Of *Seringueiros* and *Sambistas*: Occupation Mappings in Historical Text

Valeria de Paiva[1⋆], Aikaterini-Lida Kalouli[2], and Livy Real[3]

[1] Topos Institute, Berkeley, USA valeria@topos.institute
[2] CIS, LMU Munich kalouli@cis.lmu.de
[3] americanas s.a. livyreal@gmail.com
http://vcvpaiva.github.io/

Abstract. This work shows how shallow processing of available resources can help us improve the coverage of existing large-scale lexical resources like the OpenWordNet-PT, the Portuguese version of WordNet. Specifically, the work employs the Brazilian Dictionary of Historical Biographies, a dictionary whose entries are short biographies of personalities of the History of Brazil since the 1930s, and the European multilingual classification of Skills, Competences and Occupations resource (ESCO), in order to extract professions and occupations and check how many of them are already present in OpenWordNet-PT. The work also allows interesting side-observations, for example on the quality of non-English NLP tools as well as within the socio-political scenery.

Keywords: OpenWordNet-PT · occupations and professions · historical biographies · spacy processing · Portuguese NER

1 Introduction

Despite the huge successes of machine learning techniques over big data, lexical resources in the style of Princeton WordNet (PWN) are still necessary for many tasks in the applications resulting from processing natural language. Such resources are not well-developed for languages other than English and long drawn processes are many times necessary to circumvent the lack of such a resource. For Portuguese, one project for the development of a Portuguese open wordnet since 2012 is OpenWordNet-PT [7]. Creating new lexical resources is easier, but improving and maintaining the ones you have, not so much. In particular, the work of verifying accuracy and improving translations of versions of PWN requires finding specific sub-problems within the data that can be seen as closed sub-problems, where you can circumscribe a task, finish it, declare victory and then write about the sub-project.

The problem of looking at professions and occupations in a historical corpus, in a open-source lexical resource and in the OpenWordNet-PT seems a good candidate for such a closed sub-problem. For one, when considered from the viewpoint of knowledge representation (KR), it characterizes a semantic domain, a subclass of human activities

that one might hope to be able to complete to a desired level of accuracy. Additionally, given the characteristics of the corpus in question, a dictionary of historical figures, it is one of the essential pieces of information that every entry possesses. Thirdly, while some of the (political) occupations are similar in English and Portuguese (e.g president, lawyer, senator, janitor, etc..) it is clear that some, like the ones in our title (*seringueiro* is someone who extracts latex from trees, *sambista* someone who composes, dances or plays a style of Brazilian music, *samba*) only need to exist in a Portuguese wordnet.

Completing OpenWordNet-PT is a founding stone for work we want to do on creating and reasoning with logical representations for the meanings of sentences in (English and in) Portuguese. If we can automatically generate logical knowledge representations for the historical characters of a corpus, we can do many more extraction tasks: essentially, we can *reason* with this kind of information to deduce new information. For instance, if you want to know which politician was the first female governor of a state in Brazil and you ask Google for "primeira governadora do Brasil" (first female governor in Brazil), you may get what seems like contradictory information: One link provides the sentence "Em 1986, quando o governador eleito do Acre deixou o cargo para disputar uma vaga no Senado, Iolanda Fleming se tornou a primeira mulher a governar um estado da federação"[4] while another link provides "Roseana foi a primeira mulher eleita governadora"[5]. But this information is not contradictory: Fleming was the first female governor, even if she was only elected as a vice-governor. And eight years later, Roseanna Sarney became the first woman elected as a governor. Solving this kind of reasoning problem seems too much to ask of the systems we are currently developing, but discovering the issue (the apparent contradiction) should be what a logical, reasoning system is supposed to be doing. Work in this direction has been completed by [4], building on much earlier work envisaged in [5]. Necessary for such endeavours are lexical resources as complete as possible, which is why we set out to improve the coverage of OpenWordNet-PT.

2 Lexical Resources and Corpus

Wordnets are lexical databases that offer information on open class words, that is, adjectives, nouns, verbs and adverbs. Wordnets descend from Princeton's Wordnet, developed by Miller and Fellbaum [2] at Princeton University. Wordnet was designed as a dictionary and thesaurus for human use. However, researchers in AI and Natural Language Processing (NLP) have found WordNet and its taxonomic concept hierarchies useful for computational systems. Hence, WordNet has seen widespread use and it has became a "de facto" standard in NLP. The familiarity of its taxonomic structure, coupled with its extensive coverage, as well as its open source license, all helped to account for WordNet's popularity and for the proliferation of similar resources in other languages, including Portuguese.

[4] 'In 1986, when the elected governor of Acre left the position to dispute a seat in the Senate, Iolanda Fleming became the first woman to govern a state of the federation.'

[5] 'Roseana was the first woman elected governor.'

OpenWordnet-PT is a wordnet for Portuguese, open-source software. [6] Despite being in development since 2012, it is not a finished product. The original idea was that having a fairly high-quality translation from Princeton WordNet would be a good starting point for work in Portuguese semantics, which we would improve as much as we could, using mini-projects that could interest students. Devising these mini-projects is not a triviality, as they need to have well-defined evaluation criteria and lead to some clear improvement of the thesaurus, as it stands. Work on grammatical classes has been done before and it seemed a sensible idea to try to venture into semantic domains, such as occupations from now on.

To detect missing professions and occupations from the current version of Portuguese wordnet OpenWordNet-PT and be able to improve it, we decide to look at two resources: the Brazilian Dictionary of Historical Biographies (DHBB in Portuguese)[7] and the occupations list provided by the European Skills and Competences, qualifications and Occupations (ESCO) initiative[8].

The DHBB corpus was originally designed to provide researchers and scholars with organized and systematic information about personalities and themes considered noteworthy in the recent history of Brazil, from the Revolution of 1930 onwards. The corpus comprises about 7,500 biographic or thematic entries, covering people, institutions, organizations, and events. The majority of entries are biographical in nature, with over 6,500 biographies and some 1,000 thematic entries. Each entry consists a separate file. Biographical entries have a header summarizing the description of the person, with key positions held and the respective periods the position was exercised. This corpus is suitable for several reasons. First, the entries are well-written and hand-curated. They follow some carefully designed guidelines, but are not written in a *controlled language*. They use a medium register, not too erudite (as it is supposed to be useful for high school students), but not too popular or informal. Hence, we do not have too many of the problems associated with slang, regionalisms, neologisms, emoticons or out of vocabulary words of web texts. Also, the language, associated to historical biographies of personalities since the 1930s, is similar to news text, hence it does not require a large vocabulary of specific terms, as it is required in works in the fields of Law or Geology, for example. Given the historic, biographical nature of the corpus, it is suitable for sociological research. Indeed, work done by [6] shows examples of socio-political questions historians might like to be able to answer given this corpus. For example, the question is posed on whether women who lived in Rio de Janeiro at the time and occupied high positions in the Executive branch between the decades of 1960 and 1980 attended the same academic circles or the same intellectual environment as their male colleagues. Currently and without suitable logical representations, we are still unable to answer such questions. Improving the coverage of resources like the OpenWordNet-PT with important semantic categories like that of occupations can contribute to solving these questions.

[6] It can be browsed at `http://openwordnet-PT.org` and downloaded from `https://github.com/own-pt/openWordnet-PT`.

[7] `https://github.com/cpdoc/dhbb`

[8] `https://ec.europa.eu/esco/portal/download`

On the other hand, we also make use of the occupations resource provided by ESCO. ESCO is the European multilingual classification of Skills, Competences and Occupations. ESCO works as a dictionary, describing and classifying professional occupations and skills relevant for the EU labour market, education and training. The resource is constructed with the goal to be used by electronic systems that provide services like matching job seekers to jobs on the basis of their skills, suggesting training options to people who want to re-skill or up-skill etc. ESCO provides 2,942 basic occupations for 27 languages. For Portuguese, we are able to extract a total of 5,103 occupations, also including alternative names provided for the basic occupations. At this point, the question may arise of why we also use the DHBB corpus, since the ESCO resource seems to contain a solid base against which we would check OpenWordNet-PT's coverage. There are several reasons for our choice. First, the ESCO resource is focused on the EU geographic area and although most occupations will exist across the world, there are professions which are specific to the part of the world they occur and the language they are used in. Therefore, the DHBB could provide us with such occupations within the Brazilian space, especially older professions which might not exist in their original form any more. Additionally, having the historical text where the occupation is used helps with possible ambiguities. Another reason for our choice is that we would like to use this work to make preliminary sociological observations on the DHBB corpus, attempting to contribute to the open questions raised in [6].

3 Occupation Extraction

A first processing of the DHBB, described in [6] was based on FreeLing 3.0 [1]. Freeling[9] is an open source multilingual language processing pipeline, which offers modules for Portuguese processing. FreeLing's modules for named entity recognition and part-of-speech tagging were used in that first processing. However, the frameworks for named entity recognition and parsing should have improved considerably since then. In particular, spaCy[10] has proven to be a powerful NLP tool, currently supporting more than 64 languages. Note that the choice of an NLP tool is not trivial since there are only a few tools that can process languages other than English. Therefore, we choose spaCy and process the DHBB corpus with it.[11] For the processing, we use the largest model provided for Portuguese[12] and we perform tokenization, part-of-speech tagging (POS-tagging) and named-entity recognition (NER). Particularly, we extract verbs and common nouns, we list the historic figures described in the dictionary and produce separate lists of persons' names as extracted by the NER module, locations, organizations and miscellaneous named entities. Some of our comments below refer to the people presented as main entries of the dictionary, but others refer to 'people' as extracted by the NER module as mentioned within the text of entries (including relatives, colleagues, superiors, etc.)

[9] `http://nlp.lsi.upc.edu/freeling/`
[10] `https://spacy.io/`
[11] All processing available under `https://github.com/vcvpaiva/DHBBspacy`
[12] *pt_core_news_lg*

Based on the extracted nouns, we are able to query these nouns for occupations and professions. We extract all nouns that are either included in the ESCO occupations resource or end with suffixes common for professions in Portuguese, such as -or (*compositor*, 'composer'), -ista (*motorista*, 'driver'), -nte (*presidente*, 'president'). This results into a list of 1749 professions. The list needs to be further manually curated, e.g., because these suffixes do not always express professions and due to other issues. First, many nouns in our list are not really occupations or professions, but behave syntactically as if they were. For example, in a sentence such as[13] "Suplente do primeiro-secretário da mesa da Câmara entre 1967 e 1968, em 1970 passou a exercer o cargo de primeiro-secretário." the word for a political substitute *suplente* behaves like a proper occupation. In general, words such as *suplente* ('substitute'), *grevista* ('striker') and *golpista* ('putschist') that can be adjectives, but are commonly used as nouns, did not enter in the final list, since they are not real professions and our goal was to complete the list of occupations in OpenWordnet-PT. Words related to political positions such as *oposicionista* ('opposition member') were not considered either, even if they can, in some cases, syntactically stand for an occupation noun and many times they do. The expression *o deputado oposicionista* ('the opposition representative') becomes in Portuguese simply *o oposicionista*. Other nouns can be, in one of their senses, a specific profession, for example *segurança* ('security' as in 'security guard') and *liderança* (leadership as in 'the party's leadership'), but they were only listed in our most recurring nouns because they are very polysemous. The noun *segurança* refers to the abstract concept of 'security' and 'safety' in a sentence such as 'O artigo 162 previa o Conselho de Segurança Nacional, encarregado de estudar as questões de segurança.' ('Article 162 envisaged a National Security Council, in charge of studying safety issues.'). Generic positions such as *empregador* ('employer') and *comprador* ('buyer') were kept in the list, because even if they do not always refer to a specific profession they refer to important roles in the domain of workers and jobs. After the manual curation, the cleaned list of occupations contains 853 entries.

We check how many of these entries are included in the OpenWordNet-PT and make interesting observations. From the 853 entries, 282 are missing from OpenWordNet-PT, so around 33%. From these 282 missing synsets, 7 might be considered prefix issues. Words like *ex-ministro, segundo-secretário* ('ex-minister, second-secretary') in principle should be covered given some language pre-processing dealing with prefixes. Some are clearly occupations that only need to exist in a Brazilian Portuguese wordnet. Occupations such as *usineiro, cafeicultor, posseiro* ('sugar refinery owner, coffee producer, squatter') are examples. Clearly these occupations exist in other cultures, but the nuances implied by the nouns are very much Brazilian. The same way, young military men wanting a voice in politics do happen in many places, but a political movement called *Tenentismo* ('Tenentism') meaning 'related to lieutenants', is a phenomenon typical of Brazilian history of the 20th century.

[13] 'A substitute to the first secretary of the Bureau of the Assembly between 1967 and 1968, in 1970 he started to hold the post of first secretary.'

4 Additional Findings

Apart from detecting the occupations missing from OpenWordNet-PT and being able to improve the coverage of the resource, our processing of DHBB allows us further observations.

4.1 Quality and Development of non-English NLP tools

Special mention should be made to the locations list extracted from our processing. spaCy identifies some 27,000 entries as locations. The three most common ones are Rio de Janeiro with 11,894 occurrences (also found as Rio), Brasil with 9,579 occurrences and São Paulo with 6808. Rio de Janeiro appearing most often can probably be attributed to the fact that the entries cover the period when Rio was the capital of Brazil. Interestingly, the modern capital of Brazil (Brasília) only appears 1882 times, even less than United States with 2095 occurrences. (This can be partially explained by the use of DF (Distrito Federal/federal District) which presumably referred to Rio before 1965 and to Brasília after the change of the Capital.) However, such findings uncover socio-political aspects of the time, e.g., the importance of the United States to Brazilian politics, and can partly also be confirmed by the older findings by [6]. It is worth noting that, although the current work is done 8 years after the paper published by [6], the quality of the retrieved results has not improved as much as expected. spaCy has been shown as a much more powerful tool than Freeling and the advances of NLP in the last 8 years are considered tremendous. Still, our processing of the DHBB might suggest that most developments have been achieved for English and that other languages still lag behind: the quality of the locations we retrieve is unexpectedly low. We manually look at the first 300 entries of our locations list and find that many of the entities classified as locations are actually organizations. Thus, for example, among the 30 first ”locations” (with more than a thousand occurrences each) we find eight mistakes: the Republic (República), the lower chamber of the Congress (Câmara dos Deputados), the Congress (Congresso and Congresso Nacional), the Electoral College (Colégio Eleitoral), the Supreme Federal Tribunal (STF) and the Ministry of Finance (Fazenda) and another tokenization error (de São Paulo) (This is 8 errors in 30 of the most popular locations).

Further work is needed to determine quantitatively the unexpectedly low performance of the NER of locations and whether other categories or other NLP pipeline processing steps also show similar performance.

4.2 Sociological Observations

Our work analyzing the data in the DHBB discovered that women are very badly represented in the dictionary. Out of 6,750 historical personalities that have a devoted entry in the dictionary only 224 are women, around 3% of the entries. Within the list of people generally appearing in the texts of the dictionary (i.e., without having a devoted entry), there was not a *single* woman among the names that appear more than 250 times in the corpus. When we checked for female names among the people with at least 5 occurrences (a total of 4944 “persons”) only 145 were women.

According to the processing in 2014 with FreeLing, the most important woman in Brazilian history, if number of occurrences in this corpus was a sensible metric, would be Ivete Vargas, congresswoman from São Paulo State and niece of Getúlio Vargas, a former President of Brazil. Ivete Vargas name had 125 occurrences in the text of the DHBB, but she is not a very influential figure in Brazilian History. She was followed by Luísa Erundina, ex-mayor and congresswoman from São Paulo, with 104 occurrences. In third place we had Alzira Alves, with 94 occurrences in the corpus, a researcher at CPDOC (Centro de Pesquisa e Documentação de História Contemporânea do Brasil - Research and Documentation Center of History of Brazil), the center that produces the DHBB data. This was the result of some metadata been (mis)classified as textual data. Alves in that preliminary processing, was ahead of Marta Suplicy (senator and São Paulo's ex-mayor, with 85 occurrences), Roseana Sarney (congresswoman, senator and Maranhão's governor, with 75 occurrences), Benedita Silva (Rio de Janeiro's congresswoman and senator, 50 occurrences), Marina Silva (Acre's congresswoman and senator, 32 occurrences) and especially of Dilma Rousseff, former president of Brazil, which had only 23 occurrences in the corpus.

The current processing using spaCy and the cleaning of the data improved this situation considerably. Dilma Rousseff now shows up with 249 occurrences as the most cited woman. Ivete Vargas name has 131 occurrences in the text of the DHBB. Luisa Erundina has 125 occurrences, Marta Suplicy is now 4th place with 105 occurrences, followed by Roseana Sarney (89 occurrences), Marina Silva (76 occurrences), Heloísa Helena (52 occurrences), Benedita da Silva (51 occurrences), Marina (46 occurrences), and Rosinha Garotinho (42 occurrences). But these numbers are very small indeed.

It would be naive to think that number of occurrences in the text is a perfect proxy for importance/relevance in politics or even in the dictionary itself. There are several issues with using this proxy: one would need to cluster the different ways of referring to a single entity (e.g. "Vargas" appears as the first name on the list with 4810 occurrences and then in the 3rd place as "Getúlio Vargas" with 2540 occurrences, but both refer to the same dictator). Some Christian names are individual enough that they can be used by themselves (e.g. "Lula" is a very popular nickname for "Luis", but there is only one Lula in Brazilian politics). The list of person names does contain a fair number of mistakes that would need to be manually corrected. Mostly they correspond to organizations misclassified as people, showing again that NER can be at very high numbers for English, but the reality in other languages is different[14].

However, the disparity between numbers of male and female historical characters does tell us something. As does the fact that the list preserves its order eight years later, at least to a certain extent. As little as 61 female names appear between 249 to 10 occurrences. The numbers of women are not aligned with their perceived importance in Brazilian politics, while the opposite seems to happen to numbers of occurrences of names of male politicians.

We can cite some examples of names that we expected to see in the dictionary: Lueci Ramos (city representative for Cuiabá, four times re-elected), Olívia Santana

[14] A great example of this is the number of verbs, at the beginning of sentences that are misclassified as people, e.g. "Pressionado" (pressured), "Adepto" (follower), or "Impossibilitado" (not able to).

(city representative for Salvador e Education Secretary), Matilde Ribeiro, (ex-minister of Polícies for Racial Equality in Brazil, has only two occurrences), Sandra Regina Machado Arantes do Nascimento Felinto (city representative for Santos, SP and daughter of Pelé does not show), Simone Tebet, senator (since 2014) and federal deputy since 2002, has only one occurrence.

These numbers seem to reflect some implicit gender bias of historians, as there were very significant women part of this history, such as Maria da Penha Fernandes (Law Maria da Penha), Leci Brandão (state deputy from São Paulo since 2011 and musician, since the seventies), Marilena Chauí (philosopher and founder of the Workers Party) and Zuzu Angel (fashion designer presumed killed by the dictatorship), for instance, that do not have entries for themselves. Such a finding opens the way for more research in this socio-political dimension, which we hope to explore further in our future research.

5 Conclusions and Further Work

In this work, we set out to improve the coverage of OpenWordnet-PT, aiming at contributing to our ultimate goal of automatically producing knowledge representations for Brazilian Portuguese text. To this end, we conducted a small-scale information extraction task considering occupations and professions described in the Brazilian Dictionary of Historical Biographies (DHBB) and the classification of the European Skills and Competences, qualifications and Occupations (ESCO) initiative. This processing not only allowed us to detect synsets missing from OpenWordNet-PT, but also led us to further observations about the quality of the current, non-English NLP tools in Portuguese and some sociological "distant reading" observations [3] about the dictionary and the times when it was created. Future steps include making sure that the synsets we were able to detect are added to OpenWordNet-PT and have an appropriate, up-to-date mapping to the SUMO ontology[15]. Also we want to dig deeper in the side-observations that emerged out of this study.

References

1. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: An open-source suite of language analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04) (2004)
2. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
3. Higuchi, S., Santos, D., Freitas, C., Rademaker, A.: Distant reading brazilian politics. In: In Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries (Copenhagen Março de 2019 (2019)
4. Kalouli, A.L.: Hy-NLI : a Hybrid system for state-of-the-art Natural Language Inference. Ph.D. thesis, Universität Konstanz, Konstanz (2021)
5. de Paiva, V.: Bridges from language to logic: Concepts, Contexts and Ontologies. Electronic Notes in Theoretical Computer Science **269**, 83–94 (2011)

[15] `https://www.ontologyportal.org/`

6. de Paiva, V., Oliveira, D., Higuchi, S., Rademaker, A., Melo, G.D.: Exploratory information extraction from a historical dictionary. In: IEEE 10th International Conference on e-Science (e-Science). vol. 2, pp. 11–18. IEEE (oct 2014). https://doi.org/http://dx.doi.org/10.1109/eScience.2014.50
7. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In: Proc. of 24th International Conference on Computational Linguistics. COLING (Demo Paper) (2012)