Revisiting a Brazilian WordNet

Valeria de Paiva
Rearden Commerce
Foster City, CA, USA
valeria.depaiva@gmail.com

Alexandre Rademaker
Applied Mathematics School, FGV
Rio de Janeiro, Brazil
arademaker@gmail.com

Abstract

Brazilian Portuguese needs a WordNet that is open access, downloadable and changeable, so that it can be improved by researchers, such as the community interested in automated deduction. This would be very valuable to linguists and computer scientist interested in representing knowledge obtained from texts. We discuss briefly why we want a Brazilian Portuguese WordNet and how we are going about getting one. These are only first steps, though, as our project is just starting.

1 Introduction

WordNet (Fellbaum, 1998) 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.

But if there is still a lack of lexical resources for English, the problem is ten-fold more acute for other languages, which lack even easily accessible corpora and basic tools such as tokenizers, taggers and splitters. This lack of resources slows down considerably, almost stops completely any work on reasoning about knowledge obtained from language, our main goal.

We are starting a project at Foundação Getulio Vargas (FGV) in Brazil, where we want, in the long run, to use formal logical tools to reason about knowledge obtained from text in Portuguese. We are logicians, not linguists, so we want to minimize the amount of Computational Linguistics that we have to develop. Hence it would have been very sensible to use a Brazilian WordNet, if we could have one. While we originally expected to be able to use some existing Brazilian WordNet, out of the box, it turns out that these are not available. There are some attempts.

There is the project WordNet.PT (Portuguese WordNet) from the "Centro de Linguitica da Universidade de Lisboa" headed by Palmira Marrafa. But this is available online only, no download available and, as far as we can see on their webpages, little development has happened recently to this project. The WordNet.PT version available online [1] has about 19000 lexical expressions, from different semantic fields. The fragment made available online includes expressions from subdomains such as art, clothing, geography, health, institutions, living entities and transportation, but no description of other domains and/or future releases of the database are discussed. The group has also a newer version of WordNet.PT called $\texttt{WordNet.PT}_{\texttt{global}}$ (Marrafa et al., 2011) which pays attention to different varieties of Portuguese, like African variations in the language. But while this is very interesting for linguistic comparative research and useful for online queries [2], this smaller version of WordNet.PT is still not available for download and/or modifications and improvements.

Then there is also the MultiWordNet project and its Portuguese version MWN.PT, developed by Antônio Branco and colleagues at the NLX-Natural Language and Speech Group, of the University of Lisbon, Department of Informatics. According to their description [3] MWN.PT the MultiWordnet of Portuguese (version 1) spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over

[1] http://tinyurl.com/6p2vsy3.
[2] http://www.clul.ul.pt/wnglobal/.
[3] http://tinyurl.com/bum4mmh.

21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the English Princeton WordNet and to the 98 Base Concepts suggested by the Global Wordnet Association, and the 164 Core Base Concepts indicated by the EuroWordNet project. But again this wordnet is available online only and/or with a restrictive license that requires payment.

Thirdly there is a first version of a Brazilian Portuguese version of Wordnet developed by Bento Dias da Silva and collaborators (Dias-Da-Silva et al., 2000; Scarton and Aluisio, 2009). But this also cannot be downloaded, is not available online and is not being maintained in an open access basis, which is one of the strongest points of Princeton WordNet. Open access availability is one of the main reasons we would like to have our own version of WordNet-BR, which we are calling WN-BR. This is because we believe that resources like Wikipedia and WordNet need to be open and modifiable by others in order to improve over time.

Finally there is a whole batch of work on merging WordNet with Wikipedia categories and infoboxes, that tries to leverage the work already done by the Wikipedia volunteers. Amongst these we are particularly excited about YAGO/MENTA (de Melo and Weikum, 2010) (and YAGO2 for further work on temporal/spatial information), which we describe below. This kind of work fits in well with our goals of ultimately doing reasoning, in large scale, with knowledge obtained from text.

2 Global WordNet Grid

In this forum it is perhaps not necessary to recall that the Global WordNet Association aims at the development of wordnets for all languages of the world and to extend the existing wordnets to full coverage and all parts-of-speech. In 2006 the association launched a project to start building a completely free worldwide wordnet "grid". This grid would be built around a shared set of concepts, and would be expressed in terms of the original Wordnet synsets and SUMO (Niles and Pease, 2001) terms. The idea of the grid is very appealing and the suggested procedure to create wordnets looks sensible and feasible.

To recap the suggestion was to build the first version of the Grid around the set of 4689 "Common Base Concepts" and to make the Grid free, following the example of the Princeton WordNet. Now the Base Concepts are supposed to be the most important concepts in the various wordnets of different languages. The importance of the concepts was measured in terms of two main criteria: (1) A high position in the semantic hierarchy; (2) Having many relationships to other concepts. The procedure described as the "expand approach" seems to us viable: First translate the synsets in the Princeton WordNet to Portuguese, then take over the relations from Princeton and revise, adding the Portuguese terms that satisfy different relations. Then revise and revise and revise until we can guarantee the consistency of the taxonomy.

In somewhat more detail but still following the suggestions of the global wordnet grid, we think we can develop a core WordNet for Brazilian Portuguese by: first representing the 1024 core basic concepts by one or more synsets in Portuguese that are either equivalent or very closely associated to the original core concepts in the Princeton WordNet. Then adding Base Concepts that are important to Brazilian Portuguese, but not in the set of core basic concepts. (For that one could use lists of Portuguese words listed by frequency, and comparisons with other Wordnets for romance languages.) Next we need to check that this forms a closed and consistent hierarchy. Finally we should add further relations necessary to specify the semantics of the basic concepts.

While this procedure seems sensible and doable, despite being hard work, the existence of several versions of Wordnet, and the fact that the concepts uncovered as basic ones are not related to their synsets in WordNet 3.0 makes things more difficult. WordNet 3.0 has the huge advantage of disambiguated glosses, a real plus if the goal is semantics. But the mechanics of following the procedure sketched above turn out more complicated than expected.

Our solution is to use the mapping from WordNet 2.0 to WordNet 3.0 provided by (Daudé et al., 2000). The idea consists in, using the map, identify the WordNet 3.0 synsets equivalent to the 4689 WordNet 2.0 basic concepts listed in the EuroWordNet project (Stamou et al., 2002).

We plan to use the RDF version of WordNet 3.0 made freely available online [4] by Mark van Assem and Jacco van Ossenbruggen. The RDF version has the benefit of better supporting our collaborative work, facilitating the maintenance in the same data structure of both English and Portuguese versions of the glosses and lexical forms, and, once loaded in a Triple Store, providing us with a working environment to add or remove relations, comments and so on. Unfortunately, there are at least two different versions of WordNet available in RDF. The RDF representation of WordNet 2.0 is described in `http://www.w3.org/TR/wordnet-rdf/` and seems to be used as reference for the newly minted RDF representation of WordNet 3.0. Nevertheless, we still have to investigate the differences between the WordNet 2.0/3.0 mapping used in the RDF representation of WordNet 3.0 and the mapping provided by (Daudé et al., 2000).

3 Sumo and WordNet-BR

The Global WordNet Grid approach has three aspects that seem to us worth considering. First there is the work on determining which synsets (corresponding to concepts) are most popular in several languages. This work was done in the EuroWordNet projects and it would be a shame not to use it. The emphasis on many languages should help filter out personal and cultural biases. Then there is the proposed stepping up schedule, which seems to us very attractive, as a way of helping to get the "grunt" work done. Finally and in some ways more importantly there is the connection with SUMO that we discuss now.

According to Wikipedia, the Suggested Upper Merged Ontology or SUMO is an upper ontology intended as a foundation ontology for a variety of computer information processing systems. It can be downloaded and used freely and it has been available and in development since 2000. A mapping from WordNet synsets to SUMO has also been defined and maintained for several versions of WordNet. Most importantly for us, SUMO is organized for interoperability of automated reasoning engines. In particular SUMO's associated open source knowledge engineering environment, Sigma [5] runs already in Vampire (Ganzinger et al., 1999) and Leo-II (Benzmueller and Paulson, 2010), for example. Projections of SUMO into description logics, automatically available, can be run in the OWL reasoners.

In the beginning of 2010 we started an informal project of discussing how logic and automated reasoning could have a bigger impact, if coupled with natural language processing and how it would be a great thing to translate some of the advances already made for English text understanding to Portuguese text understanding and reasoning.

Since one of us (Valeria de Paiva) had worked for almost nine years in Xerox PARC, in the systems developed by the Natural Language Theory and Technology (NLTT) group, particularly on the system Bridge, we requested an academic license to the XLE (Xerox Language Engine) to try to adapt the systems to Brazilian Portuguese. However, we are both logicians, our expertise lies at the end of the long pipeline of the system Bridge and we tried to recruit, still informally, people with more expertise on the language side of the project. But at that stage we did not have any formal backing, so despite some interesting offers, nothing much happened. Recently we have been granted formal backing, although still in small scale, and one of the opportunities that presented itself was to forestall the need for the creation of a Brazilian WordNet (or perhaps to help improve the creation of such) , via the use of SUMO.

Wordnet is an important component of the XLE Unified Lexicon (UL (Crouch and King, 2005)), as the logical formulae created by the Abstract Knowledge Representation (AKR) component of the system are given meaning, in terms of Wordnet sysets. A previous version of the system used, instead of the Unified Lexicon, Cyc (Lenat, 1995) concepts as semantics. As discussed in (De Paiva et al., 2007) the sparseness of Cyc concepts was the main reason to move away from Cyc onto a version of the Bridge system based on the UL and WordNet. Since a WordNet-BR is not available, a workaround might be gotten via SUMO, if this were to be available in Portuguese. As a warming exercise we translated the basic concepts used for the basic concept descriptions in SUMO to Brazilian Portuguese [6] and this is already available on the SourceForge repository for Sigma, SUMO's knowledge engineering platform. This is not a substitute for a Brazilian Portuguese WordNet, but merely a stopping stone towards it.

[4] `http://semanticweb.cs.vu.nl/lod/wn30/`.
[5] `http://sigmakee.sourceforge.net/`.
[6] `http://tinyurl.com/bu874aq`

4 Scaling Up?

One of the ways we are considering of scaling up our proposal, from the five thousand concepts suggested in the grid page to the level that we think is necessary for our application goes via the work on YAGO (Suchanek et al., 2007) and, perhaps, YAGO2. The YAGO approach to information extraction for building a searchable, large-scale, highly accurate knowledge base of common facts goes via harvesting infoboxes and category names from Wikipedia for facts about individual entities. It reconciles these with the taxonomic backbone of WordNet in order to ensure that all entities have proper classes and the class system is consistent. The work in YAGO at the Max-PlanckInstitute has led de Melo and Weikum to work in MENTA (Multilingual Taxonomies from Wikipedia), which can be considered a multilingual version of WordNet. From this multilingual version (with 254 languages) we want to 'project' the component consisting of Portuguese synsets only. (The plan is to use the work in the Portuguese version of MENTA to complement, automatically, the five thousand concepts for WN-BR, obtained through manual translation. We believe that the MENTA (de Melo and Weikum, 2010) projection into Portuguese could give us a reasonable basis in terms of synsets in Portuguese to which we would like to compare the existing versions of Portuguese wordnets.) Since YAGO is already integrated with SUMO (De Melo et al., 2008), we hope to be able to maintain consistency of the database.

5 Conclusions

It is early days for our project and time will tell whether our decision to follow the global WordNet grid guidelines for seeding new wordnets will pay off or not, and if so, how well. We have now a master student interested in the project and more interested students are expected. One thing is clear to us, whatever kinds of resource we end up with, we hope to make them freely available in one of the numerous sites (SourceForge, GitHub, etc.) at our disposal nowadays. We are not aware of any such specialized lexicons available for Brazilian Portuguese and it is about time that we had them openly and freely accessible.

References

C. Benzmueller and L.C. Paulson. 2010. Multimodal and intuitionistic logics in simple type theory. *Logic Journal of IGPL*, 18(6):881.

D. Crouch and T.H. King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 504–511. Association for Computational Linguistics. `http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=57`.

G. de Melo and G. Weikum. 2010. Menta: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1099–1108. ACM.

G. De Melo, F. Suchanek, and A. Pease. 2008. Integrating yago into the suggested upper merged ontology. In *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, volume 1, pages 190–193. IEEE.

V. De Paiva, DG Bobrow, C. Condoravdi, R. Crouch, L. Karttunen, TH King, R. Nairn, and A. Zaenen. 2007. Textual inference logic: Take two. *Proceedings of the Workshop on Contexts and Ontologies, Representation and Reasoning*, page 27.

B. C. Dias-Da-Silva, H. R. Moraes, M. F. Oliveira, R. Hasegawa, D. A. Amorim, C. Paschoalino, and A. C. Nascimento. 2000. Construção de um thesaurus eletrônico para o português do brasil. In *Processamento Computacional do Português escrito e falado (Propor)*, pages 1–10.

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

Harald Ganzinger, Alexandre Riazanov, and Andrei Voronkov. 1999. Vampire. In *Automated Deduction CADE-16*, volume 1632 of *Lecture Notes in Computer Science*, pages 674–674. Springer Berlin / Heidelberg.

D.B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Palmira Marrafa, Raquel Amaro, and Sara Mendes. 2011. Wordnet.pt global – extending wordnet.pt to portuguese varieties. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland, July. Association for Computational Linguistics.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM.

Carolina Scarton and Sandra Aluisio. 2009. Herança automática das relações de hiperônimia para a wordnet.br. Technical Report NILC-TR-09-10, USP, Sao Carlos, SP, Brazil.

S. Stamou, K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, and M. Grigoriadou. 2002. Balkanet a multilingual semantic network for the balkan languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.