OpenWordNet-PT: Taking Stock

Valeria de Paiva Alexandre Rademaker Livy Real
Fabricio Chalub Gerard de Melo

May 2018

Abstract

This note discusses work on lexical resources for Portuguese centered around OpenWordNet-PT, an open source wordnet-like resource for Portuguese. We discuss the initial developments, the sister project Nomlex-PT and focus in particular on applications that were developed in the quest of facilitating reasoning based on Portuguese texts.

1 Introduction

This note surveys some of our recent work on lexical resources for the Portuguese language. From the beginning, the ultimate vision driving this work on lexical resources has been the goal of facilitating logical reasoning with information extracted from texts in Portuguese, as attested by the title of the original paper on the resource "OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning". However, as is well known,

> Linguistic resources are very easy to start working on, very hard to improve on and extremely difficult to maintain, as funding usually only works for new resources. (*Anon*)

Thus, we review here mostly our work on improving, patching, checking and verifying OpenWordNet-PT, the open source version of Princeton WordNet for Portuguese that we have been developing since 2011, in the hopes that this will lead to genuine reasoning in the future. The reason we have been working on OpenWordNet-PT this long is that this work is somewhat like dealing with the mythical Hydra: upon solving one small problem, two new ones seem to appear. Thus, it makes sense to stop and take stock of 1) what has been accomplished thus far, 2) what remains ongoing work, and 3) what we intend to achieve in the near future.

2 OpenWordNet-PT

Brazilian Portuguese needs a Wordnet that is open access, downloadable and updateable, so that it can be improved by the community interested in using it

for knowledge representation and automated deduction. This kind of resource is also very valuable to linguists and computer scientists interested in extracting and representing knowledge obtained from text. WordNet and its various non-English counterparts have 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. The lack of such a resource for a given language slows down considerably, and may even bring to a halt, any work on reasoning about knowledge extracted from that language, which has been our main goal for Portuguese.

OpenWordnet-PT (abbreviated as OpenWN-PT or simply OWN-PT) is an open access wordnet for Portuguese, originally developed [13] as a transformation and extension of data from the Universal Wordnet/MENTA (UWN/MENTA) project [6, 7]. Like many other open wordnet creators, we believe that lexical resources need to be open to be useful, so our data has been, from the beginning, available for download, under an open source license. Nowadays, OWN-PT is available from GitHub at `https://github.com/own-pt/openWordnet-PT` and browsable from `http://openwordnet-pt.org` or `http://openwordnet-pt.com`.

The process of building OpenWN-PT used machine learning to construct relationships between graphs representing information coming from several different language versions of Wikipedia, as well as open dictionaries. OpenWordNet-PT was created by drawing on a two-tiered methodology so as to offer high precision for the more salient and frequent words of the language, but also high recall in order to cover a wide range of words in the long tail. We thus combine manual base concept annotation with statistical cross-lingual projection techniques. Starting as a projection of the Universal WordNet (UWN) [6] at the level of the lemmas in Portuguese and their relationships, the OpenWN-PT has been continuously improved through linguistically motivated additions and removals, either manual or semi-automatic, making use of large Portuguese corpora.[1]

As described in [9], in a first step, the information in the English Princeton WordNet is projected to Portuguese by using translation dictionaries to map the English members of a synset to possible Portuguese translation candidates. In order to disambiguate and choose the correct translations, feature vectors for possible translations are created by computing graph-based statistics in the graph of words, translations, and synsets. Additional monolingual wordnets and parallel corpora are used to enrich this graph. Finally, statistical learning techniques are used to iteratively refine this information and build an output graph connecting Portuguese words to synsets. In a second step, Wikipedia pages are linked to relevant WordNet synsets by learning from similar graph-based features as well as gloss similarity scores. Such mappings allow us to attach the article titles of the Portuguese Wikipedia with WordNet synsets, thus further increasing the coverage. We refer to this whole process as a "semantic projec-

[1] This kind of construction automatically started, but manually curated and improved, is also well exemplified in the creation of our sister project, the NomLex-PT, an open access, wide coverage lexicon of nominalizations in Portuguese, described in Section 3.

tion into Portuguese" based on the Universal WordNet. It is worth noting that similar kinds of endeavours could (and perhaps should) be started for other languages, given that the Universal Wordnet provides data in numerous languages. The approach is most successful for languages sufficiently well represented on the internet (see `https://en.wikipedia.org/wiki/Languages_used_on_the_Internet`) and for which Wikipedia is reasonably large.

Our main paper on the creation of OWN-PT was [9] in 2012, which we still request people who use the data to cite. At that stage, the coverage was just 24K synsets and neither glosses nor example sentences were available in Portuguese. Two years later in [27], we had tried the human correction route, but we were not getting very far. We had added 2,498 manually entered sense-word pairs as well as some 1,299 manually written Portuguese synset glosses. Since we relied on native speakers, but not linguists, no one had a particularly clear idea of the taxonomies of PWN, which were not presented to the annotators. This was a substantial source of errors, as we discovered later on.

The main good news from the progress report in 2014 was the fact that the first applications had begun to emerge. This includes the fact that FreeLing 3.0 [23], another open source project, had decided to use our lexicon for their Portuguese analyzer. Thus, a given Portuguese text can automatically be annotated with word senses in Portuguese using FreeLing. Next, IBM Research Brazil had a project developing sentiment analysis using tweets about the 2015 Confederation Soccer Coup [3], which relied on OWN-PT as its lexicon. This was integrated into the IBM InfoSphere Streams (ISS) platform and dealt with 1 million tweets, covering 4 friendly matches featuring the Brazilian national squad, using 7 classes of positivity.

3 NomLex-PT

Nominalizations are an important phenomenon in Computational Linguistics. They are useful for linguistic research as well as for information extraction. Since one team member was already conducting research on nominals in English [16] and another one was writing a doctoral thesis on them in Portuguese [28], it made sense to investigate how nominalizations could and should be incorporated into wordnet-like resources.

We envisaged an extension of OWN-PT, incorporating links connecting deverbal nouns with their corresponding verbs. To bootstrap this extension, we manually created over 2,000 entries via translation of the English NOMLEX entries [11]. Incorporating the NOMLEX-PT data into OpenWN-PT has shown itself useful in pinpointing some issues with the coherence and richness of OpenWN-PT. For example, the word *abasement* in English corresponds (`https://nlp.cs.nyu.edu/nomlex`) to the verb *abase* according to the English resource NOMLEX, and thus we would like a similar correspondence between the Portuguese noun *aviltamento* and the verb *aviltar* (the suggested translations of the pair). OpenWN-PT simply had two synsets {*humilhar, abaixar*} (humiliate, lower) and {*humilhar, rebaixar*} (humiliate, downgrade). The more

common verb in Portuguese *humilhar/humiliate* was repeated, while the less common *aviltar* (which is similar to *abase*) was left out altogether. However, due to the integration of the nominalizations, we managed to add the not so common Portuguese verb *aviltar* to the synset with *humiliate*, as necessary.

Extending our first work introducing NomLex-PT [11], we have taken further steps 1) to expand our set of nominalizations using lexica in cognate languages [32], 2) to actual computationally incorporate the nominals into OWN-PT [26] and to augment the lexicon of nominalizations using data from Portuguese corpora [14]. This work has not been fully completed. However, at that stage we considered that the development of our web interface for browsing and collaborative editing of OWN-PT was our most important pending issue and hence efforts were directed to the interface.

4 Social Interface

In 2016, we managed to have a new social and collaborative interface implemented and deployed. This was described in the paper aptly named "Seeing is Correcting" [29], as it was now possible to investigate hypernyms and hyponyms while annotating and better communicate between different human annotators. At that stage, OpenWN-PT was already part of the collection of wordnets for various languages jointly described and distributed through the Open MultiLingual WordNet [2] and the Global WordNet Association (`http://globalwordnet.org/wordnets-in-the-world/`).

It was surprising how a simple interface could make content so much more perspicuous. Thus the title of our paper was explained, if seeing is believing, new ways of seeing the data and of slicing it, according to our requirements, were necessary for curating, correcting and improving this data. In particular, the new interface allowed us to engage into localized experiments, dedicated to improving specific facets of OWN-PT. In this regard, we improved the verb lexicon a little [8] and took a stab at a specific class of adjectives, a subset of pertainyms, the *gentilics* [31]. We also decided to try to incorporate into the network the Princeton WordNet *morpholinks*, which were created by the Princeton team a while back, but were never fully integrated into the downloads of the Princeton WordNet. We reasoned that the same kinds of morpholinks should exist in Portuguese [4] and that they would help us in our Augean task of cleaning and completing OWN-PT.

During this time, the effort seemed to be paying off, as OWN-PT was being used by several important projects, such as Google Translate[2], FreeLing [23], OMW [2], BabelNet [20], Onto.PT [15], etc. However, in 2016, the deep neural network-based machine translation revolution hit Google Translate and for a while, it mentioned no lexical resources at all. While OMW, BabelNet, Onto.PT, Freeling still use OWN-PT, there are fewer attempts at using lexical resources altogether in Computational Linguistics.

[2]We can still find references to OWN-PT at `https://translate.google.com/intl/en/about/license.html`.

5 Applications

Our first envisaged application pertains to the Brazilian Dictionary of Historical Biographies (DHBB). The dictionary could be consulted online for a long time, but the data itself only recently became available under an open license. The writers of the dictionary entries (biographical sketches of politicians in Brazilian History since the 1930's) were asked to follow a set of guidelines with respect to the information that these entries should contain. This means that the data is clean, of high quality, not in a too high register (as they wanted the entries to be accessible to high school students), and almost semi-structured.

Our preliminary work on this resource is described in [24]. Biographical data, historical or not, requires a well developed system for detecting and classifying Named Entities, such as people, places, organizations, etc. Princeton WordNet has some little information on those entities, for English, but not enough even for English. We did not think that we should reproduce named entity synsets of PWN, neither for English, nor for Portuguese entities. But even recognizing these synsets that correspond to proper nouns can be complicated, as described in [17].

Our vision of a more thorough extraction of information from the text [24] is still under consideration. However, recently, using IBM Watson Knowledge Studio[3], we undertook a small experiment annotating 50 documents with both entities and relations. The results of this annotation are still being worked on, and an online demo is available at `http://cpdoc.github.io/dhbb`.

Another way of using our data is in surveying the different kinds of wordnet-like lexical resources that exist for Portuguese. We discussed and compared the several lexical resources of this sort that have been created over the years, so that it is more clear how they relate to each other. We wrote about this both in Portuguese [21] and in English [10] together with the creators of other WordNet-like data, e.g. Onto.PT [22] and PULO [34].

More recent applications have started expanding PWN itself [19] with regard to geological time periods, for example. Another domain-specific direction, also reported in the last Global Wordnet Conference, considered a preliminary application [12] in the legal domain: expanding and using OWN-PT to develop a benchmark for the Brazilian Bar Exam (the "Ordem dos Advogados do Brasil" exam, or OAB exam for short).

We also decided to invest in the Universal Dependencies as an open source resource of syntactic information. Thus, work was put into reannotating the Portuguese Bosque corpus using Universal dependencies and joining their community [25]. So far, for these efforts, we have not been able to use open parsers developed originally for Portuguese. For the re-annotation of Bosque with Universal Dependencies, we relied on a conversion of the original parsing with PALAVRAS [1] and have developed libraries and tools to compare CoNLL structures. There is ongoing work on improving the tools to guarantee the quality of the dependency representations.

[3] `https://www.ibm.com/watson/services/knowledge-studio/`

We also want to complete OWN-PT so that it is helpful for temporal processing. Recent and preliminary work [33] shows that using original PWN temporal relations can lead to mistakes when doing Portuguese processing. For example, *Father's Day* happens in June for Anglophone cultures, but not for Lusophone ones. Thus, having the synset for Father's Day as a hyponym of *June* does not make sense in Portuguese. The issue is how to obtain local culturally relevant synsets in OWN-PT without losing the current alignment between PWN and OWN-PT. This is part of a broader discussion within the Global WordNet Association.

Towards our original goal of having a comprehensive coverage of Portuguese vocabulary, we would like to make sure that we have all (reasonable) lemmas of, say, the Bosque Corpus, perhaps the most used corpus for Portuguese processing, in OWN-PT. Since lexical resources and corpora are intrinsically related, we want to make sure that everyone who wants to process any of the several versions of Bosque will find all of its lemmas correctly placed in OWN-PT. To this end, we have started some experiments in `http://wnpt.brlcloud.com/wn/prototypes`, but much remains to be done.

6 Future Work

In an ideal world, where help was plentiful, we would like to work on three levels. First, the work on *correcting* the data is, as always, necessary. While the main goal remains the original one of using the data for reasoning, the data in question has to be high quality and correct for all the applications that now rely on OWN-PT. However, some of original WordNet data is not logically consistent and our translations in Portuguese are not fully consistent either. For these tasks of correcting OWN-PT, focused small projects, such as the one on gentilics or the one in plurality of nouns [30] appear to work well.

Then, there is the second level of *expanding* the data to more fully cover real-world Portuguese text. In this regard, we have thus far not yet experimented sufficiently with small projects, but the work on using true Portuguese corpora to extend NOMLEX-PT [14] is a blueprint to follow.

Most importantly, however, at the third level, we need to decide on how to *connect* OWN-PT with a semantic parsing system that will allow us to infer the semantics and draw inferences on larger texts. Here, the best path forward seems to be the *let all blossoms bloom* approach, exploiting different syntactic and semantic parsing frameworks, different word embeddings, different algorithms for parsing, POS tagging, named entity recognition, coference resolution, time expression resolution, multiword expression recognition, etc.

7 Conclusions

We have been working on expanding and improving our Portuguese lexical resources for more than seven years. However, much remains to be done, and

we have to pick the most effective and productive ideas to move forward. The productvity of this work is a function not only of the research working the way we envisage it (after all we are talking about research, not simply development) but also of the funding possibilities for junior researchers and students. This depends also on variables beyond our control.

Within the loosely defined group behind OWN-PT, there are different priorities and preferences. It is not always possible to align goals and methods. In particular, one of our reviewers would like to see "more general lessons that could be drawn out" of this kind of work on lexical resources. One lesson we learned is that lexical resources do not thrive in a vacuum, they need other resources to interact. Thus to make OWN-PT a really useful resource, we need to have an eco-system of different other lexica, plus corpora, plus knowledge bases, plus APIs and lower level components, all interconnected. This is somewhat related to the vision of Linguistic Linked Data [18, 5]. Thus we have worked on corpora like Bosque, DHBB, using Freeling, UDpipe, HeidelTime, SUMO, and would like to connect and experiment with many further sorts of components. However, one main difference is that we believe in mappings connecting all these resources, more than in the resources themselves.

Despite all the heterogeneity, it seems to us that all the work that has emerged from this coalition of interests has been beneficial to the advancement of NLP in Brazilian Portuguese so far and we hope that more of this 'broad church' kind of work can be accomplished in the near future.

References

[1] Eckhardt Bick. *The parsing System PALAVRAS Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Ph. D. thesis, Department of Linguistics, University of Århus, Denmark, 2000.

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

[3] Paulo Rodrigo Cavalin, Maira Athanazio C. Gatti, Tiago G. P. Moraes, Fabio S. Oliveira, Claudio S. Pinhanez, Alexandre Rademaker, and Rogerio Abreu de Paula. A scalable architecture for real-time analysis of microblogging data. *IBM Journal of Research and Development*, 59(2/3):16:1–16:10, March 2015.

[4] 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.

[5] Gerard de Melo. Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web*, 6(4):393–400, August 2015.

[6] Gerard de Melo and Gerhard Weikum. MENTA: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1099–1108. ACM, 2010.

[7] Gerard de Melo and Gerhard Weikum. UWN: A large multilingual lexical knowledge base. In *Proceedings of the ACL 2012 System Demonstrations*, pages 151–156. Association for Computational Linguistics, 2012.

[8] 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.

[9] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. Published also as Techreport, http://hdl.handle.net/10438/10274.

[10] 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.

[11] Valeria de Paiva, Livy Real, Alexandre Rademaker, and Gerard de Melo. Nomlex-pt: A lexicon of portuguese nominalizations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[12] Pedro Delfino, Bruno Cuconato, Guilherme Paulino Passos, Gerson Zaverucha, and Alexandre Rademaker. Using openwordnet-pt for question answering on legal domain. In *Global Wordnet Conference 2018*, Singapore, January 2018. to appear.

[13] Valeria dePaiva and Alexandre Rademaker. Revisiting a brazilian wordnet. In *Proceedings of Global Wordnet Conference*, Matsue, 2012. Global Wordnet Association. http://www.globalwordnet.org/gwa/gwaconferences.html.

[14] Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne Silva. Extending a lexicon of portuguese nominalizations with data from corpora. In *International Conference on Computational Processing of the Portuguese Language*, pages 114–124. Springer, 2014.

[15] Hugo Gonçalo Oliveira and Paulo Gomes. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation Journal*, 48(2):373–393, 2014.

[16] Olga Gurevich, Richard Crouch, Tracy Holloway King, and Valeria De Paiva. Deverbal nouns in knowledge representation. *Journal of Logic and Computation*, 18(3):385–404, 2007.

[17] John P. McCrae. Mapping wordnet instances to wikipedia. In *Proceedings of Global Wordnet Conference*, Singapore, 2018. Global Wordnet Association. http://www.globalwordnet.org/gwa/gwaconferences.html.

[18] John Philip McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. The open linguistics working group: Developing the linguistic linked open data cloud. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 2435–2441, 2016.

[19] Henrique Muniz, Fabricio Chalub, Alexandre Rademaker, and Valeria de Paiva. Extending wordnet to geological times. In *Proceedings of Global Wordnet Conference*, Singapore, 2018. Global Wordnet Association. http://www.globalwordnet.org/gwa/gwaconferences.html.

[20] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[21] Hugo Gonçalo Oliveira, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real, and Alberto Simões. As wordnets do português. *Oslo Studies in Language*, 7(1):397–424, 2015.

[22] Hugo Gonçalo Oliveira and Paulo Gomes. Onto.pt: automatic construction of a lexical ontology for portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, volume 222, pages 199–211, 2010.

[23] Lluis Padro 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.

[24] 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.

[25] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva Universal Dependencies for Portuguese.

Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017.

[26] Alexandre Rademaker, Valeria De Paiva, Gerard de Melo, and Livy Maria Real Coelho. Embedding NomLex-BR nominalizations into OpenWordNet-PT. In *Proceedings of the Seventh Global Wordnet Conference*, pages 378–382, 2014.

[27] Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Real, and Maira Gatti. OpenWordNet-PT: A project report. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, January 2014.

[28] Livy Real. *Nominalizações*. PhD thesis, Ph. D. thesis, Federal University of Paraná, Brazil, 2014.

[29] Livy Real, Fabricio Chalub, Valeria de Paiva, Claudia Freitas, and Alexandre Rademaker. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of Asian Federation of Natural Language Processing - Fourth Workshop on Linked Data in Linguistic Resources and Applications (LDL 2015)*, Beijing, China, July 2015.

[30] Livy Real and Valeria de Paiva. Plurality in wordnets. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, 2016.

[31] 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.

[32] Livy Real, Valeria de Paiva, and Alexandre Rademaker. Extending nomlex-pt using ancora-nom. In Laura Alonso Alemany, Muntsa Padró, Alexandre Rademaker, and Aline Villavicencio, editors, *Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp)*, São Carlos, Brazil, October 2014. Biblioteca Digital Brasileira de Computação, UFMG, Brazil.

[33] Livy Real, Alexandre Livy Real, A Rademaker, Fabricio Chalub, and Valeria V de Paiva. Towards temporal reasoning in portuguese. In *Proceedings of the LREC2018 Workshop Linked Data in Linguistics*, 2018.

[34] Alberto Simoes and Xavier Gómez Guinovart. Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 239–248. Springer, 2014.