Extending the adverbial coverage of OpenWordnet-PT

Priscila Côrtes[1], Mateus Riva[2], and Livy Real[3]

[1] Samsung Institute of Development for Informatics (SIDI) – Brazil
[2] Institute of Mathematics and Statistic (IME) – University of São Paulo, Brazil
[3] Group of Computational Linguistics (GLiC) – University of São Paulo, Brazil
po.cortes@sidi.org.br, mriva@ime.usp.br, livyreal@gmail.com

Abstract. We describe our work on extending and completing the Open-Wordnet-PT, an open wordnet for Portuguese, using data from the Bosque corpus. We used the Bosque-UD, the Universal Dependencies version of the Bosque and checked the adverbial coverage of the OpenWordnet-PT over the corpus. We found that 123 adverbs, including adverbial phrases, from the Bosque were missing in the OpenWordnet-PT, and we manually added 104 to the correct synsets. We point out the patterns we found and discuss why some of these adverbs were not added to OpenWordnet-PT. We also found minor mistakes in the Bosque-UD and discuss possible reasons for them.

Keywords: Wordnet · Adverb · Portuguese · Quality Assurance · Open Lexical Resources.

1 Introduction

The goal of this work is to make sure all of the adverbs in the Bosque Corpus are represented in the OpenWordnet-PT (OWN-PT). Although the OWN-PT has been under development since 2012, the resource still lacks many frequent words of Portuguese. Even for processing one of the most used open Portuguese corpora, namely the Bosque corpus, OWN-PT does not offer all the required Portuguese words.

Previous attempts of automatically completing wordnets have shown that it can dramatically decrease the quality of the resource [10]. It was also shown that a layperson's effort is not always helpful in completing a linguistic resource keeping the aimed quality. Given the fact that the Princeton WordNet was manually created, and the OWN-PT has been going through a manual curating process for years, we aim to make sure that the information added to the OWN-PT also has a high quality.

Completing a lexical resource is a very time consuming process, specially considering that we would like to have language specialists curating these data. Thus, we decided to focus on a small set of words to be added to the OWN-PT, following previous work such as [13]. The contribution of this project is twofold:

extending the coverage of the OWN-PT and helping to improve the quality of the Bosque-UD corpus.

In this paper, we briefly introduce the resources used, namely the OWN-PT and the Bosque-UD. Then we discuss why and how we checked the adverbs coverage of the Bosque in the OWN-PT. We present some results and discussion about the work done, in order to set some new steps to be followed so we can continue the present work.

2 The OpenWordnet-PT

The OpenWordnet-PT is a freely available wordnet for Portuguese [9]. It was originally developed as a syntactic projection of the Universal Wordnet [6], using machine learning techniques that create relations between graphs representing lexical information from Wikipedia entries and open electronic dictionaries. The OWN-PT has been continuously improved through linguistically motivated additions and removals, either manual or semi-automatic, making use of different corpora.

OWN-PT is aligned to the Princeton WordNet (PWN) and to the Open Multilingual Wordnet project[4] [1], that is to say, it has the same synsets that PWN has and is linked to more than 30 wordnets for different languages. However, many Portuguese synsets are empty or lacking information: words, examples or glosses. Several works have been done in order to complete the resource, such as [8]. Since the OWN-PT is the wordnet used in several projects, such as Freeling [7] and Google Translate[5], it makes sense to pay attention to the relevant synsets that still lack Portuguese translation. For now, we focus on completing the lemmas of adverbs.

3 The Bosque

Created by the Linguateca Team[6], the Bosque is one of the most used open Portuguese corpora. It contains around 10,000 sentences extracted from Brazilian and Portuguese newspapers. It has undergone many manual revisions, and is available with different annotations, including a version annotated within the Universal Dependencies framework. Because it is one of the most used corpora and has so many resources, it makes sense to make sure the Bosque can be processed using the OWN-PT.

We used the version Bosque-UD [11] since it is part of the data used in two shared tasks for Universal Dependencies parsing, namely CoNLL 2017 Shared Task[7] and CoNLL 2018 Shared Task[8]. This version was firstly automatically

[4] http://compling.hss.ntu.edu.sg/omw/
[5] https://translate.google.com/intl/en/about/license.html
[6] https://www.linguateca.pt/Floresta/corpus.html
[7] http://universaldependencies.org/conll17/
[8] http://universaldependencies.org/conll18/

converted from the Bosque 8.0 and then some relevant phenomena were manually revised. Therefore we also expect to help this community with a small assessment of this version of the Bosque[9] paying special attention on adverbs.

4 Adverbs within Lexical Resources

Adverbs are words that modify something other than a noun [5], they are considered open class words, thus, under the scope of wordnets. PWN pays special attention to adverbs derived from adjectives, as *extremely* and *poorly*, but some adverbial phrases and underived adverbs, as *then*, are also in the base [2, Chap. 2].

While wordnets are probably the most used lexical resources for NLP, [4] points that 'adverbs are usually neglected in wordnet: there are none in GermaNet, and less than 3% of all lexical units in [Princeton] WordNet are adverbs'. PWN 3.0 has 3,621 adverbial synsets and from those, more than 2,000 synsets do not have a Portuguese translation in OWN-PT.

Previous work by [14] has shown that the adverbial class of the OWN-PT still lacks many relevant Portuguese words. The long-term goal of that project is to be able to do temporal reasoning over Portuguese data. So a version of the Bosque corpus that is annotated with temporal information, Bosque-T, was produced. During the assessment of Bosque-T[10], the authors noted that relevant information for temporal processing was missing from the OWN-PT. Considering the necessity of having the adverbs in place in order to process temporal information present in the Bosque corpus, we set to ourselves the short-term goal of making sure that the adverbs in the corpus are represented in the OWN-PT.

5 Completing the OWN-PT

To consult the OWN-PT, we used the SparqL Endpoint[11], which offers an easy way to retrieve OWN-PT data. To add Portuguese words to synsets, we used the browsable OWN-PT interface, which enables anyone to collaborate with the database. As stated in [12], the OWN-PT interface makes the work of consulting and collaborating to the base easy and straightforward.

The Princeton WordNet has 5,580 adverbial wordforms among 3,621 adverbial synsets. OWN-PT, when we started this work, had 1,934 wordforms in 1,059 adverb synsets. The Bosque corpus contained 494 unique lemmas tagged as adverbs (that is, lemmas whose POS data column were "ADV"), and 16 adverbial multi-word expressions (MWEs) such as *às vezes* (sometimes) and *em outras palavras* (in other words). To find the MWEs with adverbial value, we consulted the field MISC (`MWEPOS=ADV`) of Bosque-UD, which has the POS tag of MWEs as tagged in previous versions of the Bosque corpus. While previous versions of

[9] https://github.com/UniversalDependencies/UD_Portuguese-Bosque
[10] https://github.com/own-pt/portuguese-time
[11] http://wnpt.brlcloud.com:10035/#/repositories/wn30/

the Bosque corpus label MWEs as a single token and assign them a single POS, within the UD project, each element of a MWE should have an independent POS, considering the primary POS tag of it. So, following the proposed guidelines, *em outras palavras* has *em* `ADP`; *outras* `DET` and *palavras* `NOUN`. As stated in [11], having the POS of a MWE explicit in a corpus is a valuable information, so they were kept in the MISC field. Although we can retrieve these 16 adverbial MWEs present in Bosque-UD using the dependency relation `advmod`, it wouldn't be straightforward: there are more than 8k `advmod` dependencies in the corpus.

All in all, we had 510 adverbs from the Bosque-UD. From these, 170 lemmas were not present in OWN-PT. However, 57 were not actually adverbs. Many of them belonged to other part-of-speech (nouns, verbs, numbers, adjectives), few are time expressions (such as *20h45*) or even prepositions (such as *a*). While one would want to annotate temporal expressions as adverbs, there is no reason for annotating the preposition *a* in *como todo o processo nada tem* ***a*** *ver com Ciência* (as the whole process has nothing **to** do with Science) as an adverb. We also found three typos in this small set of lemmas (*de**s**testavelmente*, *supreendentemente* and *enqanto*)[12]. The mistakes found in the Bosque-UD during this work were reported to the Bosque-UD team via GitHub issues.

After this first cleaning, we ended up with a list of 113 adverbs that were not in the OWN-PT. As for the adverbial phrases, 10 out of the 16 found in the Bosque-UD were not in the OWN-PT. Out of those, three adverbial phrases could not be added to the OWN-PT.Two of them were not in the PWN — we recap here that OWN-PT only has synsets present in PWN, so we can not add a Portuguese expression to OWN-PT if there is no English correspondent in PWN — (*in loco* and *nada mais nada menos* (*nothing less than*)). The other one, *a grosso modo*, is characteristic of the spoken language, thus to be avoided in a written language corpus, therefore we added to OWN-PT the standard Latin expression *grosso modo*. Among the adverbs that were not in the OWN-PT, we found very frequent words, such as *onde*, *aonde* and *tão*, to name a few, that were not part of any synset.

In order to add the Portuguese adverbs and adverbial phrases to the OWN-PT, we first looked for their translations in English. Provided the English correspondent was also an adverb, or an adverbial phrase, we would look for it in the database and then add its Portuguese equivalent to the corresponding synset. In some cases (16), the adverbs in Portuguese had an equivalent in meaning that belonged to another POS in English. We will discuss these cases in the next section. So, excluding these and the cases mentioned before (non-adverbs and typos), we were able to add 97 adverbs to the OWN-PT, and seven adverbial phrases, making a total of 104 new unique wordforms in the OWN-PT.

[12] The correct spelling of those words is, respectively, *detestavelmente*, *su**r**preendentemente* and *enq**u**anto*.

6 Discussion

During the process of completing the OWN-PT with the missing adverbs, we faced some issues related to the correspondence of English and Portuguese words. What in Portuguese is expressed by an adverb might in English more commonly be conveyed by a multi-word expression, i.e. a prepositional phrase or a noun phrase. This was the case for the Portuguese adverbs *anteontem* (the day before yesterday), *extramuros* (outside the walls), *computadorizadamente* (in a computerized way) and *impunemente* (with impunity). A similar issue has also been pointed out by [3] when analyzing the adverbial correspondence between the Sanskrit wordnet and its source, the Hindi wordnet.

WH-words such as "where", "how" and "when" are not described in the Princeton WN, probably because they were not considered to be open class terms. This points to an inconsistency in the PWN, since it has adverbs such as *whence* and *wherever*, which are words that belong to the same adverbial class of the WH-words mentioned above. The lack of those words impaired our addition process of the Portuguese adverbs corresponding to them, such as *aonde*, *quão*, *quando* and *onde*. The fact that we would like to have these words in the OWN-PT, despite their not being present in the PWN, suggests the necessity of having a Portuguese resource not totally aligned to the PWN, but with concepts that are relevant to the Portuguese processing itself.

During this work, we did also pay attention to how adverbs appear in the Bosque-UD corpus. We found 60 occurrences of other forms of the verb *ser* (to be) wrongly analyzed in the Bosque-UD corpus, such as: ***Era*** *de facto por ali que começava a surgir perigo para o Celtics.* (It was indeed there that some danger to the Celtics emerged.)

These cases reflect an annotation choice of previous versions of the Bosque corpus, that annotate focus structures composed by *ser (era/foi/é) que* as adverbs. Although this phenomenon deserves a deep discussion, the annotation guidelines of UD project are quite clear about the treatment of MWES and all the forms of the verb *ser* must have the lemma *ser* no matter their syntactic role.

The Bosque-UD was automatically converted from the Bosque 8.0 version. One of the issues its creators faced was how to deal with multi-word expressions, as discussed by [11]. While all MWEs in the Bosque 8.0 are tokenized as a single word, within UD each word that is part of a MWE should have an independent POS. Although the Bosque-UD creators performed a manual review of MWEs, we can still find some issues in its analysis. For example, the comparative phrase *do que* (than that) should be analyzed as 'de' + 'o' + 'que', and each of these words should have a POS. However, in the Bosque-UD, we still find three occurrences of *do* as a single token with the lemma *do*, and not analyzed as *de* + *o*, as expected. A mistake that clearly comes from the conversion from Bosque 8.0, as in it *do que* is a single token tagged as a conjunction. It suggests that a new review of MWEs in the Bosque-UD is needed.

Another issue we found in the Bosque-UD were several missing lemmas. This is why we had '_' in our first list of adverbial lemmas. Nevertheless, it seems to

be a mistake introduced at some point of the Bosque-UD creation, since only two sentences (`CP205-1` and `CP126-7`) have this problem.

These and other minor mistakes, such as prepositions analyzed as adverbs, were reported to the Bosque-UD team via GitHub issues. We did not correct any of those problems since the corpus is now being used as part of the data for the CoNLL 2018 Shared task, therefore modifying the data is not currently possible.

We were also able to add to the wordnet adverbs spelled following the old rules of European Portuguese, namely *subjectivamente*, *tão-pouco*, *electronicamente* and *objectivamente*, cf. the Brazilian variants *subjetivamente*, *tampouco*, *eletronicamente* and *objetivamente*. Although these forms are not under the new orthographic agreement, one wanting to process texts before the 2009 agreement would need to have them in place. Also, it is well known that the recent orthographic agreement is debatable and not fully accepted by the Portuguese community, therefore we opted for having these forms in OWN-PT.

7 Conclusion and Future Work

We described our work on the contribution to the completeness of the OpenWordnet-PT, an open wordnet for Portuguese. We focused on the OpenWordnet-PT adverbial class using data from the Bosque corpus, one of the most used open corpora for Portuguese. We used the Universal Dependencies version of the corpus, which contains 494 unique adverbial lemmas and 16 unique adverbial phrases. From these lemmas, 53 are not adverbs, three are typos and one is a merging mistake. Out of the remaining 438 adverbs, 113 were not present in the OWN-PT, and 16 were not in the PWN, leaving us with 97 adverbs. As for the adverbial phrases, 10 were not in the OWN-PT, of which three could not be added to it. Summing up the remaining 97 adverbs and 7 adverbial phrases, we were able to manually complete 110 synsets with 104 unique wordforms, thus making sure that all (real) adverbs in the Bosque are present in the OWN-PT. We also contributed to the quality of the Bosque-UD, since our careful analysis of adverbs revealed some issues in it.

Completing the OpenWordnet-PT, or at least, making sure that the most frequent Portuguese words are in the right place is a long-term goal. Considering only adverbs, we still have more than 2,000 synsets that do not have a corresponding word in Portuguese. As next steps, we would like to make sure that the OpenWordnet-PT has at least all the necessary content to process the Bosque corpus. Considering the mistakes we found in the Bosque-UD, we expect noticing them will be useful to its creators in the enhancement of its quality.

References

1. Bond, F., Foster, R.: Linking and extending an Open Multilingual Wordnet. In: ACL, 2013 (2013)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press (1998)

3. Kulkarni, M.e.a.: Adverbs in sanskrit wordnet p. 6
4. Marek Maziarz, Maciej Piasecki, E.R.S.S.P.K.: plwordnet 3.0 – a comprehensive lexical-semantic resource. COLING (2016)
5. Matthews, P.H.: The Concise Oxford Dictionary of Linguistics (01 2014), http://www.oxfordreference.com/view/10.1093/acref/9780199675128.001.0001/acref-9780199675128
6. de Melo, G., Weikum, G.: Menta: Inducing multilingual taxonomies from wikipedia. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1099–1108. ACM (2010)
7. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC 2012. ELRA, Istanbul, Turkey (May 2012)
8. de Paiva, V., Chalub, F., Real, L., Rademaker, A.: Making virtue of necessity: a verb lexicon. In: PROPOR – International Conference on the Computational Processing of Portuguese. Tomar, Portugal (2016)
9. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In: COLING 2012 (2012)
10. de Paiva, V., Rademaker, A., Real, L., Chalub, F., Freitas, C.: Openwordnet-pt: Taking stock. In: NLCS'18 (2018)
11. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). pp. 197–206 (2017)
12. Real, L., Chalub, F., de Paiva, V., Freitas, C., Rademaker, A.: Seeing is correcting: curating lexical resources using social interfaces. In: Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of Asian Federation of Natural Language Processing - Fourth Workshop on Linked Data in Linguistics: Resources and Applications (LDL 2015). Beijing, China (Jul 2015)
13. Real, L., de Paiva, V., Chalub, F., Rademaker, A.: Gentle with gentilics. In: Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016). Slovenia (May 2016)
14. Real, L., Rademaker, A., Chalub, F., V de Paiva, V.: Towards temporal reasoning in portuguese. In: Proceedings of the LREC2018 Workshop Linked Data in Linguistics (2018)