Using openWordnet-PT to improve VIVO

Alexandre Rademaker[1], Daniela Brauner[2], Glauco Munsberg[2] and André Peil[2]

[1]IBM Research and FGV/EMAp
[2]UFPel

VIVO is an open source semantic web application for research discovery. The power of VIVO relies mainly on the VIVO-ISF ontology and its expressivity to represent all information about researchers and the research domain. VIVO-ISF makes all those types of information interconnected and browsable in the VIVO application. Nevertheless, although VIVO has good support for faceted search across disciplines, it is still not anything more than a keyword-based search engine. That is; it is still not using the power of semantics for information retrieving.

WordNet [5] is an extremely valuable resource for research in Computational Linguistics and Natural Language Processing in general. WordNet has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, and dozens of other knowledge intensive projects. WordNets model the semantic relationships between words and they are crucial in all those applications because computational systems are not aware of the fact that *salário* (salary) and *contra-cheque* (paycheck) both refer to salary, or that *benefício* (benefit) is also related to these words as a common more general hypernym. Wordnets have been distributed in a wide range of different incompatible data formats. An increasingly popular way of addressing the issue of interoperability is to rely on Linked Data and Semantic Web standards such as RDF [2] and OWL [7], which have led to the emergence of a number of Linked Data projects for lexical resources [3, 1].

OpenWordnet-PT [1] is a lexical-semantic resource describing (Brazilian) Portuguese words and their relationships. It is modelled after and fully interoperable with the original Princeton WordNet for English [5], relying

[1] `http://logics.emap.fgv.br/wn/`

on the same identifiers as WordNet 3.0. This means that one can easily find Portuguese equivalents for specific English word senses and conversely. This also means that OpenWordnet-PT is part of a large ecosystem of compatible resources, including domain identifiers [8] and mappings to Wikipedia [4]. OpenWordnet-PT is encoded and distributed in RDF/OWL. Not only do these standards allow us to publish both the data model and the actual data in the same format. They also provide for instant compatibility with a vast range of existing data processing tools [6], including databases (so-called "triple stores") and semantic applications like VIVO.

In Brazil, typically implementations of VIVO will require the institutions to leverage the data available in the Lattes Platform and their Digital Libraries. The Lattes Platform [2] is an information system maintained by the National Council for Scientific and Technological Development (CNPq), which contains Brazilian researchers resumes and research groups information. On the other hand, institution's repositories (Digital Libraries) usually have data about thesis and technical reports. Those data are complementary to the researcher's resumes, as the resumes provide more information about the researchers, publications, and their participation in research projects. Given that, institutions have to collect and consolidate the data from researchers resumes and from their digital repositories for ingest into VIVO application. The collected data have duplicated resources that we can't identify as the same real entity because of the lack of identifiers and links between both datasets. To solve it, we are exploiting disambiguation techniques giving weights to VIVO-ISF properties. Once the data is consolidated, we believe that openWordnet-PT can empower VIVO providing two usable services: concepts disambiguation and query expansion.

Concepts are everywhere in this domain. VIVO use them to classify and cluster publications, people or other research activities. They can represent researchers interest, department areas of concentration, publication subject etc. Given that, it is important to have a perfectly consolidated list of concepts to help on data interconnection. In the Lattes Platform, a researcher relies on a semi-controlled vocabulary of concepts provided by CNPq. That is, a researcher may add free terms into the hierarchy of concepts provided by CNPq. Concepts that came from digital libraries records are also not free from some messy. Although librarians usually curate the catalog, we have already found some inconsistency in the use of controlled vocabularies and keywords to classify documents. We propose a novel use of openWordnet-PT to consolidate the set of concepts originated from different sources before

[2] `http://lattes.cnpq.br`

ingest data into VIVO. The synsets of openWordnet-PT play the role of canonical concepts in our approach.

Query expansion for improving information retrieving is another issue that we want to exploit with the use of Wordnets. Given a search term "robótica" (robotics), instead of using only the traditional information retrieval techniques based on the given keyword, we want to automatically expand the query using the term's hyponym (animatronics) and hypernym (artificial intelligence) extracted from the openWordnet-PT to get more relevant results to increase user experience. The results, when presented to the user, should carry a relevance indicator, to get the users feedback about the relevance of the terms used.

In this paper, we want to report our ideas in the use of openWordnet-PT's relations (hyperym and hyponym) and VIVO-ISF ontology to improve user experience in VIVO. We believe that this experiment is an important movement into making VIVO a more powerfull semantic search engine.

References

[1] Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. *Linked data in linguistics: Representing and connecting language data and language metadata*. Springer, 2012.

[2] Richard Cyganiak and David Wood. RDF 1.1 concepts and abstract syntax. Technical Report Draft 23 July 2013, W3C, 2003.

[3] Gerard de Melo and Gerhard Weikum. Language as a foundation of the Semantic Web. In *Proc. of ISWC 2008*, volume 401, 2008.

[4] Gerard de Melo and Gerhard Weikum. MENTA: inducing multilingual taxonomies from Wikipedia. In *Proc. of CIKM 2010*, pages 1099–1108. ACM, 2010.

[5] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT press, 1998.

[6] Steve Harris and Andy Seaborne. SPARQL 1.1 query language. Technical Report W3C Recommendation 21 March 2013, W3C, 2013.

[7] Pascal Hitzler, Markus Krotzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. OWL 2 web ontology language primer. Technical Report W3C Rec 11 Dec 2012, W3C, 2012.

[8] Bernardo Magnini and Gabriela Cavaglia. Integrating subject field codes into WordNet. In *Proc. of LREC*, pages 1413–1418, 2000.