Introducing OpenWordnet-PT: a open Portuguese wordnet for reasoning

Alexandre Rademaker* Valeria de Paiva† Livy Real‡
Fabricio Chalub§ Claudia Freitas¶

July 2016

Semantic relations are a key aspect when developing computer programs capable of handling language – they establish (labeled) associations between words and can be integrated into lexical-semantic knowledge bases. Available since the beginning of the 1990s, Princeton's WordNet [8], henceforth PWN, is a paradigmatic lexical resource. Originally created for English, its model is now a "de facto" standard, due to its wide use in applications and its adaptation to different languages. For Portuguese, the first resource of this kind, WordNet.PT was announced in 2001 but, unlike PWN, was never free to use. A few alternatives were created, some of which we compared in [7]. But if those alternatives proved themselves useful for some tasks, they were not enough to enable all of the standard uses of a wordnet in Natural Language Processing (NLP), including similarity computation or word sense disambiguation.

This talk introduces our project OpenWordNet-PT [6]. abbreviated to OpenWN-PT, OpenWordNet-PT is a wordnet for Portuguese originally developed as a syntactic projection of the Universal WordNet (UNW) of Weikum and de Melo. Its long-term goal is to serve as the main lexicon for a NLP system, focused on logical reasoning and based on a logical representation of knowledge. The process of creating OpenWN-PT uses machine learning techniques to build relations between graphs representing lexical information from versions in multiple languages of Wikipedia entries and open electronic dictionaries. One of the features of the resource is that it incorporates different kinds of quality data, already produced and made available for Portuguese, independent of the variant, Brazilian or European.

OpenWN-PT has been improved constantly since 2011 through linguistically motivated additions, either manually or from evidence in large corpora. This is the case for the lexicon of nominalizations, tightly integrated with the OpenWN-PT [9], the review of the coverage of the verbs [5], the mapping of

*IBM Research and FGV/EMAp, Brazil
†Nuance Communications, USA
‡IBM Research, Brazil
§IBM Research, Brazil
¶PUC-Rio, Brazil

some nominalizations to Princeton's Morpholinks [3] and the work with gentilics, a particular class of relational adjectives [12].

OpenWN-PT is available in RDF/OWL following and expanding, when necessary, the original mappings. Both the OpenWN-PT data and schema of the RDF model are freely available for download. The philosophy of OpenWN-PT is to keep a close connection with PWN, but try to fix the biggest mistakes created by the automated translation methods, through language skills and tools. A consequence of this close connection to other languages is the ability to minimize the impact of lexicographical decisions on splitting/grouping the senses in a synset. While such decisions are, to a great extent, arbitrary, the practical criterion of following the multilingual alignment behaves as a pragmatic and practical guiding solution.

OpenWN-PT's extent of coverage is constantly monitored from its homepage `http://wnpt.brlcloud.com/wn/`. OpenWN-PT currently has 43,925 synsets, of which 32,696 correspond to nouns, 4,675 to verbs, 5,575 to adjectives and 979 to adverbs. Besides being available for download, the data can be retrieved via a SPARQL endpoint and can be consulted and compared with other wordnets both through the Open Multilingual WordNet (OMWN) interface and its own interface. OpenWN-PT's quality of coverage is difficult to measure, but OpenWN-PT was chosen by the developers of the Open MultiLingual WordNet OMWW [2], Freeling [10], BabelNet and Google Translate, as the representative Portuguese wordnet in those projects, due to its comprehensive coverage of the language and its accuracy.

Lexical resources are much more useful when they are aligned with other such resources and openly shared on the web. Since no resource is truly complete and different theoretical basis lead to different strengths, both in terms of coverage and of accuracy, one of the main points of the Linguistic Linked Open Data movement [4] is that we should be able to connect our resources via mappings that make the most of their complementarity. As they say

> 'Open Data' has become very important in a wide range of fields. However for linguistics, much data is still published in proprietary, closed formats and is not made available on the web. We propose the use of linked data principles to enable language resources to be published and interlinked openly on the web[..] Furthermore, we argue that modeling and publishing language resources as linked data offers crucial advantages as compared to existing formalisms. In particular, it is explained how this can enhance the interoperability and the integration of linguistic resources. Further benefits of this approach include unambiguous identifiability of elements of linguistic description, the creation of dynamic, but unambiguous links between different resources, the possibility to query across distributed resources, and the availability of a mature technological infrastructure.

We would like to follow this model, which already connects Princeton WordNet to (English) FrameNet [1], for Portuguese resources as well. Thus we would

like to pursue a project of aligning the FrameNet-BR [13] to OpenWordNet-PT. Preliminary conversations indicate that we should, perhaps, consider first the alignment in the restricted domain of "Tourism". OpenWordNet-PT has no domains as such, but some of our corpus work [11] is concerned with historical figures in Brazilian recent history, so locations and geographical and historical events are important for us and FrameNet-BR has already produced much on this domain.

The recently launched m.knob `http://www.ufjf.br/framenetbr/m-knob/` has already a multilingual repository of knowledge that we would like to compare to the knowledge we obtain via our mappings to SUMO. Since m.knob is already connected to BabelNet, which is related to OpenWordNet-PT via the OpenMultilingualWordNet, we expect to see how BabelNet deals with our data, as well as to learn how to improve the quality of our data, using BabelNet's other resources, such as Wikipedia and Wiktionary.

A different proposed application that we envisage is to create and develop 'frames' for historical characters in the DHBB. This seems an ideal application of FrameNet and we would like to see how much of the semantic content of the our favorite DHBB corpus can be covered with a small collection of frames for being born, graduating, becoming a politician, legislating, etc.. Filling up gaps in our resource that do exist in Wikipedia, DBpedia and others will be useful, but further in the future we would like to join forces with FrameNet-BR to come up with a sizeable collection of multiword expressions (mwes) in Portuguese. WordNets are not particularly good at multiword expressions and there is a general awareness that mwes tend to depend on the domain of discourse, so this could be an ideal follow-up to the project of creating and assigning frames to the DHBB entries.

References

[1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-1998)*, Montreal, Canada, June 1998. ACL.

[2] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August 2013. ACL Press.

[3] Fabricio Chalub, Livy Real, Alexandre Rademaker, and Valeria de Paiva. Semantic links for portuguese. In *10th Edition of its Language Resources and Evaluation Conference (LREC)*, Portoroz, Slovenia, May 2016.

[4] Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors, *New Trends of*

Research in Ontologies and Lexical Resources, Theory and Applications of Natural Language Processing, pages 7–25. Springer-Verlag, 2013.

[5] Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker. Making virtue of necessity: a verb lexicon. In *PROPOR – International Conference on the Computational Processing of Portuguese*, Tomar, Portugal, 2016.

[6] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.

[7] Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Alexandre Rademaker, Cláudia Freitas, and Alberto Simões. An overview of portuguese wordnets. In *Global Wordnet Conference 2016*, Bucharest, Romenia, January 2016.

[8] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[9] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. Extending a lexicon of portuguese nominalizations with data from corpora. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil, October 2014. Springer.

[10] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[11] Valeria De Paiva, Dário Oliveira, Suemi Higuchi, Alexandre Rademaker, and Gerard De Melo. Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE, October 2014.

[12] Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. Gentle with gentilics. In *Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016)*, Slovenia, May 2016.

[13] Tiago Torrent, Maria Margarida M. Salomão, Fernanda C. A. Campos, Regina M. M. Braga, Ely E. S. Matos, Maucha A. Gamonal, Julia A. Gonçalves, Bruno C. P. Souza, Daniela S. Gomes, and Simone R. Peron. Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 10–14*, Dublin, Ireland, August 2014. ACL.