Plurality in Wordnets

Livy Real and Valeria de Paiva

[1] IBM Research, Brazil
[2] Nuance Communications, USA
livym@br.ibm.com valeria.depaiva@nuance.com

Abstract. We investigate the features of Princeton WordNet associated with nouns that are essentially plural. This means exploring the Princeton WordNet feature `classifiedByUsage: plural` that labels synsets and words commonly used in the plural. We decided to investigate how this feature works for Portuguese and here we discuss the best way to encode this kind of lexical information in OpenWordnet-PT, an open wordnet for Portuguese.

1 Introduction

Lexical resources are required for several Natural Language Processing tasks. They are canonically used on word disambiguation tasks [1], in information retrieval [6], as input to several ontologies [7], and more recently as features when building space vectors for machine learning approaches [5] or in anaphora resolution [8]. What we commonly expect from lexical resources is that they should collect all lexical information one needs for processing texts, from syntactical frames to semantic features. Wordnets are one of the most used lexical resources and they provide generally syntactic, semantic and even contextual/pragmatical information about words and their senses. Here we focus on how wordnets encode a lexical plurality feature. This is an idiosyncratic feature that tell us that a given word is often used in the plural, as for example the English words *glasses* and *manners*.

We start by looking at how Princeton Wordnet (PWN) encodes this information, through a pointer `ClassifiedByUsage` applied to a specific synset domain `plural`. Then we look at an ongoing wordnet project for the Portuguese language, the OpenWordnet-PT (OWN-PT). The Portuguese lexicon OWN-PT is a projection of PWN and has all PWN relations and features, hence OWN-PT inherits the plural classification from PWN. However, as it is expected, plurality does not hold in Portuguese for the same word senses that it does in English. The word for *glasses*, for example, in Portuguese *óculos*, can be used both in the singular or in the plural having the same real world reference, a pair of glasses.

The issue of whether objects are referred to in the plural or not is related to the question of how to encode information about mass nouns, that is, whether the countable/uncountable distinction, which distinguishes words such as *apple* (countable) and *blood* (uncountable) needs to be marked on the lexicon and how. As preparation for deciding on the best way to have the countable vs.

uncountable lexical information encoded in OpenWordnet-PT, we investigate how the PWN plurality feature fits in with Portuguese words and how this information can be useful to Natural Language Processing (NLP) tasks.

In the next section we introduce the plurality issue, from the theoretical semantics and the NLP perspectives, pointing out some tasks where lexical information on plurality is relevant. In section 3, we recall Princeton Wordnet and its plurality feature. In section 4, we briefly introduce OpenWordnet-PT and discuss how this feature appears in Portuguese. We also present some data and statistics on how English plurality fits in with Portuguese word senses in section 4.1. To finish we discuss the best way to encode lexical plurality in OWN-PT and draw some conclusions.

2 Plurality

This section offers a brief overview of semantic studies on plurality, mainly based on Chierchia's[3] assumptions. Then we discuss how plurality has been used in NLP tasks and present our motivations for the present work.

2.1 Plurality in Semantics

Plurality has been discussed in formal linguistics studies at least since Quine's seminal work [12]. Plurality and associated issues, such as, its formalization, collective readings and the well known distinction between mass-nouns and count-nouns, have been comprehensively studied and play an important role in most theories that formalize the semantic behavior of natural language.

Here we are interested in how to code plurality information in lexical resources like wordnets. For this, we first state some definitions based on [3]. Chierchia states that what distinguishes mass nouns (as *blood, water, furniture*) from count nouns (such as *boy, drop, sofa*) is an intrinsically semantic property from which many morpho-syntactic properties follow.

Common count nouns point to what Chierchia calls *singularities* in the lexicon. They can refer both to a class of objects or to a single unit, *coin, the coin*, in sentences such as *Give me the coin* or *A coin is a piece of hard material.* Mass nouns are 'generally interpreted as a mereological whole of some kind' and the domain of its minimal components is somehow more vague than singularities, as it is the case with the word *change.* It is important to note that this is an intrinsically grammatical property. That is, this property is not related to the ontological objects those words refer to. It is the word itself that dictates whether a noun will be countable or not. Examples that show that are the pairs *coin/change*, *shoe/footwear* and *virtue/honesty.* While common count nouns express singularities, these nouns in plural express a set formed by these singularities, *boys* is a set of some individual *boy*s, in which we can still see the minimal unit, a single *boy.*

For now, we are more interested in count nouns lexicalized as plurals, such as *manners* (for example in the phrase *he has the manners of a pig*) and in nouns whose only possible form are plural, such as *pants* and *quarters*. However, we expect that our discussion on how to encode plurality in lexical resources should bring insights about how to encode information on mass/count nouns. We are also interested in collective nouns, such as *group* and *committee*, that differ from mass nouns as they can be pluralized — *groups, committees*, but not **waters, *furnitures* — and also differ from common count nouns as they already refer to sets of things.

Lexical resources, such as wordnets, in general offer only the lemma of a given word, as its dictionary form. In the case of nouns, this means the masculine and singular form, when applicable, such as *boy, manner* and *group*. Lexical resources also offer semantic relevant information, as when they group words commonly used in the plural — in PWN this feature in encoded via the feature `ClassifiedByUsage:plural` — or when a noun is a collective noun — that in PWN is described via the lexicographer files. The collective nouns are part of the file `noun.group`. A researcher interested on collective nouns or in words normally used in the plural can find enough information within PWN's state of art for English. However, the mass/count distinction is not one of PWN's classification features, at the moment.

2.2 Plurality in NLP

We briefly review the motivations for this work. First, we are interested in completing and improving the lexicon of OpenWordnet-PT. Improving lexical resources is a hard and time consuming task with no laurels, but extremely necessary for lesser resourced languages, as Portuguese still is. To ensure that OWN-PT is as informative for Portuguese as PWN is for English, and to make sure that the PWN information inherited by OWN-PT is correct, we want to check all synsets and relations in OWN-PT. However, since that is a large amount of data, we have been revising OWN-PT's content in pieces, considering different features or relations of PWN and consistently checking how these are encoded in OWN-PT.

Second, we are interested in defining notions of plurality and count-mass distinctions in lexical resources, as this information is lexical and can be used in several applied tasks. For example, knowing that some pluralized expressions actually refer to unique entities is necessary for doing textual reasoning.

Recent work of [4] points out the necessity of considering pluralized word forms when building vector space models for machine learning applications. The author notices that the distribution of words and other textual features changes when a word is often used in its plural form. Thus building word vectors only considering its lemma oversimplifies the features. However, as [11] and [14] show, using wordnet relations as features in vector space modelling can improve machine learning algorithms results. Thus it seems that this PWN classification can be helpful when modelling plurality or countability.

Another use case of this feature can be seen in [9] that proposes a new method of a picture based communication, a language independent method of communication for people with disabilities. The proposed method uses the PWN

ClassifiedByUsage:plural relation to improve the possibility that a given picture corresponds perfectly to a concept. Having this feature encoded allows the system to automatically recognize that a picture that contains only one element can be correctly associated to a word in a synset classified as plural. One single pants is enough to describe 'pants', differently, for example, from 'birds' that should be described by a picture containing more than one single bird.

3 Plurality on Princeton Wordnet

Princeton Wordnet(PWN) is the mother of all wordnets and has been developed for English by the Princeton team over the last three decades. PWN offers a database of nouns, verbs, adjectives and adverbs arranged into sets of cognitive synonyms (synsets). Each synset refers to a single concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. These synsets are related through many (conceptual-semantical) relations, such as synonymy, heteronomy and meronymy. Each synset has one specific ID, one or more word forms that express its sense, a gloss (a small concept definition) and many of them also have sentences exemplifying its use.

The latest version of PWN available is version 3.1, that contains some 117,000 synsets. Released in 2006, this version of PWN is still the larger and most reliable lexical resource available for English. There are many relations in PWN classified by pointers, that can be semantic or lexical. According to the PWN documentation, lexical pointers, such as Antonym and Derivationally Related, are *normally* used to indicate lexical relations that hold between words. Semantic pointers establish semantic relations that are *generally* used for linking synsets. Examples of semantic pointers are Hypernym and Hyponym relations, that clearly should hold between synsets.

However, pointers such as the Region domain, that assigns a word form to where it is usually used (in the US or England, for example) and the USAGE domain, that specifies content as an archaism or a plurality, are hybrid between lexical and semantic pointers, an issue that had been pointed out by Eric Kafe in the 2000's and is discussed again in [13]. The definition of PWN pointers itself is vague and PWN seems to accommodate pointers that have a hybrid range, that is, that classify different kinds of objects. However it is still an ongoing discussion on PWN and Global Wordnet Association[3] communities what to do with thes hybrid pointers.

The feature ClassifiedByUsage:plural, or Domain-Usage:plural, filters nouns synsets *and* word senses that are often used in plural. This relation applies only to nouns and labels 239 elements. Some examples are {07942152-n people| any group of human beings (men or women or children) collectively | ''old people''}; {04862236-n nerves | control of your emotions | ''this kind of tension is not good for my nerves''} and synset

[3] http://globalwordnet.org/.

{`04356056-n sunglasses, shades, dark glasses | spectacles that are darkened or polarized to protect the eyes from the glare of the sun| ‘‘he was wearing a pair of mirrored shades’’`}.

The PWN list of synsets classified by plural usage can be found at `http://wordnet-rdf.princeton.edu/wn31/106230167-n`. This comes from dictionaries, since this feature is idiosyncratic and there is no general pattern that can capture only those words. Although the PWN classification of plurality appears useful, the feature is not well curated. The documentation says that the word is usually used in the plural, but some of word forms appear in the plural, others in the singular and sometimes PWN have both forms in the same synset. For example we have the synset {`06630627-n regard, wish, compliments | a polite expression of desire for someone’s welfare | ‘‘give him my kind regards’’; ‘‘my best wishes’’`}, where *wish* is in the singular, while *compliments* is in the plural.

4 OpenWordnet-PT

The OpenWordnet-PT (OWN-PT) is an open wordnet for Portuguese, in development since 2012 and modelled after and fully interoperable with the original PWN. The OWN-PT uses the same identifiers as the last released version of PWN and it is browsable at and downloadable from `http://wnpt.brlcloud.com/wn/`. The OWN-PT is also linked to the largest open source common sense ontology, the Suggested Upper Merged Ontology (SUMO)[4], described in [10] and to the Open Multilingual WordNet (OMW) project[5], again browsable and downloadable as described in [2]. Since the Open Multilingual Wordnet project merges dozens of wordnets, ways of improving each one of these wordnets might percolate to the other ones. Thus the plurality encoding issue discussed here can, in principle, affect and/or be useful to all of these other lexical resources.

For these reasons we would like to be sure that all the PWN relations and features that are inherited by OWN-PT are not too language specific to English and should be present in derived wordnets, such as ours. Our hypothesis was that the feature of plural usage is indeed idiosyncratic and we doubted the suitability of automatically percolating it through to the Portuguese synsets. Thus we decided to check all the synsets marked with this feature, completing all of them in OWN-PT and collecting candidates that should and should not keep this feature in Portuguese. The data derived from this methodology is investigated in the following.

4.1 Data and Statistics

From the 239 synsets that PWN marked as used in plural, OWN-PT had 72 nonempty synsets, that is, synsets with some corresponding Portuguese words added.

[4] `http://www.ontologyportal.org`

[5] `http://compling.hss.ntu.edu.sg/omw/`

Thus, our first step was to complete the empty synsets in OWN-PT, as for example, {`03684224-n locking pliers | pliers that can be locked in place`}, where the Portuguese word *alicate de pressão* was added.

From those 239 PWN synsets, we did not complete some 18 synsets, as we could not find lexicalized forms in Portuguese for those senses. For example we do not find Portuguese words to complete the synset {`04570532-n widow's weeds, weeds |a black garment (dress) worn by a widow as a sign of mourning`}. We left those synsets empty and marked them in the web interface of OWN-PT with the tag `en_only`, a tag proposed to identify PWN synsets verified, but with no translation to Portuguese. This list can be checked at `http://wnpt.brlcloud.com/wn/search-activities`, if one searches for the `en_only` hashtag.

From those 221 synsets that should have senses in Portuguese, 48 synsets have both plural and singular word forms used in Portuguese. For these we completed the synsets with the singular lemma and left the PWN plurality feature indicating that these synsets are in general used in the plural. Examples of this kind are: {`02854739-n pants, bloomers, knickers, drawers | calcinha, calça | underpants worn by women; ‘‘she was afraid that her bloomers might have been showing"`} and {`03041449-n cleats | chuteira | shoes with leather or metal projections on the soles; ‘‘the football players all wore cleats"`}

Around 100 synsets seem to be expressed only through singular words in Portuguese, such as {`02850552-n bleachers | arquibancada | an outdoor grandstand without a roof; patrons are exposed to the sun as linens are when they are bleached`}. We listed them as synsets that maybe should loose the PWN plurality classification, but have not changed them in our lexical base. This list is available at `https://github.com/livyreal/Singular_Synsets`.

Around 74 synsets, approximately 30% of the original PWN list, are actually often used in the plural. For those, we added the singular and the plural form in OWN-PT, as one could use this pluralized lemma information. We decided for adding also the singular form, since in many pipelines the word forms searched within wordnets are lemmas, singular forms. Synset { `07943646-n ancients | antigo, antigos | people who lived in times long past (especially during the historical period before the fall of the Roman Empire in western Europe)`} and {`00179916-n wings | asas, asa | a means of flight or ascent; ‘‘necessity lends wings to inspiration"`} are examples of this.

Finally some 27 synsets were completed with mass nouns, which are indeed singular forms. We completed them and marked all of them with the `#mass`, labelling them for future work. Examples are {`07942152-n people | população, gente, povo | any group of human beings (men or women or children) collectively; ‘‘old people"; ‘‘there were at least 200 people in the audience"`} and { `02730568-n fitting, appointment | aparelhagem | furnishings and equipment (especially for a ship or hotel)`}.

4.2 Discussion

Some interesting points came up when looking this data. Besides the fact that the feature is idiosyncratic, we can sketch some preliminary conclusions. For example, nouns related to clothing and instruments, such as *pliers, tongs, pants* and *suspenders*, that are in English lexicalized as plural forms, do not have the same behavior in Portuguese: *alicate, pinça, calça, suspensório* can be used in the plural, but do not have to be so. The singular forms are perfectly acceptable. Clothing nouns in Portuguese sometimes can refer both in the singular or the plural to the same entity. The forms *calça* or *calças* can refer to a single object, but the singular usage is more frequent, at least in Brazil. The same does not hold for instruments, as *tesouras* and *alicates* only refer to more than one object. We decided to add to OWN-PT only word forms in the singular in these cases and keep the feature `ClassifiedByUsage:plural` in the synsets of words that have the same referent both in plural and in singular.

Also in the clothing domain, we found some English words that have made their way into Portuguese: *shorts, jeans.* They are interesting because they are often used in plural in English and arrived in Portuguese already with the plural mark (*-s*), but, even with the plural mark, they are common count nouns and currently used in the singular, *o meu jeans, o shorts dela.* For those, we added only the singular word forms in Portuguese.

In English, there are also many pluralized abstract nouns, as *congratulations, felicitations, compliments, regards, wishes*, that are also, in general, used in plural in Portuguese, *congratulações, felicitações, cumprimentos, parabéns, votos.* In both languages, we can find contexts where those words are used in singular: *That was a nice compliment!/Isto foi um ótimo cumprimento!.* For those, we keep the `ClassifiedByUsage:plural` feature, but add only the word form in singular, except for the case of the word *parabéns*, that has only one form.

A general remark about this `ClassifiedByUsage:plural` PWN feature is related to its vague definition that influences what kind of objects in wordnet it classifies, following remarks in [13]. The documentation of PWN says `ClassifiedByUsage:plural` labels synsets. Those synsets come with a gloss *often in plural* (or *usually plural*) and are connected to the synset {`06295235-n plural, plural form | the form of a word that is used to denote more than one`}. However, many of those synsets have more than one word form and not all of them are used in plural. This is the case of {`03504723-n central office, main office, home office, home base, headquarters | the office that serves as the administrative center of an enterprise; ‘‘many companies have their headquarters in New York"`}, in which only the word *headquarters* are actually often used in plural. Keeping this feature related to synsets would also cause some problems in Portuguese, as some synsets bring both words in plural and in singular, e.g. {`03405265-n furnishing | móvel, mobília, mobiliário | the instrumentalities (furniture and appliances and other movable accessories including curtains and rugs) that make a home (or other area) livable`}. *Móvel* only have this sense when in plural and both *mobília* and *mobiliário* are mass nouns, which show us that we need to have a wordnet story to tell about mass nouns too. For now we decided to not include in the glosses of synsets the information *usually in plural* and collecting candidates for lose this feature.

This mass/count issue also appears when looking to synsets that refer to groups, as {`08179205-n poor people, poor |people without possessions or wealth (considered as a group; ‘‘the urban poor need assistance"`} and {`08477307-n unemployed people, unemployed | people who are involuntarily out of work (considered as a group; ‘‘the long-term unemployed need assistance")`}. For those synsets, we add mass nouns (whenever is possible) and word forms in singular when we do not have a mass noun for it. We also leave those synsets labelled with the PWN plural feature. For now, we do not have a way to mark mass nouns in OWN-PT and we leave this as near future work.

5 Conclusions

This work is an investigation on how to encode plurality in lexical resources, namely we checked how Princeton Wordnet brings this information and how is the best way to fit it in OpenWordnet-PT, an open wordnet for Portuguese language.

Princeton Wordnet has the classifier feature `ClassifiedByUsage:plural`, that labels synsets and words which are often used in the plural. However one can not expect that this idiosyncratic feature could percolate to other languages. Then we checked in OpenWordnet-PT how these PWN pluralized synsets should be stated in Portuguese. Around 55% of the pluralized synsets in English are truly used in plural in Portuguese and many of them can also be used in the singular, keeping the same meaning. We then list the remains as candidates for loosing this plurality feature inherited from PWN.

From this manual checking of OpenWordnet-PT synsets, we completed more than 200 synsets in Portuguese. Even in synsets with word forms usually used in plural, we decide for completing Portuguese synsets with its singular form. We decide for this uniform treatment mostly thinking on NLP tools that search for lemmas in wordnet. Keeping their lemma (as usual in singular) and also the plural label, we think we have the correct information encoded.

However, many concepts with this plurality feature can be translated as mass nouns, which we think is the best translation in several cases. How to encode this mass feature, however, is still an open question to us, that we leave as future work.

References

1. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. pp. 136–145. CICLing ’02, Springer-Verlag, London, UK, UK (2002), http://dl.acm.org/citation.cfm?id=647344.724142

2. Bond, F., Foster, R.: 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, ACL, Sofia (2013)
3. Chierchia, G.: Plurality of mass nouns and the notion of 'semantic parameter'. In: Rothstein, S. (ed.) Events and Grammar, pp. p. 53–103. Kluwer (1998)
4. Katz, G., Zamparelli, R.: Meaning-shifting plurality ans the count/mass distinction. In: Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4) (2011)
5. Legrand, S., Pulido, J.: A hybrid approach to word sense disambiguation: Neural clustering with class labeling. Knowledge Discovery and Ontologies (KDO-2004) workshop, 15th European Conference on Machine Learning (ECML) and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (2004)
6. Mandala, Rila, T.T., Hozumi, T.: The use of wordnet in information retrieval. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal (1998)
7. Mann, G.: Fine-grained proper noun ontologies for question answering. In: Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks. Taipei (2002)
8. Meyer, J., ', R.D.: Using the wordnet hierarchy for associative anaphora resolution. In: Proceedings of the Coling 2002 Workshop 'SemaNet'02: Building and Using Semantic Networks. Taipei (2012)
9. Narayanan, A.: Systems and methods for picture based communication (Apr 29 2014), `https://www.google.com/patents/US8712780`, uS Patent 8,712,780
10. Niles, I., Pease, A.: Toward a Standard Upper Ontology. In: Welty, C., Smith, B. (eds.) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems. FOIS-2001 (2001)
11. Patwardhan, S., Pedersen, T.: Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL (2006)
12. Quine, W.: Word and Object. MIT Press, Cambridge (1960)
13. Rademaker, A., Chalub, F.: Verifying integrity constraints of a rdf-based wordnet. In: Global Wordnet Conference 2016. Bucharest, Romenia (Jan 2016)
14. Wibowo, A., Christian, P., Handojo, A., Halim, A.: Application of topic based vector space model with wordnet. In: Proceedings of Uncertainty Reasoning and Knowledge Engineering (URKE) (2011)