OpenWordnet-PT: A Project Report

Alexandre Rademaker[1,5] Valeria de Paiva[2] Gerard de Melo[3]
Livy Maria Real Coelho[4] Maria Gatti[5]

FGV/EMAp

Nunance Comm.

Tsinghua University

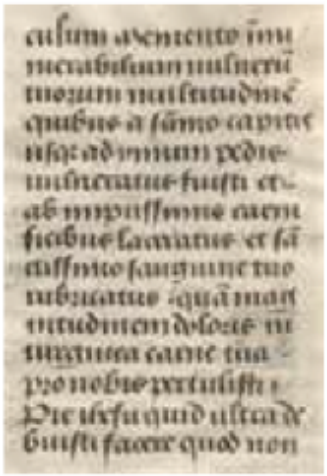
UFP

IBM Research

February 2, 2014

Why we started openWordnet-PT?

We need a Portuguese Wordnet for our work, but none of the previous projects is openly available.



Aren't all wordnets open?

Getulio Vargas Foundation (FGV)

Brazilian higher education and research institution founded in 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and Applied Mathematics. Its original goal was to train people for the country's public and private-sector management. Considered a top-5 policymaker think-tank worldwide.

`http://portal.fgv.br`

CPDOC - Center of Brazilian Contemporary History

A major center for teaching and researching in the Social Sciences and Contemporary History located in Rio de Janeiro. It holds:

- **Personal Archives (Acessus)** $\approx$ 200 archives, up to 1,8M docs or 5.2M pages (700K digitalized), among text (handwritten and printed), letters, memos, diaries, images and videos.
- **Oral History Program (PHO)** A huge set of testimonies (in audio and video) consisting of more than 2K interviews, which correspond to up to 6K hours of recordings. 90% in digital format. Almost all transcribed. Limit access, not online.
- **Brazilian Historical Biographic Dictionary (DHBB)** 7,5K entries, 6,5K are of biographical and 1K related to institutions, events and concepts of interest for the Brazilian history after 1930. Carefully revised entries by researchers. Few metadata.

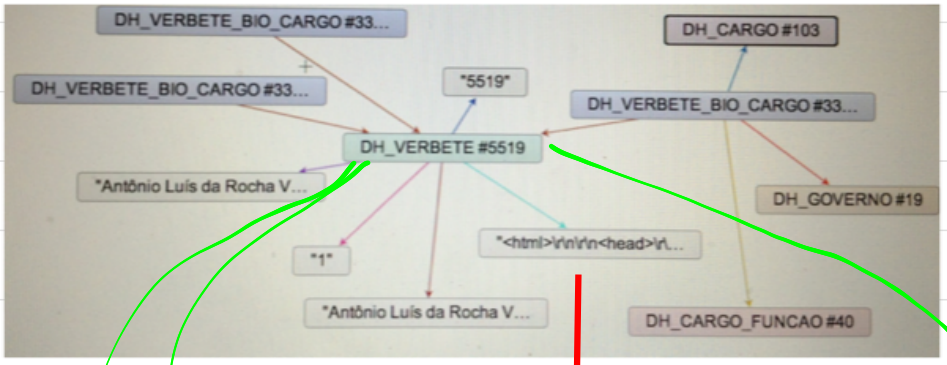
The Long Run Project

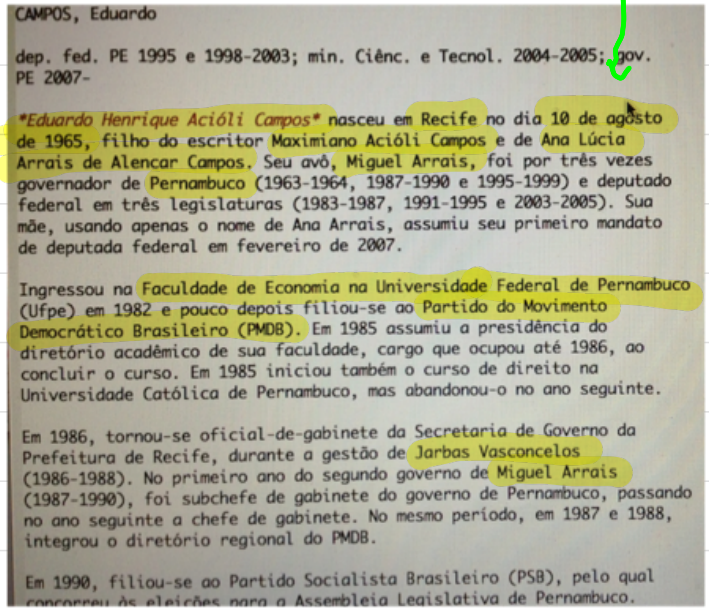
- Joint project between CPDOC and EMAp (Mathematical School);
- Enrich the structure (semantics) of CPDOC data;
- Open and expose CPDOC's data and architecture making it more maintainable and dynamic;
- Uniform and integrated data treatment (standards and interlinks between collections).

NLP of CPDOC's data

- Linking to dbpedia (Presidents of Brazil, presidents of the Senate, political parties etc)
- NLP and text mining of DHBB entries: (1) proper names; (2) word sense disambiguation using the openWordnet-PT; and (3) named entity recognition and creation of links between DHBB entries.
- 133,036 proper names identified (some few mistakes). Potentially entities (people, locations, organizations etc)
- Use grammars, lexical resources, formal ontologies, and logical tools to reason about knowledge obtained from processing text in Portuguese: QA, Knowledge Extraction, Computational Semantics (KB, KR and ATP).

NLP of CPDOC's data (cont.)





CAMPOS, Eduardo

dep. fed. PE 1995 e 1998-2003; min. Ciênc. e Tecnol. 2004-2005; gov. PE 2007-

Eduardo Henrique Acióli Campos nasceu em Recife no dia 10 de agosto de 1965, filho do escritor Maximiano Acióli Campos e de Ana Lúcia Arrais de Alencar Campos. Seu avô, Miguel Arrais, foi por três vezes governador de Pernambuco (1963-1964, 1987-1990 e 1995-1999) e deputado federal em três legislaturas (1983-1987, 1991-1995 e 2003-2005). Sua mãe, usando apenas o nome de Ana Arrais, assumiu seu primeiro mandato de deputada federal em fevereiro de 2007.

Ingressou na Faculdade de Economia na Universidade Federal de Pernambuco (Ufpe) em 1982 e pouco depois filiou-se ao Partido do Movimento Democrático Brasileiro (PMDB). Em 1985 assumiu a presidência do diretório acadêmico de sua faculdade, cargo que ocupou até 1986, ao concluir o curso. Em 1985 iniciou também o curso de direito na Universidade Católica de Pernambuco, mas abandonou-o no ano seguinte.

Em 1986, tornou-se oficial-de-gabinete da Secretaria de Governo da Prefeitura de Recife, durante a gestão de Jarbas Vasconcelos (1986-1988). No primeiro ano do segundo governo de Miguel Arrais (1987-1990), foi subchefe de gabinete do governo de Pernambuco, passando no ano seguinte a chefe de gabinete. No mesmo período, em 1987 e 1988, integrou o diretório regional do PMDB.

Em 1990, filiou-se ao Partido Socialista Brasileiro (PSB), pelo qual concorreu às eleições para a Assembleia Legislativa de Pernambuco.

Previous Portuguese Wordnets

- WordNet.PT e WordNet.PT Global (P. Marrafa) since 1999, part of EuroWordNet, 19K expressions, manually curated, online consulting only, some domains.
- MWN.PT - MultiWordnet of Portuguese (A. Branco), since 2008, part of MWN, over 17,200 manually validated concepts/synsets, not free.
- WN.Br (B. Dias da Silva) since 2000, not open, not available online. REBECA system (LREC 2010) only for "wheeled vehicles" domain, not clear the diff from Adam[1], based on WN.Pr 2.0. Some names confusion WordNet.br [2] and TEP [3].
- More recently, Onto.PT [4].

---

[1] **pease2009formal**.
[2] `http://www.nilc.icmc.usp.br/wordnetbr/`
[3] `http://www.nilc.icmc.usp.br/tep2/`
[4] `http://ontopt.dei.uc.pt`

OpenWordnet-PT: What?

- Leverage EuroWordNet, MultiWordNet, Global WordNet experience.
- Recruited Gerard de Melo for project. Leverage YAGO, UWN/Menta experience. A large-scale multilingual lexical knowledge base built using statistical methods, transforming WordNet into a massively multilingual resource.
- Portuguese "projection" of UWN/Menta is the basis of automated version of a OpenWordNet-PT, publicly available.

The basis

- Princeton WordNet 3.0 used to obtain English glosses and English terms for each synset.
- The unreleased 2010-12 version UWN and MENTA provided candidate terms in Portuguese, few candidate glosses in PT (from Wikipedia), and candidate terms in Spanish.
- The EuroWordNet base concept list (5000_bc.xml) provides the base concept numbers. The core concepts are also considered.
- The original file was mapped from WordNet 2.0 to 3.0 using the mappings from WN-Map. When multiple mappings for a WordNet 2.0 synset existed, all possible WordNet 3.0 synsets were kept.

OpenWordnet-PT: the method

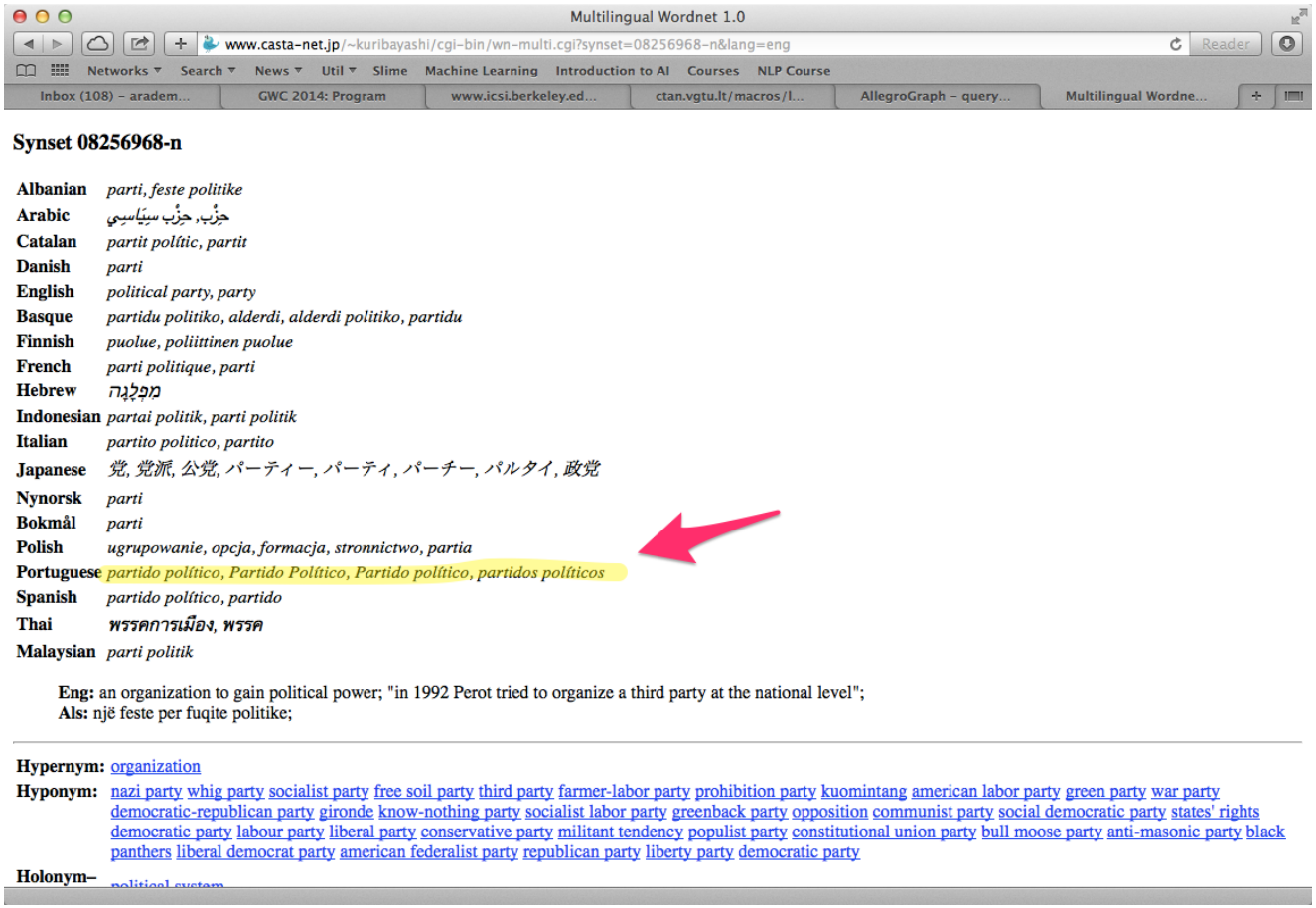
- a two-tiered methodology: high precision for the more frequent words of the language, but also high to cover a wide range of words in the long tail.
- Translation dictionaries to map the English members of a synset to possible Portuguese translation candidates. To disambiguate and choose the correct translations, feature vectors for possible translations are created by computing graph-based statistics in the graph of words, translations, and synsets. Monolingual wordnets and parallel corpora used to enrich this graph. Statistical learning techniques used to iteratively refine this information and build an output graph connecting Portuguese words to synsets.
- Wikipedia pages are then linked to relevant WordNet synsets by learning from similar graph-based features as well as gloss similarity scores.

OpenWordnet-PT: the method (cont.)

- To have high precision for the most important concepts of a language, rely on human annotators.
- Set of 4689 "Common Base Concepts" from GWA.
- 2,498 manually entered sense-word pairs as well as an additional 1,299 manually written Portuguese synset glosses. Native speakers, but not linguists. Plenty of errors.

Results

Good and bad cases: capitalized items, plurals, duplicates (6K words diff only in upper/lower case), a few gender issues, missing items (true lexical gaps?) etc. Easy and hard cases.



RDF Representation

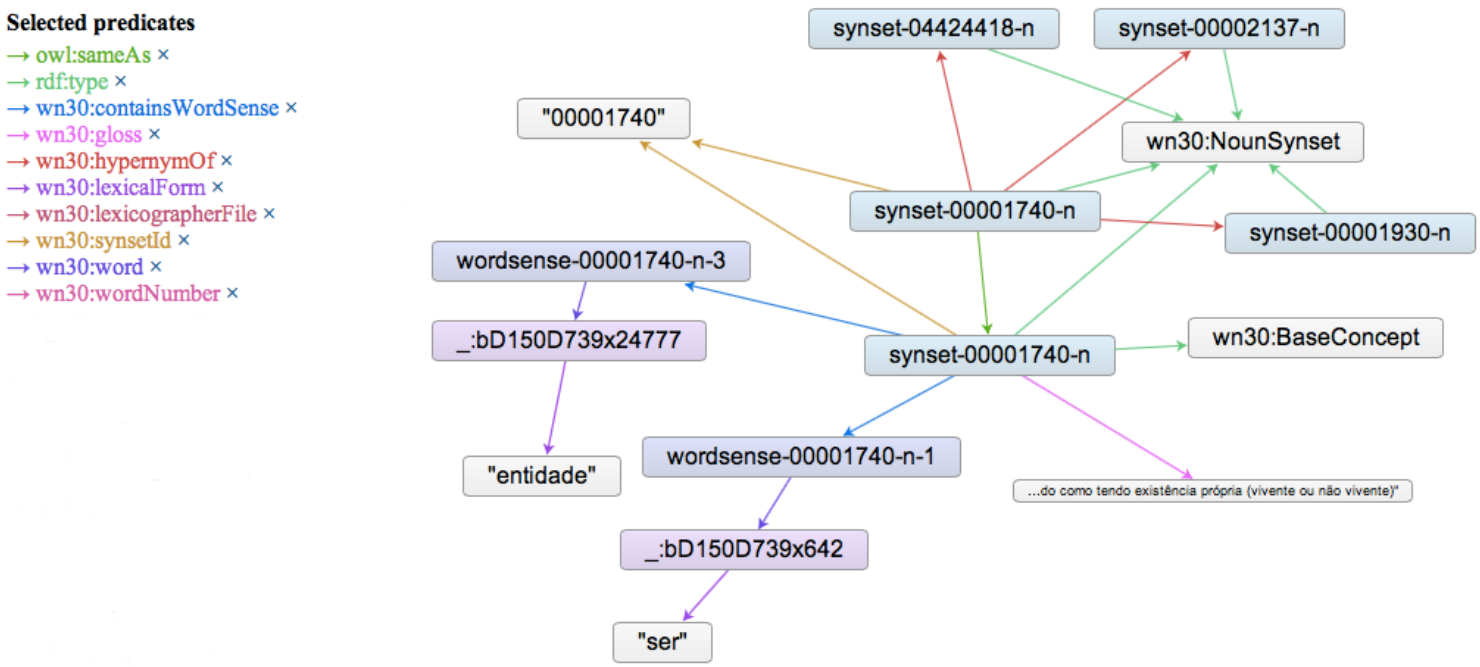
- Interoperability between wordnets. Linked Data and Semantic Web standards such as RDF and OWL.
- The emergence of Linked Data projects for lexical and reasoning resources make OpenWN-PT encoded and distributed in RDF/OWL.
- Standards allow both data model and data in the same format. Tools including databases (triple stores) with SQL-like query interfaces (SPARQL). Schema Free.
- Standard W3C encoding of WordNet in RDF since 2006[5]. OpenWN-PT is modelled after and fully interoperable with Princeton WordNet. Our own lisp parser [6].
- Part of a large ecosystem of compatible resources, including domain identifiers and mappings to Wikipedia.

[5] **wn-rdf**.

[6] `https://github.com/arademaker/wordnet2rdf`

RDF Representation (cont.)

One can easily find Portuguese equivalents for specific English word senses and vice versa. See `http://bit.ly/1aPxd7J`.



URIs for name resources

- `http://arademaker.github.com/wn30/schema/` (instead of `http://purl.org/vocabularies/princeton/wn30/` or `http://www.w3.org/2006/03/wn/wn20/schema/` or `http://wordnet.princeton.edu/wn20/schema/`)
- `http://arademaker.github.com/wn30/instances/`
- `http://arademaker.github.com/wn30-br/instances/`

We are still thinking in better and stable URIs!

Progress Report

- Checking is much easier than starting from scratch.
- But long and tedious work to check even the initial 5k synsets suggested by GWA (not done, yet!), let alone all synsets in OpenWN-PT.
- Necessary? YES! Lexical gaps of all sorts.
- But resource is being used.
- Improving the resource: new data from Bond[7] and some manual additions (NOMLEX-BR project).

	2011	2013	increase
synsets	41,810	43,895	5%
words	52,220	54,125	3%
senses	68,285	74,054	8%

[7] **bond-foster:2013:ACL2013**.

Synsets missing PT words by type

Edit query

Query language: SPARQL Query planner: default Result limit: 100

```
1 select ?type (count(?s) as ?total) {
2   ?s wn30:synsetId ?id .
3   ?s a ?type .
4   FILTER NOT EXISTS { ?s wn30:containsWordSense ?senes . }
5 }
6 group by ?type
7
```

Execute Save as Add to repository

Result

Download as SPARQL JSON

type	total
wn30:VerbSynset	"9095"
wn30:AdjectiveSynset	"12581"
wn30:BaseConcept	"750"
wn30:AdverbSynset	"2642"
wn30:NounSynset	"49431"

Synsets missing PT words by lexicographer File

See `http://bit.ly/1fm6fUC`.

lexFile	total_PT	total_Pr	percent
adj.ppl	5	60	8
verb.competition	100	459	22
noun.possession	271	1061	26
verb.creation	184	694	27
adv.all	979	3621	27
⋮	⋮	⋮	⋮
noun.phenomenon	324	641	51
noun.feeling	223	428	52
noun.object	908	1545	59
noun.location	2096	3209	65
noun.Tops	51	51	100

Use cases: FreeLing[8]

- Word Sense Disambiguation via FreeLing 3.0 An Open Source Suite of Language Analyzers.
- OpenWN-PT has been incorporated into FreeLing.
- A given Portuguese text can automatically be annotated with word senses

[8] **freeling**.

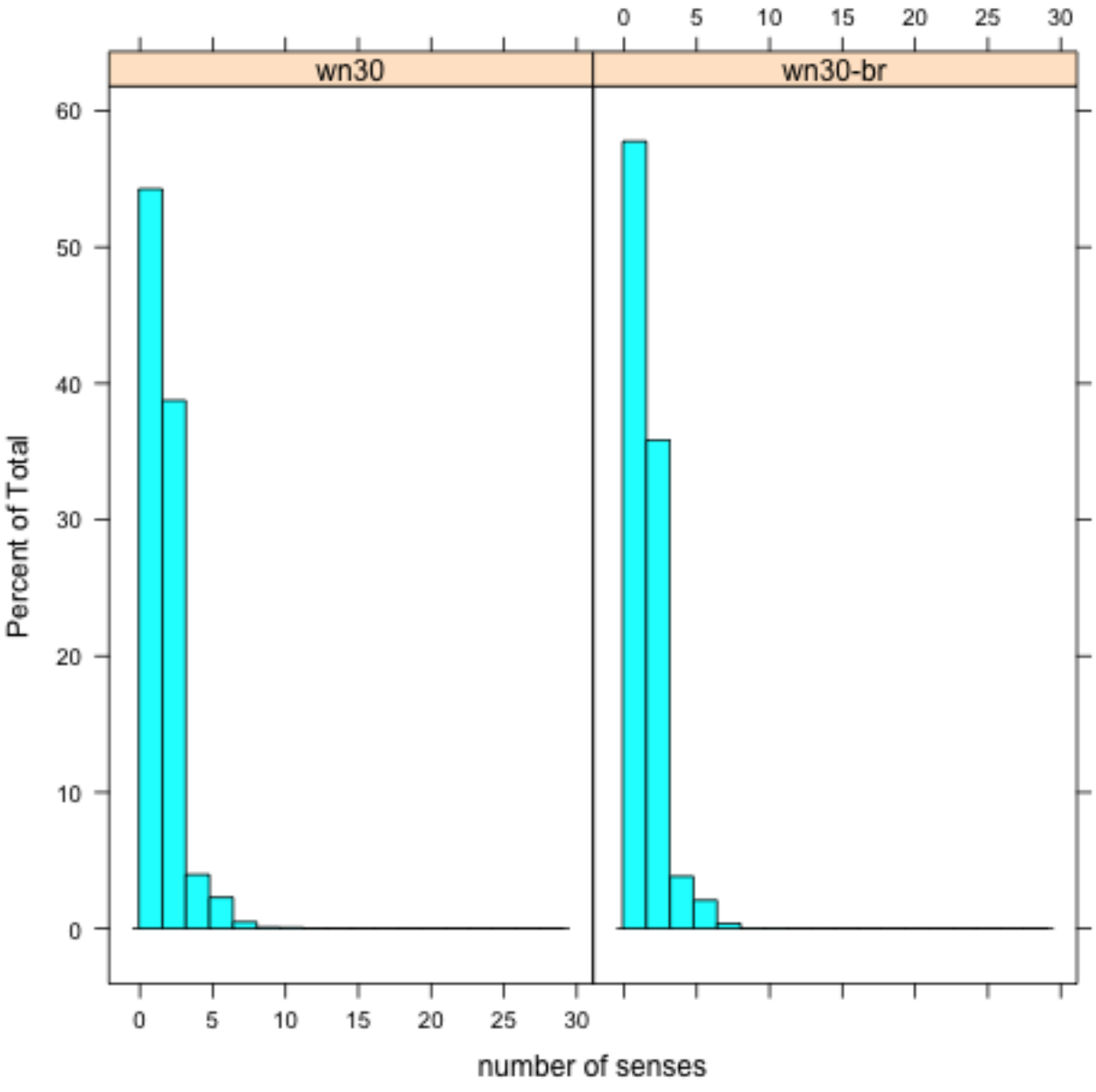
Use Cases: Sentiment Analysis

- Sentiment Analysis, using tweets about 2013 Confederation Coup games.
- OpenWN-PT and SentiWordNet to compare/develop the MachineLearning-based sentiment analysis integrated into IBM InfoSphere Streams (ISS) platform.
- 1 million tweets, 4 friendly matches Brazilian team in 2013, 7 classes of positivity
- IBM Research Brazil Project.

Use cases: Nomlex-BR

- Extension of OpenWN-PT aims at incorporating links to connect deverbal nouns with their corresponding verbs.
- We have created over 2,000 entries integrated into OpenWN-PT, will facilitate their use for linguistic research as well as information extraction
- Incorporating NOMLEX-BR data into OpenWN-PT has shown itself useful in pinpointing some issues with the coherence and richness of OpenWN-PT.
- the word **abasement** corresponds in NOMLEX to the verb **abase**, and thus we would like a similar correspondence between the Portuguese noun **aviltamento** and the verb **aviltar** (suggested translations). OpenWN-PT simply has two synsets "humilhar, abaixar" and "humilhar, rebaixar". The more common verb humilhar is repeated, while the uncommon aviltar was left out.
- More about Nomlex-BR in the last day of GWC 2014!

Miscellaneous Experiments: adding antonoym relations



OpenWordnet-PT: accuracy

- But how good are these entries? How to measure? How to improve?
- Following[9], from 6 relations (hypernymOf, memberHolonymOf, instanceOf, substanceHolonymOf, entails and causes) we randomly picked 30 pairs of synsets and then random words from each synset.
- From 180 sentences, 150 sentences marked as correct (83% of the sentences), 17 marked as wrong (one of the two words used to fill the template is probably placed in a wrong synset), and 13 marked as dubious.
- More experiments must be done. E.g. remove trivial pairs with same words.
- Some data mining could help. Synsets with an uncommonly high number of senses or words with an unexpected number of senses should be reviewed.

[9] **cruse1986**.

Conclusion

- We discussed the implementation and some applications of OpenWordNet-PT, an open Wordnet for Brazilian Portuguese.
- Recent improvements include better coverage and nominalization links connecting nouns and verbs.
- Used in high-throughput commercial system, cultural heritage project, hopefully more soo.
- Freely available from `http://github.com/arademaker/openWordnet-PT/` and a SPARQL Endpoint at `http://logics.emap.fgv.br:10035`.
- Browsing via Open Multilingual Wordnet is fun.

Next steps

- We are developing our own web interface for browsing and collaborative editing. Most important pending issue!
- First finish translating the "core" synsets in the Princeton WordNet to Portuguese.
- Finish to embed Nomlex-BR into OpenWN-PT (anchor floating words, `http://bit.ly/1aQdpkr`).
- Adding the Portuguese terms that satisfy different relations?
- Since we have a first target corpus, DHBB, we can also calculate word frequency to prioritize expansion of the OpenWN-PT and go back to the ontology building.
- Use and test the accuracy of the resource! More applications!
- OpenVerbNet-PT?
- **FOIS 2014 [10] Workshop, "Logics and Ontologies for non-English NLP". Website coming soon.**

[10] `http://fois2014.inf.ufes.br/`

Thanks! Obrigado!