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 resource, 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 ”simple” 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 [1], and not in any way an attempt to replace or compete with CiteSeer [2]. SUPeRB aims at proving the publication catalogue with organized data, which can later be updated and allows also better means accessing that bibliographic data.

2 SUPeRB, a (digital) library helper

SUPeRB is thus 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 provide 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 a author’s name and a title (complete or partial) and, in this case, SUPeRB’s objective is to retrieve the complete bibliographic reference, using online resources and present it in a format that can be handled by the publications catalogue, and 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. And **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. It also allows the user to intereact with the results in each task, supervising intermediate results in a way that they can edit incorrect data or simple remove irrelevant information. Finally, we have been very careful in making available the multilingual capabilities of the whole system, 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.

We have therefore structure the following modules:

WebSearch Woking in a similar way to [3], this module is capable of generating keyword-based query searches in the Web, using services such as as Google's and Yahoo's search APIs, that retrieve related content, focusing in the bibliographic domain.

DocumentHandler Is responsible for extracting text from different document formats(HTML, Microsoft Word and PowerPoint, Open Office Word and Presentation, Adobe PDF, PostScript, Rich Text Format) and converting it to text format, wich can later be manipulated.

ReferenceExtractor This module extracts bibliographic references from text content. It uses a secondary module, **DocumentClassifier**, to match the its structure to an academic written work, a list of references or even a presentation. Upon determining the document strucuture, it uses the best suited set of heuristics on parts of the document that are most likelly to contain bibliographic content.

ReferenceParser This module takes each textual reference and in return gives its bibliographic elements (title, author, place of publication, journal title, etc.) , taking special attention to the Brazilian and Portuguese bibliographic cases [4, 5]. It uses a Perl package described by Jewell in [6], together with heuristics for tokenizing the text reference and a gazetteer-inspired module called *REB* (Portuguese acronym for Bibliographic Elements Repository). The output format is a structured format described in [7, 8].

ReferenceMerger This module's goal is to provide a mean to complete references by identifying possible duplicates, not 100% identical referencesm and combining their elements in a single reference.

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

ReferenceTagger provides an interface that allows users to tag stored references with keywords, providing important information that can be used in searches and presentation of results by allowing grouping of related references;

ReferenceConverter is a module that provides conversion methods between several know formats that include the internal format used in our publications catalogue, BibTeX, RIS, EndNote and Refer. This facilitates users to suggest new bibliographic data as well as provide the methods to export the stored data;

REB Mentioned before, is a database of authors,, conferences, editors, collected from the catalogue and from the newly introduced data. It contains about 2000 authors and editors, 550 conferences names and 185 editors. It contains relations of equivalence, identifing names variations used by the same author, or even misspellings. It allows not only to validate data automatically (matching authors and conferences) but it also can be used in the users interface, providing autocomplete features for names.

Also, 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. We have thus catered for periodic (or scheduled) updates by SUPeRB, as a particularly relevant feature of automated help.

3 Partial component evaluation of SUPeRB

We have previously evaluated the *ReferenceExtractor* module, using a methodology inspired by the HAREM evaluation setup [9, 10], as follows:

Here we choose to evaluate the *ReferenceParser* component, one of SUPeRB's core modules. Table 1 presents the result of this evaluation. We used 33 real bibliographic references manually extracted from 33 different homepages of researchers who recently sent messages to the Corpora List and researchers 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

We have ignored distinctions between some fields, such as authors and editors (trying to assess if it identified correctly as person's name), conference title and conference short title or location and address (place of conference, publishing place). Globally, out of 239 expected elements, 102 were correctly identified, 47 were incorrectly identified and 84 were missing.

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

- Detection of authors has a good precision (for Portuguese names) with a few exceptions with non-Portuguese names;
- Numeric elements are handled rather well, with the exception of the year;
- The REB module can produce noise for some types, leading to over-generation;
- On the other hand REB still lacks knowledge when it regards other types such as publisher;
- 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 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, not withstanding 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.

As future work, we intend to improve SUPeRB with keywords and abstracts for increasing its power regarding subject ontologies and at least some kinds of running text. 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 [11].

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. Feitelson, D.G.: Cooperative indexing, classification and evaluation in bow. In: Proceedings of the 7th International Conference on Cooperative Information Systems. (2000)
2. Lawrence, S., Giles, C.L., Bollacker, K.: Digital libraries and autonomous citation indexing. IEEE Computer Society Press **32**(6) (1999)
3. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R., Silva, R., eds.: Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation(LREC’2004, Lisboa, Portugal). (26-28 May 2004)
4. Associação Brasileira das Normas Técnicas: NBR 6023: Norma Brasileira. (Agosto 2002)
5. Instituto português da Qualidade: NP 405-2: Norma Portuguesa: Documentos electrónicos. (2003)
6. Jewell, M.: ParaTools Reference Parsing Toolkit-Version 1.0 Released. D-Lib Magazine **9**(2) (2003)
7. Rocha, P.A.: Gestão das Páginas do projecto: Processamento Computacional do Português. Technical report, Departamento de Informatica, Universidade de Braga (10 November 2001)
8. Cabral, L.M.: Documentação online do SUPeRB. http://adamastor.linguateca.pt/super/docs/help.html (July 2007) Last updated 26 March 2008.
9. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., Tapias, D., eds.: Proceedings of the 5th International Conference on Language Resources and Evaluation(LREC’2006, Genoa, Italy). (22-28 May 2006) 1986–1991
10. Santos, D., Cardoso, N., eds.: Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, Linguateca (12 Novembro 2007)
11. Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y.: Learning to harvest information for the semantic web. In: 1st European Semantic Web Symposium (ESWS-2004, Heraklion, Greece). (May 2004)