Introducing OpenWordnet-PT: a open Portuguese wordnet for reasoning

Alexandre Rademaker[1,3] Valeria de Paiva[2] Fabricio Chalub[1]
Livy Real[1] Claudia Freitas[4]

[1]IBM Research, Brazil

[2]Nuance Communications, USA

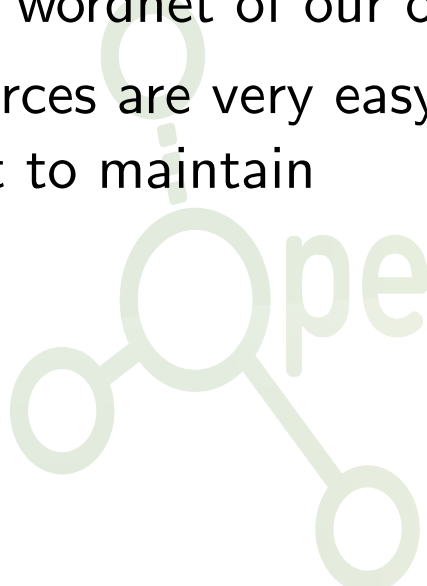
[3]FGV/EMAp, Brazil

[4]PUC-Rio, Brazil

FrameNet Workshop 2016, Juiz de Fora

Lexical Resources are Important

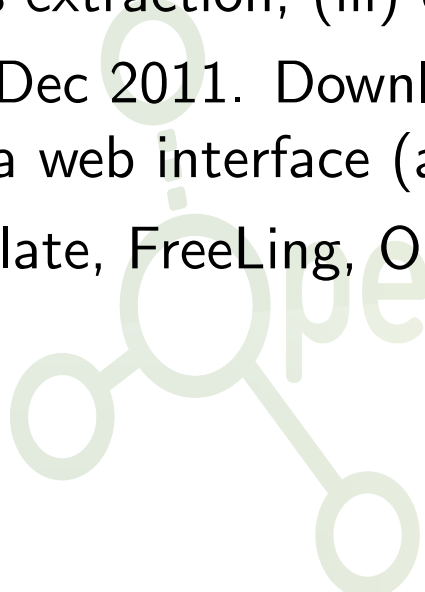
- Possibly do not need to explain it here, but...
- Semantic relations are a key aspect when developing computer programs capable of handling language
- Princeton WordNet very useful in many applications
- Want a free and open wordnet of our own
- However, lexical resources are very easy to start, very hard to improve and extremely difficult to maintain



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`

`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.
- Tokenization, lemmatization and WSD are not solved tasks! Errors propagate, i.e., “foi” to “ser” instead of “ir”.

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.

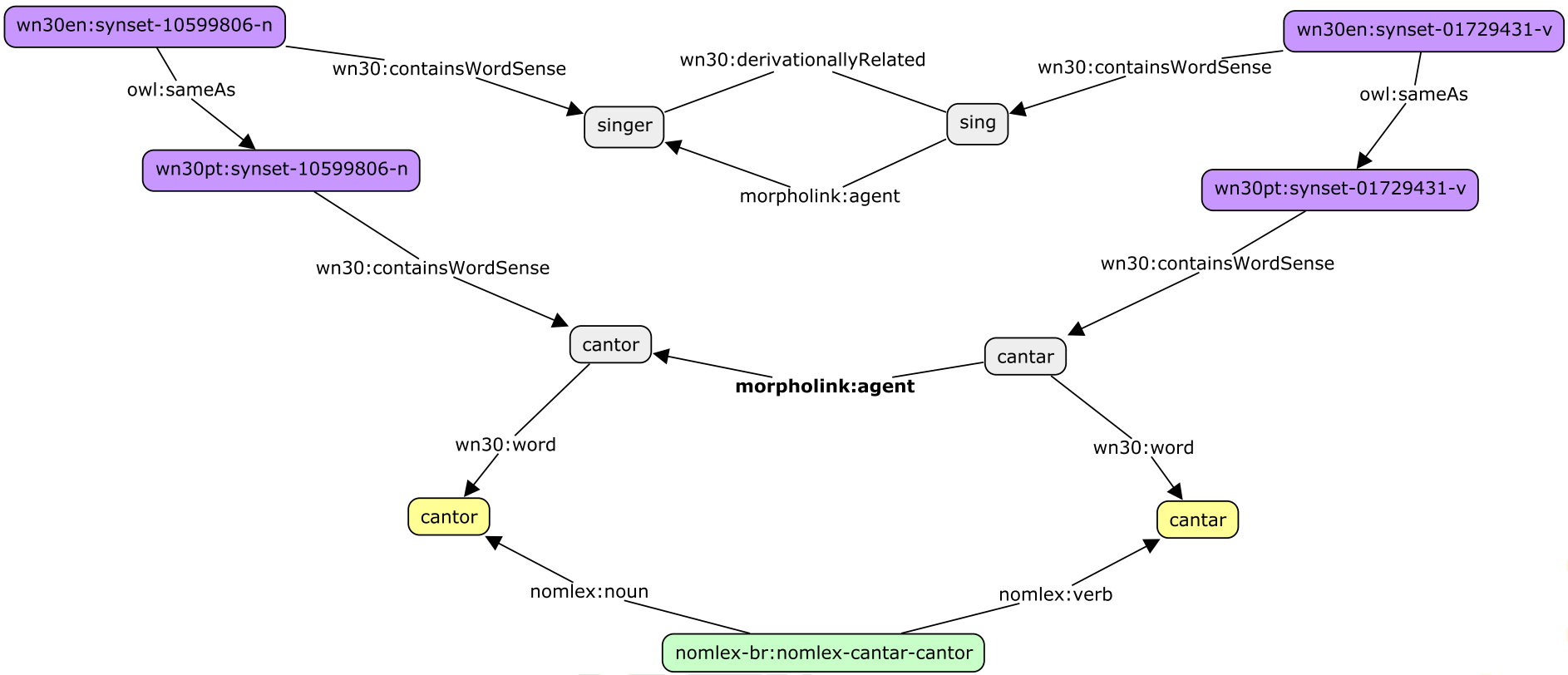
- Nominals have a clear semantic relation with the verb, but their meanings are not automatically derivable from the meaning of the base verb.
- . . . nor are they directly obtainable from the composition between the meaning of the base verb and its suffix.
- *Government*, i.e., has suffix *-ment* which, 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.



Portuguese Verbs

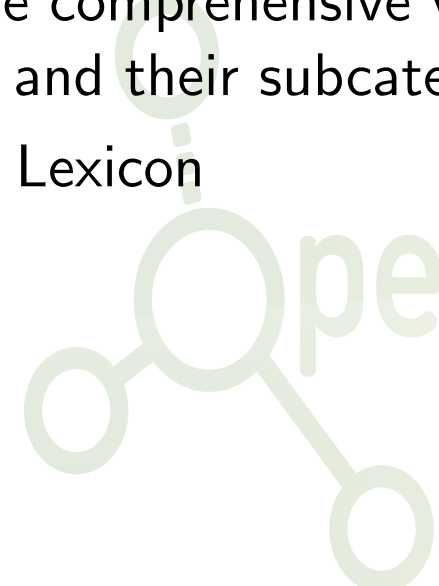
Motivation

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

- 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; 3rd most used in Facebook! (source '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



Portuguese Verbs

Some numbers

- 5902 verbal synsets in Portuguese
- 4511 verbal lemmas
- 7865 synsets in English, empty in Portuguese
- which ones are clear missing cases? “popularize/popularizar, dribble/driblar” (both already in suggestions!)
- which ones shouldn't be in PWN? “apaulistar”, “sambar” etc.
- How to go about it? It is always easier to check whether one has coverage of a lexical resource than accuracy.

Portuguese Verbs

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'.

Portuguese Verbs

Corpus DHBB

- We still have 51 such verbs missing (considering the verbs with at least +10 ocurrences)
- 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.

Demo

openWordnet-PT Demo

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

OWN-PT and FrameNet

collaboration possibilities

- Use FrameNet-BR frames to check OWN-PT's coverage (ongoing)
- Create 'Historical Frames' for DHBB: what's in each biographical entry? birth place, time? graduation frame? occupation frame? etc.
- How to connect to locations/people/organizations?
- m.knob/BabelNet and SUMO? How FrameNet.BR is using? What is the best approach for linking lexical resource to world knowledge?
- Perhaps MWEs?

A concern: Law is very different in English vs. Portuguese. Same problem with Legislation? (The Limits of Using FrameNet Frames to Build a Legal Ontology)

FrameNet-BR and OpenWordnet-PT

Lexical Intersection

- Basic first step for start any collaboration.
- 23 verbs, 480 nouns and 1 adj missing? Not bad!
- most missing verbs are compounds such as: “queimar a largada”, “perder gol”, “pegar rebote”, “marcar falta” etc.
- Only two really missing “single word” verbs? “driblar” and “quicar” (neologism).

FrameNet-BR and OpenWordnet-PT

Lexical Intersection

senses	none	a	adv	c	n	prep	v
0	1	1	0	1	480	1	23
1	1	2	0	0	244	1	35
2	0	7	2	0	176	0	24
3	3	4	1	0	106	1	48
4	0	2	0	0	63	0	27
5	0	3	0	0	36	0	15
6	0	1	0	0	38	0	35
7	0	0	0	0	35	0	19
8	0	1	0	0	12	0	14
9	0	0	0	0	7	0	32
10	1	0	0	0	2	0	9
11	0	0	0	0	2	0	13
12	0	2	0	0	4	0	14
13	0	0	0	0	4	0	7
14	0	0	0	0	5	0	8
15	0	0	0	0	3	0	4
16	0	0	0	0	0	0	6
17	0	0	0	0	0	0	7
18	0	0	0	0	1	0	4
20	0	0	0	0	0	0	5
21	0	0	0	0	0	0	3
23	0	0	0	0	0	0	3
24	0	0	0	0	0	0	5
26	0	0	0	0	0	0	4
27	0	0	0	0	0	0	1
41	0	0	0	0	0	0	3

FrameNet-BR and OpenWordnet-PT

Some english words in the FrameNet-BR PB LUs. Missing in OpenWordnet-PT and PWN:

LU	senses
back half twist	0
back swing	0
back three quarter	0
back	0
backhand clear	0
backhand	0
badminton	0
wazari	0
whipback	0
wipe-out	0
withdraw	0
wurst	0
yuko	0

FrameNet-BR and OpenWordnet-PT

Terms releated to sports:

LU	senses
jogador de badminton	0
jogador de basquete	0
jogador de handball	0
jogador de hóquei sobre grama	0
jogador de pólo	0
jogador de rúbi	0
jogador de vôlei	0

Some terms related with brazilian food: “buchada de bode” and “goiabada”. New synsets!

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, we reckon that showing is informative for users both in *en*/*pt*. Following OWN for the time being.
- we need to come up with principled ways of extending (new synsets) OpenWordNet-PT.
- on a different direction, we would like to find ways of verifying the Portuguese glosses and quality.
- OpenWordnet-PT maybe provide "type" for lexical units of FrameNet-BR, avoiding the "frames as LU sets"?

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

Thanks!