Experiments on Enlarging a Lexical Ontology

Alberto Simões[1,2] & José João Almeida[2]

[1] Centro de Estudos Humanísticos
[2] Centro Algoritmi
Universidade do Minho, Braga, Portugal
{ambs@ilch,jj@di}.uminho.pt

Abstract. This paper presents two simple experiments performed in order to enlarge the coverage of PULO, a Lexical Ontology, based and aligned with the Princeton WordNet. The first experiment explores the triangulation of the Galician, Catalan and Castillian wordnets, with translation dictionaries from the Apertium project. The second, explores Dicionário-Aberto entries, in order to extract synsets from its definitions. Although similar approaches were already applied for different languages, this document aims at documenting their results for the PULO case.

1 Introduction

Recently, a huge effort has been done to boost the development of wordnet clones for different languages. Portuguese is not an exception. There are different initiatives to create lexical ontologies, linked or not with the original Princeton WordNet [9] (WordNet.Pr). Examples of such initiatives are Onto.PT [4], PAPEL [5], TeP [8] or Open WordNet-PT [10]. Along with these, another initiative born some months ago: the Portuguese Unified Lexical Ontology (PULO) [12]. It aims at integrating different existing resources into a structure aligned with WordNet.Pr. Recently a joint effort on comparing these projects' history, goals and statuses [7], lead some teams in the direction of cooperation. Nevertheless, each project team continues their own initiatives, enriching and enlarging their resources.

The same happens with PULO. This document describes two experiments performed with the objective of enlarging the number of variants[3]. The kind of experiments are, somehow, similar to some of the previous work, done in order to bootstrap PULO [12] (as we also triangulated three different wordnets, but using probabilistic translation dictionaries), to some of the approaches used to expand GalNet [3], and to create Onto.PT [4]. Although the idea is not new, the thorough description of the process and it's brief evaluation is relevant for future initiatives with other languages.

This short article includes two main sections: section 2 describes the experiment approaches and used resources, while section 3 gives some measures on the quality of the methods application. Finally, it concludes with some brief discussion of the results and future work.

[3] This article will use the term *variant* to refer to one of the synonyms of a synset.

2 Experiments Description

Before running these experiments, PULO included a total of 18.689 variants, distributed by 17.871 synsets (meaning most synsets include only one variant). Table 1 shows how these variants are distributed by morphological category.

Table 1. Distribution of the 18.689 variants prior to the enlargement experiments.

	Nouns	**Adjectives**	**Verbs**	**Adverbs**	**Total**
Variants	10.421	3.441	4.283	544	18.689

The next subsections describe the two experiments. The first one is based in the triangulation of the Catalan, Galician and Castillian wordnets using translation dictionaries. The second one explores Dicionário-Aberto [11], an open and free definitions dictionary.

2.1 Experiment I: Triangulating Iberian Wordnets

This first experiment uses the wordnets available through Multilingual Central Repository [6], and some translation dictionaries obtained from the Apertium [2] project. Given the reduced number of dictionaries including Portuguese, only the Catalan, Galician and Castillian languages were used. Table 2 shows the sizes for these three wordnets.

Table 2. Summary of sizes for the three used wordnets.

		Nouns	**Adjectives**	**Verbs**	**Adverbs**	**Total**
Galician	**Synsets**	18.850	5.092	1.541	349	25.832
	Variants	25.205	8.050	4.145	420	37.820
Catalan	**Synsets**	36.460	4.148	5.424	1	46.033
	Variants	51.606	7.679	11.577	2	70.864
Castillian	**Synsets**	26.594	5.180	6.251	677	38.702
	Variants	39.142	6.967	10.829	1.051	57.989

Regarding the translation dictionaries, Table 3 summarizes their sizes. As can be seen, these are quite small dictionaries. This fact was the main reason why the bootstrapping approach [12] used probabilistic translation dictionaries that have a broader coverage. Also, note that most entries in this dictionary have only one translation, reducing the translation ambiguity (which is somewhat desired for a machine translation dictionary, but reduces its applicability for other tasks).

The used algorithm is quite simple. For each synset in the database, that includes at least one variant in any of the three languages, it:

Table 3. Translation dictionaries sizes.

Lang. Pair	Nr. Entries	Max Nr. Trans.	Avg. Nr. Trans.
GL–PT	11.003	4	1.07
CA–PT	6.510	7	1.11
ES–PT	12.742	6	1.07

1. Creates a multiset $S_{\mathcal{L}}$ that includes all translations obtained by the translation of all variants for language $\mathcal{L}$. Note that different variants can translate to the same word in Portuguese, so, the multiset tracks the number of times that word was obtained.
2. Compute the multiset $S = S_{GL} \cup S_{CA} \cup S_{ES}$. This means that, if a Portuguese word was obtained by translating just one variant for each of the source languages, it would have a multiplicity of three. On the other hand, if three variants for just one language generated a Portuguese word, that was not obtained from any of the other languages, its multiplicity would be, as well three. Not giving extra weight if the word was obtained from different languages or every time from the same language was decided in order to keep the algorithm simple.
3. Filter the multiset S, removing all Portuguese variants with a multiplicity of just one. To define this cut line, each variant was checked against current variants in PULO. Figure 1 shows this test. Bars at the left represent variants found in PULO, while bars at the right represent new variants. Given the huge amount of new variants with a multiplicity of 1, it was decided to ignore them (trying to improve accuracy).

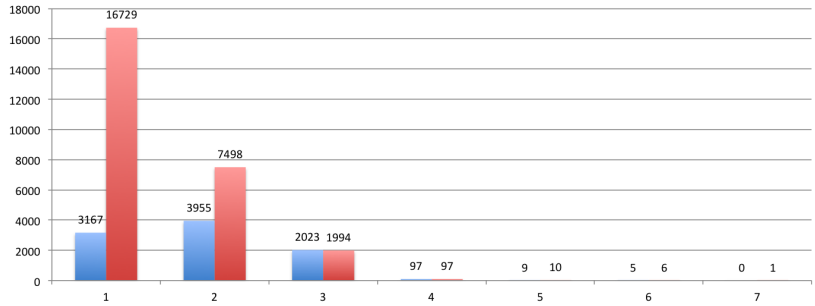

Fig. 1. Number of candidate variants already existing in PULO (left bars) against the new candidates (bars at the right), distributed by their multiplicity in multiset S.

4. The bootstrapping approach for PULO used dictionaries obtained from European Portuguese corpora with its old orthography[4]. The dictionaries from

[4] Orthography prior to the 1990 agreement, that was officiated in 2008 by the Portuguese Government, and still being, progressively, adopted in Portugal.

Apertium used, essentially, Brazilian orthography that, curiously, is now the correct form for European Portuguese. With that in mind, a simple tool was used to remove variants written in the old orthography, and adding the respective new orthography in case it was not yet present. This process was performed using JSpell morphological analyzer [1].

This process created a total of 7.229 new variants, and removed 261 of existing variants with the old orthography. Table 4 summarizes the distribution of PULO variants by morphological category after this experiment.

Table 4. Distribution of the 25.657 variants after the first enlargement experiment.

	Nouns	**Adjectives**	**Verbs**	**Adverbs**	**Total**
Variants	14.062	4.825	6.172	598	25.657

2.2 Experiment II: Synset Extraction from Definitions Dictionary

This second experiment was prepared already with the expectation of a big amount of false positives. Nevertheless, there was interest on confirm that expectation. The main idea was to use Dicionário-Aberto (DA) [11] definitions to construct synsets. DA is partially encoded in TEI[5].

DA definitions are stored in `def` XML elements, with the new line signaling the change of sense[6]. Although XML should ignore spaces and new lines, this decision was taken during the dictionary encoding process for simplicity. Each sense line can include very different types of information. The most common is a standard definition, explaining the concept. In other cases, there are examples, or *see also* references. But there is another kind of definition that is quite interesting for the PULO enlargement process. Some lines include a set of synonyms separated by a semicolon (see an example in Figure 2). Thus, this second experiment finds lines in DA that are only a sequence of terms separated by a semicolon. For each of these sequences, the list of synonyms, together with the entry head word, are stored.

Exploring the 128.521 entries in DA, 4.842 synsets were found. These synsets have from 3 to 7 synonyms, with an average of 3.14 synonyms per synset. Follow some examples of such synsets:

acobertar, encobrir, dissimular
açôfar, pechisbeque, latão
acordança, melodia, consonância
acôrdo, convenção, ajuste
acoroçoado, animado, incitado

[5] Text Encoding Initiative XML schema, that includes notation to encode different kind of resources from simple books to corpora or dictionaries.
[6] This distinction is, of course, of the responsibility of the original lexicographer.

Fig. 2. Example of an entry from Dicionário Aberto with a line of synonyms.

In order to map these synsets to PULO synsets, a simple heuristic was used: find an intersection between the synonyms from the two sources that includes, at least, two variants. This means that for a synset obtained from DA $\langle s_1, s_2, s_3 \rangle$, s_i will be suggested as a candidate if there is a synset S in PULO that contains s_j and s_k with $i \neq j \neq k$.

Table 5 show some synsets from PULO (left column) and the aligned synset from DA (right column). In italic are the terms that were used for the alignment.

cima, cimeira, *cimo*, cumbre, *cume*	vértice, *cimo*, *cume*, culminância
lista, *relação*	tabela, *relação*, catálogo, *lista*
alegria, *prazer*	*prazer*, *alegria*, jovialidade, satisfação, delícia, aprazimento, agrado

Table 5. Synsets from PULO at the left, and aligned synset from DA at the right.

This process suggested 1.150 additions. Given this dictionary is quite noisy, and includes a lot of words with old orthography (previous to the 1945 agreement), these suggestions were not added automatically to PULO.

3 Experiments Evaluation

Both evaluations reported here were performed by sampling, given there is no gold standard that can be used to evaluate these candidates, neither the manual power needed to fully (manually) evaluate all candidates from both experiments.

For the second experiment, all suggestions need to be evaluated before being added to PULO. Nevertheless, there was no time to complete that task yet.

3.1 Experiment I

For the first experiment, 200 of the added variants were chosen randomly. This sample included 101 nouns, 39 adjectives, 2 adverbs and 58 verbs.

The evaluation divided these variants into three different categories:

– **Correct Variants**
152 of the obtained variants were classified as correct. This evaluation was performed looking to the word and the sense gloss. When in doubt, a standard dictionary was used, in order to check if that specific sense was present in the definition.

Follows some examples of variants evaluated in this class, together with its gloss[7]:

- progredir — get better
- corrupção — the state of being corrupt
- aguentar — hang on during a trial of endurance

– **Incorrect Variants**
40 of the variant candidates were marked as incorrect. Most of these were easy to spot, looking to the synset gloss. Examples of such entries are:

- pegar — take away to an undisclosed location against their will and usually in order to extract a ransom
- remeter — make less fast or intense
- bola — a statement that deviates from or perverts the truth

– **Ambiguous Variants**
There were 8 of the proposed variants that the authors feel they are not incorrect, because there are some situations in which they can be used to represent the synset concept. Nevertheless, as this decision might not be consensual, the variants were classified as ambiguous. Some examples:

- desnudar — take away possessions from someone
- puro — spotlessly clean and fresh

Table 6 present these numbers distributed by morphological category, with an accuracy (by sampling) of 76%[8].

Table 6. Distribution of correct, incorrect and ambiguous variants distributed by morphologic category for first experiment.

	Nouns	**Adjectives**	**Verbs**	**Adverbs**	**Total**
Correct	82 (81%)	27 (69%)	41 (71%)	2 (100%)	152 (76%)
Incorrect	17 (17%)	9 (23%)	14 (24%)	0 (0%)	40 (20%)
Ambiguous	2 (2%)	3 (8%)	3 (5%)	0 (0%)	8 (4%)
Total	101	39	58	2	200

[7] In these and next examples, the authors decided not to translate the variant itself, as a direct translation will lose part of the cultural/usage meaning.

[8] Given the obtained accuracy and the lack of human resources for a through validation, the authors decided to include the obtained variants without further analysis.

3.2 Experiment II

Again, for this second experiment, 200 of the candidate variants where chosen randomly and classified in the three classes defined in the previous section. This evaluation resulted in only 115 variant candidates marked for acceptance, while 74 were marked as wrong, and 11 as ambiguous. Table 7 shows the distribution of these candidates by morphologic category. The accuracy[9] on this experiment was 58%.

Follow some examples of entries obtained throw this experiment for each of the three classes:

- **Correct Variants**
  - constância — persistent determination
  - sólido — securely in position; not shaky
  - truculento — very unpleasant
- **Incorrect Variants**
  - carraceno — very small
  - eduzir — make a subtraction
  - sisudez — a solemn and dignified feeling
- **Ambiguous Variants**
  - bom — to a complete degree or to the full or entire extent
  - aquentar — spur on or encourage especially by cheers and shouts

Table 7. Distribution of correct, incorrect and ambiguous variants distributed by morphologic category for second experiment.

	Nouns	**Adjectives**	**Verbs**	**Adverbs**	**Total**
Correct	56 (62%)	28 (56%)	31 (52%)	0	115 (58%)
Incorrect	28 (31%)	19 (38%)	27 (45%)	0	74 (37%)
Ambiguous	6 (7%)	3 (6%)	2 (3%)	0	11 (5%)
Total	90	50	60	0	200

4 Conclusions

This article reports two experiments on expanding PULO coverage. Although the used methods are not new, the experiments have shown that these methods can get acceptable accuracy. Even the second method, that used a very noisy and old dictionary (from 1913), could suggest a good set of new variants. Nevertheless, when dealing with semantics, decisions are not consensual, and probably other researchers would accept or reject different number of entries.

[9] Given the low accuracy and the small number of proposed variants, the authors decided to perform a manual validation prior to their incorporation into PULO.

Acknowledgements: Thanks to Nuno Carvalho for the proofreading. This work has been partially supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

1. Almeida, J.J., Pinto, U.: Jspell – um módulo para análise léxica genérica de linguagem natural. In: Actas do X Encontro da Associação Portuguesa de Linguística. pp. 1–15. Évora 1994 (1995)
2. Forcada, M.L.: Apertium: traducció automàtica de codi obert per a les llengües romàniques. Linguamática 1(1), 13–23 (May 2009)
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. Gonçalo Oliveira, H., Gomes, P.: ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation Journal 48(2), 373–393 (2014)
5. Gonçalo Oliveira, H., Santos, D., Gomes, P., Seco, N.: PAPEL: A dictionary-based lexical ontology for Portuguese. In: Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR). vol. 5190, pp. 31–40. Springer (2008)
6. Gonzalez-Agirre, A., Laparra, E., Rigau, G.: Multilingual Central Repository version 3.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12). pp. 2525–2529. ELRA (2012)
7. Gonçalo Oliveira, H., de Paiva, V., Freitas, C., Rademaker, A., Real, L., Simões, A.: As Wordnets do Português. In: Simões, A., Barreiro, A., Santos, D., Sousa-Silva, R., Tagnin, S. (eds.) Linguística, Informática e Tradução: Mundos que se Cruzam, vol. 7, pp. 397–424 (March 2015)
8. Maziero, E.G., Pardo, T.A.S., Felippo, A.D., Dias-da-Silva, B.C.: A Base de Dados Lexical e a Interface Web do TeP 2.0. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana. pp. 390–392 (2008)
9. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM 38, 39–41 (1995)
10. Rademaker, A., Paiva, V.D., de Melo, G., Coelho, L.M.R., Gatti, M.: OpenWordNet-PT: A Project Report. In: Proceedings of the 7th Global WordNet Conference. pp. 383–390 (2014)
11. Simões, A., Farinha, R.: Dicionário Aberto: um recurso para processamento de linguagem natural. Vice-Versa 16, 159–171 (December 2011)
12. Simões, A., Guinovart, X.G.: Bootstrapping a portuguese wordnet from galician, spanish and english wordnets. Advances in Speech and Language Technologies for Iberian Languages 8854, 239–248 (2014)