Universal POS tagging for Portuguese: Issues and Opportunities

Valeria de Paiva and Livy Real

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

Abstract. Part-of-Speech (POS) tagging consists of labeling every token of a text with its correct morphosyntactic category and is considered by many a solved task in NLP. However, there are many tag systems in use, tags are not very easy to compare, there is no official golden standard and hence comparing performance of different systems is a nightmare, even for English. Much more so for less resourced languages. Recently a collective of researchers decided to tackle this issue and there is a new initiative, the Universal Dependencies project, that is developing cross-linguistically consistent treebanks and annotations for many languages. We look at how the coarse categories of POS tags defined by the Universal Dependencies project would work for Portuguese and describe the issues of aligning them with the POS tags produced by FreeLing, the open source NLP system we use.

1 Introduction

Part-of-Speech (POS) tagging consists of labeling every token of a text with its correct morpho-syntactic category and is considered by many a solved task in NLP, for English, at least. Supervised POS tagging accuracies for English, measured on the Wall Street Journal portion of the PennTreebank, have converged to an impressive 97% [15]. But for languages other than English the situation is not so rosy. For a start, for most languages there are not as many open source POS tagging systems as there are for English. And actually, even for English, the situation is not as good as this simple number might indicate (see [8]).

Nevertheless, work on supervised and unsupervised multilingual tagging is progressing and there is a new initiative, the project Universal Dependencies[3] (UD), that is developing cross-linguistically consistent treebanks and annotations for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning and multilingual parsing research. They aim to produce truly universal POS tags, based on the idea that there is a set of (coarse) syntactic POS categories that work in similar fashion across many, perhaps all, languages. The project is ongoing, having had its first official release (with ten languages) in January 2015. Version 1.1 with eight additional languages was released in May 2015 and subsequent releases are expected every six months, with

[3] http://universaldependencies.github.io/docs/

the next one schedule for May 2016. The guidelines for the UD were released October 1, 2014 and were kept stable for a year. It is expected that guidelines, tags and features may be revised as the discussions unfold and the empirical basis for generalization increases. A 'laundry list' of 17 issues (similar to the ones discussed here) was discussed at the Uppsala meeting, as part of Depling 2015, and can be found in `http://universaldependencies.org/issues.html`. It is worth noticing that many remain open questions, as of this writing.

A basic assumption of the Universal Dependencies project, in the words of Nivre [10] is

> [...]that dependency relations hold primarily between content words, while function words are pushed to the bottom of the trees and attached in a flat structure to the content word with which they are most closely associated. This principle is enforced to maximize parallelism across languages, since content words and their relations are more likely to be similar across languages, while function words in one language often correspond to morphological inflection (or nothing at all) in other languages.

While the general principle seems sound and very useful, there are too many details that are not clear cut and seem to deserve a more detailed discussion, in the specific settings of different languages. In this note we look at how these coarse categories of POS tags would work for Portuguese and describe the issues of aligning them with the POS tags produced by FreeLing [12], the open source NLP system which we have been using so far. We are not interested in the POS tagging task in NLP per se, but on whether the tag system proposed by Universal Dependencies project is adequate for Portuguese and if not, how to make it so.

We are also interested in the converse task, the use of pos-tagging to improve lexical resources such as the OpenWordNet-PT [4]. Thus we investigate the state of the existing tags, and then discuss possibilities of implementing new coarser tags similar to the ones in the Universal Dependencies project.

2 Google and Universal tags

To facilitate research in unsupervised induction of syntactic structure and to help standardize best-practices, Petrov, Das and MacDonald [13] proposed a tagset that consists of 12 universal POS categories. As they explain, their reasons were pragmatic: there might be some controversy about what the exact tagset should be, but these categories cover the most frequent parts of speech that seem to exist in most languages. They also developed a mapping from finer grained POS tags for 25 different treebanks to this universal set, showing some level of universality of their tagset. They made the tagset plus mappings[4] available in 2012.

Their universal tagset grew out of the cross-linguistic error analysis based on the CoNLL-X shared task data by [9]. It was initially used for unsupervised part-of-speech tagging by [3] and has since been adopted as a widely used standard

[4] `https://code.google.com/p/uni-dep-tb/`

for mapping diverse tagsets to a common standard, as explained in the Universal Dependencies website.

After extensive discussion, the original set of Google tags was improved to make some distinctions that were missing in the original proposal, but were perceived to be of importance by many. The universal part-of-speech tags (UPOS) are based on the Google universal tagset, which has been extended and redefined from the original 12 to the current 17 tags. The additional 5 tags added are: auxiliary verb (AUX), interjection (INTJ), proper noun (PROPN), subordinating conjunction (SCONJ), and symbol (SYM). In addition, UD also defines a set of 17 universal features that can be used to describe lexical and inflectional properties of words. These features are especially useful for morphologically rich languages. The core feature set is based on Interset [16], an interlingua for morphosyntactic tagsets. It is likely that new features or new feature values will be identified as new languages are added; therefore, the UD format allows additional language-specific features. The full set of 17 tags is listed in Table 1.

open class words	closed class words	other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Table 1. Universal Dependencies tag set

It is worth noticing that Princeton WordNet (PWN) does not list interjections in their open class words. Also it does not deal with any of the closed class words. Since we are mostly interested in the open class words that PWN has content on, we we restrict ourselves to nouns, verbs, adjectives and adverbs in some of our discussion.

Given the weight of the proponents of this suggested lingua franca, Google Research and the Stanford NLP group, it seems very likely that these will become the *de facto* standard in the description and annotation of corpora and hence it makes sense to see how difficult it would be to construct mappings to this standard set from other tagsets. This is a necessary step before defining the universal dependencies for Portuguese, which we also would like to do soon.

3 POS-Tagging Portuguese

There is a considerable amount of work in pos-tagging in Portuguese. In particular recently Garcia and Gamallo have worked exactly on pos-tagging in Portuguese using FreeLing [6,7]. Garcia et al. experiments show that consistency between the training corpus and the dictionary used has a major effect in the

POS tagger performance, at least for the taggers they used. Given the variation between European Portuguese and Brazilian Portuguese, this consistency can be somewhat problematic to attain. Garcia and Gamallo [6] used Brants' tagger to adapt FreeLing for European Portuguese (and for Galician), achieving precision results of up to 96.3%. For Brazilian Portuguese, the work in [1] compared several POS taggers, with best results of 90.25% using the MXPOST algorithm of Ratnaparkhi. Further development, with simplified tagsets, improved the precision up to 97%, but the authors warn that this figure should be taken with some care. As they say, it must be remembered that the corpus used during the training is small, and it is not a representative model of the Portuguese language in general.

FreeLing has a careful discussion of the code it uses for POS tagging on its online documentation. It says it has two different modules to perform POS tagging: a developer needs to decide which method is to be used for a specific application and instantiate the right class. The first POS tagger is the hmm_tagger class, which is a classical trigam Markovian tagger, following the work of Brants [2]. The second module, named relax_tagger, is a hybrid system capable of integrating statistical and hand-coded knowledge, following [11]. The hmm_tagger module is somewhat faster than relax_tagger, but the later allows you to add manual constraints to the model. The manual describes its tagsets for Portuguese in `https://talp-upc.gitbooks.io/FreeLing-user-manual/content/tagsets/tagset-pt.html`. We repeat the 12 tags in Table 2.

open class words	closed class words	other
adjective	adposition	punctuation
adverb		
interjection	conjunction	
noun	determiner	
	number	
verb		
	pronoun	
	date	

Table 2. FreeLing tag set

It is easy to see that FreeLing tagset misses the Universal tags SCONJ (subordinating conjunction), AUX (auxiliary verbs), PROPN (proper nouns), PART (participles), SYM (symbols) and X. In compensation FreeLing has an extra tag for dates, as FreeLing offers a Date Detection Module that already tags time expressions.

We have tried a simple experiment checking a small collection of sentences of the *Floresta Sintáctica* corpus [5], to see which issues we would need to deal with, both with old and new tags. We discuss some of these issues below, treating them as questions, as we have not decided on how to proceed yet.

4 Issues with POS tagging

We have taken a small sample of sentences in Portuguese, from Brazilian newspaper articles and analysed them. We want to use the lexical resource OpenWordNet-PT as a lexicon for further processing, thus we check which words FreeLing tags as nouns, verbs, adjectives and adverbs in these sentences and we try to find them in OpenWordNet-PT. We check which ones of these words are present in the OpenWordNet-PT, with the right part-of-speech and the right meaning and which ones are not and why not.

Some researchers will say that "POS tagging is a mostly (if not purely) syntactic task". We disagree, the task is syntactic for sure, but it has a huge component of semantic information involved. As Zeman [16] explains most of the time a tag is a "compressed representation of a feature-value structure", hence the use of the term "morphological tag" for them. The goal of POS tagging task for us is to make sure that the expected semantics of the sentences is respected by the segmentation/tagging interplay.

The idea in this note is both to improve the lexical resource, by checking that it has the required words with appropriate parts of speech and meanings, but also to verify the quality of the POS tagging code, by checking how many correct tags it gets, for each sentence. Thirdly and most importantly, we want to check the adequacy of the proposed Google tags for Portuguese. This implies reviewing and discussing the relevant issues that are still undecided on that project. Some issues are practically very important, even if theoretically not so. For instance, it is recognized that having the full sentence without annotations as part of the treebank is very useful: for machine learning and linguists. Standardizing on having such and with a single, uniform label is easy, needs to be done, but does not reflect any theoretical insights. However many of the issues under discussion do reflect theoretical differences (e.g. how to annotate light verb constructions, how to annotate pronominal verbs, etc).

For the sentence *Aqui era o quarto, pobre, limpo, simples e acolhedor*[5] we would like FreeLing to detect that *quarto* ('room') is a noun, that *era* ('be') is the verb, that *aqui* ('here') is an adverb and that *pobre, limpo, simples, acolhedor* are adjectives. FreeLing recognizes *quarto* as the adjective 'fourth', not as a noun[6], but the other content words are properly tagged. All the content words are in the lexicon with the appropriate parts of speech.

For the sentence *Os jogadores se dividem pelos dez quartos do alojamento, equipados com frigobar, ar condicionado, televisão e telefone*[7] we would like FreeLing to detect that *jogador, quarto, alojamento, frigobar, televisão e telefone* ('player', 'room', 'lodge', 'minibar', 'television', 'telephone') are nouns, that *dividir, equipar* ('share', 'equip') are verbs and that *dez* ('ten') is a numeral. But

[5] 'Here was the room, poor, clean, simple and cozy.'

[6] FreeLing does have *quarto* as a noun in its dictionary, it just prefers the adjective part of speech in this example.

[7] 'The players are sharing ten rooms in the lodge, equipped with minibar, air conditioning, television and telephone.'

the POS tagging only recognizes *dez quartos* as a unit in this example. We would also want the tagging to see *ar condicionado* as a multi-word expression (mwe). If the tokenization is wrong and *ar condicionado* comes as two tokens, how do we measure the error? Is it one error or two? Lastly, we could want the tagger to know the determiners and the prepositions in the sentence, but for the purpose of the exercise in this note and for checking the lexical resource, we only need to check the open class words of nouns, verbs, adjectives and adverbs. So we restrict our attention to these.

Several questions present themselves, when we start to look at this set of sentences. Some of these questions are language specific, but mostly they are about the POS tagging state of art and how to define it, so that it is parallel in many languages.

What should we do with out of vocabulary (OOV) words? Which is the most perspicuous tag for them? They can be of several kinds, for instance colloquialisms (*cê tá indo aonde?*/ 'u going where?'), foreign words used in their original language (*teens, blues*), regionalisms (*piracema*/'a natural phenomenon when fish swim up river'), neologisms (*frigobar*, 'a hotel small refrigerator'; *tuitar* 'to tweet'), acronyms (IBM, FSE, OIC), etc. Most dictionaries would not have these words, but they do show up in corpora and we need to decide how to deal with them. Taggers usually have defaults and one needs to check that they are appropriate. Tagging *tá* (the verb *estar* can be used for *Yes!*) as an interjection is very reasonable, but not always. FreeLing's 'Unknown Word Guesser Module' seems to do a good job most of the time.

More importantly, there is also the out-of-vocabulary issue that is truly a failing of the lexical resources and these should be counted separately, perhaps. A word might be missing from the processing dictionary (and be treated as a unknown word) and/or can be known by the processing, but be missing the semantic meaning in the OpenWordNet-PT. In the previous example the word *frigobar* (for a refrigerator in a hotel room) was missing both from the FreeLing dictionary (it was guessed as a verb), and from OWN-PT. The word *vão* ('hole', 'opening') was missing in the OWN-PT, as a noun, in the sentence *Para melhorar a ventilação, podem ser criadas janelas nos telhados ou pequenos vãos.com telas para evitar a entrada de insetos*[8] but it also did not show up in the Freeling processing, due to a tokenization error.

What should we do with Named Entities? Should they be tagged as proper nouns or nouns? The universal tags have proper nouns, and FreeLing does have the subcategory, so making the change is not difficult. Some named entities are present in our lexicon at the moment, e.g. *Charles de Gaulle*, many will not be, e.g. *Barak Obama*. Some might be abbreviations, such as IBM and NY; some might look like multiword expressions, like *Ministério da Fazenda* (Department of Finance). Some abbreviations are fairly well-known, such as ONU (Organização das Nações Unidas or UN, for United Nations), and OMS (Organização Mundia da Saúde or 'WHO', World Health Organization). Others,

[8] 'To improve ventilation, windows or small openings can be created on rooftops, with screens to prevent the entry of insects.'

like *FSE*[9] in the sentence *Na época, o então ministro da Fazenda, Fernando Henrique Cardoso, fez um pronunciamento em cadeia nacional para anunciar a intenção do governo de destinar o FSE a investimentos sociais*[10], are not so well known.

Recognizing named entities is, of course, a problem on its own, but they have to be classified as well. Which types of named entities should we have as a bare minimum? Most systems have types for *person, location, organization* and a category *other* seems sensible. But there is also the discussion of which of these named entities should you have in your lexical resource. Since our lexical database OpenWordnet-PT comes from Princeton's Wordnet, only a few named entities are available in that resource. We need to address the issue of how to deal with named entities, since this kind of information could also be extracted from an encyclopedic resource, such as Wikipedia, DBpedia or GeoNames, as discussed in [14].

Which kinds of numbers in the same tag? Most of the tag systems have numerals, like the *dez* ('ten') in *dez quartos* ('ten rooms') in the sentence above. But which other kinds of mathematical entities should be in the same tag? FreeLing has a special tag *date* which is not in the UD tagset. A recent discussion in the issues tracker for the Open Dependencies project showed that Germanic languages differ from Romance languages as to how they refer to dates, for instance. The discussion and (preliminary) conclusions are recorded at `https://github.com/UniversalDependencies/docs/issues/210`. A similar, but not finalized discussion, is going on about hours: should *20:30* in *he met me at 20:30* be tagged as a noun or as a numeral? What if you write it as *20h30*? Does it matter if you say the 'hours/horas' or not when you read the sentence? Similarly the UD tagset has the tag SYM (symbol) to be used for percent signs and other mathematical symbols, but FreeLing has not.

What to do with what are clearly typos in the text? For instance the full period in the example *Para melhorar a ventilação, podem ser criadas janelas nos telhados ou pequenos vãos.com telas para evitar a entrada de insetos*[11] that should perhaps be a comma. For the Portuguese corpus *Floresta Sintáctica* there were guidelines that enforced the non-modification of the sentences in the corpus. Corpora in general will have typos and mistakes and normally this is not an issue. But when the corpus is supposed to be used as the golden standard from where all the community will learn its annotations, it can be frustrating. Especially when it has many words that do not exist in the original language, that are simply misspellings of true words.

[9] *Fundo Social Europeu*, 'European Social Fund' (ESF).

[10] At that time, the then finance minister, Fernando Henrique Cardoso, made a statement on national television to announce the government's intention of allocating the EFS to social investments.'

[11] 'To improve ventilation, windows or small openings can be created on rooftops, with screens to prevent the entry of insects.'

What to do with MWEs? How to deal with them minimally? As the Universal Dependencies site explains, when discussing tokenization[12] "in principle, the lexicalist view could also be taken to imply that certain multiword annotations should be treated as single words in the annotation. So far, however, multiword expressions are annotated as such using special dependency relations, rather than by collapsing multiple tokens into one." While following their lead is the easiest option, given this work's origin in using OpenWordnet-PT and Princeton's WordNet, many MWEs are already lexicalized, like "air conditioning", for example. Not using such MWEs seems a step backwards, semantically. Particularly when it comes to adverbial expressions, not to treat them as MWE seems a seriously bad idea. Do we need to be able to separate noun-noun compounds, like *assessor de imprensa* at this level or not?

but experience shows that coarser tags get better numbers.

What to do about reported speech and quotations? Many other grammatical issues are still being discussed. As far as verbs are concerned, the working group decided that marking auxiliar verbs as distinct from main lexical verbs was important. But many questions remain: how to mark light verbs? What is the extent of the auxiliary verbs?

To start to survey these issues and determine reasonable ways of measuring precision and recall for POS tagging, a small corpus of twenty five short sentences was extracted from the manually corrected portion of the Bosque corpus and analyzed. The main conclusion, so far, is that the questions discussed above need addressing and that more experimentation with adapters for FreeLing is necessary.

5 Experiment and numbers

So far we have performed a very small experiment, devising our own golden standard, where we disagreed with the Bosque tags, whose numbers can be summarized thus:

	FreeLing	Bosque	Golden
sentences	25	25	25
tokens	720	716	714
nouns	131	151	142
verbs	84	82	82
adjectives	36	43	42
adverbs	18	20	20
proper nouns	57	43	46
numbers	22	21	20
dates	7	0	0
symbol	0	0	3

Table 3. Comparing tags

FreeLing does not have the 5 new tags added to the Google Tags by the Universal Dependencies project. We would like to have them. FreeLing has one

[12] `http://universaldependencies.org/u/overview/tokenization.html`

tag that both Bosque and the UDs do not consider, a special tag for dates, which we think is not necessary as a morphological tag.

The fact that dates are separate in FreeLing explains some of the differences in number of nouns as in, e.g. the date *31 de janeiro*, *janeiro* is a noun. Temporal expressions are also treated differently and are a topic under discussion in the Universal Dependencies forum.

A well known issue occurs with participles: sometimes they are tagged as verbs, sometimes as adjectives and the difference is not so easy to detect. World knowledge can play a part even on this shallow level of processing: the sentence *BRASÍLIA Pesquisa Datafolha publicada hoje revela um dado supreendente: recusando uma postura radical, a esmagadora maioria (77%) dos eleitores quer o PT participando do Governo Fernando Henrique Cardoso*[13], FreeLing tags *quer* as a conjunction, when it is clearly a form of the verb *querer* (to want).

Another small difference between tagsets is treating the percent sign `%` as either a noun or as a symbol. We follow the UD tags and think this should be a symbol, just as the dollar sign `$`. Altogether FreeeLing's performance is very good, as we are comparing it to humans and these already have differences amongst themselves.

But most of the disagreement is on how to tokenize multi word expressions (MWEs) and especially entity names, both in UDs and in FreeLing. (There are also many differences on how to tokenize and classify prepositions and determiners, but we are not interested in those, for the time being.) There is one adjective (*italiano*) that the Bosque treats as a noun, maybe simply an oversight. There are two nouns that our golden standard considers proper nouns (*Lua, Terra/Moon, Earth*) while Bosque thinks of them as common nouns.

6 Conclusion

This preliminary note puts forward the idea of adapting FreeLing to use the POS tags of the project Universal Dependencies and discusses some of the issues involved. While it seems clear that POS tagging, named entity recognition and tokenization are inter-related tasks, it is not so clear to us which ways will lead to better performance. The ever present problems of recognizing MWEs, compounds and OOV words, as well as the ambiguity issues are still plaguing us very much, but some progress seems to have been made and more of it can be made, if multilingual corpora, tags, and dependencies can be aligned. As a next step we want to run FreeLing in the whole Bosque corpus and adapt the UD dependencies scripts to check for the inconsistencies between Zeman's conversion of the Bosque dependencies in `https://github.com/UniversalDependencies/UD_Portuguese` and our own results, as well as the official guidelines. Aligning tags and dependencies, with our aim firmly set on semantics, is our goal.

[13] Brasilia Datafolha research published today reveals a surprising fact: refusing a radical posture, the absolute majority of the electors wants the PT participating in the Government of Fernando Henrique Cardoso.

References

1. Rachel V. Xavier Aires, Sandra M. Aluísio, Denise C. S. Kuhn, Marcio L. B. Andreeta, and Osvaldo N. Oliveira. Combining Multiple Classifiers to Improve Part of Speech Tagging: A case study for Brazilian Portuguese. In *Proceedings of the Brazilian AI Symposium (SBIA2000)*, pages 20–22, 2000.
2. Thorsten Brants. TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
3. Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL*, 2011.
4. Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An open Brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.
5. Cláudia Freitas, Paulo Rocha, and Eckhard Bick. Floresta sintá (c) tica: Bigger, thicker and easier. In *Computational Processing of the Portuguese Language*, pages 216–219. Springer, 2008.
6. Marcos Garcia and Pablo Gamallo. Análise morfossintáctica para português europeu e galego: Problemas, soluçoes e avaliaçao. *Linguamática*, 2(2):59–67, 2010.
7. Marcos Garcia, Pablo Gamallo, Iria Gayo, and Miguel A. Pousada Cruz. Postagging the web in Portuguese. national varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53(0):95–101, 2014.
8. Christopher D Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer, 2011.
9. Ryan T McDonald and Joakim Nivre. Characterizing the errors of data-driven dependency parsing models. In *EMNLP-CoNLL*, pages 122–131, 2007.
10. Joakim Nivre. Universal dependencies for swedish. In *In Proceedings of the Fifth Swedish Language Technology Conference (SLTC) Uppsala University, 13-14 November 2014*, 2014.
11. Lluís Padró. A hybrid environment for syntax-semantic tagging. *arXiv preprint cmp-lg/9802002*, 1998.
12. Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.
13. Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.
14. Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. Gentle with gentilics. In *Proceedings of the Workshop on LREC*, 2016.
15. Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Procs. of the 2003 North American Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. ACL, 2003.
16. Daniel Zeman. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*, 2008.