Seeing is Correcting: the affordances of editing interfaces

Fabricio Chalub Valeria de Paiva Livy Real
Alexandre Rademaker Claudia Freitas

April 2015

Abstract

This note describes OpenWordNet-PT, an automatically created, manually curated wordnet for Portuguese and introduces the new interface we are using to speed up its manual curation. OpenWordNet-PT is part of a collection of automatically created wordnets for various languages, jointly described and distributed by Bond and others through the project Open MultiLingual WordNet. We contend that the creation of such large, distributed and linkable resources, through the use of semantic web tools, such as RDF and SPARQL endpoints is on the cusp of revolutionizing multilingual language processing to the next level, the truly semantic one. But to get there, there is a need for interfaces that allow ordinary users and (not really computational) linguists to help on the construction, checking, cleaning up and verification of the quality of these resources.

1 Introduction

Lexical knowledge bases are organized repositories of lexical items. These resources typically include information about the possible meanings of words, relations between these meanings, definitions and phrases that exemplify their use and maybe some numeric grades of confidence in the information provided. The Princeton WordNet model, with English as its target language, is probably the most popular model to represent this type of resource. Our main goal is to provide good quality lexical resources for Portuguese, making use, as much as possible, of the effort already spent creating similar resources for English.

There is a theory which states linguistic resources are very easy to start, very hard to improve and extremely difficult to maintain, as the last two tasks do not get much in the way of recognition and kudos that the first task gets. Given this intrinsic barrier, many well-funded projects, with institutional or commercial backing cannot keep their momentum. So it is rather pleasant to see that a project like ours, without any kind of official infra-structure, has been able to continue development and improvement, so far, re-inventing its tools and

methods, to the extent that it has been chosen by Google Translate to be used as their source of lexical information for Portuguese.

This note reports on a new interface for consulting, checking and collaborating on the improvement of OpenWordNet-PT[4], an automatically created, but manually verified wordnet for Portuguese, fully compatible and connected to Princeton's paradigmatic WordNet[5]. To some of us it has been surprising how a simple interface, based in, by now, traditional tools like Lucene and NoSQL databases, can make content much more perspicuous. This is what is meant by our title: if seeing is believing, new ways of seeing the data and of slicing it, according to our requirements, are necessary for correcting and improving this data.

Correcting and improving linguistic data is a hard task, as the guidelines for what to aim for are not set in stone and known in advance. While the WordNet model has been paradigmatic in modern computational lexicography, this model is not without its failings and shortcomings, as far as specific tasks are concerned. Also while it is easy and satisfying to provide copious quantitive descriptions of numbers of synsets, for different parts-of-speech, of triples associated to these synsets and of intersection with different subsets of Wordnet, for example, the whole community dedicated to creating wordnets in other languages that not English, has not come up with criteria for accuracy of these new resources. Thus qualitative assessment of a new wordnet is, at the moment, a matter of judgement and art, more than a commonly agreed practice. Believing that this qualitative assessment is important, and so far rather elusive, we propose in this note that having many eyes over the resource, with the ability to shape it in the directions wanted, is a main advantage. This notion of volunteer curated content, as first and foremost exemplified by Wikipedia, needs adaptation to work for lexical resources. This note describes one such adaptation.

`description of paper here`

2 OpenWordNet-PT

The OpenWordNet-PT [12], abbreviated as OpenWN-PT, is a wordnet originally developed as a syntactic projection of the Universal WordNet (UNW) of de Melo and Weikum [2]. Its aim is to serve as the main lexicon for a system of natural language processing focused on logical reasoning, based on representation of knowledge, using an ontology, such as SUMO [11].

The process of building OpenWN-PT, in its original source in the universal wordnet UWN, uses machine learning techniques to build relations between graphs representing lexical information coming from versions (in multiple languages) of Wikipedia entries and open electronic dictionaries. For details one can consult [2]. Then a projection targeting only the synsets in Portuguese is calculated. Despite starting as a projection only, at the level of the lemmas in Portuguese and their relationships, the OpenWN-PT has been constantly improved through *linguistically motivated* additions and removals, either manually or by making use of large corpora.

The philosophy of OpenWN-PT is to maintain a close connection with Princeton's wordnet, but try to remove the biggest mistakes created by the automated methods, using linguistic skills and tools. One consequence of this close connection of OpenWN-PT with the English wordnet is the ability to minimize the impact of lexicographical decisions on separation or grouping of senses in a synset. Such decisions are inherently arbitrary [9], thus the criterion of following the multilingual alignment produces a pragmatic and practical solution.

This lexical enrichment process of OpenWN-PT employs three language strategies: (i) translation; (ii) corpus extraction; (iii) dictionaries. Regarding translations, glossaries and lists produced for other languages, such as English, French and Spanish are used, automatically translated and manually revised. The addition of corpora data contributes words or phrases in common use which may be specific to the Portuguese language or which do not appear (too rare?) via the automatic construction. The first corpora experiment in OpenWN-PT was the integration of the nominalizations lexicon, the NomLex-PT [6]. Use of a corpus, while helpful for specific conceptualizations in the language, brings additional challenges for mapping alignment, since it is expected that there will be expressions for which there is no synset in the English wordnet. As for the information of dictionaries, this was used both for the original creation of Portuguese synsets but also indirectly through the linguists' use of PAPEL[8]. This resource is not without errors, but has the advantage of containing only relations between words of the language, including those that are not very popular or frequent. Using PAPEL, we constructed extra pairs of words of shape (verb,nominalization) when there was no morphological clue available.

2.1 Challenges of lexical enrichment

The task we set ourselves was to build a wordnet for Portuguese, based on the Princeton wordnet model, which is not the same thing as building the Princeton wordnet in Portuguese. Thus we do not propose to simply translate the original wordnet, but to create a wordnet for Portuguese based on Princeton's architecture and, as much as possible, linked to it at the level of the synsets.

The task of building a wordnet in Portuguese imposes many challenges and choices. For example, we needed to decide which variants of Portuguese we were going to treat, how to include existing senses in Portuguese that do not exist in English or at least in the Princeton project and how to deal with senses that apparently do not have a straight correspondent in Portuguese.

The simple translation of other lexical resources for comparison and further extension of our resources is already a complicated task that requires different techniques and theoretical decisions that have direct consequences on the type and quality of the new resource. For example, starting from Princeton Wordnet and NOMLEX [1], we realize that the synsets of OpenWN-PT were automatically provided with very erudite and rare words. It happens not only because translator tools do not provide colloquialisms, but mainly because we have parted from corpora and tools for English, French and Spanish, languages

whose words have a direct relation to Portuguese roots and suffixes. For example, in the first version of NomLex-PT, *arbitration* was translated to *arbitração*, despite the existence of the Portuguese form *arbitragem*, which seems a more frequent form. Actually, NomLex-PT is completed integrated to OpenWN-PT and also presents *arbitragem* as a possible word to this synset.

Since our goal is to produce a wordnet that considers all variants of Portuguese, building a structured lexicon that has only uncommom lexical itens does not help with the general idea of the project.

First of all, we have decided that OpenWN-PT must include all variants of Portuguese, not only Brazilian Portuguese. Based on that, senses that can be expressed through words that have different spellings on different Portuguese dialects must include all these variants. As we master only Brazilian Portuguese, we have extracted many variants of European Portuguese from PULO [13], abbreviation for Portuguese Unified Lexical Ontology, and Onto.PT (originally presented by [7]. A future work to include African variants of Portuguese is to consider Corpus África[10], online corpora of five different dialects of Portuguese.

Our strategy to enrich and improve OpenWN-PT is to choose specific points of it to promptly correct and upgrade them. During these strategic fixes, the need for guidelines to annotators which dictate the format of examples, glosses and variations of the words in synsets. The guidelines for annotators are available in https://github.com/arademaker/openWordnet-PT/wiki.

The need of an online and searchable version of OpenWN-PT comes for two reasons: (i) to have an accessible tool for ordinary users, (ii) to improve our strategy to enrich the resource. How the interface helps to enrich the tool is discussed here in Section 5 and the guidelines for users are available on https://github.com/arademaker/openWordnet-PT/wiki/How-to-use.

One of the main problems we have is the lack of many senses from Princeton wordnet in Portuguese. Even if OpenWN-PT does have the same synsets that the original wordnet has, many original senses are close related to the American use of English and it seems have no reason to keep them in a Portuguese wordnet. The synset 13390244-n, which brings a specific word ('quarter') for the concept of "a United States or Canadian coin worth one fourth of a dollar" and the synset 08139795-n, which corresponds to the United States Department of the Treasury, are examples of senses which are not general enough to participate in an international and really universal ontology of senses. For now, we have chose to keep all the Princeton synsets and translate them for the closer hyponym existent in Portuguese. In the exemplified cases, we keep the synsets, the gloss and the examples, but we translated words for commom expressions in Portuguese, as *moeda de 25 centavos* '25 cents coin' and *Ministério da Fazenda* 'department of treasury'.

Other very relevant problem is the senses that we do have in Portuguese and do not have a correspondent in Princeton wordnet or even in English, as *jogo de cintura*, which means a property of someone who can easily and friendly adapt his/her aims and feelings to a certain situation. Of course, this issue is already expected, as the original wordnet was designed based on English, but it requires an important choice from scholars who decide to build new wordnets related to

Princeton project: is it better to keep only synsets linked to Princeton wordnet or it is better creating new synsets? To us, it seems the last option is much more interesting, given our goal, but this choice also means that we would lose closeness from other wordnets. From this perspective, it arises a new problem: being the Princeton wordnet the standard wordnet, how to create new senses and to be sure that pre-existent wordnets have not created the same senses before us? And more, how to be sure that next wordnets will have access to those senses that are present in Portuguese, but not in English? For now, this issue remains as future work to this group.

3 Linked Data Rationale

Like many other recent lexical resources, e.g. Onto.PT[7], OpenWN-PT is also available as an RDF/OWL download, following and expanding, when necessary, the original mappings proposed by [14]. Both the data for the OpenWN-PT and the template settings for the RDF model (classes and properties) are freely available for download.

4 Current status

The 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. It is nowhere as comprehensive as Princeton's wordnet with 117K synsets or the Finnish or the Thai wordnets, but it is not too small either. It's more than twice the size of the Russian wordnet, bigger than the Spanish and just a little smaller than the French wordnet. But as discussed in the introduction, the quality of these resources is much harder to compare.

Besides being able to be downloaded, the data on Portuguese can be retrieved via a SPARQL endpoint [1]. The multilingual base can be consulted and compared with other wordnets using the OMWN interface (`http://compling.hss.ntu.edu.sg/omw`) and changing preferences to the desired languages, assuming the lexical item is found.

But if the ability of comparing senses in several languages was already useful to try to judge meanings in Portuguese, it did not allow us to compare with the collection of other words with the same meaning, or with different shades of meaning, appearing both in English and Portuguese. This also changed, since we started developing a new search and editing interface in September 2014.

5 The New Interface

The main reason for the new interface was the need to edit the entries of OpenWN-PT, as they existed. The first design decision was that before adding

[1] http://logics.emap.fgv.br:10035/repositories/wn30

new synsets corresponding to the Portuguese reality, we should clean up the network from its most egregious mistakes, caused by the automatic processing of the entries.

Thus we have had to remove Spanish, Galician or Catalan words that were misjudged as Portuguese and we have had to make sure that the part of speech classification was preserved: many times the popularity driven automatic process prefers the noun meaning of a verb that can be both. We also have several problems with the lemmatization of entries, as criteria for the use of capitalization is different in English and Portuguese and our entries were not lematized. We decided to follow the Portuguese tradition in dictionaries and mostly only list the masculine singular form of nouns and adjectives, for example.

Then we have the problem with clitics or reflexive verbs, which are much less frequent in English than in Romance languages. Many design decisions were needed as well. Finally there is still the big difficult problem of when to 'lexicalize' a verbal expression as a 'word' on a synset, or as a synset of its own. This problems are still being decided, via a lazy strategy of cleaning up what is clearly wrong first, and collecting subsidies for the more intrincate lexicographic decisions later on. Some of these discussions and decisions were described in our [3].

But apart from phenomena that we realize have to be dealt with in a uniform way, we have also problems of disambiguation that are one-offs, like the verb *to date=datar*, that in Portuguese is only to put a date on to a document, or a monument or a rock or a tree, when in English it also means *to go out with*. Thus the automatic processing ended up with a synset meaning both "finding the age of" and "going out with" (00619183-v), which makes no sense. To see and check this kind of situation it was decided that the interface would allow linguists to accept or remove a word, to accept or remove a gloss and to accept or remove an example of the use of the synset.

But the new interface was much more useful than simply offering the possibilty of local rewrites, as it allowed us to search for different classes of synsets and of words, both in English and in Portuguese. It clearly shows the synsets that have no words in Portuguese, which allows us to target these synsets and to decide whether they are simply missing a not very popular word (e.g. 00308534-v is missing the not terribly interesting verb *hidrogenar*, an exact correspondent to *hydrogenate*) or they correspond to a sense that does not work exactly the same way in English and Portuguese. For example, back to the verb *to date* as in *romantically going out with someone*, English seems to leave underspecified whether it is a habitual event or a single one, while in Portuguese we use different verbs, *namorar* or *sair*, but if we want to not commit ourselves to either kind of engagement, we use the verbal expression *sair com*

The new interface also makes use of social networking techniques to help achieve common sense by encouraging discourse and diversity of opinions. The system authenticates users via Github and keeps track of suggestions and comments made by those users. We encourage the patterns of communication very frequently associated with social networks such as Twitter and Reddit where users can 'tag' other users in comments (thus asking for the attention on that

particular topic). Comments have also have 'hash tags' that are used, for instance, to tag particularly difficult synsets for later consideration. We also allow users to vote positively or negatively on the suggestions and use this score to automatically accept or reject those suggestions. Once a suggestion is accepted, it is then committed to the database, effectively modifying OpenWN-PT with the proposed suggestion.

5.1 Testing and Verifying

6 Future Work

As the task of having a complete, universal and totally integrated wordnet seems to be an infinite labor, here we focus on our next steps to enrich our lexical resource.

For now, we still need to complete gloss and examples from many synsets and this is our main future task. Also we would like to consider and decide how to integrated African Portuguese variants to OpenWN-PT. We plan to do it using Corpus África [10].

Also the decision of how to integrate Portuguese senses that lack from English and Princeton wordnet's senses remains as future work, as this choice will draw an important design feature of OpenWN-PT.

References

[1] Adam Meyers Leslie Barrett Ruth Reeves Catherine Macleod, Ralph Grishman. Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium, 1998.

[2] Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA, 2009. ACM.

[3] Valeria de Paiva, Cláudia Freitas, Livy Real, and Alexandre Rademaker. Improving the verb lexicon of openwordnet-pt. 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, oct 2014. Biblioteca Digital Brasileira de Computação, UFMG, Brazil.

[4] 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.

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

[6] 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, oct 2014. Springer.

[7] 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 of *Frontiers in Artificial Intelligence and Applications*, pages 199–211. IOS Press, 2010.

[8] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal, September 2008. Springer.

[9] Adam Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, (31):91–113, 1997.

[10] José Bettencourt Antónia Estrela Sancho Oliveira Rui Santos Maria Fernanda Bacelar do Nascimento, Luísa Alice Santos Pereira. Corpus África: as cinco variedades africanas do português. In *Textos Seleccionados - XXIII Encontro Nacional da Associação Portuguesa de Linguística*, pages 373–384, Lisboa, 2008.

[11] Adam Pease and Christiane Fellbaum. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press, 2010.

[12] Alexandre Rademaker, Valeria De Paiva, Gerard de Melo, Livy Maria Real Coelho, and Maira Gatti. Openwordnet-pt: A project report. In *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan 2014.

[13] Alberto Simões and Xavier Gómez Guinovart. Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, volume 8854 of *LNCS*, pages 239–248. Springer, 2014.

[14] Mark van Assem, Aldo Gangemi, and Guus Schreiber. RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium, June 2006.