Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets*

Alberto Simões[1] and Xavier Gómez Guinovart[2]

[1] Centro de Estudos Humanísticos
Universidade do Minho, Portugal
ambs@ilch.uminho.pt
[2] Seminario de Lingüística Informática
Universidade de Vigo, Spain
xgg@uvigo.es

Abstract. In this article we exploit the possibility on bootstrapping an European Portuguese WordNet from the English, Spanish and Galician wordnets using Probabilistic Translation Dictionaries automatically created from parallel corpora.
The process generated a total of 56 770 synsets and 97 058 variants. An evaluation of the results using the Brazilian OpenWordNet-PT as a gold standard resulted on a precision varying from 53% to 75% percent, depending on the cut-line. The results were satisfying and comparable to similar experiments using the WN-Toolkit.

Keywords: WordNet, Portuguese, probabilistic translation dictionaries, parallel corpora, knowledge acquisition

1 Introduction

For the Portuguese community there is a lack of a good, complete and free accessible WordNet. There are lot of different projects whose main goal is to construct such a resource, but most are incomplete, not free, or heavily based on machine translation.

We propose and evaluate a method to bootstrap an European Portuguese WordNet using the Galician, Spanish and English wordnets as guidance, and using Probabilistic Translation Dictionaries (PTDs) for their Portuguese translation. The main difference on this approach when compared with others, namely the Unified Wordnet [12] or the WN-Toolkit [15], is the use of probabilistic translation dictionaries that, being probabilistic and automatically generated, give a wider set of translations, rather than the small set of possible translations usually presented on standard bilingual lexicons.

* This research has been carried out thanks to Portuguese National Funds, through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project PEst-OE/EEI/UI0752/2014, and the Project SKATeR (TIN2012-38584-C06-01 and TIN2012-38584-C06-04) supported by the Ministry of Economy and Competitiveness of the Spanish Government.

Also, we want to exploit the proximity of the Portuguese language with the Galician language. Although we do not have access to a big bilingual lexicon for these two languages, a rewriting method to bring Portuguese closer to Galician was used [19].

This bootstrapped version will be incorporated into Multilingual Central Repository [8, 1]. From this base work we plan corrections and different expansion works, using similar approaches to the ones being taken by GalNet [3].

This document is structured as follows: first we discuss briefly similar approaches to this task (Section 2), followed by a presentation on the resources that were used in our experiments (Section 3). In Section 4, the algorithm used in this research is explained using a specific synset example. Section 5 evaluates the obtained results, and in Section 6 we draw some conclusions and point some directions in our future work.

2 Similar Approaches

There are different initiatives on the creation or enlargment of wordnets and similar lexical databases. In this section we will focus mainly three of these initiatives.

Onto.PT [6] includes more than 117,000 synsets. These synsets were computed using different mechanisms, and incorporating data from different sources. A big amount of relations were computed using patterns over conventional electronic dictionaries [7]. Some other were extracted processing Wikipedia [5]. The authors also incorporated data from other lexical resources, like TeP 2.0 [11], OpenWordNet-PT [17] or OpenThesaurus.PT.

WN-Toolkit [15] is a lexical extraction tool for the creation of wordnets from bilingual resources, including lexical resources (such as bilingual dictionaries) and textual resources (such as parallel corpora), which has already been used for expanding the Catalan, Spanish and Galician wordnets [4].

Universal WordNet [12] initiative used the Princeton WordNet, other monolingual wordnets, bilingual dictionaries and parallel corpora to bootstrap wordnets in more than 200 languages. The resulting resource, looking to the Portuguese language, includes only 23,500 synsets. Unfortunately the resulting resource was only used for some cross-lingual text classification, and no results are presented for the Portuguese language.

3 Used Resources

This section describes the resources used in our experiments, namely the English, Galician and Spanish wordnets, a set of probabilistic translation dictionaries,

a dynamic Portuguese-Galician dictionary and the *Vocabulário Ortográfico do Português* (Portuguese Orthographic Vocabulary or VOP).

3.1 English, Spanish and Galician Wordnets

The English WordNet, known as Princeton WordNet [13], was used to guide our extraction as other languages wordnets usually rely on the Interlingual Index (ILI) to synchronize concepts with it. We used the English WordNet version 3.0 as it is the base for the current Spanish and Galician wordnets.

The Spanish [2] and Galician [3] wordnets were obtained from the Multilingual Central Repository [8, 1]. This project aims to integrate the wordnets for the languages of Spain in a similar repository, together with the English WordNet 3.0.

Table 1 presents some statistics on English, Galician and Spanish wordnets.

		Nouns	Verbs	Adjectives	Adverbs	Total
English	Syn.	82 889	13 769	18 156	3 621	118 435
	Var.	147 358	25 051	30 004	5 580	207 993
Galician	Syn.	16 812	1 413	4 962	223	23 419
	Var.	22 186	3 996	7 884	253	34 319
Spanish	Syn.	26 594	6 251	5 180	677	38 702
	Var.	39 142	10 829	6 967	1 051	57 989

Table 1. Number of synsets and variants of English, Galician and Spanish wordnets, distributed by part-of-speech.

3.2 Probabilistic Translation Dictionaries

PTDs (or probabilistic translation dictionaries) are dictionaries obtained with NATools [20] by the word-alignment of parallel corpora. Unlike other tools, like Giza++ [14] that aims on extracting a relationship between each occurrence of a word and a specific occurrence of its translation, NATools extracts a single mapping for each source-language word. Each mapping associates a set of possible translations (in the parallel corpora) together with a probability measure of it being a correct translation.

Consider the following example of a PTD entry:

$$\mathcal{T}\,(\text{codificada}) = \begin{cases} \text{codified} & 62.83\% \\ \text{uncoded} & 13.16\% \\ \text{coded} & 6.47\% \\ \dots & \end{cases}$$

This example states that the Portuguese word *codificada* is usually co-ocurrent with the English words *codified*, *coded* and *uncoded*. Other than this co-ocurrence

information, the dictionary adds a probability measure to each possible co-ocurrent word. When these resources are extracted from aligned parallel corpora, it is usual that this co-ocurrence can be seen as a translation measure. Nevertheless, and as presented in the example, it might happen that some relations are not really translations of each other: they might be related (like the fact of *uncoded* being the antonym of *codificada*), or not related at all. Nevertheless, as a statistical measure, we expect it to be have a small probability for these situations.

The PTDs used in this experiment were extracted both from the Per-Fide corpora[3] and the CLUVI [9] corpus. From the Per-Fide project we extracted dictionaries between Portuguese–Spanish and Portuguese–English. From the CLUVI corpus we extracted a Spanish–Galician dictionary. Using the composition [18] of PT–ES and ES–GL dictionaries we obtained a PT–GL dictionary.

Given that wordnets do not include word forms, the corpora were lemmatized and tagged using FreeLing [16]. Thus, the corpora words were replaced by the pair `lemma/pos` before the PTD extraction. This results on a better translation dictionary.

3.3 Portuguese–Galician Dictionary

There is not a wide translation dictionary for the PT–GL languages. Nevertheless, given the big proximity of the two languages, and despite the fact of existing some false friends, it is possible to rewrite, with a reasonable precision, Portuguese words in their Galician counterparts [19].

This kind of approach can be seen as a dynamic dictionary, as new words, as far as they follow the usual pattern, can be translated by the tool without the need of a lexicon.

3.4 Portuguese Orthographic Vocabulary

The *Vocabulário Ortográfico do Português* (VOP)[4] is a list of 182 012 lemmas of Portuguese words, together with their part-of-speech. For our work we just considered the list of lemmas, discarding all other information.

4 Algorithm

The bootstrapping algorithm uses a "score" approach. Different variants are generated, together with an associated score. As other languages or heuristics are analyzed, the system adapts the variant score accordingly. The variants with higher score are then returned.

[3] The corpora available in the Per-Fide project includes most of the free available corpora in the Web, like EuroParl, JRC-Acquis or the DGT Translation Memories, as well as corpora computed from the Vatican or European Central Bank websites.

[4] `http://www.portaldalinguaportuguesa.org/vop.html`

In order to explain the approach we will explain the process for a specific synset, reference `00008007-r`, that corresponds to an adverb.

The first step is to search, in each wordnet, for this synset. The Spanish WordNet does not include this synset, but it exists for the other two languages:

$$variants_{EN} = \{all, altogether, completely, entirely, totally, whole, wholly\}$$
$$variants_{GL} = \{completamente, totalmente\}$$

For each language the probabilistic translation dictionary is queried, searching for translations for these words, although maintaining the same part-of-speech. Table 2 and table 3 show the contents of the probabilistic translation dictionaries for each original variant[5]

completamente	totalmente
completamente 0.48523	simplemente 0.00746
totalmente 0.14573	totalmente 0.53396
plenamente 0.06537	completamente 0.10343
absolutamente 0.01693	plenamente 0.04045
simplemente 0.00204	absolutamente 0.01739
definitivamente 0.00025	no 0.00002

Table 2. Result of translating the Galician synset.

completely	totally
completamente 0.350345	totalmente 0.728418
totalmente 0.332604	inteiramente 0.056268
inteiramente 0.096318	completamente 0.052408
plenamente 0.091360	plenamente 0.022330
absolutamente 0.045507	absolutamente 0.012403
perfeitamente 0.005288	perfeitamente 0.008233
ainda 0.004253	não 0.002670
definitivamente 0.000979	ainda 0.002226
integralmente 0.000152	integralmente 0.000173

Table 3. Result of translating part of the English synset.

For each different Portuguese variant candidate (for each source language) the maximum value is chosen.

If the translation is symmetric, the score is incremented by 0.5. For example, consider the Portuguese variant *completamente* obtained using the English–Portuguese dictionary, by translating the English variant *completely*.

[5] The described process needs to be performed to every word of the presented set. Nevertheless, for simplicity, we chose only two of the seven English variants.

If the Portuguese–English dictionary also maps the word *completamente* into the English word *completely*, this variant score is incremented. So, for example, the new score for *completamente* would be 0.850345, but for *ainda* it will be maintained, as the word *completely* does not occur as its translation. The words with a star ($\star$) in table 4 are reflexive.

Galician			**English**		
Variant	Max	Sym.	Variant	Max	Sym.
completamente	0.48523	0.98523 $\star$	completamente	0.350345	0.850345 $\star$
simplesmente	0.00746	0.50746 $\star$	totalmente	0.728418	1.228418 $\star$
totalmente	0.53396	1.03396 $\star$	inteiramente	0.096318	0.596318 $\star$
plenamente	0.06537	0.56537 $\star$	plenamente	0.091360	0.591360 $\star$
absolutamente	0.01739	0.51739 $\star$	absolutamente	0.045507	0.545507 $\star$
definitivamente	0.00025	0.50025 $\star$	perfeitamente	0.008233	0.508233 $\star$
no	0.00002	0.00002	ainda	0.004253	0.004253
			não	0.002670	0.002670
			definitivamente	0.000979	0.500979 $\star$
			integralmente	0.000173	0.500173 $\star$

Table 4. Result after scoring the Portuguese candidate words.

The next step is to find out how many language wordnets generated each of the Portuguese variants. For example, looking to Table 4 we can notice that the word *no* is generated only in the Galician side. In the other hand, the word *totalmente* is generated from both languages.

So, for each Portuguese variant candidate we will multiply the maximum probability found (1.728418 for the *totalmente* word) by the number of wordnets that generated this candidate: $1.228418 \times 2 = 2.456836$. Table 5 show the result of this step ($score_1$).

This next step tries to take advantage of the Portuguese–Galician proximity. It uses the Portuguese–Galician dictionary (not the probabilistic dictionary) for each one of the Portuguese variant candidates. If any of its translations occurs in the Galician set of variants, then the variant candidate score is incremented by 1. Table 5 shows what happens to our candidate set (column $score_2$).

Finally the VOP is used to decrease by 1 point all words that are not part of the Portuguese vocabulary. In our example nothing changes, as all words are present in VOP.

The top classified variant candidates are then returned. At the moment the number of words to return is the size of the biggest set of variants for the source languages (English, Spanish or Galician). Therefore, considering the example above is complete, the first seven candidates would be returned.

Note that we do not define any kind of cut-line, other than suggesting that it is not probable to compute more variants than the number of existing variants for other languages. The threshold should be defined later, accordingly with a

Variant	Max Score	$Score_1$	$Score_2$
totalmente	1.228418	2.456836	3.456836
completamente	0.985230	1.970460	2.970460
plenamente	0.591360	1.182720	1.182720
absolutamente	0.545507	1.091014	1.091014
inteiramente	0.596318	0.596318	0.596318
perfeitamente	0.508233	0.508233	0.508233
definitivamente	0.500979	1.001958	1.001958
integralmente	0.500173	0.500173	0.500173
simplesmente	0.507460	0.507460	0.507460
ainda	0.004253	0.004253	0.004253
não	0.002670	0.002670	0.002670
no	0.000020	0.000020	0.000020

Table 5. Portuguese variants, the maximum score obtained from the English or Galician wordnets, $score_1$ and $score_2$.

specific goal. In the evaluation section different cut-lines will be used, and the results analyzed.

5 Evaluation

This section presents an automatic evaluation of the obtained results, first using OpenWordNet-PT [17] as a gold standard, and then a comparing with the results obtained by WN-Toolkit in a similar experiment [4].

Given one of our work motivation is the lack of a good wordnet for the Portuguese language, it gets hard to have a gold standard to evaluate our work with[6]. Nevertheless, and although it contains only half the synsets created by our tool, we used the Brazilian OpenWordNet-PT [17] as our gold standard. The OpenWordNet-PT version used in these experiments contains 43 895 synsets, with a total of 74 012 variants.

After running the tool we obtained 56 770 synsets (33 275 nouns, 10 803 verbs, 10 733 adjectives and 1 959 adverbs) with a total of 97 058 variants. From these synsets, only 49.6% are available on OpenWordNet-PT (28 156 synsets).

These candidates synsets were tested using different heuristics to select which variants to test.

Heuristic A: Evaluate variants with a score greater or equal to 2.5;
Heuristic B: Evaluate variants with a score greater or equal to 2.0;
Heuristic C: Evaluate variants with a score greater or equal to 1.5;
Heuristic D: Evaluate the higher score variant for all synsets, and any other variant with a score greater or equal to 2.0;

[6] Unfortunately all freely available wordnets follow the automatic generation of synsets, without any real, thorough, manual evaluation. They do not, even, have some kind of score to be used to know each synset variant expected quaity.

Table 6 summarizes the obtained results. For each test it includes the number of variants tested, the average score for these variants, the number of correct variants, accordingly with OpenWordNet-PT, and, finally, the percentage of the tested variants that are correct[7].

Heuristic	Nr. Variants	Average Score	Correct Variants
A	9 307	3.0005	6 813 (73.20%)
B	13 785	2.8501	9 426 (68.38%)
C	19 315	2.7360	11 189 (57.93%)
D	31 526	2.1180	16 424 (53.37%)

Table 6. Result of the four approaches for the synsets evaluation.

When analyzing the results we noticed that there were some false negatives. As OpenWordNet-PT is based on Brazilian Portuguese, and the corpora used by us if from European Portuguese, there were some words that did not match (for example, "*açã̃o*" vs "*acçã̃o*"). Therefore, in top of the previously described heuristics we used the Levenshtein algorithm [10], and decided to accept candidate variants as if they are at an edit distance of 1. Table 7 show the values obtained with this approach.

Heuristic	Nr. Variants	Average Score	Correct Variants
A'	9 307	3.0001	6 996 (75.17%)
B'	13 785	2.8477	9 747 (70.71%)
C'	19 315	2.7312	11 662 (60.38%)
D'	31 526	2.0970	17 726 (56.23%)

Table 7. Result of the four approaches for the synsets evaluation, with Levenshtein distance of 1.

Finally, we did an extra manual evaluation on the obtained variants, both by an author of this paper, and an external researcher[8]. Table 8 summarizes the results.

E$_1$ In the first evaluation we selected 100 variants not present in OpenWordNet-PT, but which respective synsets are. When looking for the number of variants approved by both evaluators there is a correctness of 41%.

E$_2$ For the second evaluation we selected 100 variants which synsets are not present in OpenWordNet-PT. In this case we selected the higher scoring variant for each of these synsets. There is a correctness of 45% when looking to the two evaluators agreement.

[7] Note that it does not make much sense to compute the recall, as we are not trying to generate the complete Gold standard.

[8] Our deepest thanks to Hugo Gonçalo Oliveira for his evaluation of our candidate variants.

Evaluation	Internal Evaluator	External Evaluator	Common Overally
E_1	46 OK	54 OK	41 OK
E_2	54 OK	53 OK	45 OK

Table 8. Manual evaluation for 200 variant candidates.

These results are very similar to those obtained with the WN-Toolkit [4] using data for candidates acquired from only one resource. In this case, the automatic precision value was 77.02%, with a real precision (calculated with human revision) of 70% for new variants for empty synsets and 53% for the candidate variants for not empty synsets.

6 Conclusions and Future Work

We presented a quick way to bootstrap a wordnet for Portuguese. Although the initial results are not satisfactory, we were still able to extract about 56,700 synsets. An automatic evaluation to part of these synsets measure a correctness of 54.87%. If this ratio is maintained for every extracted synset, it means there are 31,000 correct synsets, which is already 3,000 more than the total number of synsets in OpenWordNet-PT.

Nevertheless, this was a primary study on the process. We will perform an evaluation on the set of synsets not present on OpenWordNet-PT, as well as the variants from synsets on OpenWordNet-PT that are not recognized. Finally, the use of TeP or Onto.PT will allow the automatic enlargement of the synsets.

References

1. Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. In: Second International WordNet Conference. pp. 80–210 (2004)
2. Fernández Montraveta, A., Vázquez, G.: La construcción del wordnet 3.0 en español. In: Castillo, M.A., Platero, J.M.G. (eds.) La lexicografía en su dimensión teórica. pp. 201–220. Universidad de Málaga, Málaga (2010)
3. Gómez Guinovart, X., Clemente, X.M.G., Pereira, A.G., Lorenzo, V.T.: Galnet: WordNet 3.0 do galego. Linguamática 3(1), 61–67 (2011)
4. Gómez Guinovart, X., Oliver, T.: Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. Procesamiento del Lenguaje Natural 53, 43–50 (2014)
5. Gonçalo Oliveira, H., Costa, H., Gomes, P.: Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In: Proceedings of INFORUM 2010, Simpósio de Informática. Braga, Portugal (September 2010)
6. Gonçalo Oliveira, H., Gomes, P.: Towards the automatic creation of a wordnet from a term-based lexical network. In: Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing. pp. 10–18. ACL Press (July 2010)

7. Gonçalo Oliveira, H., Gomes, P.: Automatic discovery of fuzzy synsets from dictionary definitions. In: Proceedings of 22nd International Joint Conference on Artificial Intelligence. pp. 1801–1806. IJCAI 2011, AAAI Press, Barcelona, Spain (July 2011)
8. González, A., Laparra, E., Rigau, G.: Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In: 6th Global WordNetConference. Matsue, Japan (2012)
9. Gómez Guinovart, X.: A hybrid corpus-based approach to bilingual terminology extraction. In: Fandiño, I.M.S., Crespo, B. (eds.) Encoding the Past, Decoding The Future: Corpora in the 21st Century. pp. 147–175. Cambridge Scholar Publishing, Newcastle upon Tyne (2012)
10. Levenshtein, V.I.: On the minimal redundancy of binary error-correcting codes. Information and Control 28(4), 268–291 (1975)
11. Maziero, E.G., Pardo, T.A.S., Di Felippo, A., Dias-da Silva, B.C.: A base de dados lexical e a interface Web do TeP 2.0: Thesaurus eletrônico para o português do brasil. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. pp. 390–392. WebMedia '08, ACM, New York, NY, USA (2008)
12. de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 513–522. CIKM '09, ACM, New York, NY, USA (2009)
13. Miller, G.A.: WordNet: A lexical database for English. Commun. ACM 38(11), 39–41 (Nov 1995)
14. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
15. Oliver, A.: Wn-toolkit: Automatic generation of wordnets following the expand model. In: Proceedings of the 7th Global WordNetConference. Tartu, Estonia (2014)
16. Padró, L.: Analizadores multilingües en FreeLing. Linguamática 3(2), 13–20 (December 2011)
17. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An open Brazilian WordNet for reasoning. In: Proceedings of the 24th International Conference on Computational Linguistics (2012)
18. Simões, A., Almeida, J.J., Carvalho, N.R.: Defining a probabilistic translation dictionaries algebra. In: Correia, L., Reis, L.P., Cascalho, J., Gomes, L., Guerra, H., Cardoso, P. (eds.) XVI Portuguese Conference on Artificial Inteligence - EPIA. pp. 444–455. Angra do Heroismo, Azores (September 2013)
19. Simões, A., Guinovart, X.G.: Dictionary Alignment by Rewrite-based Entry Translation. In: Leal, J.P., Rocha, R., Simões, A. (eds.) 2nd Symposium on Languages, Applications and Technologies. OpenAccess Series in Informatics (OASIcs), vol. 29, pp. 237–247. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2013)
20. Simões, A.M., Almeida, J.J.: NATools – a statistical word aligner workbench. Procesamiento del Lenguaje Natural 31, 217–224 (September 2003)