NOMLEX-BR: A Five Finger Exercise in Lexical Resource Creation

Valeria de Paiva Alexandre Rademaker

Jun, 2011

Abstract

We describe a Brazilian Portuguese version of the original English NOMLEX lexical resource created by the project Proteus [8]. We first describe NOMLEX, then we explain why we believe a Brazilian Portuguese version (which we call NOMLEX-BR), should be useful, in the context of a recently started project on Knowledge Representation in the FGV. Then we outline some of the other, more substantial work that we plan to engage in for the project.

1 Introduction

Our aim is to discuss the production and distribution of an electronic, small, lexicon of nominalizations in Brazilian Portuguese, as well as a semantically annotated corpus of examples of these deverbal nouns.

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 be easily 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 much harder to deal with. It describes the same event of destruction, with the same semantic arguments, but these arguments are harder to obtain automatically from a syntactic parsing of the sentence.

Nominalizations are well-studied in English, with the NOMLEX project ([8]) providing an well-established, open access baseline for corresponding results in other languages. Our work in NOMLEX-BR builds up from previous work on nominalizations in English [6]. This work extends the coverage of NOMLEX English norminalizations, via the use of Xerox PARC's state-of-the-art natural language processing system XLE [9] and some simple, but effective heuristics and compared it to NOMLEX-PLUS [10], the state-of-the-art in 2004. Our work is here is an attempt at building the basic blocks underlying that, for Brazilian Portuguese. We hope that the work done for English can be suitably adapted and re-used for Portuguese, if we keep the languages comparisons

in place. We also hope to learn and adapt from the French experience with nominalizations, described in the Nomage project [2].

The original version of NOMLEX, is a small resource, only around a thousand nominalizations, which seemed ideal to kick off a new collaborative project between the investigators, who are trying to work together in a field (lexical resource creation) that turns out to be new to both.

The original NOMLEX was constructed starting out with nominalizations with the -ion, -ment and -er suffixes, taking samples of the most frequent words first 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 suffix tend to be erudite words and these tend to work similarly in different (but related) languages, was the working hypothesis, which seems confirmed, to some degree by our (admittedly very small) prototype.

2 NOMLEX

NOMLEX is a lexicon of English nominalizations developed by the group at New York University for many years. It relates the arguments of a nominalization to the predicate argument structure of its associated verb, but it does not require exactly the same structure for the nominal and the verbal lexical item. It also records details of the syntactic realization of the arguments, including prepositions associate with the arguments. For example, the entry for ”promotion” in NOMLEX reads:

```
(nom :orth promotion
     :verb promote
     :nom-type((verb-nom))
     :verb-subj ((n-n-mod) (det-poss))
     :verb-subc ((nom-np :object ((det-poss)(n-n-mod)(pp-of)))
 (nom-np-as-np :object ((det-poss) (pp-of)))
 (nom-possing :nom-subc ((p-possing :pval (of))))
 (nom-np-pp :object ((det-poss) (n-n-mod) (pp-of))
    :pval (into from for to))
 (nom-np-pp-pp:object ((det-poss) (n-n-mod) (pp-of))
      :pval (for into to) :pval2 (from))))
```

Our Brazilian Portuguese version keeps the original structures of the original English version of NOMLEX, but adds an extra field corresponding to some usage example in Portuguese. This is usually called a ‘gloss’ in WordNet[5]. Glosses for NOMLEX-BR were obtained using the ‘Corpus do Português’ [4].

3 Why NOMLEX-BR?

We recently started a project called “Logics and Ontologies for Brazilian Portuguese” whose ultimate aim is to represent, in a suitably described logic, the meanings of sentences in Brazilian Portuguese. Given that this work is to be conduced at distance (one

of us is in California, the other in Brazil) and given that this is somewhat a labor of love and not many resources are allocated to it, it makes sense to build up our systems in smaller chunks.

Lexical resources for languages other than English are notoriously difficult to come by. The fact that there is not even a version of a Portuguese WordNet freely available for download and modification by anyone is a clear indication of the difficulties ahead. To follow somewhat the traditional pipeline for logic based systems based on language, e.g. the one described by the Bridge system of PARC ([3]) 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 Brazilian Portuguese and while we are pursing leads in this direction ([1]), this may take a while to get, as hand-crafted large coverage grammars are very labor intensive. Meanwhile we thought we would experiment with off the box Portuguese parsing in the style of the Stanford parser, adapted by the the group at the University of Lisbon, led by Prof Antonio Branco ([12]). At the same time, it seems sensible to construct ourselves some of the resources that we are more familiar with, and a small version of NOMLEX for Portuguese, NOMLEX-BR seems just the ideal small project to get things rolling.

Another somehow indirect route that we are taking towards our goal is to consider a generic, open source ontology like SUMO/Sigma [11] and trying to adapt it to Portuguese concepts.

4 Knowledge Representation and Textual Inference

Knowledge representation languages tend to be based on concepts, usually denoted by nouns. Verbs, which tend to represent processes or events, are as important as nouns, if not more so, but somehow it is less clear how to deal with them in traditional logic-based knowledge representation. The ontologists or model constructors normally decide which concepts they consider the most productive ones for a given domain. They also decide how to represent the most productive concepts via a judicious mixture of processes and events. Eventually whether using first-order or higher-order logic or description logic or modal logics or a combination of all the above, one ends up with a collection of predicates that seem very *ad hoc*.

Given that there are many different frameworks for knowledge representation and that the reasons for choosing particular features of frameworks are very varied too, it is a hard task to compare frameworks. To decide whether a representation in logical form of a sentence in one framework is better than another is also a difficult task. Of course once a whole theory has been formalized in a given framework as a mathematical object, one can do the usual things that one does with a logic, compare it to other logics, prove the usual traditional theorems, etc. But measuring the *adequacy* of your logical formalization when compared to the raw phenomena you started with is a hard job. If the phenomena you've started from is described in a collection of sentences, maybe it is easier to consider some notion of *textual inference* in the original collection of sentences.

Textual inference is an informal relationship between two pieces of text where the

first text (the premiss P) is supposed to entail the second text (the hypothesis H).That is, the second text H follows *logically*, but not necessarily formally from te the first text (the premise P). Once the content of the texts has been formally rendered as pieces of logic, we expect the corresponding logical expressions to entail in the same direction $P \rightarrow H$.

The same way using textual entailment bypasses some of the problems of deciding whether a particular rendering in logic of a sentence is or is not adequate, the use of large scale knowledge representation systems together with textual entailment tasks should help us decide which predicates are the most useful ones. The comparison between deverbal notions of concepts and their corresponding verbal versions also should help with the required choice between predicates.

5 Further Work

This small lexicon of deverbals is just a first step. Certainly it would be useful to have similar versions of nominalizations of adjectives and adverbs, which also need a common concept mapping. Also in the immediate plans is a version of VerbNet-BR, as the syntactic alternances captured by the original VerbNet [7] correspond to useful semantical information. Again there is some hope that some of the original Levin classes used for the construction of VerbNet are also valid in Portuguese, but this is mostly a hope, so far.

Summarizing the creation of linguistic resources requires openness of programs and of code. The only way to keep alive some/any resource is make sure that people can modify it to their own purposes, be they commercial or not. If one wants the enterprise of automatic language understanding to flourish one must make sure that lexical resources exist, are freely available and easy to use.

References

[1] L. F. de Alencar. Complementos verbais oracionais – uma anlise lxico-funcional. *Lingua(gem)*, 1(1):173–218, 2004.

[2] A. Balvet, L. Barque, M. H. Condette, P. Haas, R. Huyghe, R. Marn, and A. Merlo. Nomage: an electronic lexicon of french deverbal nouns based on a semantically annotated corpus. In *Proceedings of the First International Workshop on Lexical Resources*, Ljubljana, Slovenia, 2011.

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

[4] Mark Davies and Michael Ferreira. Corpus do português: 45 million words, 1300s-1900s. available at `http://www.corpusdoportugues.org`.

[5] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT press, 1998.

[6] O. Gurevich, D. Crouch, T. H. King, and Valeria de Paiva. Deverbal nouns in knowledge representation. *FLAIRS: The Florida Artificial Intelligence Research Society*, May 2006.

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

[8] Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barret, and Ruth Reeves. Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex 1998*, Liege, Belgium, 1998.

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

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

[11] I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19, Ogunquit, Maine, October 2001. See also `http://www.ontologyportal.org`.

[12] João Silva, Antônio Branco, Sérgio Castro, and Ruben Reis. Out-of-the-box robust parsing of portuguese. *Lecture Notes in Artificial Intelligence*, pages 86–89, 2010.