An overview of Portuguese WordNets

Valeria de Paiva Livy Real Hugo Gonçalo Oliveira
Alexandre Rademaker Cláudia Freitas Alberto Simões

Nuance Communications, EUA

IBM Research, Brazil

Univ. Coimbra, Portugal

Univ. Minho, Portugal

January 30, 2016

Introduction

- Portuguese WordNets
- Closed WordNets: WordNet.PT [Marrafa, 2001], WordNet.BR [Dias-da-Silva, 2006], MultiWordNet.PT [Pianta et al., 2002]
- Open WordNets: Onto.PT [Gonçalo Oliveira and Gomes, 2014], OpenWordNet-PT [de Paiva et al., 2012], PULO [Simões and Guinovart, 2014], UFES-WordNet [Gomes et al., 2013]
- Description: origins, creation, sizes, and usage restrictions
- Quantitative comparison
- Potential collaboration between open WordNets

Previous Review: Santos et al. 2010

Name	License	Size	started
PWN	Open	117K	1986
WordNet.PT (Marrafa)	Closed	19K	1998
TEP, Wordnet.BR	Open?	19K	2000
PAPEL	Open	191K triples	2007
Port4NooJ	Open	10K	2008
MWN.PT	Closed	17K	2008

"Although there appears to be enough material... we are still far from having well documented and a consensus..."

Multilingual Wordnets

- EuroWordNet (Vossen) and MultiWordNet (Pianta and Bentivogli)
- Multilingual Central Repository (Gonzalez-Agirre at el)
- Open Multilingual WordNet (Francis)
- Align Wordnets with other resources: YAGO, BabelNet, SUMO, DOLCE etc.

Closed Portuguese Wordnets

- WordNet.PT (Marrafa) follow EuroWordNet. Not available online.
- WordNet.BR (Dias-da-Silva), manually from corpora and dictionaries. Second version not available. First version not aligned with PWN.
- MultiWordNet.PT.

Open Portuguese Wordnets

Open wordnets for Portuguese appeared in the early 2010s.

Onto.PT

http://ontopt.dei.uc.pt/

- Initially developed in the scope of Hugo Gonçalo Oliveira's PhD (started in 2008)
- Created **automatically** through the exploitation of **Portuguese** public lexical resources
  - PAPEL lexical-semantic network
  - Broad range of relation types
  - Other dictionaries (so far, Dicionário Aberto and Wiktionary.PT)
  - Thesauri (TeP, OpenThesaurus)
  - Other wordnets (OpenWN-PT)
- Available in RDF/OWL
- Not aligned to any other wordnet

Onto.PT

http://ontopt.dei.uc.pt/

- Synset boundaries + relation attachments from scracth
- ECO approach, tailored for this project
  - **Extraction** of relations from text
    - Using the grammars of PAPEL
    - e.g. (*animal* member-of *gado*), (*rebanho* synonym-of *gado*)
  - **Clustering** synonyms as synsets
    - e.g. {rebanho, gado, manada}, {animal, bicho}
  - **Ontologising**: selecting the most suitable synset for each the arguments of each extracted relation
    - e.g. {rebanho, gado, manada} member-of {animal, bicho}

Onto.PT v0.6

http://ontopt.dei.uc.pt/

- ≈169,000 lexical items
- ≈117,000 synsets
    - nouns (≈68,000), verbs (≈26,000), adjectives (≈21,000), adverbs (≈2,000)
- ≈174,000 direct connections between synsets
    - hypernymy, part-of, member-of, causation-of, purpose-of, property-of, manner-of, location-of,. . .

Onto.PT

Future Developments

- CONTO.PT: a **fuzzy** wordnet for Portuguese
    - Adapt ECO
    - Compute fuzzy memberships automatically
        - Membership of each word in a synset
        - Membership of each synset connection
    - **Redundancy** across open Portuguese lexical resources
        - Can be used as **confidence measures**!

Examples
condição(0.97), *disposição*(0.92), *situação*(0.88) hypernym-of(0.82) *crispação*(0.8), *tensão*(0.73), *contração*(0.6)
enfeite(1.0), *adorno*(0.98), *ornato*(0.80) fazSeCom(0.42) *jarro*(1.71), *jarra*(1.29), *vaso*(0.63)
pressentir(1.73), *prognosticar*(1.73), *prever*(1.61) accaoQueCausa(0.45) *prognóstico*(2.0), *presságio*(1.77), *vaticínio*(1.74)

PULO

http://www.wordnet.pt/

- Stands for *Portuguese Unified Lexical Ontology*;
- Bootstrapped automatically using:
  - Wordnets from other languages:
    English, Galician, Spanish, Catalan;
  - probabilistic translation dictionaries from parallel corpora;
  - translation dictionaries from Apertium MT system;
- Aligned with Princeton WordNet:
  - Uses Multilingual Central Repository (MCR) structure;
  - Directly aligned with the Spanish Official Languages wordnets;
  - Allows researchers from each one of these languages to integrate Portuguese in their systems;
- Glosses translated by MT using MyMemory.translated.net

PULO - Current contents

http://www.wordnet.pt/

	Variants	**Synsets**
Nouns	14.084	9.994
Adjectives	4.837	3.556
Adverbs	596	521
Verbs	6.184	3.770
Total	25.701	17.841

Variants are the “lexical items”.

PULO

Future Developments

- Develop other methods, and exploit other resources to:
  - enlarge the number of synsets and variants;
  - mark variants for review or deletion;
- Include *crowdsourcing* tools in the website:
  - allow users to rate variants;
  - allow users to suggest variants;
- Perform manual evaluation and revision:
  - variants marked for revision;
  - low rated variants;
  - suggested new variants;
- Cross data with other resources (Onto.PT, OWN.PT);
- Review gloss translations;
- Perform example translations;

OpenWordnet-PT

http://wnpt.brlcloud.com/wn/

- Goal : not a simple translation of PWN, based on PWN architecture.
- originally created from a (PT) projection of the Universal WordNet (Gerard de Melo)
- Three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries.
  - Corpora: AC/DC project, DHBB CPDOC/FGV etc.
  - LR: morphosemantic links, nominalizations from NomLex, Nomage and Wiktionary etc.
- Freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
- Different vocabularies for RDF encoding!
- used by "Google Translate", FreeLing, OMW, BabelNet and Onto.PT.

OpenWordnet-PT - recent additions

http://wnpt.brlcloud.com/wn/

- Translations of glosses and examples.
- Work on using morphosemantic links from PWN to improve our resource, the nominalizations, synsets revisions.
- Improvements in the web interface for votes and suggestions.
- IBM BlueMix API improvements (undergoing) for application developers (mobile apps etc)

OpenWordnet-PT - challenges

http://wnpt.brlcloud.com/wn/

- Variants of Portuguese to include? How?
- Limits of the lexicon: colloquialisms, coarse language?
- PWN concepts with no direct translation into Portuguese?
- Adding new synsets without losing interoperability and consistency?
- PWN has a number of A-BOX synsets (about entities) and US-centred concepts. Demonyms should be in the lexicon, but related geografical names may not.
- Fellbaum lexicon-grammar example: Portuguese particle “-se”
- Other related issues address by Fellbaum and ILI initiative.

“a word to identify residents or natives of a particular place, which is derived from the name of that particular place.”
https://en.wikipedia.org/wiki/Demonym

Comparing

Name	Creation		Update	Usage
	Synsets	Relations		
WN.PT	manual	manual	manual	closed
WN.BR	manual	transitivity	manual?	free synsets
MWN.PT	manual? trans.	transitivity	?	paid license
Onto.PT	RE,*clustering*	RE,*clustering*	automatic	free
OpenWN-PT	UWN project.	transitivity	semi-autom	free
UfesWN.BR	MT	transitivity	?	free
PULO	triangulation	transitivity	semi-autom	free

Comparing

- Fully automatic construction approach leads to a larger resource.
- An intrinsic trade-off between the size of a wordnet and the accuracy and usefulness of the resource under scrutiny.

Conclusions

- Linguistic resources are very easy to start, hard to improve and extremely difficult to maintain.
- Size of lexical resources are easy to compare, quality is hard.
- GWA *must* be more active. Data is not shared, can be improved. code and tools are not easily shared.
- Portuguese is becoming less of a problem of finding a workaround solution, and increasingly more one of choosing the most suit- able within the available alternatives.

References

de Paiva, V., Rademaker, A., and de Melo, G. (2012).
OpenWordNet-PT: An Open Brazilian WordNet for Reasoning.
In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Dias-da-Silva, B. C. (2006).
Wordnet.Br: An exercise of human language technology research.
In *Proceedings of 3rd International WordNet Conference (GWC)*, GWC 2006, pages 301–303, South Jeju Island, Korea.

Gomes, M. M., Beltrame, W., and Cury, D. (2013).
Automatic construction of brazilian portuguese WordNet.
In *Proceedings of X National Meeting on Artificial and Computational Intelligence*, ENIAC 2013.

Gonçalo Oliveira, H. and Gomes, P. (2014).
ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically.
Language Resources and Evaluation, 48(2):373–393.

Marrafa, P. (2001).
WordNet do Português: uma base de dados de conhecimento linguístico.
Instituto Camões.

Pianta, E., Bentivogli, L., and Girardi, C. (2002).
MultiWordNet: developing an aligned multilingual database.
In *Proceedings of 1st International Conference on Global WordNet*, GWC 2002.

Simões, A. and Guinovart, X. G. (2014).
Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets.
In *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, volume 8854 of *LNCS*, pages 239–248. Springer.