or URL of the source page.
- Translate rules: The format rules search certain tags, such as <TITLE> or <BODY>, and put the contained text into the target document (Figure 3). Source sections not matching any translate rule are not included in the target document. In addition, the translate rule can match and access attributes of tags or obtain the value of ill-formed tags (i.e. tags not closed, as is usual with <P> or
, for example).
- Ignore rules: by default, WebReader eliminates all tags contained in source documents. Translate rules obtain the value contained in matched sections, but the section start/end tags and the rest of the tags in the section are eliminated. If we wish to keep a tag on source document, we must include an ignore rule. Ignore rules can be included to preserve basic formatting text tags such as
 or <P> (Figure 3).

The WebReader user is able to insert rules in every level of the specification document (WebReader, site, section or link). Because WebReader CSD has a hierarchical structure, rules are inherited from a level to their sublevels. For example, in Figure 1, the append rule "date" at WebReader level ensures the date appears at the end of each generated document. In the same way, for each document generated from "www.elpais.com" a *<SOURCE>* section is added containing "El Pais" text. When two rules can be applied at the same time over the same document, the most specific one is applied. The propagation of the rules is the main mechanism that WebReader provides to obtain homogeneous document style from heterogeneous document sources. The next section shows how WebReader is used to create a comparable corpus extracted from the Web. The corpus is document-level aligned and is

used to create a multilingual similarity thesaurus.

```
<!DOCTYPE webreader SYSTEM "webreader.dtd">
<webreader sgml_style="yes" sgml_root="DOC"> <!-- target corpus will be sgml structured -->
    <site name="Diario El Pais" url="http://www.elpais.es" process_links="true" process_content="false">
<section name="Seccion Internacional" url="/diario/internacional">
            <link>articulo.html</link> <!-- every hiper-link containing "articulo.html" text will be downloaded -->
            <format> <!-- specific rules for El Pais international section -->
                <append label="SECTION" value="International" />
                <append label="DOCNO" value="ELPAIS-INTERNACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
                <append label="DOCID" value="ELPAIS-INTERNACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
            </format>
        </section >
        <section name="Seccion Nacional" url="/diario/nacional">
            <link>articulo.html</link>
            <format> <!-- specific rules for El Pais national section -->
                <append label="SECTION" value="National" />
                <append label="DOCNO" value="ELPAIS-NACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
                <append label="DOCID" value="ELPAIS-NACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
            </format>
        </section>
            <format> <!-- common rules for every page from "El Pais" -->
                <translate>
                    <tag_in>TITLE</tag_in>
                    <tag_out>TITLE</tag_out>
                </translate>
                <!-- This is a more complex "translate" rule. Tag label and attr "name" value must be matched.
                    The translated value  is the value of "CONTENT" attribute -->
                <translate>
                    <tag_in closed="false" attr_name="NAME" attr_value="author" attr_get="CONTENT">META</tag_in>
                    <tag_out>AUTHOR</tag_out>
                </translate>
                <translate>
                    <tag_in>FONT</tag_in>
                    <tag_out>TEXT</tag_out>
                </translate>
                <append label="SOURCE" value="El Pais" />
            </format>
    </site>
    <site name="Diario ABC" url="http://www.abc.es" process_links="true" process_content="false">
        <section name="Seccion Internacional" url="/internacional/index.asp">
            <link>pa00</link>
            <format>
                <append label="SECTION" value="Internacional" />
                <append label="DOCNO" value="ABC-INTERNACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
                <append label="DOCID" value="ABC-INTERNACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
            </format>
        </section>
        <section name="Seccion Nacional" url="/nacional/index.asp">
            <link>pa00</link>
            <format>
                <append label="SECTION" value="Nacional" />
                <append label="DOCNO" value="ABC-NACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
                <append label="DOCID" value="ABC-NACIONAL-$YYYY$$MM$$DD$-$COUNTER$" />
            </format>
        </section>
        <format>        <!-- Rules for the whole of ABC site -->
            <translate> <!-- HTML text of ABC site has xml tags, so we use it -->
                <tag_in>titulo</tag_in>
                <tag_out>TITLE</tag_out>
            </translate>
            <translate>
            <tag_in>entradilla</tag_in>
                <tag_out>ABSTRACT</tag_out>
            </translate>
            <translate>
                <tag_in>firma</tag_in>
                <tag_out>AUTHOR</tag_out>
            </translate>
            <translate>
                <tag_in>texto</tag_in>
                <tag_out>TEXT</tag_out>
            </translate>
            <append label="SOURCE" value="ABC"/>
        </format>
    </site>
    <format>  <!-- common rules for the whole of downloaded pages -->
        <append label="LANG UAGE" value="SPANISH" />
        <append label="DATE" value="$YYYY$$MM$$DD$"/>
        <append label="LINK" value="$URL$" />
        <ignore>P</ignore>
        <ignore>BR</ignore>
        <ignore>STRONG</ignore>
    </format>
    <target append="true">
        <path>/home/dofer/projects/webreader/data</path>
        <file>spanish_news.data</file>
    </target>
</webreader>
```

Figure 1. A Conversion Specification Document for www.abc.es and www.elpais.es

Applied rule	*HTML source*	*Target document*
<translate> <tag_in> SPAN </tag_in> <tag_out>TEXT<tag_out> </translate> <ignore>P</ignore>	 This text <B>is putted into target document</B> this text is <P> eliminated </B> this is <P>not eliminated	<TEXT> This text is putted into target document this is <P>not eliminated</P> </TEXT>
<translate> <tag_in closed="false" attr_name="NAME" attr_value="TITULO" attr_get="CONTENT"> META </tag_in> <tag_out>TITLE</tag_out> </translate>	<META name="TITULO" content="Finaliza la cumbre europea"> <META name="FECHA" content="12/11/2000">	<TITLE> Finaliza la cumbre europea </TITLE>

Figure 2. Samples of application of rules

```
<DOC>
<LANGUAGE>
SPANISH
</LANGUAGE>
<DATE>
12-08-2001
</DATE>
<LINK>
http://www.ABC.es/ABC/fijas/nacional/006pa00.asp
</LINK>
<SOURCE>
ABC
</SOURCE>
<SECTION>
Nacional
</SECTION>
<DOCNO>ABC-NACIONAL-20011208-13" </DOCNO>
<DOCID>ABC-NACIONAL-20011208-13" </DOCID>
<AUTHOR> </AUTHOR>
<TITLE>
El PSOE busca apoyos en el Congreso para exigir al Reino Unido la retirada del 'Tireless'
</TITLE>
< ABSTRACT >
Crisis diplomatica con el Reino Unido por el submarino Tireless anclado en Gibraltar<br></br>
</ABSTRACT>
<TEXT>
<BR><BR>
...
El PSOE busca apoyos en el Congreso para exigir al Reino Unido la retirada del submarino </BR></BR>
<P> Zapatero afirma que la actuación del Gobierno "es un monumento a la incompetencia" </P>
<P>A.DÍEZ/P.EGURBIDE,
Madrid/Niza </P>
<BR> La revelación del presidente del Gobierno, José María Aznar, ... tiene ahora que obtener resultados". </BR> ...
</TEXT>
</DOC>
```

Figure 3. Document extracted from ABC newspaper using CSD of Figure 1

4. MULTILINGUAL SIMILARITY THESAURUS GENERATION

Several Web sites have been selected and processed with the proposal tool, WebReader, to generate a comparable corpus for natural language processing from the Web. Such a corpus is the training collection for the generation of a multilingual similarity thesaurus.

4.1. Description of the comparable corpus

A comparable corpus is required before the generation of the multilingual similarity thesaurus. The generated corpus is composed of

English and Spanish news published in several online newspaper sites since 2001. The selected Spanish sites are "El Pais", "El Mundo" and "ABC" online editions. English news has been extracted from "The Observer", "CNN" and "Washington Post" online editions. Only national and international news has been extracted. The actual size of the corpus is about 200.000 news published for 16 months, about 65 news for each site and day (Table 1). Since Spanish and English news have been published for the same time period in 2001, both corpora are alike, but not fully comparable. For each article published in a given newspaper, we cannot be sure the same article is published in the other given language. In addition, the creation of a multilingual similarity thesaurus corpus requires document-level alignment. This is a trivial operation in a parallel collection but requires some effort in comparable collections. Therefore, our comparable corpus requires one further step: the comparable documents must be identified and aligned. A measure of similarity is given by proper nouns, which are not usually translated. The documents to be aligned are newspaper articles, where proper nouns are very frequent. To improve the alignment precision, we have compared only news published on the same day.

Site	Pages
ABC	28,173
CNN	36,691
El Mundo	29,828
El Pais	31,863
The Observer	29,153
W. Post	32,683

Table 1. Sites and downloaded pages

The document similarity has been calculated by using SLINK clustering algorithm with the restrictions presented above. The comparable corpus is composed of clusters with English and Spanish documents. For each multilingual cluster, we have obtained an aligned pair of English/Spanish documents. Each aligned document is composed of all the news in the same language and contained in the same cluster.

4.2. Creating a multilingual similarity thesaurus from a comparable corpus

Once the comparable corpus is available, the construction of the multilingual similarity thesaurus is made in two ways:

- In the standard way, the term similarity is measured by using the document index. That is, the terms are indexed by the documents. The more similarly two terms are indexed by the documents, the greater the degree of similarity between them. More formally, given a term, the *tf·idf* weighting formula for a given document is interpreted as follows:

  $$w(d_i,t_k) = df(d_i,t_k) \cdot idf(d_i,t_k)$$

  where d_i is a document from the corpus, t_k is a term from the corpus and $df(d_i,t_k)$ is document frequency for the term t_k (how many times t_k occurs in d_i). This value is equal to $tf(d_i,t_k)$.

  $$idf(d_i,t_k) = \log((1+m)/(1 + |d_i|))$$

  where m is the number of different terms in the whole document collection and $|d_i|$ is the number of different terms in document d_j (the length of the document).
- We have tested a second way: the term similarity is measured by using term index. The documents are indexed by the terms. The more similarly two documents are indexed by the terms, the greater the degree of similarity between them. With this approach, *tf·idf* formula is interpreted as usual in traditional IR [16]:

  $$w(t_k,d_i) = tf(t_k, d_i) \cdot idf(t_k, d_i)$$

  the interpretation is like above, except that *idf* part is calculated as follows:

  $$idf(t_k,d_i) = \log((1+N)/(1 + |t_k|))$$

  where N is the number of documents in the whole corpus and $|t_k|$ is the number of documents which are indexed by t_k, the number of documents in the collection that contain the term (document frequency).

5. EVALUATION

In order to test the entire process, we have used the corpus corresponding to the 1994 news of Los Angeles Times (LAT). This contains more than 100,000 documents and a assembly of 40 queries in English and Spanish corresponding to relevance judgments of CLEF'2000. The set of queries must be translated from Spanish to English. The task is retrieve relevant documents from LAT for each translated query. In order to measure the effectiveness of our similarity thesaurus, we have carried out three experiments:

- Query translation is performed by using the system of automatic translation SYSTRAN[1].

- EuroWordNet is used by translating each query word for word [18]. EuroWordNet is a multilingual database with wordnets for several European languages. We have used the synonymy relation in that lexical database to translate each query.
- Every original Spanish term is expanded to similar English terms through the multilingual similarity thesaurus.

Due to the reduced size of own multilingual similarity thesaurus, certain terms could not be translated. In order to measure the effect of this factor, we have used two sets of queries. The first is formed by the 40 original queries, while the second is a subset of the 40 original queries: it is composed of only the queries that do not contain untranslated terms (Table 2). Altogether, there are 86 non-empty words in 31 different queries.

	Queries	Terms
Queries Set I	40	95
Queries Set II	31	86

Table 2. Sets of queries

A summary of the result average precision is provided in Table 3. Despite the limited size of the comparable corpus we have created, our results highlight its quality. Obviously, the performance of a multilingual similarity thesaurus depends fully on the resource used in its construction.

95 terms (Table 2) are translated by using the multilingual similarity thesaurus, so the obtained recall is about 90%. Although this recall is quite high, because certain terms have no translation, the difference in precision between the two sets of queries is 5% for the case of multilingual similarity thesaurus. Using other resources (Machine Translation – MT, Machine Readable Dictionary – MRD), the difference between both collections is negligible: 1% and 2% for MT and MRD, respectively.

	I	II
SYSTRAN	0.25	0.26
EuroWordNet	0.17	0.19
Multilingual Similarity Thesaurus I	0.11	0.16
Multilingual Similarity Thesaurus I	0.14	0.18

Table 3. Average precision

In addition, the precision obtained by similarity thesaurus is 3% below that obtained with EuroWordNet, when queries do not contain untranslated terms. These data lead is to believe that a sufficiently extensive comparable corpus would even allow us to obtain better results than an MRD.

The data shown in Table 3for the case of multilingual similarity thesaurus are obtained extending each term in a query with similar terms given by the thesaurus, specifically by the terms with a similarity coefficient greater than or equal to 0.5.

6. CONCLUSION AND FURTHER WORK

We have shown a new tool, WebReader, used to generate corpora from the Web. Once WebReader is configured, it obtains relevant and structured information from selected Web sites. We have used this tool to create an English/Spanish corpus from news Web sites. Then, we have created a multilingual similarity thesaurus. We have tested this thesaurus in CLIR tasks. In spite of the limited size of the corpus, the performance of the thesaurus is similar to that obtained with an MRD.

Our next step must be to generate larger corpora from the Web, and obtain multilingual comparable corpora with the most of the European Community languages. Once the corpora are created, document alignment is required in order to create the thesaurus. However, instead of document alignment, we think that paragraph alignment will improve performance. The current version of WebReader is based on free-context rules. Further versions will have more expressive rules, such as chain rules: some rules will only be applied where intermediate ancestor rules were applied previously. In addition, XSL-T functionality will be explored to obtain more complex target document structures.

REFERENCES

[1] AGIRRE, E. and MARTINEZ, D. (2000): "Exploring automatic word sense disambiguation with decision lists and the web". *Proceedings COLING Workshop on Semantic Annotation and Intelligent Content*

[2] BERNERS-LEE, T. (1998): "Semantic Web Road Map". Available from http://www.w3.org/DesignIssues/Semantic.html

[3] BERNERS-LEE, T., HENDLER, J. and LASSILA, O. (2001): "The Semantic Web". Available from

http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

[4] FUJII, A. and ISHIKAWA, T. (2000): "Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts". *Proceedings 38th Meeting of the ACL*

[5] GOLDBARG, C. F. and PRESCOD, P. (2000): "Why XML?". Chapter 1 in [6]

[6] GOLDBARG, C. F. and PRESCOD, P. (2000): *The XML Handbook*. Prentice-Hall

[7] GREFENSTETTE, G. and NIOCHE, J. (2000): "Estimation of English and non-English Language Use on the WWW". *Proceedings RIAO (Recherche d'Informations Assistee par Ordinateur)*

[8] GROSS, M. and LYNCH, J. (2000): "Planning for document conversion". Chapter 39 in [6]

[9] JONES, R. and GHANI, R. (2000): "Automatically building a corpus for a minority language from the web". *38th Meeting of the ACL, Proceedings of the Student Research Workshop*

[10] LEVESQUE, H.J. and BRACHMAN, R.J. (1984): *Readings in Knowledge Representation*. Brachman Levesque Editors

[11] MARTÍNEZ, F., UREÑA, A. and GARCÍA, M. (2001): "WWW como Fuente de Recursos Lingüísticos para su Uso en PLN". *Proceedings of SEPLN*

[12] PIERRE, J.M. (2001): "On the Automated Classification of Web Sites". *Computer and Information Science.* 6.

[13] QIU, Y. and FREI, H. (1993): "Concept Based Query Expansion". *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*

[14] Resnik, P.(1999): "Mining the web for bilingual text". *Proceedings 37th Meeting of ACL*

[15] ROMESBURG, H.C. (1984): *Cluster Analisys for Researchers. Lifetime Learning Publications*

[16] SALTON, G. and McGILL, M.J. (1983): *Introduction to Modern Information Retrieval.* McGraw-Hill

[17] SHERIDAN, P. and SCHÄUBLE, P. (1997): "Cross-language information retrieval in a multilingual legal domain". *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*

[18] VOSSEN, P. (1997): "EuroWordNet: A Multilingual Database for Information Retrieval". *Third Delos Wokshop Cross-Language Information Retrieval*

[1] SYSTRAN is available form http://babelfish.altavista.com