Part-of-Speech Tagging of Portuguese Using Hidden Markov Models with Character Language Model Emissions

Marcelo Rodrigues de Holanda Maia[1]**, Geraldo Bonorino Xexéo**[1]

[1]Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ

{mmaia,xexeo}@cos.ufrj.br

Abstract. *This paper presents a probabilistic approach for POS tagging that combines HMMs and character language models being applied to Portuguese texts. In this approach, the emission probabilities for each hidden state in a HMM are estimated by a proper character language model. The tagger built has been trained and tested on Bosque, a subset of Floresta Sintá(c)tica treebank, reaching 96.2% accuracy with a 39-tag tagset and 92.0% with a 257-tag tagset extended with inflexion information.*

1. Introduction

Some advances in part-of-speech (POS) tagging techniques still have not been fully explored for other languages than English. Works in this field with Portuguese include Bick's rule-based tagger, with over 99% accuracy, which is a very impressive performance. However, probabilistic approaches applied to Portuguese have not reached the same levels achieved with English (96-97%) yet.

In this paper we present the results obtained using a probabilistic approach that combines hidden Markov models (HMMs) and character language models for POS tagging of Portuguese.

The remainder of this paper is structured as follows: in Section 2 the background for the addressed problem is discussed; in Section 3 we introduce the approach taken in this work; Section 4 presents the corpus used for training and testing the tagger; in Section 5 we define the experiment setup; Section 6 displays the key results; in Section 7 conclusions and future works are discussed.

2. Background

Part-of-speech tagging is the task of assigning parts-of-speech (morphological classes) to words in a sentence. As in most other NLP fields, the biggest problem faced by POS taggers is ambiguity. Many algorithms have been applied to this problem, generally following rule-based, probabilistic or hybrid approaches.

Rule-based approaches use hand-written rules for disambiguation. In probabilistic approaches, ambiguity is resolved by models induced from training corpora. Hybrid is based on disambiguation rules, but in this case the rules are induced from tagged corpora. For Portuguese, rule-based approaches are used in [3] (>99% accuracy) and [10] (~98.6% accuracy), probabilistic approaches in [12] (84.5% accuracy) and [8] (88.7% accuracy), and hybrid approaches in [2] (~89.4% accuracy), [9] (~90% accuracy), [7] (95% accuracy) and [6] (97.2% accuracy).

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 159–163, Cuiabá, MT, Brazil, October 24–26, 2011. ©2011 Sociedade Brasileira de Computação

The latest taggers developed following probabilistic approach, like HMM, have reached accuracy levels about 96-97% with English. Its advantages against rule-based and hybrid approaches include: (1) it does not require so much manual effort or linguistic knowledge to be employed on tagger development; (2) probabilistic taggers are not constrained by rule set coverage (due to the complexity of natural languages, it is extremely hard to build a rule set covering all cases and exceptions); and (3) since it does not depend on language-specific rules, it can be easily adapted for many languages.

3. Extension to HMM with Character Language Models

In this work, we use LingPipe's HMMs [5] for POS tagging. They are much similar to typical HMM tagger implementations, where hidden states correspond to tags and contextual probabilities are estimated for bigrams (first-order HMMs). The point in which they differ from usual HMMs is the way they estimate emission probabilities: instead of using relative frequency distribution of unique tokens, they use bounded character language models, one for each hidden state.

The character language models define probability distributions over strings to be emitted from their respective hidden states. They are bounded language models based on character-level n-grams, where probabilities are normalized over strings of a fixed length. This approach brings the advantages of implicitly including morpheme information in the model and defining a proper probability distribution normalized over the infinite set of possible string emissions (so it can estimate probabilities even for words not seen on training). Further details on LingPipe's character language models are presented in [4].

4. The Bosque Corpus

For training and testing the tagger, we chose to use Bosque corpus, a subset of the Floresta Sintá(c)tica treebank [1] composed of sentences from newspaper texts, with over 180 thousand words tagged with PALAVRAS parser [3] and fully revised by linguists.

4.1. Corpus Preprocessing

In this work we used Bosque version 8.0 in SimTreeML [11] format. Since it is a treebank and not just a POS tagged corpus, it includes much information that is beyond our scope. So the first step was to remove all information but the sentences and the POS tags assigned to each of their tokens.

For the purposes of this work two tagsets were considered: one including only word class tags (named TS_c) and another combining word class and inflexion tags (named TS_i). Additionally, punctuation tags were added to both tagsets (each corresponding to one of the punctuation tokens found in the corpus).

4.2. Final Tagsets

After preprocessing the corpus, the final version of TS_c tagset presented the following 20 word class tags: {***adj***, ***adv***, ***art***, ***conj-c***, ***conj-s***, ***ec***, ***intj***, ***n***, ***n-adj***, ***num***, ***pp***, ***pron-det***, ***pron-indp***, ***pron-pers***, ***prop***, ***prp***, ***v-fin***, ***v-ger***, ***v-inf***, ***v-pcp***}. The TS_i tagset presented

238 tags, each one being a combination of one word class tag and one or more of the following inflexion tags (no more than one of each type): {***M***, ***F***, ***M/F***} for gender; {***S***, ***P***, ***S/P***} for number; {***NOM***, ***ACC***, ***DAT***, ***PIV***, ***ACC/DAT***, ***NOM/PIV***} for case; {***1***, ***2***, ***3***, ***1/3***} for person; {***PR***, ***IMPF***, ***PS***, ***MQP***, ***FUT***, ***COND***, ***PS/MQP***} for tense; and {***IND***, ***SUBJ***, ***IMP***} for mood. Additionally, as stated before, both tagsets presented 19 punctuation tags.

5. Experiment Setup

The aim of this work was to evaluate the presented approach for POS tagging regarding accuracy (number of correctly tagged tokens / total number of tokens). In order to do so we have randomly partitioned the corpus sentences in two fixed sets: a training set containing approximately 90% of the corpus tokens and a test set containing the remaining tokens. The first was used for training, testing and tuning the models during development with 10-fold cross-validation, and then the second (unseen during development) was used for testing the final models (trained on the entire training set).

Since tagsets may differ between works we firstly defined a lower-bound baseline for comparison purposes. This baseline was a unigram POS tagger.

As stated in Section 3, our approach presents HMMs with character language model emissions and those language models are based on n-grams. So, there was one parameter to be tuned in our model: max n-gram size for HMM emissions. This tuning was done by varying max n-gram size on a range of values and, for each value, training and testing a corresponding model. This range has the lower bound 1 (unigram) and some upper bound defined by max length of words in the language (setting max n-gram sizes greater than max word length would be useless).

Additionally, confusion matrices were built on each test. They were helpful for error analysis, revealing most common sources of tagging mistakes.

6. Results

6.1. Development Phase

The first models trained were the unigram baselines. For TS_c the accuracy achieved was approximately 85.8% while for TS_i the accuracy achieved was approximately 82.3%.

The accuracies obtained for TS_c and TS_i with the HMMs were plotted in the chart shown in Figure 1. Both present a similar logarithmic growth in accuracy as max n-gram size values are increased until they reach a convergence point. As it can be seen, a max n-gram size 2 is enough to overcome the baselines and with max n-gram size 6 accuracy for TS_c surpasses 96%, the same level achieved with English probabilistic POS taggers. For TS_i, max accuracy reached is approximately 91.9%, which is not a bad result if compared to other Portuguese probabilistic POS taggers and represents a relatively low decrease since it contains more than 6 times the number of tags in TS_c.

6.2. Test Phase

After the analysis performed in development phase, we have found that a max n-gram size 10 should be enough to achieve the highest accuracy levels: approximately 96.2% with TS_c and 91.9% with TS_i.

In test phase we validate our results by testing our models on unseen data (the test set mentioned in Section 5). So we have re-trained the models with max n-gram size 10 on the entire training data set and tested it on the test set. The accuracy rates achieved were approximately 96.2% with TS_c and 92.0% with TS_i, which validates the results obtained during development.

From the confusion matrix built from the test ran on TS_c, we have found that the most common sources of tagging mistakes – summing to approximately 25% of all erroneous tag assignments – involved the tags ***adj*** (adjective), ***n*** (noun) and ***prop*** (proper noun). This means those are the hardest word classes for the POS tagger to distinguish.

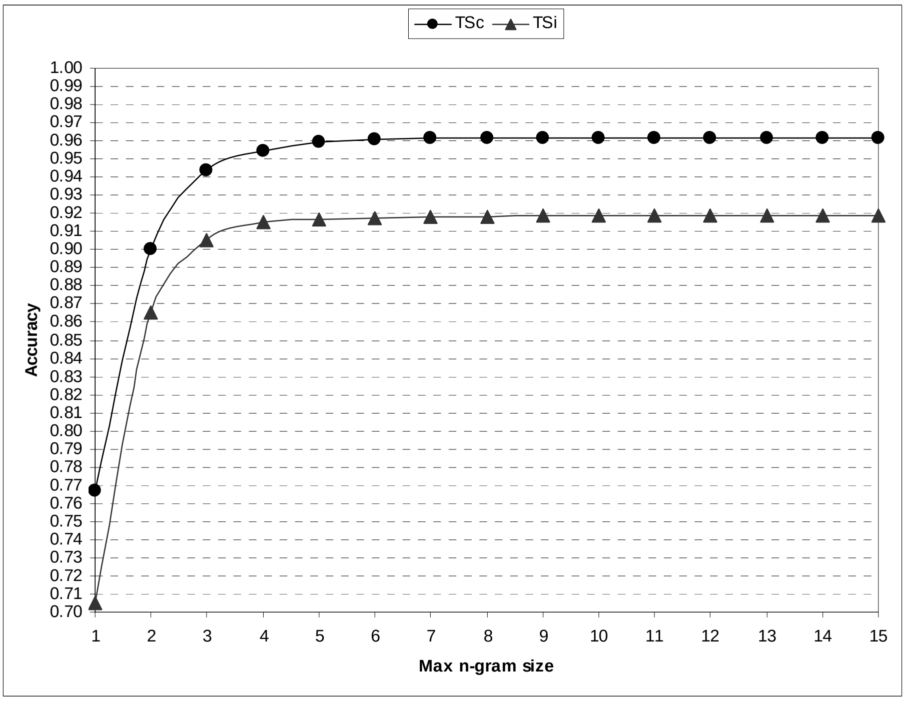

Figure 1. HMM accuracy versus max n-gram sizes for TS_c and TS_i

7. Conclusions and Future Works

The results obtained in this work show that the presented approach overcomes probabilistic POS tagging methods previously tried with Portuguese. The accuracy achieved with TS_c (~96.2%) is in the same range of those presented by English probabilistic POS taggers.

With TS_i we wish to evaluate the POS tagger with an extended tagset including inflexion information, which could be helpful in applications for which POS tagging is a basic task, like syntactic parsing. The accuracy obtained (~92.0%) is also higher than previous results on POS tagging of Portuguese and relatively good if we take into account the hardness generated by the high increase in number of tags.

We believe that improving interpolation for the character language models is a possible strategy to raise those accuracy levels. In future works we intend to perform further investigation on this POS tagging approach in order to reduce the error rate and apply it as a basis for a probabilistic parsing system framework.

References

1. Afonso, S., Bick, E., Haber, R. and Santos, D. (2002). *"Floresta sintá(c)tica": a treebank for Portuguese*, In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, pages 1698-1703.

2. Aires, R. V. X., Aluísio, S. M., Kuhn, D. C. S., Andreeta, M. L. B. and Oliveira Jr., O. N. (2000). *Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese*, In: Proceedings of the 15th Brazilian Symposium on Artificial Intelligence (SBIA'2000), Atibaia, Brazil.

3. Bick, E. (2000). *The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, PhD Thesis, Aarhus University Press.

4. Carpenter, B. (2005). *Scaling high-order character language models to gigabytes*, In: Proceedings of the ACL 2005 Workshop on Software, Ann Arbor, USA, pages 86-99.

5. Carpenter, B. (2007). *LingPipe for 99.99% recall of gene mentions*, In: Proceedings of the 2nd BioCreative Challenge Evaluation Workshop, Madrid, Spain, pages 307-309.

6. Domingues, M. L., Favero, E. L. and Medeiros, I. P. (2007). *Etiquetagem de Palavras para o Português do Brasil*, In: Proceedings of the 5th Workshop in Information and Human Language Technology (TIL'2007), Rio de Janeiro, Brazil, pages 1721-1724.

7. Kinoshita, J., Salvador, L. N. and Menezes, C. E. D. (2006). *CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus*, In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy.

8. Marques, N. C. and Lopes, G. P. (1996). *A neural network approach to part-of-speech tagging*, In: Proceedings of the 2nd Meeting for Computational Processing of Spoken and Written Portuguese (PROPOR'1996), Curitiba, Brazil.

9. Menezes, C. E. D. and Neto, J. J. (2002). *Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos*, In: Proceedings of the 2nd Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI'2002), Orlando, USA.

10. Seara, I. C., Pacheco, F. S., Kafka, S. G., Seara Jr., R. and Seara, R. (2010). *Morphosyntactic Parser for Brazilian Portuguese: Methodology for Development and Assessment*, In: Extended Activities Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (PROPOR'2010), Porto Alegre, Brazil.

11. Vilela, R., Simões, A., Bick, E. and Almeida, J. J. (2005). *Representação em XML da Floresta Sintáctica*, In: Actas da 3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA'2005), Braga, Portugal, pages 351-361.

12. Villavicencio, A. (1995). *Avaliando um rotulador estatístico de categorias morfo-sintáticas para a língua portuguesa*, Master's Thesis, UFRGS.