Morphoanalysis of Spanish Texts: Two Applications for Web Pages

Octavio Santana Suárez, Zenón José Hernández Figueroa, Gustavo Rodríguez Rodríguez

Departamento de Informática y Sistemas de la Universidad de Las Palmas de Gran Canaria, Edificio Departamental de Informática y Matemáticas, Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain
{osantana, zhernandez, grodriguez}@dis.ulpgc.es
http://www.gedlc.ulpgc.es

Abstract. The applications described here follow up the works performed in the recent last years by the Data Structures and Computational Linguistics Group at Las Palmas de Gran Canaria University. These works have been developed about computational Linguistics and, as one of their results, some tools for morphologic identification and generation have been released. This work presents the use of those tools as parts of new applications designed to benefit from the great linguistic information flow from Internet. Two kinds of applications are identified, both according to the interactive grade of the linguistics studies to be done, and two prototypes, named DAWeb and NAWeb, are developed with special attention to their architecture in order to maximize the efficiency of both. Analysis modes include: neologism detection, word use (qualitative and quantitative measurements) and some syntax aspects like lexical collocations or prepositional regimes.

1 Introduction

This work describes two computer applications developed by the Data Structures and Computational Linguistics Group at Las Palmas de Gran Canaria University. These applications follow up the previous works about computational linguistics performed by the Group in the last few years. As a result of those works some tools for morphological identification and generation of Spanish words have been released; several of these tools can be used on-line from http://www.gedlc.ulpgc.es. This work intends to apply those tools as part of new developments designed to get at the big flow of linguistic information of Internet documents. The results can be named as "tools for morphological analysis of web pages.

The study of the use of the language can be done in two ways: intensive, of a single or few documents, in order to identify concrete characteristics, or extensive, of a great mass of documents in order to obtain patterns frequently used. So two applications were developed, named DAWeb (Web Downloader and Analyzer, in

Spanish “Descargador y Analizador de Web”) and NAWeb (Web Browser and Analyzer, in Spanish “Navegador y Analizador de Web”).

2 DAWeb

DAWeb is oriented to the massive analysis of documents from one or various websites. Formally, it can be compared to a downloader program; with the difference that it does not save the retrieved documents locally, but only the results of the performed analysis. Internally, DAWeb is organized into three modules: Configuration Module, Document Retrieval Module and Analysis Module. The Configuration Module serves to define an “Analysis project”, specifying the set of URLs to explore, criteria to discard URLs, morphological analysis options, number of threads to use and, optionally, times to running the project.

2.1 Document Retrieval Module

The Document Retrieval Module consists of one Distribution Module and a variable number of Retrieval Modules interacting with Internet. The Distribution Module coordinates the work of the Retrieval Modules and prepares the results obtained by them to be used by the Document Analysis Module. It takes URLs from the “pending URLs list” and assigns them to idle Retrieval Modules, if any. Each Retrieval Module tries then to get the document tied to the URL assigned to it. When all Retrieval Modules are busy or all pending URLs are assigned, the Distribution Module waits for results. When a Retrieval Module gets a document, the Distribution Module adds it to the “retrieved document queue” to be processed by the Analysis Module, extracts the URLs contained in the document and adds them to the “pending URLs list”, but only when they are not there, they were not there previously and they do not match the critera to discard URLs. If a Retrieval Module reports a fail, the Distribution Module analyzes the problem and decides whether to repeat getting the document later or discard the URL, annotating the case in the “log list”. The work of the Document Retrieval Module ends when the “pending URLs list” is empty and all Retrieval Modules are idle.

Each Retrieval Module is a HTTP (HiperText Transfer Protocol) component running on an independent thread. The number of active Retrieval Modules can be fixed in the range one to ten (default number is five). This configurability produces a great adaptability to changing circumstances of the net.

2.2 Document Analysis Module

The first step when analysing a web page is to extract the text contained in it. Text extraction task is performed by a one pass syntactic parser that consists of an automaton driven by the characters sequence in the document. The automaton

searches the text for the HTML (HiperText Mark-up Language) tags to discard or change them in order to obtain the text of the document free of marks but with a minimum structural information.

The core of the Document Analysis Module is a Morphological Recognition Tool that takes a word and returns a list of possible base forms, inflections, and grammatical values for it. The Morphological Recognition Tool was originally designed to identify single words. When the target is the analysis of texts this ratio can be improved taking advantage of the fact that a normal text is mainly composed using a small set of words that are repeated many times. Consequently, a special Morphological Recognition Improvement Tool using an ultra speed hashing structure to save information about each word when it first appears during the document analysis process has been developed for this work.

The results of performed analysis are structured into a hierarchy derived from the pair domain-page and organized by the different kind of programmed analysis. This hierarchical structure is saved to disk in a suitable format to be easily studied later.

3 NAWeb

The major differences between the architectures of DAWeb and NAWeb are the substitution of the Document Retrieval module by a TWebBrowser component that encapsulates the features of Microsoft Internet Explorer, and the inclusion of modules to classify and show the results in an interactive way.

Due to its interactive orientation, NAWeb looks like a typical web browser with additional features. Its main window is divided into three zones: the menu and toolbars zone, up, the views & annotation zone, in the middle, and the results zone, in the bottom. Both the views & annotations and the results zones are organized by means of multiple pages to allow a better analysis.

There are six pages in the views & annotations zone. All but the last one show a different view of the document (web page) retrieved from Internet. The first page shows the web page as it can be seen on any web browser. The second page shows the "pure" text of the document, free of HTML tags. The third page shows the HTML code for the web page. The fourth page shows the "lemma" view of the text, a list view of all the words in the text joined with the results of its morphoanalysis identification. the fifth page shows the information header of the document. The sixth page is an edition area where the user can do annotations and transfers information about the document.

The results zone shows results of analysis arranged by different criteria, each in a different page. As in the views & annotations zone there are six pages. The first page shows the words in the document distributed into six lists by their grammatical category (verbs, nouns, adjectives…). The second page shows the base forms of the word sorted by frequency, direct and invert alphabetical order and length. The third page shows the words in the document sorted by the same criteria and for distance respect to the selected one; two possible metrics of distance can be chosen: Levenshtein Distance (DL) or Largest Common Subsequence Distance (SCML). The

fourth page shows, on the left, general information about the words in the document such as repetition grade and global category distribution; on the right, the outline view of the document is showed, this is a graphic representing the ratio of new words introduction. the fifth page shows complex and useful measurements of the spatial distribution/concentration of the words in the document and permits to compare similar patterns of distribution. The sixth page permits to locate repeated sequences of words by length and minimum frequency of apparition; this option is very useful for coocurrence studies.

4 Conclusions

The "Year 2000 annals" of the series "Spanish in the world" published by the "Instituto Cervantes" [8] points out the necessity of tools for intelligent information retrieval and analysis in order to increase the presence of the Spanish language on Internet. On the other hand, linguistic researchers need tools suitable to do larger and more accurate studies of the language and look at the web as a big source of documents for research. This work presents two tools that intend to fit these requirements: they are useful for linguistic analysis of information from the web, they can be used in the future as part of intelligent information retrieval systems adapted for Spanish documents, and further, they can serve other objectives as, for example, assistant tools for students of Spanish.

References

1. Santana, O., Hernández, Z. J., Rodríguez, G.: Conjugaciones Verbales. Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, Nº 13, (feb. 1993) 443–450
2. Rodríguez, G., Hernández, Z., Santana, O.: Agrupaciones de Tiempos Verbales en un Texto. Anales de las II Jornadas de Sistemas Informáticos y de Computación, Quito, Ecuador, (Apr. 1993) 132-137
3. Santana, O., Hernández, Z. J., Rodríguez, G.: Reconocedor de Conjugación en Formas Verbales que Trata Pronombres Enclíticos, Lingüística Española Actual, Ed. Arco-Libros, Nº 16, (1994) 125-133
4. Alameda, J. R., Cuetos, F.: Diccionario de Frecuencias de las Unidades Lingüísticas del Español, Servicio de Publicaciones de la Universidad de Oviedo, (1995)
5. Santana, O., Pérez, J., Hernández, Z., Carreras, F., Rodríguez, G.: FLAVER: Flexionador y Lematizador Automático de Formas Verbales, Lingüística Española Actual, Ed. Arco-Libros, Nº XIX, 2, (1997) 229-282
6. Santana, O., Pérez, J., Carreras, F., Duque, J., Rodríguez, G.: FLANOM: Flexionador y Lematizador Automático de Formas Nominales, Lingüística Española Actual, Ed. Arco-Libros, Nº XXI, 2, (1999) 253-297
7. Millán, J. A.: Estaciones Filológicas, Filología e Informática, Seminario de Filología e Informática de la Universidad Autónoma de Barcelona, (1999) 143-164
8. Anuario 2000 del Español en el Mundo, Centro Virtual Cervantes, http://cvc.cervantes.es/obref/anuario/anuario_00