Improving the Verb Lexicon of OpenWN-PT *

Valeria de Paiva[1], Claudia Freitas[2], Livy Real[3], and Alexandre Rademaker[4]

[1] Nuance Communications, USA
[2] PUC-Rio, Brazil
[3] IBM Research, Brazil
[4] IBM Research and FGV/EMAp, Brazil

Abstract. This preliminary account of our work on improving the verb lexicon of OpenWordNet-PT describes some of the issues that one faces when manually cleaning up a semi-automatically constructed lexical resource and some of the lessons we learned while doing it.
Keywords: wordnet, verb lexicon, openwordnet-pt

1 Introduction

This note discusses and implements a series of small improvements to the verb lexicon that is part of OpenWN-PT [1, 2], our open source and freely available Portuguese WordNet. The OpenWordNet-PT is automatically created using multilingual lexica, hence it is not surprising that it has some strange expressions within its synsets. Some are just mistakes or typos; some are design decisions that are sensible when constructing multilingual thesauri, but that linguistic work in a single language can improve upon; and some we simply do not know their origin.

The Universal WordNet (UWN) [3] approach we followed has been to be permissive and include everything that can be found in electronic dictionaries: it prefers not to distinguish support verb constructions from full verbs, or single word verbs from compound verbs, thus *estar de cócoras/to be in a squat* is considered a perfectly valid verb. Since the UWN deals with around 200 languages, the difference between single word verbs and compound verbs is not relevant and there is no clear criteria for what is to be considered a compound verb, except the algorithmic one. This produces compound 'verbs' that do not look like or behave like verbs in Portuguese.

In this note we look at these problematic examples of expressions in verb synsets, then we discuss how to solve the issues, to improve the verb lexicon to, at least, the stage of a clean baseline.

2 Problems with Verbs

To uncover the problems besetting the verb lexicon of OpenWN-PT, we listed the full collection of expressions within verbal synsets in a spreadsheet and had three

* We would like to thank Bianca Freitas for help with the translation and analysis of the verbs in VerbOcean.

native speakers look over the list, marking issues that needed to be discussed. In previous work [2] we noted that a main goal of the OpenWN-PT is to allow reasoning with language, and this perspective influenced some of our decisions on what to consider a problem. We first discuss these issues informally, then we explain our 'solutions', concluding with further work.

2.1 What's in a Verbal Synset?

The first issue is the presence of parts of speech other than verbs within a verbal synset. For example the synset 01978576-v, corresponding to the English verb *come down*, in Portuguese has as representatives *aceso/lit, descer/get down, desembarcar/alight, pousar/land*. While *descer, desembarcar, pousar* are good verbs, the synset has an extra word, the past participle *aceso* that should not be there. This is easy to correct, as we can simply remove the extra word from the synset.

In a verbal synset we expect only verbs in Portuguese and only in the infinitival form, but many synsets in the automatically created lexicon do not satisfy this condition. Some are really verbs, but inflected, e.g the synset for the verb *run away* (02073714-v) in Portuguese consists of four expressions *fugir, evada-se, esconder-se, escapar*. While *fugir/run away, esconder-se/hide, escapar/escape* are verbs in their infinitive form, *evada-se* is the imperative of *evadir-se*. Again this is easy to correct, as we can transform inflected forms of verbs into their infinitival version. But it points to a more complex problem, when should we list verbs with the pronominal particle *-se*?

2.2 Verbs and the *-se* particle

We noticed that some verbs that can be used with the *-se* particle were *only* appearing as such in OpenWN-PT. This is not sensible, as when searching for words, people would not think of adding particles/prepositions to the words. Thus looking for *insurgir/insurrect* would give no results in our web interface, but the verb *insurgir-se* is in the lexicon. Verbs should appear in the lexicon in their lemmatized form, even if we know that they are mostly used with a given particle. The information that the verb can be mostly pronominal is very valuable, and we want to keep it, but it is also important to list these verbs in the lexicon as single lemmas. Luckily there were only a few verbs in this situation. Besides fitting these few verbs into our hierarchy, it would be nice to check whether there are meaning differences between the pronominal and the non-pronominal forms of a given verb.

There is a substantial difference of meaning between *admirar* and *admirar-se* – the first means to admire something, and it may be reflexive, while the second means to be surprised by something. However, in many other cases the difference is not so clear. In *Pedro mudou de emprego* vs. *Pedro mudou-se*, the latter seems to mean *Pedro moved [house]* while the former seems to say something like *Pedro changed jobs*. Is this difference clear enough to support splitting the verb

mudar/mudar-se into two synsets? We believe so, but postpone this task for the time being.

To postpone this large task of looking over all verbs with *-se* and deciding whether the meaning with the particle is 'different enough' from the meaning without it, we opted for keeping all verbs that the automatic procedure produced with the particle *-se*, but we also add to the lexicon these same verbs, without the particle, whenever they were not present. (There were 34 verbs in this situation.) The guideline is clear: all verbs that are present in pronominal form, should also be present without the (constitutive or not) particle *-se*. This means that verbs like *queixar/complaint* were added to synsets, and with this decision we are in line with regular dictionaries (at least, Portuguese ones). This is a preliminary step, in the future we will distinguish meanings, informed by corpora.

2.3 Light Verbs

We also noticed the need for a theory of support verbs or light verbs or *copula verbs*, in order to deal with verbal expressions such as the ones in the synset for *squat* (e.g synset 02725562-v,*estar de cócoras, estar agachado*). In this case, since we have also synset 01545314-v *acocorar, baixar-se, ficar de cócoras, acocorar-se, agachar* (sit on one's heels) the meaning is already covered via a single word verb:*estar de cócoras* is the same as *acocorar*. But if you consider the verb 'to campaign' in English, synset 01094086-v *run, campaign*, the definition reads *(run, stand, or compete for an office or a position)*. As a single verb this does not exist in Portuguese, instead one tends to use the expression *fazer campanha/to run a campaign*. Maybe this kind of compound should be included in the lexicon?

There are many of these compound verbal expressions in Portuguese and while we do not want to list every single one in our lexicon, we would like to have the ones that correspond to well-identified synsets. Many of the compound verbs automatically obtained (e.g. *beijar ruidosamente/kiss with a big noise*) do not seem to deserve being considered a verb.

2.4 Typos and Idioms

Typographical mistakes are always present in automatically created resources, and we found 28 mistakes that were relatively easy to correct. Examples include *annimar-se, applaudir*, whose correct spelling is *animar-se, aplaudir*. Some 'errors' are more complicated, as they might correspond to European Portuguese usage that we are not familiar with or they might correspond to Catalan or Spanish that the automated system mistook for Portuguese.

Idioms are even more complicated. The criteria for defining idioms are not very clear and the way to represent them in the lexicon seems even less so. But the automatic process of creation of the OpenWN-PT produces them, so we need to find a way of marking them within the synsets and of removing the multiple versions that arise due to lack of morphological processing of the input. For example we have verbal expressions *bater as botas, bater a bota, bater_as_botas*

all meaning the same, to die, clearly a processing mistake: we should have only *bater as botas*, the usual idiom.

2.5 Completing the Verb Lexicon?

A reasonably complete verb lexicon is essential for reasoning with language. We would like to make sure that we possess coverage of the verbs used in colloquial Portuguese, but would also like to make sure that verbs associated with simple inferences such as the ones described in [4] are present in our lexicon.

To this end we automatically translated the list of verbs provided by VerbOcean and compared the inventory of Portuguese verbs from VerbOcean, manually translated, to OpenWN-PT.VerbOcean is a broad-coverage semantic network of verbs, which uncovers semantic relations (similarity, strength, antonymy, enablement, and temporal relations) between verbs in English. Since we are mostly interested in semantic relations between verbs, having a translated version of VerbOcean seems an useful resource. But clearly we are just starting our work on that.

We also decided to compare our OpenWordNet-PT verb lexicon with a verb lexicon obtained from a Portuguese corpus, the PropBank.BR verbs. The PropBank.BR verbs are based on the Brazilian portion of the Bosque corpus, a subset of Floresta Sintá(c)tica [5]. The verbs in PropBank.BR are around 16K and the ones in OpenWN-PT are around 4K, so we have plenty of work to do to complete our verb lexicon.

3 Discussion: methods and results

The main objective of this small experiment with verbs was to examine the verb lexicon, to clean up egregious mistakes of the automatic processing and to decide on criteria for identification of better verb synsets. Our methods and solutions were different for each of the different issues uncovered.

First we identified expressions in the verbal synsets that are not verbal expressions (either simple or compound). This was relatively easy to solve, as few synsets were involved.

Second the issue of Catalan or Spanish 'bleeding' into Portuguese synsets is also relatively easy to solve, via search in online dictionaries. Solving the issue is, however, time-consuming and somewhat error-prone, as we strive for vocabulary that is common usage in Brazil. Detecting what is common usage, as opposed to what is 'dictionarese' (exists in the dictionary but no one uses) is hard: an example would be *mondar*, a portuguese verb that exists in the dictionaries, but which would be useless for a common speaker of Portuguese, which happen not to know the meaning of *weed*.[5]

Third our list of 'problematic synsets' showed us that the pronominal form of verbs in Portuguese is a controversial issue. We decided to take a conservative

[5] *To weed* means to remove the *weeds/ervas daninhas* from the garden.

stance and instead of facing the difficult problem of classifying the various phenomena involved with the particle *-se* in Portuguese verbs, we opted for keeping all the information provided by the automatic processing, whether or not we agreed with the specific instances. As described for instance in Duran et al [6], there are various issues intertwined when trying to classify *-se* in Portuguese and given that the statistically based methods for the creation of OpenWordNet-PT gave us some particle *-se* usage information, we decided to only add all the verbs that were not listed without a *-se* particle, but keep all the infinitival verbs that appeared in the lexicon with a *-se*. Later on we plan to use our work with a Portuguese corpus to try to detect pronominal uses of *-se* as compared to 'reflexive' uses of it. We are also investigating the cases where the use of the particle *-se* changes the meaning of the verb, as compared to the version without it.

The fourth and most serious problem we encountered was the problem of deciding which compound expressions should be considered a composite verb in our lexicon. Again our conservative stance was to try and keep as many of the composite expressions as we could justify for ourselves. We focused on expressions headed by light verbs, due to their high frequency in the Portuguese language. We looked at the verbal multiword expressions (MWE) lists provided by Garrão [7], created using a statistically-based corpus analysis. However, since we did not want to inflate our lexicon, we devised the following strategy: if the verbal MWE in question could be paraphrased by a single verb either in Portuguese (preferably) or in English, it should be considered a composite verb and it should be added to the corresponding synset, together with the single-word synonymous verb. As a result, combinations such as *fazer um teste/to take a test* and *fazer a diferença/to make a difference* are aligned with *testar/to test* and *sobressair/to highlight*, respectively; while *fazer amigos/to make friends* and *fazer justiça/to do justice* are not part of a synset (these four combinations are listed in Garrão [7]). It is worth noting that, with this decision, we are not claiming that only the first combinations are multi-word expressions. At this stage, we are just focusing on verbal combinations that fit well with the existing synsets. There is a rich literature on support verbs in Portuguese, but instead of solving the problem of representing these verbs, we merely want to make sure that we have a minimum criteria for acceptability of composite expressions in our derived lexicon.

Finally, having compared verbs translated from VerbOcean [8] to the verbs from OpenWordNet-PT, we realized that we were missing a considerable number of verbs in OpenWordNet-PT. We must now devise ways of fitting these missing verbs into our lexical ontology.

References

1. Rademaker, A., de Paiva, V., de Melo, G., Real, L.: OpenWordnet-PT: a progress report. In Orav, H., Fellbaum, C., Vossen, P., eds.: Proceedings of the 7th Global WordNet Conference, Tartu, Estonia (2014) 378–382
2. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An open Brazilian wordnet for reasoning. (2012)

3. de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: Proc. of CIKM 2009, New York, USA, ACM (2009) 513–522
4. de Melo, G., de Paiva, V.: Sense-specific implicative commitments. In: Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014). LNCS, Springer (2014)
5. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, Thicker and Easier. In Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P., eds.: Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008). Volume Vol. 5190., Springer Verlag (Sep 2008) 216–219
6. Magali Sanches Duran, Carolina Evaristo Scarton, S.M.A.C.R.: Identifying pronominal verbs: Towards automatic disambiguation of the clitic 'se' in portuguese. In: Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), Atlanta, Association for Computational Linguistic. (2013) 93–100
7. Garrão, M.: O corpus não mente jamais: sobre a identificação e uso de combinações multivocabulares do tipo verbo mais sintagma nominal
8. Chklovski, T., Pantel, P.: Verbocean: Mining the web for fine-grained semantic verb relations. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04), Barcelona, Spain (2004)