Gentle with the Gentilics

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

[1]IBM Research, Brazil

[2]Nuance Communications, USA

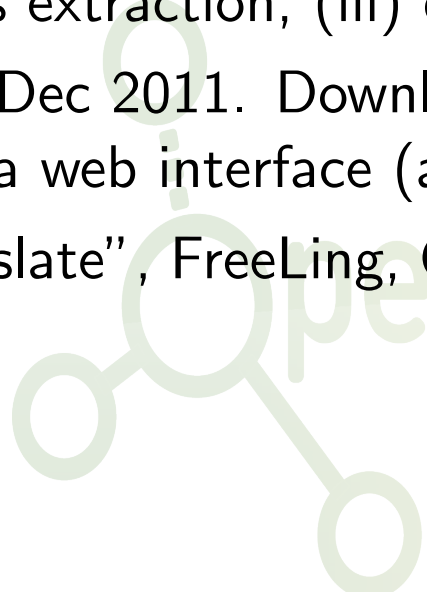
[3]FGV/EMAp, Brazil

May 23, 2016

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.
- 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 and Onto.PT.

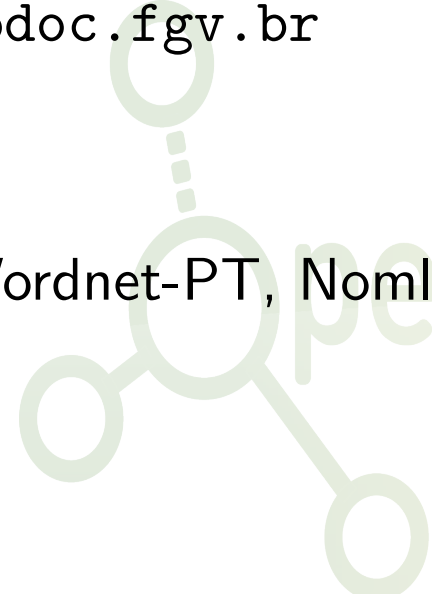


OpenWordnet-PT and DHBB

Motivation

We started in 2010 a project of extracting information from an dictionary of historical biographies, the "Dicionário Histórico-Biográfico Brasileiro" (the Brazilian Historical and Biographical Dictionary, shortened as DHBB), a longstanding project at the *Centro de Pesquisa e Documentação de História Contemporânea do Brasil* (CPDOC) of the *Fundação Getulio Vargas* (FGV). `http://cpdoc.fgv.br`

We use: FreeLing, OpenWordnet-PT, Nomlex-PT etc.



Gentilics

- Inferring from **Brasília is the Brazilian capital** that **Brasília is the capital of Brazil** is an obvious task for a human, but doing it automatically in NLP system requires some effort.
- Having this kind of information encoded in a lexical resource can help in several tasks.
- Deciding which kind of ontological information should be present in lexical resources, or specific knowledge bases, such as DBpedia, Wikidata, or Geonames is a complex decision.
- We deal in this paper mostly with *gentilics*, a class of pertainym adjectives that sits in between lexical and ontological knowledge and whose proper linguistic treatment requires access to ontological resources such as linked geo-spatial data and formal ontologies.

Relational, Pertainym and Gentilics

- We decided to investigate relational adjectives; as adjectives, they should appear in a lexical resource . . . But closely related to ontological knowledge;
- Pertainyms are adjectives that are associated with a base noun – *Brazilian/Brazil* and *fictional/fiction*. Defined as 'of pertaining to' another word.
- PWN has a separated lexicographer file ADJ.PERT (pertainym adjectives); 3661 adj.pert, of which 2617 had no translation to Portuguese in our OpenWordNet-PT (May 2015).
- But discovered that *gentilics*, a subclass containing adjectives pertaining only to *locational nouns*, offered enough challenges.

Pertainyms, Demonyms and Gentilics

- 'demonym' is a word created to identify residents or natives of a particular place; usually derived from the name of that particular place.
- Examples: *Chinese* (China), *Brazilian* (Brazil), *American* (United States of America or Americas as a whole).
- Just as a single demonym may refer to two different groups of natives, a particular group may be referred to by multiple demonyms, e.g. natives of the United Kingdom are the *British* or the *Britons*.
- The word *gentilic* comes from the Latin, the word *demonym* was derived from the Greek word meaning populace (*demos*) with the suffix for name (*-onym*). For English and Portuguese there is a generalized, but principled ambiguity.

Pertainyms, Demonyms and Gentilics

cont.

- *Brazilian/brasileiro*, without any context, we mean either the noun or the adjective.
- Natural ambiguity: `http://wnpt.brlcloud.com/wn/search?term=slovenian`
- We call *gentilics* the adjectives (pertainyms) and *demonyms* the associated location (noun).
- Finally, *toponyms* are place's names: United Kingdom, Brazil, Slovenia, Portorož etc.

Main question

What is linguistic knowledge vs. world knowledge? How much of world knowledge needs to be present in a lexical-ontological resource such as a wordnet?

GeoWordNet is a resource that fully merges the GeoNames database, Princeton WordNet 1.6 and the Italian portion of MultiWordnet.

But perhaps a wordnet does not need to have much geographical information, there are many geographic databases, they could be used instead of growing the number of synsets referring to locations. Language is tied up to culture and clearly when discussing the meanings of words in Portuguese we need to deal with meanings that do not exist in other languages. Mostly to places but also to religions, styles of philosophy, music etc.

DHBB use cases

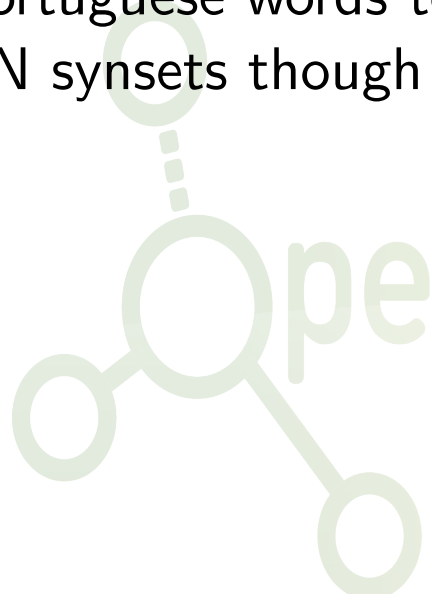
"...o deputado federal **pernambucano** Fernando Lira ...votou a favor da emenda da reeleição [...]" *The congressman* **from Pernambuco** *Fernando Lira voted in favor of the reelection amendment."*

See "paulista" (Paulo de Maio), "carioca" (O Nacional), "amazonense" (Partido Trabalhista Amazonense).

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

Completing and Expanding OpenWordnet-PT

- Before starting creating new synsets for the gentilics of the states and cities in Brazil (e.g. *paulistano, amazonense*) we needed to complete the gentilics present in PWN synsets with no Portuguese words in the corresponding OWN-PT synset.
- Adding the missing Portuguese words to the OWN-PT synsets equivalent to the PWN synsets though is a manual labor (many suffixes to consider).



Many suffixes in Portuguese

ês	portugu**ês** (Portuguese)
ano	haiti**ano** (Haitian)
ino	argent**ino** (Argentinian)
eiro	brasil**eiro** (Brazilian)
ão	afeg**ão** (Afghan)
ense	angol**ense**, (Angolan)
ista	sul-african**ista** (South-African)
enho	carib**enho** (Caribbean)
-	bósnio (Bosnian) or Búlgaro (Bulgarian)

Some not morphologically related to the location nouns that they refer to, such as *barriga-verdes* ('green-bellies'), state of Santa Catarina and *capixabas*, state of Espírito Santo.

Completing and Expanding OpenWordnet-PT

cont.

- Given our choice of encoding OpenWordnet-PT in RDF, simple SPARQL queries were used to find the pertainym synsets with no Portuguese words.
- Retrieves all pairs of synsets (s_1, s_2) that have senses related by ADJECTIVEPERTAINSTO, with s_1 corresponds to the gentilic and the second synset s_2 is the place it is associated with (PWN lexicographer file NOUN.LOCATION).
- A preliminary list of verified entries was obtained from Portuguese DBpedia.

Completing and Expanding OpenWordnet-PT

cont.

- As expected PWN does not have most of the gentilics related to Brazilian culture and language. Only one demonym carioca.
- the long list of gentilics from the Dictionary of Gentilics and Toponyms provided by the Portal of the Portuguese Language, many are not very important and regular.
- What should be the criteria to decide on the 'notoriety' of words that justify creating a synset for them? We used Wikipedia.

Gentilics extracted from Wikipedia

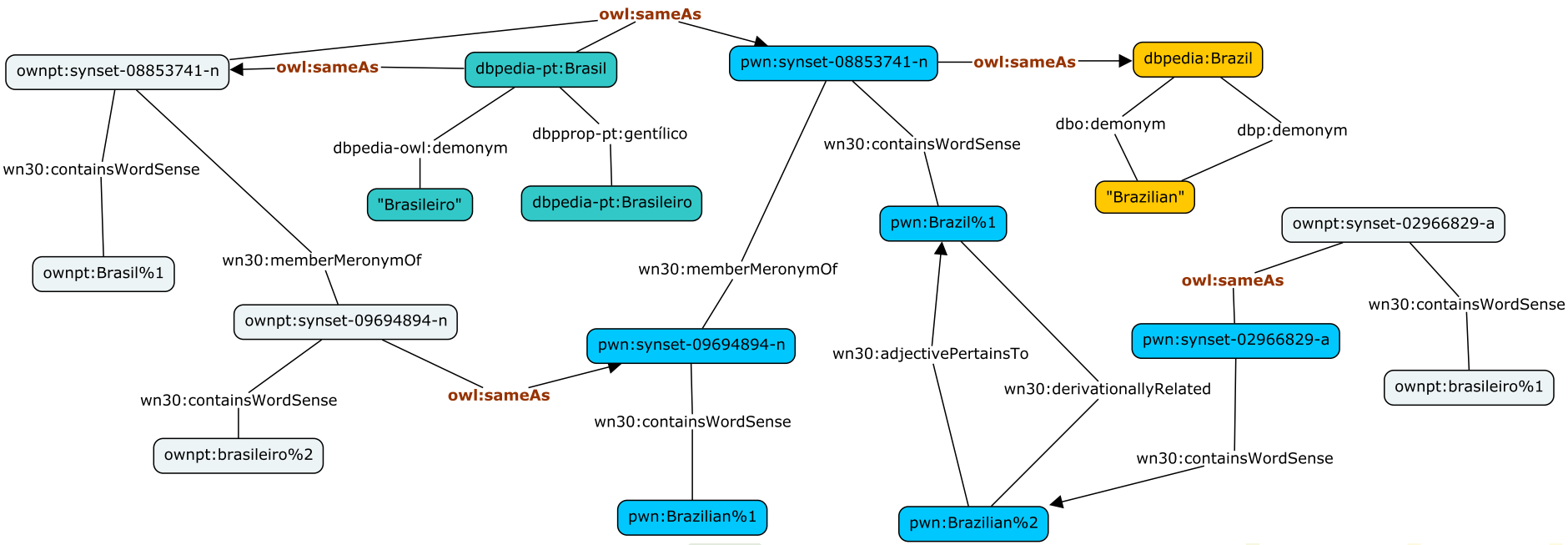
Number of Gentilics	Locations
27	States of Brazil
455	World countries
532	Brazilian cities
288	cities in the state of Minas Gerais
93	cities in the state of Rio de Janeiro
274	cities in the state of São Paulo

Completing and Expanding OpenWordnet-PT

Cont.

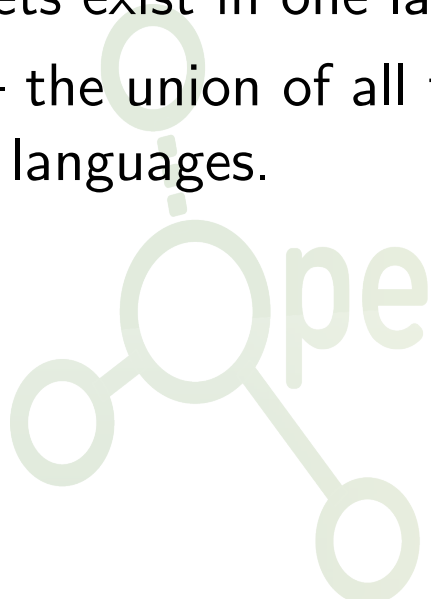
- Adding Brazilian gentilics to OpenWordnet-PT was a good way to start adding synsets for Portuguese specific concepts.
- Regular relations to their related nouns and are easily inserted in PWN's hierarchy.
- Lexical entries of gentilics (and demonyms) is easily retrievable from DBpedia, as it links location articles to its demonym via a OWL:DEMONYM relation.
- We started to investigating how to link (better than merge) DBpedia-EN, PWN, DBpedia-PT and OWN-PT.
- Wikipedia infoboxes still lack an uniform treatment for gentilics and demonyms — some of them actually bring plurals, *Brasileiros*, and feminine and masculine forms in different patterns, as *Australiano, Australiana vs Espanhol(a)*.

Connecting DBpedia with PWN and OWN-PT



SUMO and World Knowledge

- Given our use of linked data and given the easy access to the mappings of PWN into SUMO, how the mapping of new possible synsets to SUMO would proceed?
- While it is desirable to link all languages via OMW, there some difficulties, when synsets exist in one language but not in another.
- An Interlingua index – the union of all the concepts that are lexicalized in different languages.



Mappings from PWN synsets to SUMO concepts

SUMO Concept	PWN Gentilic	PWN noun.location
Nation	172	20
'Specific Places'	7	199
GeographicArea	21	35
LandArea	27	64
GeopoliticalArea	33	10
City	30	37
Island	14	45
EthnicGroup or Human	13	0
Others	92	0

Mappings from PWN synsets to SUMO concepts

- The synset for *Paris* is mapped to `ParisFrance` concept, but the synset for *Venice* is mapped into `PortCity`.
- Even when a precise SUMO concept exists, its corresponding WordNet mapping may not have been updated (mapped to a general definition).
- Almost half of the mappings of the gentilics go to an instance of the concept `Nation`.
- One might expect that gentilic adjectives (e.g. 'Brazilian' in *Brazilian cuisine*) would be mapped to a relation, relating the type of the object it applies to (`Cuisine` is a class in SUMO) to the generic property of being associated with that place.
- Instead, the gentilic adjectives are mapped to the geographical concepts they are associated with, such as `Nation`, `Island` and `LandArea`.

Mappings from PWN synsets to SUMO concepts

Cont.

- These mappings are somewhat inconsistently done as well.
- The actual mapping implicitly tells us that gentilic is a relation between an entity and a location.
- There are many cases where this seems wrong. Examples include nomadic people like ‘gypsies’ or ‘Bedouins’, not to mention all the Brazilian native tribes.
- We would prefer not to be too specific, as demonyms and gentilics do not carry only the meaning of the place where someone lives or was born, as a preliminary view suggests.

Conclusions

- Gentilics are an interesting and useful phenomenon to investigate, the frontiers of lexical resources and world ontologies.
- Lexical, but related to locations, which are named entities and hence more akin to world knowledge than lexical knowledge.
- Easier adjectives to deal with? As one does not have to worry too much about scales of being *paulista* 'of São Paulo', for example.
- Then they are slightly more amenable to Knowledge Representation methods and tools, as one can, as in the SUMO mapping available, use the location itself as a proxy for the adjective.
- To the corpus of biographies they seem very useful, as historical data needs to be geographically located.
- As a way of starting creating new synsets, they seem a safe bet (all in the class of pertainyms and all related to locational nouns).

Conclusions

Cont.

- We leave as future work the task of adding the most relevant Portuguese gentilics for other lusophone cultures.
- Fixing bugs in the SUMO-WN mappings and adding definitions to SUMO itself.
- We may need full expressivity of higher order logic to use modal and temporally qualified expressions.
- Functions are also heavily employed so we would like to create PERSON-OF-REGION-FUNCTION with a geographical argument, without having to reify not only every country or region but also the notion of being from a region or typical of a region.
- Evaluate improvement of the IE on our corpus DHBB with the relational information in the OWN-PT lexical base.