Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 95-101 recibido 15-04-14 revisado 06-06-14 aceptado 12-06-14

PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems [*]

Anotación morfosintáctica de la Web en portugués. Variedades nacionales, tipologías textuales y sistemas ortográficos

Marcos Garcia and **Pablo Gamallo**
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
{marcos.garcia.gonzalez, pablo.gamallo}@usc.es

Iria Gayo
Cilenis Language Technology
iria.gayo@cilenis.com

Miguel A. Pousada Cruz
Universidade de Santiago de Compostela
miguelangel.pousada@usc.es

Resumen: La gran cantidad de texto producido diariamente en la Web ha provocado que ésta sea utilizada como una de las principales fuentes para la obtención de corpus lingüísticos, posteriormente analizados utilizando técnicas de Procesamiento del Lenguaje Natural. En una escala global, idiomas como el portugués —oficial en 9 estados— aparecen en la Web en diferentes variedades, con diferencias léxicas, morfológicas y sintácticas, entre otras. A esto se suma la reciente aprobación de una ortografía unificada para las diferentes variedades del portugués, cuyo proceso de implementación ya ha comenzado en varios países, pero que se prolongará todavía durante varios años, conviviendo por lo tanto también diferentes ortografías. Una vez que los etiquetadores morfosintácticos existentes para el portugués están adaptados específicamente para una variedad nacional concreta, el presente trabajo analiza diferentes combinaciones de corpus de aprendizaje y de léxicos con el fin de obtener un modelo que mantenga una alta precisión de anotación en diferentes variedades y ortografías de esta lengua. Además, se presentan diferentes diccionarios adaptados a la nueva ortografía (Acordo Ortográfico de 1990) y un nuevo corpus de evaluación con diferentes variedades y tipologías textuales, disponibilizado libremente.

Palabras clave: anotación morfosintáctica, portugués, Web as Corpus, Acordo Ortográfico

Abstract: The great amount of text produced every day in the Web turned it as one of the main sources for obtaining linguistic corpora, that are further analyzed with Natural Language Processing techniques. On a global scale, languages such as Portuguese —official in 9 countries— appear on the Web in several varieties, with lexical, morphological and syntactic (among others) differences. Besides, a unified spelling system for Portuguese has been recently approved, and its implementation process has already started in some countries. However, it will last several years, so different varieties and spelling systems coexist. Since PoS-taggers for Portuguese are specifically built for a particular variety, this work analyzes different training corpora and lexica combinations aimed at building a model with high-precision annotation in several varieties and spelling systems of this language. Moreover, this paper presents different dictionaries of the new orthography (Spelling Agreement) as well as a new freely available testing corpus, containing different varieties and textual typologies.

Keywords: PoS-tagging, Portuguese, Web as Corpus, Spelling Agreement

[*] This work has been supported by the HPCPLN project – Ref: EM13/041 (Galician Government) and by the Celtic – Ref: 2012-CE138 and Plastic – Ref: 2013-CE298 projects (Feder-Interconnecta).

ISSN 1135-5948

1 Introduction

In recent years, the Web has turned the main source of corpora for extracting information about different topics in many languages. Thus, Natural Language Processing (NLP) applications take advantage of web crawlers and data mining strategies in order to analyze large amounts of textual data.

In this respect, one of the main NLP tasks to be performed is Part-of-Speech (PoS) tagging, which consists of labeling every token of a text with its correct morphosyntactic category. Specifically for Portuguese, there are several state-of-the-art PoS-taggers for different varieties, such as the European (EP) (Bick, 2000; Branco and Silva, 2004) and the Brazilian one (BP) (Aires, 2000).

However, using the Web for obtaining corpora in Portuguese involves the crawling of texts from different varieties, including the African ones, as well as the use of sources which contain a mixture of national varieties, such as the Wikipedia.

Apart from that, a Spelling Agreement (Acordo Ortográfico de 1990, AO90) for Portuguese, which unifies the spelling system of the different national varieties, has been recently approved, and its implementation process has already started in some countries. The chronology of the process differs in each country, but it is expected that the new orthography will be mandatory in Brazil and Portugal before 2016, as well as in the other countries with Portuguese as official language (ending in Cape Verde in 2020).[1]

Furthermore, some of the main journals of Brazil and Portugal adopted the AO90 spelling system since 2010 (e.g., *Diário de Notícias* and *Jornal de Notícias* in Portugal, or *Folha de São Paulo* in Brazil), while others did not (e.g. *Público*, in Portugal), so large amounts of texts are published every day using this new orthography.

Taking the above facts into account, this paper evaluates the use of different PoS-taggers, trained with several combinations of European and Brazilian resources, for analyzing the Web in Portuguese, including various linguistic varieties, textual typologies and spelling systems.

In order to carry out the evaluation, this paper also presents new lexica of the AO90 and a manually revised corpus which represents in some way the (journalistic and encyclopedic) Web in Portuguese. The corpus includes European, Brazilian and African (from Angola and Mozambique) texts with samples before and after the AO90, and is freely distributed.

The experiments show that the consistency between the training corpus and the dictionary has the major effect in the PoS-tagger performance. Concerning the lexica, it is shown that the new dictionaries can be combined to better analyze texts using the AO90 orthography, without losing precision when PoS-tagging documents in different spelling systems.

Apart from this introduction, Section 2 includes the Related Work. In Section 3, the different resources used for training the PoS-taggers are presented. Then, Section 4 describes the performed experiments and their results, while Section 5 outlines the main conclusions of this paper.

2 Related Work

Several PoS-taggers were developed for Portuguese language, namely for the European and Brazilian varieties. Some of them are statistical models trained with specific resources for each variety, while others use rule-based approaches.

Among the latter ones, PALAVRAS uses large sets of rules and a lexicon of about $50,000$ lemmas for PoS-tagging (and also parsing) European Portuguese.

Marques and Lopes (2001) presented a neural-network approach for PoS-tagging, which obtain high-precision results ($\approx 96\%$) with small training corpora.

Ribeiro, Oliveira, and Trancoso (2003) compared Markov models and a transformation-based tagger (based in Brill (1995)) for PoS-tagging EP, focused on pre-processing data for a Text-To-Speech system.

In Branco and Silva (2004), the authors compare different algorithms (transformation-based (Brill, 1995), Maximum Entropy (Ratnaparkhi, 1996), Hidden Markov Models (HMM) (Tufis and Mason, 1998) and second order Markov models (Brants, 2000)) for analyzing EP. The best results (97.09%) were obtained with the transformation-based system.

In Garcia and Gamallo (2010), the HMM

[1] `http://pt.wikipedia.org/wiki/Acordo_Ortografico_de_1990`

tagger (Brants, 2000) of FreeLing (Padró and Stanilovsky, 2012) was adapted for European Portuguese (and also for Galician), achieving results up to 96.3%.

For Brazilian Portuguese, Aires (2000) also compared several PoS-taggers, with the best results of 90.25% with the MXPoST algorithm (Ratnaparkhi, 1996). Further development (with simplified tagsets) improved the precision up to 97%.[2]

MXPoST also obtained the best results for BP in (Aluísio et al., 2003), with a precision of about 95.92%.

Concerning annotated corpora for Portuguese, the Bosque corpus[3] contains about 138,000 tokens (20,884 unique token-tag pairs) for European Portuguese. For the Brazilian variety, the Mac-Morpho[4] corpus has 1,167,183 tokens (73,955 unique token-tag pairs) and it was used for training different models in Aluísio et al. (2003).

Finally, some lexica for Portuguese are available, and were also used for building PoS-taggers for this language. For EP, LABEL-LEX (SW)[5] Eleutério et al. (2003) includes 1,257,000 forms from about 120,000 different lemma-tag pairs.

In Brazilian Portuguese, Muniz (2004) presented the DELAF_PB[6] lexicon, which contains 878,651 forms from 61,095 lemmas.

Recently, some projects started to compile new resources for analyzing the new spelling system (AO90), such as Almeida et al. (2013) or the Portal da Língua Portuguesa.[7]

3 Linguistic Resources

In order to evaluate various models for PoS-tagging different varieties of Portuguese, the following resources were used.

3.1 Corpora

For European Portuguese, the Bosque corpus (footnote 3) was used. In particular, a version based on the one used in Garcia and Gamallo (2010) was adapted for a new tagset (see Table 1).

Category	**Tag**
Adjective	AD
Adverb	AV
Coordinating Conjunction	CC
Subordinating Conjunction	CS
Determiner (Definite/Indefinite)	DT
Demonstrative Determiner	DD
Possessive Determiner	DP
Preposition	PS
Verb	VB
Participle	VP
Common Noun	NC
Proper Noun	NP
Demonstrative Pronoun	PD
Exclamative Pronoun	PE
Indefinite Pronoun	PI
Personal Pronoun	PP
Relative Pronoun	PR
Interrogative Pronoun	PT
Possessive Pronoun	PX
Interjection	I
Numbers	Z
Contractions with preposition *de*	DC
Contractions with preposition *por*	PC
Punctuation	F*
Dates/Hours	W
Numerical Expressions/Quantities	Z*

Table 1: Tagset used in the experiments. Top categories are the main ones (both in lexica and in corpora). Bottom categories appear in some corpora, but not in the lexica. "*" indicates that there are several PoS-tags both for Punctuation (24 types) and for different Numerical Expressions (5 types).

For Brazilian Portuguese, the Mac-Morpho (footnote 4) was also adapted to the same tagset as the EP resource.

Finally, a new corpus (Web) containing several varieties and text typologies of Portuguese was used for evaluating the different PoS-taggers (Garcia and Gamallo, 2014). The corpus has about 52,000 tokens, and includes the following sources: three Portuguese journals, two Brazilian journals, a journal from Angola, a journal from Mozambique, and texts from the Wikipedia in Portuguese, containing texts from different varieties. Table 2 shows the details of this new resource.

[2] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

[3] http://www.linguateca.pt/Floresta/corpus.html#bosque

[4] http://www.nilc.icmc.usp.br/lacioweb/corpora.htm

[5] http://label.ist.utl.pt/pt/labellex_pt.php

[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

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

Variety	Size	Vocab
Brazil	11,460	3,137
Portugal	13,987	3,637
Angola	4,180	1,403
Mozambique	5,517	1,700
Wikipedia	17,187	4,003
Total	52,331	9,873

Table 2: Size (in number of tokens) and vocabulary (*Vocab*, number of different token-tag pairs) of the Web corpus (and subcorpora).

The journals from Mozambique and Angola, and one from Portugal do not use the AO90 orthography, while the other Portuguese corpora and the Brazilian ones use this new spelling system. Also, Mozambique and Angola have traditionally used the EP orthography (even though they have lexical and syntactic variations). Wikipedia corpus contains texts from Brazil and Portugal, with both pre-AO90 and post-AO90 spellings.

The PoS-tags were manually corrected and also converted to the same tagset as the above mentioned corpora.

3.2 Lexica

Concerning the lexica, different resources were also used:

The version of LABEL-LEX (SW) (Eleutério et al., 2003) for FreeLing suite[8] was adapted as the lexicon for EP. This lexicon was already used in Garcia and Gamallo (2010), and it has a strong consistency with the Bosque corpus.

The DELAF_PB (footnote 6) was the lexicon used for BP.

Apart from that, two new lexica were created for evaluating their influence when PoS-tagging different varieties:

PEB_Dict: PEB_Dict is a lexicon built by merging the EP and BP ones. In order to do that, all the token-lemma-tag triples of the EP and BP dictionaries were selected. Then, every triple in the EP dictionary was added to the PEB_Dict. After that, BP triples not included in EP were also added to PEB_Dict. For functional words, which sometimes had different PoS-tags in EP and BP, the European version was preferred.

The resulting dictionary contains about $1,254,000$ token-lemma-tag triples and $1,179,000$ token-tag pairs, from $112,000$ different lemmas.

It is worth noting that this fusion may increase the ambiguity of the PoS-tagger, since some entries belong to a higher number of token-lemma-tag triples.

AO+_Dict: AO+_Dict is another merged resource containing lexical units from different varieties. In order to create it, the strategy described for the PEB_Dict was followed, merging in this case a dictionary of the AO90 (developed by the authors) and the PEB_Dict. This way, AO+_Dict consists of a PEB_Dict enriched with the new forms of the Acordo Ortográfico de 1990. Note that AO+_Dict includes entries which are not correct in AO90. AO+_Dict has about $1,277,000$ token-lemma-tag triples and $\approx 1,200,000$ token-tag pairs. The number of lemmas of this resource is about $119,000$.

The tagsets of these two lexica were also unified (Table 1). As the tagset is simpler than the original one, the number of triples was reduced in about $50,000$ in each dictionary.

4 Experiments

In order to evaluate the performance of several PoS-taggers for analyzing different varieties of Portuguese, the following experiments were carried out.

First, both European Portuguese and Brazilian Portuguese resources were used for training and testing specific EP and BP taggers (`EPtag` and `BPtag`). The performance of these models was also evaluated with the Web corpus, which contains different varieties of Portuguese before and after the AO90.

Then, several training corpora and dictionaries were combined in order to evaluate (i) how they behave with the new corpora and (ii) whether they increase or decrease the PoS-tagging precision in EP and BP corpora before the AO90.[9]

4.1 Models

The HMM PoS-tagger of FreeLing (Padró and Stanilovsky, 2012) was the selected al-

[8] `http://nlp.lsi.upc.edu/freeling/`

[9] Both training and testing corpora, labeled with the different dictionaries are freely available at `http://gramatica.usc.es/~marcos/pt_tag_corpora.tar.bz2`

gorithm for doing the experiments. It is a state-of-the-art PoS-tagger algorithm implemented in an open-source suite of linguistic analysis which also contains other modules for previous and further NLP tasks.

The European Portuguese model (EPtag) was trained with ≈ 83% of the EP corpus (120, 007 tokens and 18, 035 unique token-tag pairs), and tested in the remaining 17% (with 23, 102 tokens and 5, 873 unique token-tag pairs).

The Brazilian tagger (BPtag) uses ≈ 79% for training (1, 000, 044 tokens and 62, 762 unique token-tag pairs) and ≈ 21% (267, 845 tokens and 30, 848 unique token-tag pairs) for testing. As the BP corpus is much larger than the EP one, two sub-corpora were extracted from the former, in order to obtain balanced datasets for doing more tests: (i) a short version of the training (with ≈ 150, 000 tokens and 16, 395 unique token-tag pairs) and (ii) a reduced version for testing (≈ 23, 000 tokens and 5, 690 uniq token-tag pairs). Thus, these short BP datasets have a similar size than the EP ones. Every extracted sub-corpus for both EP and BP were randomly selected, and the testing datasets were never used for training.

ALLtag model uses for training both the EP and BP training corpora, and the PEB_Dict lexicon. ALLtag+ was trained with the same corpora than ALLtag, but with the AO+_Dict.

Finally, the PEBtag taggers use the EP training corpus and the short version of the BP one, thus having a more balanced dataset. PEBtag and PEBtag+ models also differ in the dictionary: the former uses the PEB_Dict while the latter was trained with the AO+_Dict.

The tagset (Table 1) contains 23 tags, apart from punctuation (24 tags), dates and hours (1 tag), and numerical expressions (5 tags). During the experiments, only the FreeLing PoS-tagger was used, so other modules (Recognition of Dates, Numbers, Currencies, etc.) were not applied.

For testing the performance of the PoS-taggers with different varieties, the new Web corpus was used (Section 3.1). Different experiments were carried out using the sub-corpora from Angola (AN), Mozambique (MO), Brazil (BP_AO), Portugal (EP_AO) and from the Wikipedia (Wiki).

The total micro-average of the evaluation was computed by replacing the BP testing corpus with the shorter version of the same dataset, in order to reduce bias in the results.

4.2 Results and Discussion

Table 3 contains the results of the different PoS-taggers evaluated. Here, precision is the number of correctly labeled tokens in the test set divided by the total number of tokens in the same dataset.

BPtag and EPtag models obtained 95.96% and 97.46% precision values in their respective corpora, but their results are 1.4% and 0.6% (respectively) worse when analyzing the other variety. On these (EP and BP) corpora, the performance of the ALLtag and PEBtag models depends on the distribution of the training corpora. Thus, ALLtag models (with more BP data) analyze better the BP corpus, while the precision of PEBtag models is higher when tagging EP.

When comparing both versions of ALLtag and PEBtag models with the BPtag and EPtag ones, the combined taggers achieve a better tradeoff in the annotation of BP and EP corpora.

Apart from that, the impact of the AO+_Dict lexicon is null, because BP and EP corpora do not contain texts with the AO90 spelling.

Concerning the Web corpus, EPtag model is still the best in every sub-corpora, except for the Wikipedia one. In this respect, it is worth noting that the annotation consistency between the EP training corpus and the EP dictionary is higher than the other varieties, and that a large part of the Web corpus follows the EP orthography. Also, remember that AN, MO and one EP_AO sub-corpora use the EP spelling system, so the results follow similar tendencies than those in the EP corpus.

In general, PEBtag models behave slightly better than ALLtag ones (except in the Wikipedia dataset), but they do not overcome the performance of the EPtag model.

The results in the Web corpus show that using the AO+_Dict has low (but positive) impact in the annotation. Its effect is only perceived in some texts whose spelling system had more changes due to the use of the AO90 orthography (EP_AO and Wikipedia), with small improvements (≈ 0.3) when using the larger version, which includes the AO90 entries.

Model	BP	EP	AN	MO	BP_AO	EP_AO	Wiki	Web	Total
BPtag	95.96	96.03	97.06	96.39	96.35	96.88	95.52	96.28	96.13
EPtag	95.35	**97.46**	**98.18**	**97.76**	**97.29**	**97.80**	96.25	**97.20**	**96.85**
ALLtag	**96.07**	96.94	97.30	96.91	96.68	97.18	96.50	96.83	96.64
ALLtag+	**96.07**	96.94	97.30	96.91	96.68	97.21	**96.53**	96.86	96.65
PEBtag	95.74	97.04	97.37	97.06	96.97	97.28	96.43	96.92	96.65
PEBtag+	95.74	97.04	97.37	97.06	96.97	97.31	96.45	96.93	96.66

Table 3: Precision of 6 PoS-taggers on different testing corpora. *Web* is the micro-average of the *AN*, *MO*, *BP_AO*, *EP_AO* and *Wiki* results. *Total* values are the micro-average of all the results, except for *BP*, replaced by the shorter version (see Section 4.1) in order to avoid bias.

However, even though these new dictionaries increase the ambiguity of the PoS-tagging (since they contain more token-lemma-tag triples), their influence was always positive in the tests.

In conclusion, it must be said that the consistency between the training corpus and dictionary was crucial in these experiments, with the EPtag models achieving the best results in almost every dataset. Apart from that, the bias between different linguistic varieties in both training and test corpora has also impact in the results. Finally, the experiments also showed that the new dictionaries have a positive influence when PoS-tagging both pre-AO90 and post-AO90 corpora in Portuguese.

5 Conclusions

Natural Language Processing tools for languages with different varieties and spelling systems —such as Portuguese–, are often built just for one of these varieties. But current NLP tasks often use a Web as Corpus approach, so there is a need of adaptation of tools for different varieties and spelling systems of the same language.

This paper has evaluated the use of several combinations of lexica and corpora for training HMM PoS-taggers aimed at analyzing different varieties of the Portuguese language.

The combinations have been focused on the analysis of Web corpora, including different text typologies (journalistic and encyclopedic), national varieties (from Portugal, Brazil, Angola and Mozambique) and spelling systems (before and after the Spelling Agreement of Portuguese: Acordo Ortográfico de 1990).

Moreover, new resources has been presented: (i) manually revised corpora for the above mentioned varieties and text typologies and (ii) two different dictionaries for Portuguese, with various combinations of European and Brazilian forms before and after the Acordo Ortográfico de 1990.

The results of the different evaluations indicate that models built with consistent training data (both corpora and lexica) achieve the highest precision.

Concerning the lexica, it has been shown that using dictionaries enriched with AO90 entries allows PoS-taggers for Portuguese to better analyze corpora from different varieties.

Finally, using a balanced training data from different varieties also helps to build a generic PoS-tagger for different linguistic varieties and text typologies.

References

Aires, Raquel V. Xavier. 2000. Implementação, adaptação, combinação e avaliação de etiquetadores para o Português do Brasil. Master's thesis, Instituto de Ciências Matemáticas, Universidade de São Paulo, São Paulo.

Almeida, Gladis Maria de Barcellos, José Pedro Ferreira, Margarita Correia, and Gilvan Müller de Oliveira. 2013. Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *ESTUDOS LINGUÍSTICOS*, 42(1):204–215.

Aluísio, Sandra M., Gisele M. Pinheiro, Marcelo Finger, M. Graças Volpe Nunes, and Stella E. Tagnin. 2003. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In *Proceedings of Corpus Linguistics*, volume 2003, pages 14–21.

Bick, Eckhard. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, University of Aarhus, Denmark.

Branco, António and João Silva. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th edition of the Language Resources and Evaluation Conference (LREC 2004)*, pages 507–510, Paris. European Language Resources Association.

Brants, Thorsten. 2000. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*. Association for Computational Linguistics.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota, and Paula Carvalho. 2003. Dicionários Electrónicos do Português. Características e Aplicações. In *Actas del VIII Simposio Internacional de Comunicación Social*, pages 636–642, Santiago de Cuba.

Garcia, Marcos and Pablo Gamallo. 2010. Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática*, 2(2):59–67.

Garcia, Marcos and Pablo Gamallo. 2014. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 3229–3233, Reykjavik. European Language Resources Association.

Marques, Nuno and Gabriel Lopes. 2001. Tagging with Small Training Corpora. In *Proceedings of the International Conference on Intelligent Data Analysis*, volume 2189 of *Lecture Notes on Artificial Intelligente (LNAI)*, pages 63–72. Springer-Verlag.

Muniz, Marcelo Caetano Martins. 2004. A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Paulo.

Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 1996)*, volume 1, pages 133–142. Association for Computational Linguistics.

Ribeiro, Ricardo, Luís C. Oliveira, and Isabel Trancoso. 2003. Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. In *Proceedings of the 6th Workshop on Computational Processing on the Portuguese Language (PROPOR 2003)*, pages 143–150, Faro. Springer-Verlag.

Tufis, Dan and Oliver Mason. 1998. Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger. In *Proceedings of the 1st edition of the Language Resources and Evaluation Conference (LREC 1998)*, volume 1, pages 589–596. European Language Resources Association.