Gentle with the Gentilics

Livy Real, Valeria de Paiva, Fabricio Chalub, Alexandre Rademaker
IBM Research Brazil, Nuance Communications USA, IBM Research Brazil, IBM Research and FGV/EMAp, Brazil
livym@br.ibm.com, valeria.depaiva@gmail.com, fchalub@br.ibm.com, alexrad@br.ibm.com

Abstract

To get from 'Brasília is the Brazilian capital' to 'Brasília is the capital of Brazil' is obvious for a human, but it requires effort from a Natural Language Processing system, which should be helped by a lexical resource to retrieve this information. Here, we investigate how to deal with this kind of lexical information related to location entities, focusing in how to encode data about demonyms and gentilics in the Portuguese OpenWordnet-PT.

Keywords: lexical resources, ontology, gentilics, Wordnet

1. Introduction

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 an Natural Language Processing (NLP) system requires some effort. Having this kind of information encoded in a lexical resource can help in several tasks, such as information retrieval, lexical disambiguation and textual entailment. However, deciding which kind of ontological information should be present in lexical resources, be they wordnet-like, or specific knowledge bases, such as DBpedia[1], Wikidata[2], OpenStreetMap[3], or Geonames[4] 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. Thus we investigate how to deal with lexical information closely related to location entities, mainly focusing in how to encode these data in the Portuguese OpenWordnet-PT (de Paiva et al., 2012).

OpenWordNet-PT (OWN-PT) is a freely available wordnet for Portuguese.[5] OWN-PT was originally developed as a syntactic projection of the Universal WordNet (UNW) (de Melo and Weikum, 2009). Just like EuroWordNet (Vossen, 1998), OWN-PT was as much as possible built merging 'existing resources and databases with semantic information developed in various projects'. One reason to pay special attention to this wordnet for Portuguese is its connection to several other lexical resources based on Princeton WordNet (PWN) and on Linked Open Data (LOD) principles (Chiarcos, 2012). OWN-PT is available as an RDF/OWL download, following and expanding, when necessary, the original mappings proposed by (van Assem and Schreiber., 2006). Also OWN-PT is linked to the Suggested Upper Merged Ontology (SUMO)[6] (Niles and Pease, 2001) and to the Open Multilingual WordNet (OMW) project[7] (Bond and Foster, 2013). Since OMW merges dozens of wordnets, ways of improving each one of these wordnets might percolate to the other ones. Moreover, the issues we discuss in this work affect all of these other (merged or not) lexical resources, as we hope it is made clear in the sequel.

To start thinking about this kind of lexical geo-related information, we decided to investigate relational adjectives. Since these are traditional adjectives, they should appear in a lexical resource, but they are closely related to what we understand as ontological knowledge. Information about this kind of adjectives comes in PWN in a separated lexicographer file called ADJ.PERT (pertainym adjectives). Pertainyms are a class of adjectives that are associated with a base noun by the relation ADJECTIVEPERTAINSTO, such as the pairs *Brazilian/Brazil* and *fictional/fiction*. Thus a pertainym is an adjective, which can be defined as 'of pertaining to' another word.

The PWN lexicon has 3661 adjective pertainyms, of which 2617 had no translation to Portuguese in our OpenWordNet-PT lexical database in May 2015. We started working on pertainyms, but discovered that *gentilics*, a subclass containing adjectives pertaining only to *locational nouns*, offered enough challenges. Thus this note describes the work we did to produce the necessary translations to complete the Portuguese OpenWordNet-PT, as well as the work on improving the theoretical understanding of pertainyms and gentilics in this resource.

2. Pertainyms, Demonyms and Gentilics

Wikipedia tells us that 'demonym' is a word created to identify residents or natives of a particular place. A 'demonym' is also usually derived from the name of that particular place. Examples of demonyms include *Chinese* for the natives of China, *Swahili* for the natives of the Swahili coast, and *American* for the natives of the United States of America, or sometimes for the natives of the Americas as a whole. Just as *Americans* may refer to two different groups of natives, some particular groups of people may be referred to by multiple demonyms. For example, the natives of the United Kingdom are the *British* or the *Britons*.

The word gentilic comes from the Latin *gentilis* ('of a clan') and the English suffix *-ic*. 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: when we say *Brazilian/brasileiro*, without any context, we mean either

[1] http://wiki.dbpedia.org/
[2] https://www.wikidata.org/
[3] https://www.openstreetmap.org
[4] http://www.geonames.org/
[5] http://wnpt.brlcloud.com/wn/
[6] http://www.ontologyportal.org
[7] http://compling.hss.ntu.edu.sg/omw/

the noun or the adjective: {09694894-N BRAZILIAN — BRASILEIRO — A NATIVE OR INHABITANT OF BRAZIL} or {02966829-A BRAZILIAN — BRASILEIRO — OF OR RELATING TO OR CHARACTERISTIC OF BRAZIL OR THE PEOPLE OF BRAZIL}. To clearly distinguish pertainyms, which are adjectives, from the nouns (associated with a location), here we call adjectives *gentilics* and the associated location specific relational nouns *demonyms*. We are interested in discussing the adjectives, more than the nouns, but both bring to the fore one of the important issues that we grapple with: 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 (Giunchiglia et al., 2010) is a resource that fully merges the GeoNames database, Princeton WordNet 1.6 and the Italian portion of MultiWordnet (Pianta et al., 2002), but perhaps a wordnet does not need to have much geographical information. Since there are many geographic databases, they could be used instead of growing the number of synsets referring to locations within the lexicon itself. This is the approach taken for instance in (Frontini et al., 2013), which transforms the GeoNames ontology into GeoDomainWN, a linguistic linked open data resource, linking both PWN and the Italian WordNet to GeoNames.

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 English (and, in general, are not present in general knowledge bases). The most obvious of these meanings are related to pertainyms, mostly to places (gentilics) but also to religions, styles of philosophy, music, etc. Examples in English include *Buddhist, Socratic, Wagnerian, Darwinian*, etc. Examples in Portuguese include *macumbeiro* (someone who practices *macumba*, a Brazilian religious practice, a mixture of African religions and Catholicism); *machadiano* (from Machado de Assis, one of the greatest Brazilian novelists); *tropicalista* (from *Tropicália*, a musical movement).

A few of the essentially Brazilian words have made their way into English. The word *samba* for instance appears in PWN within three synsets: {01896881-V SAMBAR — DANCE THE SAMBA}, {00537192-N SAMBA — A LIVELY BALLROOM DANCE FROM BRAZIL} and {07056895-N SAMBA — MUSIC COMPOSED FOR DANCING THE SAMBA}. Most Brazilians would agree that these three kinds of senses (the kind of music, the kind of dance, and the action of dancing) exist in Portuguese, however some might disagree with the glosses: samba is not necessarily for dancing. Also we need derived words like *sambista* (someone who dances or composes samba), and compounds like *samba-choro, samba canção, escola de samba*, etc.[8]

The issue of making the Portuguese wordnet culturally relevant to Brazilians is of paramount importance to us. Given that the development of OpenWordnet-PT was motivated by its use in information extraction from the Brazilian Dictionary of Historical Biographies (DHBB) (de Paiva et al., 2014), it needs several gentilics that are not present in PWN. For example, to process a very typical sentence from DHBB, as "[...] 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.*", gentilics information is required. For this kind of corpus the work of adding Portuguese gentilics seems a manageable task and an easy introduction to creating our own essentially Portuguese synsets, that we knew from the beginning we would need in the fullness of time.

3. Completing OWN-PT

Before starting creating new synsets for the gentilics of the states in Brazil (e.g. *paulistano, amazonense*) we needed to complete the gentilics present in PWN synsets, but with no Portuguese words in the corresponding OWN-PT synset.

Given our choice of encoding OpenWordnet-PT in RDF (Rademaker et al., 2014), SPARQL (Prud'hommeaux and Seaborne, 2008) queries can be created to find the pertainym synsets with no Portuguese words and relate them to gentilics and demonyms. That is, we can formulate a query that retrieves all pairs of synsets (s_1, s_2) that have senses related by the relation ADJECTIVEPERTAINSTO, where the first synset s_1 corresponds to the gentilic and the second synset s_2 is the place it is associated with (originally defined in the PWN lexicographer file NOUN.LOCATION).

Searching for Portuguese empty gentilic synsets and completing them was the first step of our methodology. Adding the missing Portuguese words to the OWN-PT synsets equivalent to the PWN synsets though is a manual labor. Some 400 gentilics had to be added, as the semi-automatic construction process had not found them. There is no general affix rule that captures in Portuguese all possible (and the right ones) gentilics, since this morphological process can occur via, at least, six main different suffixes — namely *-ês, português, 'Portuguese'*, *-ano, haitiano, 'Haitian'*, *-ino, argentino, 'Argentinian'*, *-eiro, brasileiro, 'Brazilian'*, *-ão, afegão, 'Afghan'* and *-ense, angolense, 'Angolan'* — with no standard syntactic-semantic pattern to be followed. There are also suffixes that can produce gentilics in a non regular way, e.g. *-ista, sul-africanista, 'South-African'* and *-enho, caribenho, 'Caribean'*. Moreover, in Portuguese, the zero-suffix (also called regressive morphological process) is highly productive and gives us gentilics, such as *bósnio, 'Bosnian'* and *búlgaro, 'Bulgarian'*. There are still some lexicalized forms, which are not morphologically related to the location nouns that they refer to, such as *barriga-verdes* ('green-bellies'), for the natives of the state of 'Santa Catarina' and *capixabas*, for the natives of the state of 'Espírito Santo'. All these issues turn the automatic processing of detecting or creating gentilics a challenge. The work here takes the alternative route of checking and completing the required synsets with the right gentilics, as suggested by PWN and Wikipedia. A preliminary list of verified entries was obtained from Portuguese DBpedia Sparql Endpoint [9].

[8] Both *samba-choro* and *samba canção* are not for dancing, mostly. *Escola de samba*, 'Samba School', is a group of people that practices samba and performs it once a year in *sambódromos*, huge spaces prepared to receive samba schools during Carnival.

[9] `http://pt.dbpedia.org/sparql`

4. New synsets

As expected PWN does not have most of the gentilics related to Brazilian culture and language. Actually PWN does have only one gentilic specific to Brazil, the word *carioca*, which is in the appropriate demonym synset {09695019-N CARIOCA — CARIOCA — A NATIVE OR INHABITANT OF RIO DE JANEIRO} but it does not have all the other 26 demonyms for the other Brazilian states, for example. The English PWN does not list Brazilian gentilics, since they are not part of the English language, but clearly they ought to be in the OWN-PT, a Portuguese wordnet, as they are an important part of our lexicon.

Despite a long list of gentilics to be found in the "Dicionário de Gentílicos e Topónimos" ('Dictionary of Gentilics and Toponyms'), kindly provided by the "Portal da Língua Portuguesa" ('Portal of the Portuguese Language'),[10] we do not want to have all of these gentilics in our knowledge base, as mostly, they are regular and would not be very useful for our main task of reasoning with language. We needed to come up with useful criteria to decide on the 'notoriety' of words that justify creating a synset for them, to borrow a concept from Wikipedia.

So we started investigating the Wikipedia list of gentilics for nations and the list of Brazilian gentilics which includes all adjectives related to states, capitals and most important cities in Brazil. Wikipedia actually offers a reasonable amount of Brazilian relevant terms that could be linked to OWN-PT. Table 1 presents some numbers.

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

Table 1: Table of Gentilics extracted from Wikipedia/DBpedia

So far our work in OpenWordnet-PT has been focusing on adding Portuguese words to OpenWordnet-PT synsets related to PWN synsets, postponing the creation of new synsets. Adding Brazilian gentilics to OpenWordnet-PT seems a good way to start adding synsets for Portuguese specific concepts since they have established and regular relations to their related nouns and are easily inserted in PWN's hierarchy. This information (lexical entries of gentilics, and also, of demonyms) is easily retrievable from DBpedia, as it links location articles, as for example *Brazil*, to its demonym (*Brazilian*), via a OWL:DEMONYM relation. DBpedia-PT offers all gentilics we think we need at the moment, thus we are now investigating how to link DBpedia-EN, PWN, DBpedia-PT and OWN-PT. We believe it is better to link all those resources than try to merge and disambiguate their actual state within OWN-PT. In the one hand, DBpedia has most of the Wikipedia content we are interested in and is often updated. On another hand, 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)* — which we do have in OWN-PT. We expect to improve the present state of both resources by linking them. A preliminary proposal of how to link those resources is found in Figure 1.

[10] http://www.portaldalinguaportuguesa.org/.

5. SUMO and World Knowledge

Most of our work in lexical resources is directed towards using these resources in Knowledge Representation. A question then poses itself, how many locations, countries, cities, etc should we have in the lexicon; how many of these should be in other ontologies or gazetteers or taxonomies? The promise of the Open Linked Data project is that we can outsource some of the work related to the ontologizing of these locations to others, for example GeoNames or DBpedia. But demonyms and gentilics still have to be in the lexicon. They can not always be derived from their related nouns and they are not named entities that one could keep track of in other instance-based ontological resources. At least the Academia Brasileira de Letras (Brazilian Academy of Letters), the official keeper of the Brazilian Portuguese language, lists gentilics and demonyms, but does not list names of places in its official vocabulary list.

Given our use of linked data and given the easy access to the mappings of PWN into SUMO (Niles and Pease, 2003), we have decided to investigate 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. One possible approach is to create new synsets in PWN, but creating synsets that are not used in the language is problematic, since a wordnet is supposed to be a representation of language as actually used. A more principled approach might be to create an Interlingua index (Pease and Fellbaum, 2010; Bond and Fellbaum, 2016) that can be the union of all the concepts that are lexicalized in different languages. While demonym noun synsets in PWN are mostly mapped to SUMO as an instance of the `EthnicGroup` concept, gentilic adjectives are not consistently mapped. Table 2 shows some numbers of the mappings from PWN of gentilics and associated noun synsets into 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+Human	13	0
Others	92	0

Table 2: Mappings from PWN synsets to SUMO concepts

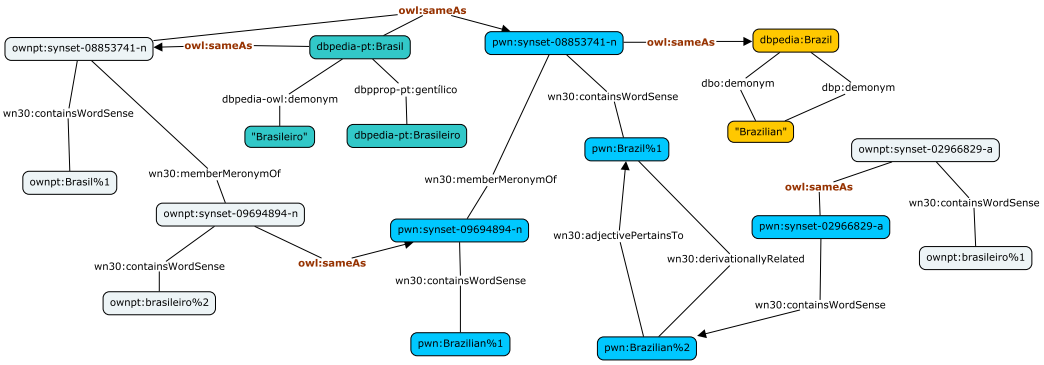

Figure 1: Connecting DBpedia with PWN and OWN-PT

In the column 'SUMO Concept', the label 'Specific Places' stands for specific places that are also specific concepts, such as `Paris`, `Brazil` and `SouthAfrica`. We can see that almost half of the nouns that we deal with are mapped, as expected, to their specific place concept: the synset {08853741-N BRAZIL — BRASIL — THE LARGEST LATIN AMERICAN COUNTRY AND THE LARGEST PORTUGUESE SPEAKING COUNTRY IN THE WORLD} is correctly mapped to concept of `Brazil`. However, while the synset for *Paris* is mapped to the concept `ParisFrance`, the synset for *Venice* is mapped into `PortCity`, a city which has a port. The PWN to SUMO mappings (as well as SUMO itself) have been constructed over a period of 15 years. Even when a precise SUMO concept is available, its corresponding WordNet mapping may not have been updated. Although SUMO has a proper treatment of many concepts, many are also missing and some are mapped to an overly general definition. Almost half of the mappings of the gentilics go to an instance of the concept `Nation`, as they are related to nouns that are instances of nations. 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, in this case, Brazil. Instead, the gentilic adjectives are mapped at the moment to the geographical and abstract concepts they are associated with, such as `Nation`, `Island` and `LandArea`. These mappings are somewhat inconsistently done as well. Were they to be more consistent, one could perhaps argue that the ontology itself did not need to have relational concepts, that the location is meaningful enough. However the consistency of the mappings itself is complicated, for example, gentilics related to island places are not necessarily mapped into `Island`: the adjective *Seychellois* is mapped into `LandArea` (as the Seychelles are an archipelago), while *Tobagonian* is mapped into `Island` but *Mauritanian* into `Nation`, even if these three places are island-like.

The actual mapping implicitly tells us that gentilic is a relation between an entity and a location. While this seems generally correct, 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.

6. Conclusions

Gentilics are an interesting and useful phenomenon to investigate, when considering the frontiers of lexical resources and world ontologies. First they are clearly lexical, but related to locations, which are named entities and hence more akin to world knowledge than lexical knowledge. Then they are somewhat 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, relaying in some other language processing.

For our own driving application to the corpus of biographies in (de Paiva et al., 2014) they seem very useful, as historical data needs to be geographically located. Finally, as a way of starting creating new synsets, they seem a safe bet, as they are sandboxed, as they ought to be all in the class of pertainyms and all related to locational nouns.

We leave as future work the task of adding the most relevant Portuguese gentilics for other lusophone cultures different from the Brazilian one, that is the gentilics most relevant for places in Portugal or Mozambique, say. We would also like to discuss with the SUMO team the best way of improving the mapping of gentilics to SUMO. This includes fixing bugs in the SUMO-WN mappings but more importantly, adding definitions to SUMO itself. While we will save a detailed treatment for a latter paper, this might require using the 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 laboriously reify not only every country or region but also the notion of being from a region or typical of a region. As a result, these definitions may have to use SUMO's expressive logic rather than a simpler language like a description logic. Finally, we would like to evaluate how much the quality of the treatment of meanings on our historical corpus Brazilian Dictionary of Historical Biographies (DHBB) increases if we have relational information in the OWN-PT lexical base.

7. Bibliographical References

Bond, Francis, P. V. J. M. and Fellbaum, C. (2016). Cili: the collaborative interlingual index. In Christiane Fellbaum Piek Vossen Verginica Barbu Mititelu, Corina Forăscu, editor, *Proceedings of the Eighth Global WordNet Conference*, Romania.

Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, Sofia. ACL.

Chiarcos, C. (2012). *Linked Data in Linguistics*. Springer.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

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

de Paiva, V., Oliveira, D., Higuchi, S., Rademaker, A., and de Melo, G. (2014). Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE, oct.

Frontini, F., del Gratta, R., and Monachini, M. (2013). Linking the geonames ontology to wordnet. In *6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Giunchiglia, F., Maltese, V., Farazi, F., and Dutta, B. (2010). Geowordnet: A resource for geo-spatial applications. In Lora Aroyo et all., editor, *The Semantic Web: Research and Applications*, volume 7th Extended Semantic Web Conference, ESWC 2010. Springer.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Chris Welty et al., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*. FOIS-2001.

Niles, I. and Pease, A. (2003). Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages pp 412–416.

Pease, A. and Fellbaum, C. (2010). Formal ontology as interlingua: The sumo and wordnet linking project and globalwordnet. In C. R. et al Huang, editor, *Ontologies and Lexical Resources*. Cambridge University Press, Cambridge.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: Developing and aligned multilingual database. *Proceedings of the First International Conference on Global WordNet*, pages pp. 293–302.

Prud'hommeaux, E. and Seaborne, A. (2008). Sparql query language for rdf. w3c recommendation, january 2008. Technical report, W3C.

Rademaker, A., de Paiva, V., de Melo, G., Coelho, L. M. R., and Gatti, M. (2014). OpenWordNet-PT: A project report. In Heili Orav, et al., editors, *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan.

van Assem, Mark, A. G. and Schreiber., G. (2006). Rdf/owl representation of wordnet. *W3c working draft, World Wide Web Consortium*.

P. Vossen, editor. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.