propor 2008

International Conference on Computational Processing of Portuguese Language

Applications of Portuguese Speech and Language Technologies

Applications of Portuguese Speech and Language Technologies - Propor 2008 Special Session

Hosted by:

Universidade de Aveiro

Promoted by:

Microsoft Language Development Center

Propor 2008 Special Session Commitee

Special Session Chair

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Organising Committee

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Programme Committee

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal
- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Vera Strube de Lima,** Pontifícia Universidade Católica do Rio Grande do Sul, Brasil
- **Luís Caldas de Oliveira,** INESC-ID/IST, Portugal

Editorial Board

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **Luanda Braga Batista,** Microsoft Language Development Center, Portugal

SUPeRB: Building bibliographic resources on the computational processing of Portuguese

Luís Miguel Cabral, Diana Santos, Luís Fernando Costa

Linguateca, Oslo node, SINTEF ICT, Norway
{Luis.M.Cabral, Diana.Santos, Luis.Costa}@sintef.no

Abstract

SUPeRB is a digital library helper that aims at updating and maintaining specific publication repositories, and assisting in the publishing of publication records, for institutions and individual actors. It gathers bibliographic data from Web pages and documents and integrates that data into a local repository of bibliographic data on a specific domain. By collecting information from these resources, SUPeRB also assists in building a bibliographic database with the specific domain intervenients such as authors, conferences and scientific journals. The computational processing of the Portuguese language has been the considered domain .

1 .Introduction

Since 1999, Linguateca has been offering a portal about the computational processing of Portuguese aiming at a reasonable complete overview of the field. Linguateca's goal is to provide a place that helps researchers and developers not to start from scratch and keep them informed of the work of their peers.

One of the resources we maintain is a publication catalogue surveying published work in this field. From 1999 to 2003, we manually gathered approximately 750 items, including, if available, their electronic version.

Although our team routinely screens mailing lists and lists of accepted papers in calls for participation for relevant conferences, it is hard to maintain this catalogue updated. It is especially troublesome to find accurate and complete information about papers and other works, since researchers often fail to keep their publications pages up to date. Furthermore, it is frequent to find barriers that difficult processing the information, such as:

- Incomplete citing by omitting the conferences' full names, the volume editors, conference edition or place of conference;
- Several bibliographic styles employ author's initials, making it hard to identify them;
- Electronic version is not exactly the same as the published one (at least in what formatting is concerned).

It should be added that virtually none of the authors we survey in our catalogue uses meta-data or any kind of categorization of their own works. Usually, their publications list is a web page presenting only their textual references, in some cases, without links to the electronic versions.

This lack of data can make it difficult to decide, only by the title, whether or not to include the item as relevant. Furthermore, users are rarely motivated enough to help us catalogue more publications by suggesting their own publications or others that they could find relevant.

In any case, with the overwhelming increase of information on the Web it is consensual that one needs digital methods to help to organize and make useful the distinct information.

We have therefore tried to address the need for an automated helper to support searches and to obtain bibliographic data from Web documents, as well as evaluating their relevance for our catalogue and organize it accordingly. Our goal was not to provide a fully automated system, but rather deploy a supervised approach to help humans obtain better results in the task of aiding an expert to create a meaningful and coherent publication list, and help maintain it with contributions from the particular community of interest. Our goal is thus similar to the one of Feitelson [**Error! Reference source not found.**], and not in any way an attempt to replace or compete with CiteSeer [**Error! Reference source not found.**]. SUPeRB aims at providing the publication catalogue with organized data, which can later be updated and allows also better means of accessing that bibliographic data.

2.SUPeRB, a (digital) library helper

SUPeRB, as described in detail in [**Error! Reference source not found.**], is a semi-automatic system whose purpose is to help searching and processing bibliographic references from the Web, with a specific contextual bias, as well as aid an expert to construct and maintain bibliographic meta-data collections from information given by several users.

SUPeRB is intended to serve as a tool which provides means to gather information from online data and to insert and validate this bibliographic data into a publication catalogue. The data is supplied by a user in several possible methods:

- a textual reference;
- a set of keywords or expressions that are to be used to find web pages with relevant bibliographic content;
- a URL that contains one or more relevant bibliographic references.

For example, a user can provide an author's name and a title (complete or partial) and, in this case, SUPeRB's task is to retrieve the complete bibliographic reference using online resources and present it in a format that can be handled by the publications catalogue, together with links for the online documents if possible.

In fact, SUPeRB was designed for two kinds of users, that interact with SUPeRB through a Web interface:

1. **Repository users**, who may use SUPeRB, searching and classifying references according to their interests and knowledge;
2. **Repository managers**, who ultimately decide what is to be kept in the repository by validating the **repository users** actions within the publications catalogue.

In order for these steps to take place, SUPeRB was conceived as a set of stand-alone modules, each addressing a specific task. This modular structure allows the modules to interact together or to work independently, allowing each to be implemented on its own in third parties applications. We have therefore set to handle each of the following tasks:

- Compose keyword-based query searches on the Web, that retrieve related content, focusing on the bibliographic domain;
- Extract text from different document formats;
- Extract bibliographic references from text;
- Decompose a text bibliographic reference into its bibliographic elements (title, author, place of publication, journal title, etc.) paying special attention to the Brazilian and Portuguese bibliographic cases [**Error! Reference source not found.**, **Error! Reference source not found.**];
- Convert references among different bibliographic styles and formats such as BibTeX or EndNote export formats;
- Compare and merge apparently different references, improving bibliographic related ontologies concerning publishers, conferences, authors, places of publication and so on;
- Provide means for storing the gathered information as well as provide further information (such as tagging).

Figure 1 describes the architecture of SUPeRB, composed of several modules which we will describe in some detail in what follows.

2.1 The WebSearch module

This module uses words or expressions given by the user, in a way similar to [**Error! Reference source not found.**], to generate queries that are then used in Web services such as Google's and Yahoo's search APIs. The result is a set of URL candidates that should be relevant to the expressions given. The data provided by the user can also be structured data, where bibliographic attributes such as the author's name, title, or year can be specified, allowing to generate several combined queries that contain the most relevant attributes. Furthermore, the produced queries are complemented with specific keywords. These keywords are indicative of bibliographic content. Some examples might be "article", "reference" or "publications". Portuguese words are also used anticipating articles written in Portuguese.

As it is not feasible to process all the obtained links, which are a considerable amount since several queries are made to external APIs, the results are ordered according to the returned ranking and the number of ocurrences of each link (and occurences of its base path) in the different query results.

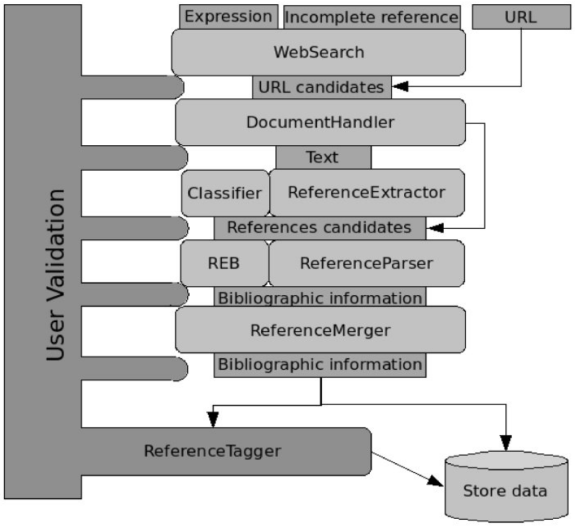

Figure 6. **SUPeRB architecture**

2.2 The DocumentHandler module

This module receives a document's file name or a URL as input and returns the document in plain text. This module extracts the text content of several document types (HTML, Microsoft Word and PowerPoint, Open Office Word and Presentation, Adobe PDF, PostScript, Rich Text Format), converting it into a common format, plain text, because this was a transformation feasible for all these document types. Apart from the HTML conversion which is done using a built-in method, all other document conversions are done using third party programs, easily available in Linux. The module can also be easily configured to use other programs that extend the format types handled.

2.3 The ReferenceExtractor module

This module receives a document in plain text and returns a list of textual bibliographic references. For this task, it relies on heuristics inferred by another module, the *DocumentClassifier*, which gathers information on the textual data that indicates the structure of the document, trying to match its structure to classes like: academic written work, a list of references, a presentation or even irrelevant formats such as a blogs, relying on the occurrence of specific words, the size of document, the distribution of text (words per line, lines per page).

Upon identifying the structure type of the document, the most likely blocks of text are analised to retrieve references. This may include:

1. Process the beginning of the text to obtain the authors, title, abstract and other bibliographic data from an academic work;
2. Process the document's ending block to retrieve references in cases of an academic work;
3. Process the entire text to obtain a list of references.

As for separating lists of references it is necessary to determine where that block starts, usually looking for specific expressions such as "References", or looking for the enumeration of references. This process is made harder by the limitations of the text extraction application and methods used in the *DocumentHandler* module, which may not convert the newlines properly, spread a reference through several lines or place two references (or the ending part of one and the start of another) in the same line. A battery of heuristic tests is therefore applied, looking for the reference enumeration, line breaks and looking to specific bibliographic style marks (such as starting with authors, authors initials, starting or ending with the year, enclosed title in quotes, ending with page numbers). All gathered candidates are later analysed, assuring that they are not too small or too big, and their position in the text is compared to assure that no overlap has occurred.

2.4 The ReferenceParser module

This module parses individual textual references, returning its bibliographic elements, properly separated and identified. The output format is consistent with our publications catalogue's storing format, described in [**Error! Reference source not found.**, **Error! Reference source not found.**]. This module uses several methods, such as a Perl package described by Jewell in [**Error! Reference source not found.**], together with heuristics for tokenizing the text reference and a gazetteer-inspired module called *REB* (Portuguese acronym for *Repositório de Elementos Bibliográficos*, which means *Bibliographic Elements Repository*). The *REB* module has two parts:

- An ontology of authors, editors, places and conferences, gathered from our publications catalogue and from the newly introduced data. Currently *REB* contains about 2000 authors and editors, 550 conferences names and short conference names, 185 publishers and 132 locations. It contains also relations between different names for the same entity (identifying abbreviations such as *Bento C. Dias da Silva* with *Bento Carlos Dias da Silva* and *Dias da Silva, Bento*, translations or even misspellings).
- A set of programs that provide validation and update methods.

2.5 The ReferenceMerger module

This module aims at identifying duplicates but also to add missing fields in one reference which are present in others. It compares two or more references and searches for identical fields in both. As this module handles formatted references, it processes individually each field, author, title, year, and so on. It requires several specific combinations of fields to be identical on both references to consider them as duplicates. Also, if one of the references evaluated is already included in the publications catalogue, it allows the updating of stored references with the new information.

2.6 Other modules and design options

There are other modules that provide useful methods to perform several tasks that improve the user interaction with the catalogue data.

1. The *ReferenceTagger* provides an interface that allows users to tag stored references with keywords, providing important information that can be used in searchs and presentation of results by allowing grouping of related references;
2. The *ReferenceConverter* is a module that provides conversion methods between several known formats that include the internal format used in our publications catalogue, BibTeX, RIS, EndNote and Refer. This makes it easier for users to suggest new bibliographic data.

The temporal axis of maintenance was contemplated in our design from the start: often, several relevant pieces of information such as page numbers, when an article appears finally in print, or the URL (when the publisher allows public release on the Web) are missing when a publication is first registered. It is also possible that papers are republished, and then cross-links should be added. URLs keep changing and an automatic helper should take care of that, periodically checking the availability and correctness of the information. We have thus catered for periodic (or scheduled) updates by SUPeRB, as a particularly relevant feature of automated help.

Also, as each module provides results, these intermediate data can be supervised by the user before feeding it to the following modules, allowing a human filtering of the processed data. This allows a better control of the data being processed, allowing one to remove or edit irrelevant data that would reduce the efficiency and quality of the results of the tasks performed by the remaining modules.

Finally, we have been very careful to make available multilingual capabilities to allow citing of the very same publications in a Portuguese, English or other language context, which implies the need for keeping different names/alias for different locations, publishers and even dates. Currently, there is full support for Portuguese and English and we are considering extending it to other languages.

3.Partial component evaluation of SUPeRB

We have previously evaluated the *ReferenceExtractor* module, using a methodology inspired by the HAREM evaluation setup [**Error! Reference source not found.**, **Error! Reference source not found.**]. In the present paper we choose to evaluate the *ReferenceParser* component, one of SUPeRB's core modules. Given a set of right elements and the set of elements suggested by SUPeRB, we count the following cases:

- #*c* - Number of elements correctly classified and delimited;
- #*w* - Number of elements correctly delimited, but incorrectly classified;
- #*i* - Number of elements correctly classified but the data is incomplete;
- #*e* - Number of elements with excess of information, correctly classified considering only part of the data;
- #*m* - Number of elements with which were was not returned nor identified;
- #*tf* - Total number of elements found;
- #*te* - Total number of elements retrieved.

We then compute the following measures:

$$Precision = \frac{\#c}{\#tf} \tag{1}$$

$$Recall = \frac{\#c}{\#te} \tag{2}$$

$$Loose - Precision = \frac{\#c + \#e}{\#tf} \tag{3}$$

$$Loose - Recall = \frac{\#c + \#e}{\#te} \tag{4}$$

$$Under - generation = \frac{\#i + \#m}{\#te} \tag{5}$$

$$Over - generation = \frac{\#w}{\#tf} \tag{6}$$

We used 33 real bibliographic references manually extracted from 33 different homepages of researchers who have either recently sent messages to the Corpora List and/or are researchers listed as actors in the computational processing of Portuguese.

Table 1. Evaluation of the *ReferenceParser* module

	Precision	Recall	F Measure	L-Precision	L-Recall	Under-Gen.	Over-Gen.
author	0.72	0.40	0.26	1.00	0.56	0.44	0.00
year	0.41	0.50	0.23	0.80	0.97	0.03	0.21
title	0.39	0.57	0.23	0.50	0.73	0.27	0.43
conference	0.36	0.44	0.20	0.45	0.56	0.44	0.39
location	0.75	0.40	0.26	0.75	0.40	0.60	0.00
pages	0.83	0.77	0.40	0.92	0.85	0.15	0.08
volume	1.00	0.33	0.25	1.00	0.33	0.67	0.00
institution	0.33	0.40	0.18	0.50	0.60	0.40	0.50
Total avg.	0.60	0.427	0.25	0.74	0.62	0.38	0.20

Globally, out of 239 expected elements, 102 were correctly identified, 47 were incorrectly identified, 53 where either incomplet or exceding elements (14+39) and 84 were missing as displayed in Table 1 (We have ignored distinctions between some fields, such as authors and editors, conference title and conference short title or location and address).

This study showed not only the global results but also the analysis of several elements in particular. Closer analysis showed that:

- Detection of authors has a good precision (for Portuguese names) with a few exceptions with non-Portuguese names;
- Numerical elements are handled rather well, with the exception of the year;
- The REB module can produce noise for some types, leading to over-generation;
- Overlap of data in REB can occur (universities acting as publishers, authors are also editors) and therefore we need to improve the contextual analysis of the reference context to single out these properties.

4.Concluding remarks

From a concrete problem in the daily life of our project, which we set out to solve using our own tools and resources for the computational processing of Portuguese, we arrived at a more general system that we hope can help researchers or expert librarians in their work with references in other specific areas.

In fact, most citations and ranking, even Portuguese and Brazilian, are done on ”international” publication, which means English. There was, therefore, very little going on on this subject in and about Portuguese, notwithstanding the fact that there are also international publications written in Portuguese.

Especially, we were not able to find any system developed specifically to deal with Portuguese references. Our system was thus geared towards publication of Portuguese native speakers – in Portuguese, English, or other languages.

In addition to supporting the management of a medium-sized publication catalogue (ca. 2080 publications and 120 conferences, books or journals), SUPeRB modules are publicly available as open-source software, to be used in other projects dealing with references in the Portuguese-speaking world, being distributed as Perl modules. For more information, visit `http://www.linguateca.pt/SUPeRB`.

As future work, we intend to improve SUPeRB with keywords and abstracts. This will allow the use of subject ontologies and allow better methods of text searches. We note that references are a kind of semi-structured text, which has been rather neglected in Portuguese but constitutes an important area of (scientific) information extraction [**Error! Reference source not found.**].

Acknowledgements This work was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

References

1. Associação Brasileira das Normas Técnicas. *NBR 6023: Norma Brasileira*, August 2002.
2. Marco Baroni and Silvia Bernardini. BootCat: Bootstrapping corpora and terms from the web. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation*, pages 1313–1316. ELDA, 2004.
3. Luís Miguel Cabral. Documentação online do SUPeRB. July 2007. Last updated 5 August 2008 http://adamastor.linguateca.pt/super/help.html.
4. Luís Miguel Cabral. SUPeRB - Sistema Uniformizado de Pesquisa de Referências Bibliográficas. Master's thesis, Faculdade de Engenharia da Universidade do Porto, March 2007.
5. Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks. Learning to Harvest Information for the Semantic Web. In Christoph Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *1st European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, Proceedings*, Lecture Notes in Computer Science, pages 312–326. Springer, 2004.
6. Dror G. Feitelson. Cooperative indexing, classification and evaluation in bow. In Opher Etzion and Peter Scheuermann, editors, *CooplS '02: Proceedings of the 7th International Conference on Cooperative Information Systems*, pages 66–77, London, UK, 2000. Springer-Verlag.
7. Instituto Português da Qualidade. *NP 405-2: Norma Portuguesa: Documentos electrónicos*, 2003.
8. M. Jewell. ParaTools Reference Parsing Toolkit-Version 1.0 Released. *D-Lib Magazine*, 9(2), 2003.
9. Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer Society Press*, 32(6):67–71, 1999.
10. Paulo Alexandre Rocha. Gestão das Páginas do projecto Processamento Computacional do Português. Technical report, 19 November 2001.
11. Diana Santos and Nuno Cardoso, editors. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, November 2007.

12. Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik, and Daniel Tapias, editors, *Proceedings of LREC2006, the 5th International Conference on Language Resources and Evaluation*, pages 1986–1991. ELD