Making Virtue of Necessity: a Verb Lexicon

Valeria de Paiva[2] Fabricio Chalub[1] Livy Real[1]
Alexandre Rademaker[1,3]

[1]IBM Research, Brazil

[2]Nuance Communications, USA

[3]FGV/EMAp, Brazil

PROPOR 2016, Tomar

OpenWordnet-PT

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

- Not a simple translation of PWN. Based on PWN architecture, a true thesaurus and dictionary for the Portuguese language, based on lexical relations
- Three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries.
- Freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
- Used by Google Translate, FreeLing, OMW, BabelNet, Onto.PT, etc.

OpenWordnet-PT and DHBB

Motivation

- Side project on historical information extraction from 2014.
- Using highly regarded by Brazilian historians "Dicionário Histórico-Biográfico Brasileiro" (DHBB).
- This is Brazilian Historical and Biographical Dictionary – entries on Brazilian History from 1930 onwards.
- long running project (since 1978) of Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) of the Fundação Getulio Vargas (FGV).
- Data available via `http://cpdoc.fgv.br`, `github.com/cpdoc`
- Previous publication on Digital Humanities Conference.

`http://wnpt.brlcloud.com/kb-extraction/search?db=dhbb&term=*`

DHBB

Cont.

- nice corpus for information extraction, the writers of the entries were asked to follow a set of guidelines with respect to the information that these entries about the historical figures should contain.
- processing this corpus we needed to deal with named entities (NER), and dates for events extraction.

Nominalizations

Previous Work

Nominalizations, nouns formed from other POS words, i.e. "construction" and "government", are one of most well known polysemous and problematic issues of formal theories in Linguistics.

We developed a smaller lexical resource, a lexicon of nominalizations in Portuguese called NomLex-PT, embedded into OpenWordnet-PT, with aprox. 4,240 pairs verb/noun.

Semi-automatically translated the original English NomLex, the French Nomage, the Spanish AnCora-Nom and manually verified.

Worrying about the missing truly Portuguese deverbals, we also used Portuguese corpora (the AC/DC corpora) to complete our collection of nominalizations.

Nominalizations

Cont.

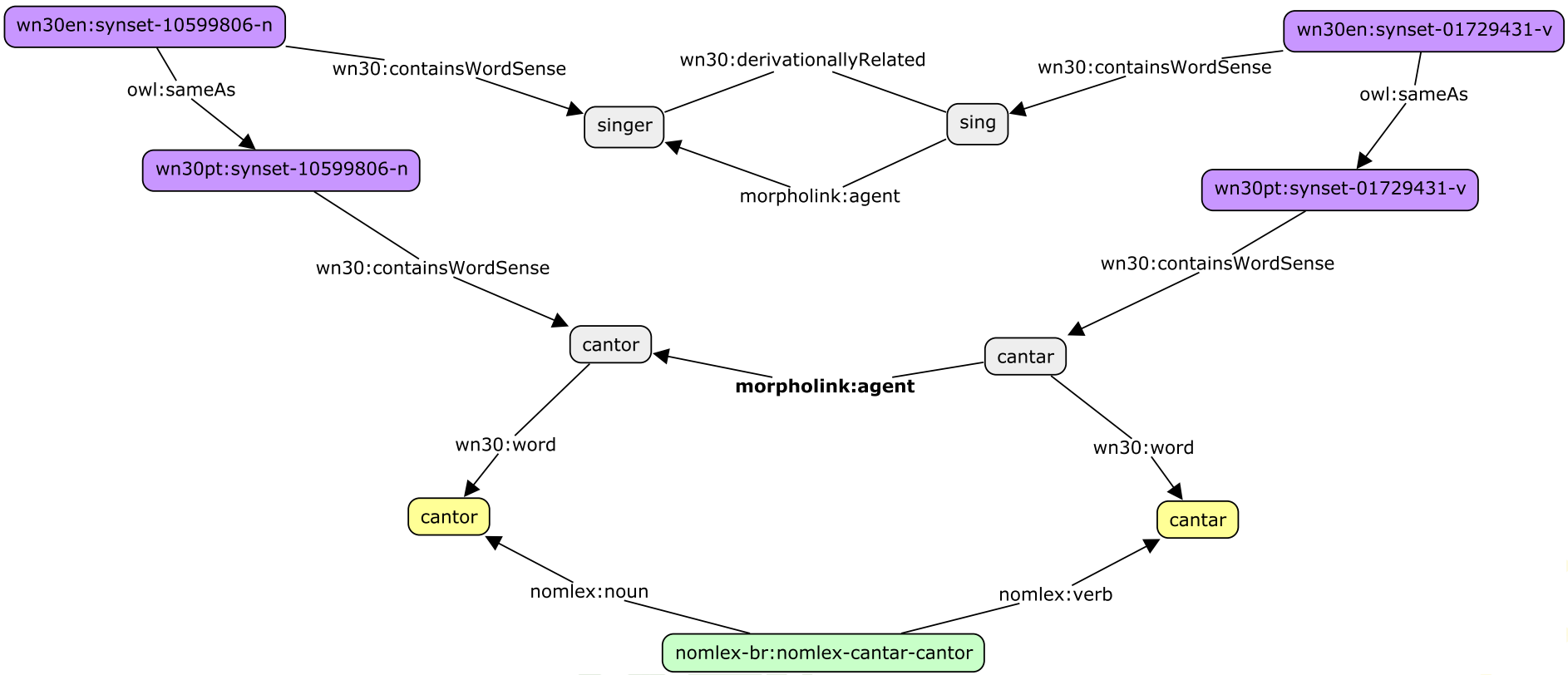
- Nominals have a clear semantic relation with the verb, but their meanings are not automatically derivable from the meaning of the base verb.
- . . . nor are they directly obtainable from the composition between the meaning of the base verb and its suffix.
- *Government*, i.e., has suffix *-ment* which, in general means "the event of doing X", but *government* (and the Portuguese *governo*) has several meanings: the event of governing, the result of governing, the period of time some governing happened, the people that govern, etc.
- We want the nominalization meanings encoded in the lexicon, as their formation can provide more semantic information.
- We started Nomlex without knowing about the PWN semantic links.

Morphosemantic links from PWN

Relation	Example
agent	*employ-employer*
body-part	*abduct-abductor*
by-means-of	*dilate-dilator*
destination	*tee-tee*
event	*employ-employment*
instrument	*poke-poker*
location	*bath-bath*
material	*insulate-insulator*
property	*cool-cool*
result	*liquefy-liquid*
state	*transcend-transcendence*
undergoer	*employee-employ*
uses	*harness-harness*
vehicle	*kayak-kayak*

Projecting the morphosemantic links

Cont.



A Portuguese Verb Lexicon?

Goal: investigate gaps and extend coverage of the verb lexicon of OpenWordNet-PT

- Why worry about verbs?
- How to go about it?
- Solved task?

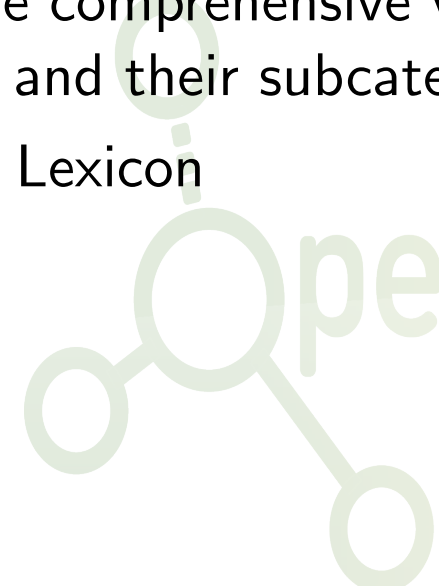
Portuguese Verb Lexicon

Motivation

- Verbs are the main bearers of meaning in sentences.
- Primary vehicle for describing events and expressing relations between entities
- Canonicalization of natural language statements requires predicates and its arguments
- Derivation of (plausible) inferences from such predicates requires lexicon markings
- Complete and improve OpenWordnet-PT's lexicon

Portuguese Verbs

- For the verbs already in OWN-PT, we can provide some indication of meaning, by giving other words related to the verb, and in the SUMO ontology.
- 4th most spoken language in the world; 3th most used in Facebook! (invited speaker from 'Instituto Camões')
- Still no freely available comprehensive verb lexicon that provides verbs, their meanings and their subcategorization frames
- We need such a Verb Lexicon
- Here are first steps



Related Work

- VerbNet.BR: computational work, very encompassing, but it has not been verified for consistency or accuracy.
- Viper: not open source.
- TeP: unclear licensing status and its definitive version is, apparently, not available yet.
- Catalog of Brazilian Portuguese Verbs
- others?

OpenWordNet-PT

Some numbers

- 5902 verbal synsets in Portuguese
- 4511 verbal lemmas
- 7865 synsets in English, empty in Portuguese
- Example
- which ones are easy missing cases? “popularize”
- which ones are impossible cases? “apaulistar”
- how to go about it? It is always easier to check whether one has coverage of a lexical resource than accuracy.

Modus Operandi

- To find where to fit in the PWN network the 'missing' Portuguese verbs from the golden VerbNet.BR.
- we translate the desired Portuguese verbs using machine translation and then we manually verify the translation.
- A list of words in Portuguese and corresponding words in English is then fed to an algorithm that looks for strict matches both of Portuguese and English words, in synsets and in glosses and then suggests these synsets to the human annotators.
- Finally at least two human annotators have to agree on the appropriateness of the word sense and its placement into the network to make it part of the official resource.

Golden VerbNet.BR

- manually verified golden subset.
- 50 verbs were found to be missing from OpenWordNet-PT from the 604 verbs in the golden subset of VerbNet.BR. Added.
- exception of two verbs, we did not find perfect synsets for them.
  - *entreabrir* 'to partially open' – conceptualization that seems to be done via an adverb in English
  - *rebolar* 'to move your hips in a rolling way'.
- typos and misspellings: captura/capturar
- different ways of writing: adjectivar/adjetivar, we can't ignore them in spite of the Portuguese Language Orthographic Agreement.

Golden VerbNet.BR

Cont.

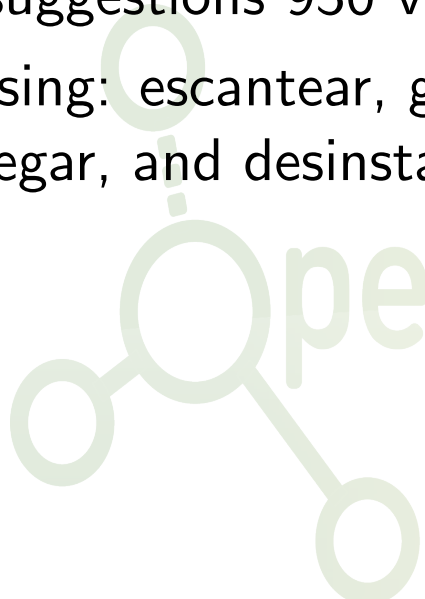
- many English verbs ‘pack in’ an adverb or two.
  - to jog is to run slowly or walk fast, hence between *correr* and *andar* in Portuguese, for the fun of it.
  - In Portuguese we have no verb between running and walking, we need the adverbs slowly, quickly and we need to indicate that the purpose is fun.
- different kinds of affixes: auto-excluir/self-exclude.
- one of the main problems, the lack of frequency/popularity of lexical items. We have no reliable frequency data, it is hard to decide on the level of coverage that is required.

Basic Coverage

- First we used a list of the thousand most common Portuguese verbs as collected by the 'Corpus do Português'
- Then we investigated a Swadesh list of the most important Portuguese words: based on meanings he presumed would be available in as many cultures as possible
- We used the Open Language Archives Community (OLAC) of the University of Pennsylvania.
- We found two verbs that we did not have (fender/'to split', desamolar/'blunt'), which we added in, but that are not that common in Brazilian Portuguese.

VerbOcean

- Textual entailment (traditional kind), using logical forms, can benefit from relations of entailment and causation between verbs. PWN does not have many of these relations.
- 2119 verbs in VerbOcean, we already had in OWN-PT 1182 verbs. Now we also have in suggestions 930 verbs.
- only six verbs still missing: escantear, gazetear, prototipar, reconfigurar, subempregar, and desinstalar.



VerbOcean

Cont.

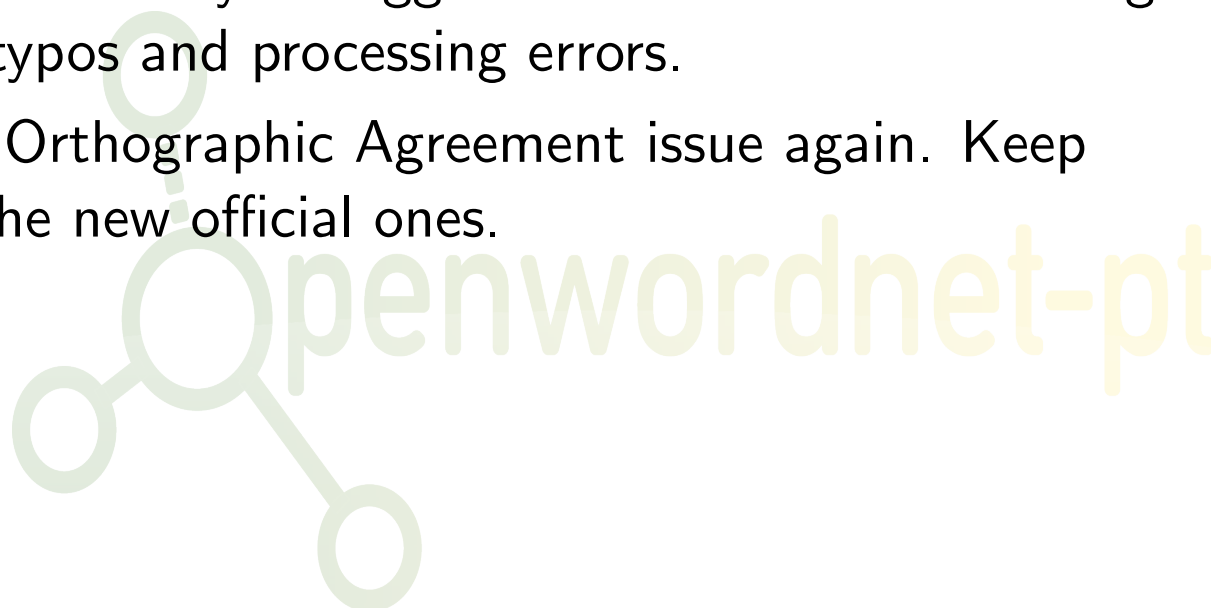
- Even if morphologically related, sometimes words can have very different meanings, the so-called semantic drifting.
- gazette in English means to publish in a gazette, in Portuguese the verb gazetear means to play truant.
- prototipar 'to prototype', desinstalar 'to uninstall', and reconfigurar 'to reconfigure' are from technology and hence should exist in English, but they are not in PWN.
- In English *underpay* for the practice of paying less than customary to workers, but in Portuguese we prefer to say subempregar, or 'under-employ'.
- issues with different national sports. In PWN many related with baseball, American football, golf and basketball (e.g. to tee in golf). In Brazilian PT expressions derived from soccer, as *escantear*.

Corpus Bosque

- news sources, reviewed by trained, native speaker linguists.
- a massive number of verbs were not available in OpenWordNet-PT, in any of their senses.
- we have 1981 verbs in Bosque-UD. We had already in OWN-PT 1043 of these. We added suggestions to 831 synsets.
- misspellings and typos (theoretical decision not to touch the contents of the texts themselves).
- While meaning can be translated from language to language, different languages will conceptualize different realities: abrasileirar, aportuguesar, apaulistar etc.
- Most of the cases of the missing from OWN-PT: differences in prefixes used, and cases of adjectives and nouns that are made into verbs in Portuguese, but not in English: indeterminar/'not determining something'. biografar/'to write a biography'.

Diário Gaúcho

- popular newspaper from the south of Brazil, hoping to find colloquial verbs not in OWN-PT. Aprox. 5 millions of tokens and the news were extracted from newspaper issues from 2008.
- Actually out of all the 2042 verbs in the corpus, 1044 were in OWN-PT and 937 were already in suggestions. Most of the missing 61 verbs are actually typos and processing errors.
- Portuguese Language Orthographic Agreement issue again. Keep both forms: old and the new official ones.



DHBB

- We still have 51 such verbs missing.
- some specific items from the politics domain (e.g. the verb subsecretariar, 'to act as a subsecretary') and some oddities that need investigation (e.g verbs pedrar, extremar and bondar).
- together with the other corpora, 150 verbs that we think deserve new Portuguese synsets.
- interesting social differences: several different verbs in Portuguese for graduating from college bacharelar, graduar, formar, doutorar, mestrar, while there is simply graduate in PWN.
- three different ways of expressing the meaning of separate from your spouse in Portuguese, with different legal status, descasar, desquitar, divorciar, of which only the last one exists as such in PWN.

Viper

- Thanks to Jorge Baptista and Nuno Mamede.
- 307 verbs in OWN-PT not in Viper: low frequency verbs.
- some erros and some with prefixes in OWN-PT.
- aprox. 10-20 cases of missing in Viper. Nice to contribute with other resources.

Viper

Cont.

# entrires	# verbs
0	307
1	2130
2	476
3	186
4	82
5	25
6	15
7	4
8	4
9	1
10	1
12	1
13	1

Conclusions

- PWN has 13767 verbal synsets. More than half of these synsets have no words in Portuguese. How many of these really constitute synsets that should not exist in a Portuguese wordnet?
- we do not have, as yet, an worked-out measure for accuracy or adequacy of our resource. Quality is difficult to measure.
- Finish to add the morphosemantic links can help wordnets to correct:
  - mistakes and omissions
  - failings of sparsity of linking between synsets
  - too fine-grained character of some synsets (GWA is working on the ILI)

Conclusions

Cont.

- bootstrap a comprehensive lexicon of subcategorization frames from both the minimal frames already present in Princeton WordNet and the annotated corpora available. Features for machine learning of semantic roles.
- still debating how to best present information, as PWN info is not available in their interface and we reckon that showing is informative for users both in *en* and in *pt*. Following OWN for the time being.
- we need to come up with principled ways of extending OpenWordNet-PT.
- on a different direction, we would like to find ways of verifying the Portuguese glosses
- Acknowledge the helpful work of Alberto Simões with the automatically translated glosses from PULO.

Linguistic resources are very easy to start working on, very hard to improve and extremely difficult to maintain.

Thanks!