Embedding NomLex-BR nominalizations into OpenWordnet-PT

Livy Maria Real Coelho
Univ. Federal do Paraná
Curitiba, Brazil
livyreal@gmail.com

Alexandre Rademaker
IBM Research Brazil and EMAp/FGV
Rio de Janeiro, Brazil
alexrad@br.ibm.com

Valeria de Paiva
Nuance Communications
Sunnyvale, CA, USA
valeria.depaiva@gmail.com

Gerard de Melo
IIIS, Tsinghua University
Beijing, China
gdm@demelo.org

Abstract

This paper presents NomLex-BR, a lexical resource describing Brazilian Portuguese nominalizations, and its integration with OpenWordnet-PT. We first describe the original English NOMLEX lexical resource and how we used it to bootstrap a Portuguese version. Subsequently, we describe how this lexicon can be embedded into OpenWordnet-PT, which facilitates its use and helps spot-checking both the bigger integrated resource and the original lexicon. Lastly, we outline some of the other, more substantial work that we plan to engage for the project of using linguistic insights for knowledge representation in Portuguese.

1 Introduction

To help investigate the semantics of deverbal nominalizations, and its implications for Natural Language Processing applications such as electronic ontologies, question answering, or information retrieval, it is useful to have a lexicon of such nominalizations. Our aim, in this paper, is to describe the production and distribution of an open-source, fully available RDF-packaged lexicon of deverbal nominalizations in Brazilian Portuguese, as well as a (still in progress) semantically annotated corpus of examples of these deverbal nouns. More generally we are interested in producing lexical resources for Portuguese that allow us to reason about the semantics of sentences in natural language.

We focus on nominalizations in this work, for several reasons. Deverbal nouns, or nominalizations, can pose serious challenges for knowledge-representation systems. A sentence like "Alexander destroyed the city in 332 BC" can easily be parsed and its semantic arguments, such as the agent of destruction (Alexander), the thing destroyed (the city), and the time (332 BC), are readily obtained for a proposed logical representation of the sentence. By contrast, a sentence like "Alexander's destruction of the city happened in 332 BC" is typically much harder to deal with. It describes the same event of destruction, with the same semantic arguments, but these are much harder to obtain automatically by syntactically parsing the sentence, for most parsers.

Nominalizations have been studied for more than four decades (Chomsky, 1970; Grimshaw, 1990; Alexiadou, 2001). While most of these works describe nominalizations' behavior through a syntactic or morphological point of view, recently, the study of nominalizations has focused also on semantic and ontological phenomena (Hamm and Kamp, 2009; Real and Retoré, 2013). With regard to lexical studies, deverbal nouns are particularly well-studied in English, with the NOMLEX project (Macleod et al., 1998) providing a well-established, open access baseline for corresponding results in other languages. Our work on NomLex-BR builds up from previous work on nominalizations in English (Gurevich et al., 2006). This previous work extended the coverage of NOMLEX's English norminalizations, via the use of Xerox PARC's state-of-the-art NLP system XLE (Maxwell and Kaplan, 1996) and some simple, but effective heuristics, and compared it to NOMLEX-PLUS (Meyers et al., 2004), the state-of-the-art in 2004. Our work here is an attempt at building the basic blocks underlying that work on nominalizations, for Portuguese. Our assumption was that the work done for English can be suitably adapted and re-used for Portuguese, if we keep the language comparisons in place. Additionally, we hoped to learn and adapt from the French experience with nominalizations, described in the No-

mage project (Balvet et al., 2011).

The original version of NOMLEX is a small resource of only around a thousand nominalizations, which seemed ideal to be used as a basis for our project. The original NOMLEX was constructed starting out with nominalizations with the suffixes *-ion*, *-ment* and *-er*, taking samples of the most frequent words in a list of nouns from a combination of the Brown Corpus and the Wall Street Journal (about 1 million words of each). Words with these kinds of suffixes tend to be erudite words and these tend to work similarly in different (but related) languages. This was the original working hypothesis, which seems confirmed, to some degree, by our prototype.

To construct our lexicon, we first translated the easiest nominalizations into Portuguese, such as *construction/ construção* and *writer/escritor*, in order to keep, to the extent possible, the "same" lexical items from NOMLEX into NomLex-BR. Our methodology provided for a fast and reliable creation of a lexical resource for Portuguese, which hereafter can work as a basis to discuss the behavior of nominalizations in general. In English, for example, many of the nominalizations with eventive readings cannot be pluralized, *confusion* and *abandonment*, e.g., lack plurals. This does not occur in Brazilian Portuguese with *confusão/confusõe**s*** and *abandono/abandono**s***. Using a lexical resource that covers the same range of words as a previously existing English one, insightful comparisons, as the inter-language relation of the nominalizer morphemes and the syntactic behavior of those nominals, can be observed more easily.

However, for comparative studies a simply text file is not very easy to use. We thus finally embedded NomLex-BR into a lexical-semantic resource called OpenWordnet-PT (de Paiva et al., 2012a), greatly facilitating search and experiments.

2 NomLex-BR

The original NOMLEX is a lexicon of English nominalizations developed at New York University over many years. It relates the arguments of a nominalization to the predicate argument structure of its associated verb, without requiring exactly the same structure for the nominal and the verbal lexical items. It also records details of the syntactic realization of the arguments, including prepositions associated with the arguments.

Unfortunately, lexical resources for languages other than English are notoriously difficult to come by. We are involved with the creation of a Portuguese WordNet freely available for download and modification by anyone OpenWordnet-PT (de Paiva et al., 2012b). To follow the traditional pipeline for natural language understanding systems, e.g. the one described by the Bridge system of PARC (Bobrow et al., 2007), we need a collection of lexical resources as well as (much as possible) off-the-shelf systems. Ideally we would want to have a broad coverage, deep processing LFG grammar of Portuguese and while we are pursuing leads in this direction (de Alencar, 2013), this may take a while, as hand-crafted large coverage grammars are very labor-intensive. In the meantime, it seemed sensible to construct some of the resources that we are most familiar with, and a small version of NOMLEX for Portuguese, NomLex-BR, seemed to be an ideal starting point.

Our Portuguese version keeps the original structures of the English version of NOMLEX, but apart from the translated nominal and verb, adds an extra field to capture usage examples in Portuguese.

Our initial pass of translating word pairs in NOMLEX by two linguists was enough to yield direct translations of around 90% of the original resource. This high rate of correspondence resulted not only from the words in NOMLEX being somewhat erudite, but also from the fact that the inter-language relations established by the nominalizer morphemes are quite straightforward. For example, *adjourn**ment**/adia**mento***, *benefici**ary**/benefici**ário***, *corro**sion**/ corro**são***.

The NomLex-BR lexicon has more than a thousand entries, mostly nominalizations formed by *-ção*, *-mento* and *-or*, as these are the corresponding suffixes to the ones adopted by the NOMLEX project. One next goal is to introduce nominalizations formed by *-ura*, considering the description proposed by (Real, 2008). We also aim to add nominalizations formed by the suffix *-ada*. These seem to require more analysis as *-ada* produces nominals from verbs (*cutucada*) and from nouns (*pedrada*) and the semantics of these nominalizations is far from obvious, as discussed by (de Medeiros, 2008).

3 Evaluation of NomLex-BR

We considered several ways of evaluating our resource and different criteria to do so. Since it is hand-constructed, via two experts, its accuracy is high enough for the nominal-verb pairs that it covers. The main challenge is its coverage and representativeness of the nominalizations in place. This was addressed by increasing the coverage and by checking the representativeness using a small corpus of biographical data. As we are working with a corpus of biographies of Brazilian historical figures (DHBB) (Abreu et al., 2010), we listed the most frequent nouns in the texts of this corpus and marked them as being nominals or not. This revealed a lack of *agentive* nominals in our initial resource that was then corrected.

A second kind of evaluation and extension we performed was comparing our resource with Nomage (Balvet et al., 2011), a similar project for the French lexicon. The Nomage corpus covers 736 nominals and 679 verbal lexemes extracted from the French Treebank (Abeille et al., 2003). Its nominalizations, annotated syntactically and semantically, are formed by the suffixes *-ade*, *-age*, *-ance/ence*, *-ee*, *-ion*, *-ment* and *-ure*. The Nomage project and ours share a similar goal, to study the inheritance of semantic and aspectual features from the verbal bases, but the Nomage lexicon was produced combining two different methodologies – "one based on transformation tests applied on real-life sentences by naive annotators, the other based on forged sentences applied by linguistically trained annotators" (Balvet et al., 2011, p. 04). We compared all of Nomage's entries with NomLex-BR. It turned out that many of the nominalizations in Nomage had to be added to NomLex-BR, somewhat contrary to our expectations that NOMLEX and NOMAGE had a big intersection.

In the future, we would like to compare the structural descriptions on each nominalization/verb pair (for example *construction/construção*) to check in which way the linguistic relations established by nominalizations with their verb bases are the same in English and in Portuguese. This kind of evaluation requires further annotation to describe the kinds of nominalizations in Portuguese. These are interesting problems on their own and we hope to report interesting results as the project progresses.

4 Embedding NomLex-BR into OpenWordnet-PT

Finally, we integrated NomLex-BR into OpenWordnet-PT, a version of WordNet for Brazilian Portuguese. Its main characteristics are its open-source license, its direct correspondence with Princeton WordNet, and, given its origins in the Universal WordNet(de Melo and Weikum, 2009), both a high recall and a high precision for the more salient words in the language.

Our choice of encoding OpenWordnet-PT in RDF makes the merging of these resources very straightforward. The details of this encoding are described elsewhere in great detail (Rademaker et al., 2014). In order to incorporate NomLex-BR into this encoding, we extended the RDF-based vocabulary to additionally describe relevant parts of the NOMLEX syntax (Macleod et al., 1999). Figure 1 presents a subgraph for the nominalization entry *promover/promoção* and its connection to OpenWordnet-PT. Note that the link between NomLex-BR and OpenWordnet-PT is achieved through the properties `noun` and `verb`. Both properties have as domain an instance of `Nominalization` and as co-domain an instance of `WordSense` (from the OpenWordnet-PT vocabulary).

This embedding of NomLex-BR into the open version of a Portuguese wordnet was helpful in multiple respects. First, it solved some minor problems with handling diacriticals, as OpenWordnet-PT has a consistent treatment of these. Secondly by checking how the nominalizations from NomLex-BR were related to the corresponding verbs in the wordnet version, we realized that some synsets were missing in the OpenWordnet-PT. These are in the process of being added manually. Finally, by re-checking the original English NOMLEX connections, we hope to spot-check the consistency of OpenWordnet-PT with respect to other specific phenomena, for example, the phenomenon of diminutivization of nominals.

5 Preliminary Conclusions

This lexicon of deverbals is just a first step for our lexical resources. It would be useful to include nominalizations of adjectives and of other nouns, which also need a common concept mapping for knowledge representation. Examples here would be the nominals "selvageria" or "bruxaria"

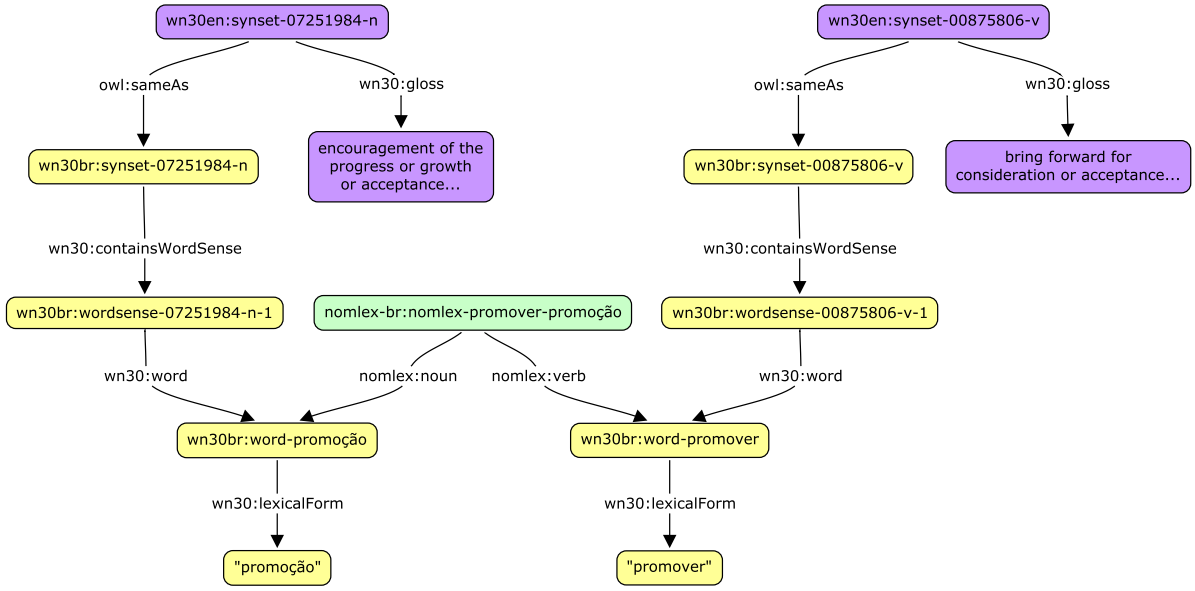

Figure 1: Entry *promover/promoção*

from the adjective "selvagem" (wild) and the noun "bruxa" (witch). Another future plan is to produce capture verb semantics and in particular verb alternations, as covered by VerbNet (Kipper et al., 2006) for English. Again, there is hope that some of the original Levin classes used for the construction of VerbNet are also valid for Portuguese.

In summary, we believe that the creation of linguistic resources requires openness of programs and of code. The only way to keep alive any resource is to make sure that people can modify it for their own purposes. If one wants the enterprise of automatic language understanding to flourish, especially in languages with fewer resources, one must make sure that the lexical resources we develop are freely available, freely modifiable and easy to use. Making our small lexicon NomLex-BR part of OpenWordnet-PT and having it downloadable, freely available from `http://github.com/arademaker/wordnet-br/` and easy to consult, we hope to make it more interesting for researchers interested in nominalizations in Portuguese.

We also wish to develop other resources for Portuguese along these same lines and we hope to work both from small hand-crafted lexica and from big machine learned ones (like OpenWordnet-PT and FreeLing-PT) to try to obtain better quality resources. Keeping these resources usable and as much as possible theory-neutral is our challange.

References

Anne Abeille, Lionel Clément, and Francois Toussenel. 2003. Building a treebank for french. In Anne Abeille, editor, *Treebanks, Building and Using Parsed Corpora*, pages 165–187. Kluwer, Dordrecht.

Alzira Alves Abreu, Fernando Lattman-Weltman, and Christiane Jalles de Paula, editors. 2010. *Dicionário Histórico-Biográfico Brasileiro pos-1930*. CPDOC/FGV, 3 edition. `http://cpdoc.fgv.br/acervo/dhbb`.

Artemis Alexiadou. 2001. *Functional structure in nominals. Nominalization and ergativity*. John Benjamins.

Antonio Balvet, Lucie Barque, Marie-Hélène Condette, Pauline Haas, Richard Huyghe, RafaelMarín, and Aurélie Merlo. 2011. Nomage: an electronic lexicon of french deverbal nouns based on a semantically annotated corpus. In *Proceedings of the First International Workshop on Lexical Resources*, pages 8–15, Ljubljana, Slovenia.

Daniel G. Bobrow, Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy H. King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. 2007. PARC's bridge and question answering system. In *Proceedings of Grammar Engineering Across Frameworks*, pages 26–45.

Noam Chomsky. 1970. Remarks on nominalization. In *Readings in English transformational grammar*. Blaisdell.

Leonel Figueiredo de Alencar. 2013. Brgram: uma gramática computacional de um fragmento do português brasileiro no formalismo da lfg. In *Proceed-*

ings of the 9th Brazilian Symposium in Information and Human Language Technology, Sociedade Brasileira de Computação, pages 183–188, Fortaleza, CE, Brazil.

Alessandro Boechat de Medeiros. 2008. *Tracos Morfossintáticos e Subespecificacão Morfológica na Gramática do Português: Um Estudo das Formas Participiais*. Ph.D. thesis, UFRJ.

Gerard de Melo and Gerhard Weikum. 2009. 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. ACM. http://doi.acm.org/10.1145/1645953.1646020.

Valeria de Paiva, Alexandre Rademaker, and de Gerard Melo. 2012a. OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 353–360. See at http://www.coling2012-iitb.org (Demo Paper). Published also as Techreport http://hdl.handle.net/10438/10274.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012b. Openwordnet-pt: an openbrazilian wordnet for reasoning. Technical report, FGV - EMAp.

Jane Grimshaw. 1990. *Argument structure*. The MIT Press.

Olga Gurevich, Richard Crouch, Tracy Holloway King, and Valeria de Paiva. 2006. Deverbal nouns in knowledge representation. In *Proceedings of the 19th International Florida AI Research Society Conference (FLAIRS'06)*, pages 670–675, Melbourne Beach, FL, May. AAAI Press.

Fritz Hamm and Hans Kamp. 2009. Ontology and inference: The case of german ung–nominals. *Disambiguation and Reambiguation*, 6:1–67.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending verbnet with novel verb classes. In *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1027–1032, Genoa, Italy, June.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barret, and Ruth Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex 1998*, pages 187–193, Liege, Belgium.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barret, and Ruth Reeves, 1999. *Manual of NOMLEX: The Regularized Version*.

John Maxwell and Ron Kaplan. 1996. An efficient parser for LFG. In *Proceedings of the First LFG Conference*, CSLI Publications.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronkia Zielinska, and Brian Young. 2004. The cross-breeding of dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.

Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Real, and Maira Gatti. 2014. OpenWordNet-PT: A project report. In *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan. to appear.

Livy Real and Christian Retoré. 2013. A generative montagovian lexicon for polysemous deverbal nouns. In *Handbook of the 4th World Congress and School on Universal Logic*.

Livy Real. 2008. Uma análise do sufixo -ura com base na morfologia categorial. *Revista InterteXto*, 1.