MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-959

Performance Analysis of a POS Tagger applied to Discharge Summaries in Portuguese

Michel Oleynik[a], Percy Nohama[a], Pindaro Secco Cancian[a], Stefan Schulz[b]

[a] *Exact Sciences and Technology Center, Pontifical Catholic University of Paraná, Curitiba, Brazil*
[b] *Institute for Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany*

Abstract

Part of speech taggers need a considerable amount of data to train their models. Such data is not readily available for medical texts in Portuguese. We evaluated the accuracy of a morphological tagger against a gold standard when trained with corpora of different sizes and domains. Accuracy was the highest with a medical corpus during the complete training process, achieving 91.5%. Training on a newswire corpus achieved 75.3% only. Furthermore, an active learning technique has been adapted to the POS tagging task. The algorithm uses a POS tagger committee to isolate the sentences with the highest disagreement indexes for manual correction. However, the method was not able to reduce training and tagging times when compared to a random selection strategy. We encourage that future works employ some effort in order to annotate a small amount of random data in the domain of study, which should be enough for higher accuracy rates.

Keywords:

Medical records, Natural language processing, Part of speech tagging.

Introduction

Natural language processing (NLP) has been studied and applied on a broad range of domains and languages across the world. An important use case is the automatic mapping of narrative EHR content to biomedical terminologies such as SNOMED CT [1], which provides the basis of clinical decision support systems. NLP tools are used for sentence and token boundary detection, acronym recognition and expansion and syntactical analysis of sentences. POS tagging is essential in this process to identify each token part of speech class and to assign the corresponding tag (e.g., V for verb, ADJ for adjective, etc.) to the term. This task can easily be addressed by a set of language-independent tools publicly available via the internet. These tools need only to be trained on a previously tagged corpus in order to build a language-dependent model.

English medical corpora are widely available, but Portuguese corpora are less common. Although pilot projects (like the one conducted by the Interinstitutional Center for Research and Development in Computational Linguistics in São Paulo University [2]) built tagged newswire corpora, the usage viability of these models on different domain, e.g. medicine, is still under discussion.

Hahn and Wermter [3] assessed the performance of POS taggers trained on the NEGRA corpus of German newspaper texts when applied on a German database of medical documents called FRAMED. They concluded, “POS taggers cannot only be immediately reused for medical NLP, but they also — when trained on medical corpora — achieve a higher performance level than for the newspaper genre”.

However, they also cite the work of Campbell and Johnson [4], which came to an opposite conclusion: POS taggers trained on newspaper data cannot be used on the medical domain without a new training process on a medical corpus.

In this paper, we want to give our own answer to this question, focusing on discharge summaries written in Portuguese, which is the focus of our main project. We also evaluated an active learning approach to accelerate the construction of a gold standard for the medical domain.

Materials and Methods

Newspaper and Medical Corpora

We used both a newspaper and a medical corpus in order to study this question. The newspaper corpus, named *MAC-Morpho*, is a compilation of 1,167,183 texts published in the *Folha de São Paulo* newspaper during the year of 1994. The corpus was compiled by the *Lácio-Web* [2] project from NILC-USP.

The medical corpus contains two collections of discharge summaries from the Clinical Hospital of Porto Alegre, Brazil. 2,453 discharge summaries cover the whole range of clinical specialties during one month (June 2007) and 5,617 discharge summaries were taken from the cardiologic department, covering a time span between June 2002 and May 2007. The corpus has been obtained from a partnership program between the Pontifical Catholic University of Paraná and the Federal University of Porto Alegre and has not been made publicly available due to privacy concerns.

In order to minimize efforts and keep compatibility between the corpora, we used only a fraction of the entire corpus. Lezius [5] argued that many German taggers were successfully trained on corpora variable from 20,000 (more common) to 200,000 (rarer) tokens. Based on that, we used a fraction of randomly acquired 120,000 tokens from each corpus.

We then extracted one out of ten sentences available in the corpora to produce a gold standard for each domain. After that, we collected some statistics from the final corpora, which we show on Table 1.

Table 1 - Corpora sizes

Corpus	Corpus Fraction	Sentences	Tokens
Newspaper	Gold	495	13,810
	Training	4,533	123,019
Medical	Gold	595	12,451
	Training	5,964	123,018

In the next step, we evaluated the newspaper and medical models on the medical gold standard over a set of iterations where training corpus size was continually increased.

Active Learning Strategy

We further analyzed an active learning strategy proposed by [6] to build a tagged medical corpus in Portuguese. The active learning approach, commonly applied during the training process of syntactical classifiers, is used there as a method for corpus construction. In our work, we employed a *committee-based* approach, which uses a committee of POS taggers to select the sentences with the highest disagreement indexes and thus indicate priority in manual correction.

In order to accomplish that goal, we first calculated the *sentence disagreement index $D_{sent}(s)$* for each sentence without manual correction on a given iteration. That index equals to the average of the *token disagreement indexes $D_{tok}(t)$* seen on a sentence. Ranged 0 to 1, those indexes show how inconsistent are the results of a tagger committee on a token or on a sentence view. Equations 1 and 2 were proposed by [7] and show those relations mathematically using a measure called *vote entropy*. Here, $\frac{V(l_i,t)}{k}$ is the ratio of k taggers that gave the tag l_i to a token t and $|s|$ is the sentence size.

$$D_{tok}(t) := -\frac{1}{\log k}\sum_{l_i}\frac{V(l_i,k)}{k}\log\frac{V(l_i,k)}{k} \tag{1}$$

$$D_{sent}(s) := \sum_{j=1}^{|s|}\frac{D_{tok}(t_j)}{|s|} \tag{2}$$

We created an *OpenNLP* [8] model for each training iteration and evaluated it on a gold standard that has been previously tagged, revised by domain specialists and isolated from the training corpus. We stored the measured accuracy and an average of $D_{sent}(s)$ indexes observed on the non-revised sentences in a relational database. We also stored the number of tokens that had been manually corrected for graphical analysis.

We ran the entire procedure three times, one for each methodological direction: (a) heterogeneous tagger committee, (b) homogeneous committee, and (c) homogeneous committee with optimal initial set. The use of different approaches aims at reducing errors due to a set of distinct taggers (approach b) and a non-optimal initial set of sentences (approach c).

In the first approach (a), we used a heterogeneous committee that combines a rule-based tagger (*Brill Tagger*) with four statistical taggers (*OpenNLP*, *MXPOST*, *TreeTagger* and *QTag*). Performance has been evaluated based on the model created for *OpenNLP*, which is the tool used in our other projects. We chose the initial collection of sentences randomly.

In the second approach (b), we applied a homogeneous committee composed of *OpenNLP* tagger trained on five different subsets of data. Nevertheless, we measured accuracy based on a model trained on all data available per iteration. The initial set was randomly chosen as in the first approach.

Finally, in the third approach (c) we also employed selection by homogeneous committee, but with an optimal non-real initial set. We chose the best sentences for training based on a model trained on data available on the entire corpus. The aim of this approach is to test the hypothesis that the initial set of sentences influences all subsequent iterations.

Results

Newspaper and Medical Corpora

Figure 1 shows *OpenNLP* accuracy evaluated on the medical gold standard according to the number of tokens used in the training process. The red and the blue lines express the results for the medical and the newspaper corpus used for training, respectively.

Active Learning Strategy

Similarly, Figure 2 shows *OpenNLP* accuracy for different training set sizes. Now, however, we compare active learning strategy (red line) with a simple strategy, like random selection (blue line). In both cases, the gold standard and the training corpus belong to the medical corpora.

Figure 3 shows the $D_{sent}(s)$ average observed in the subset of non-trained sentences for the heterogeneous committee-based approach. As in Figure 2, blue and red lines express, respectively, the active learning and random selection strategies. Flattening seen on the line beginning convey the need of a minimal amount of data to correctly train the taggers. Additionally, the abrupt dip in the end indicates process conclusion, when there were no more sentences available for training.

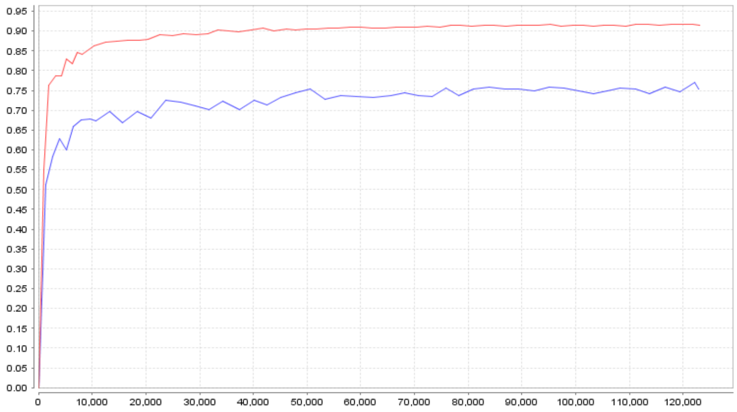

Figure 1 - Learning curves on the newspaper corpus (blue line) and on the medical corpus (red line), evaluated on medical data.

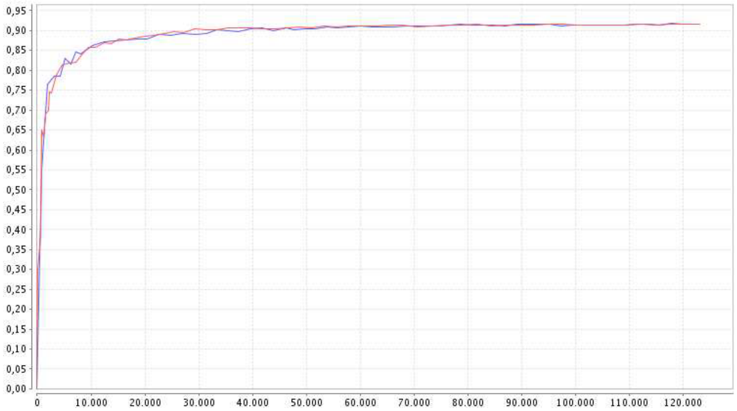

Figure 2 – Learning curves of active learning-based selection (red line) and random selection (blue line), trained and evaluated on medical data.

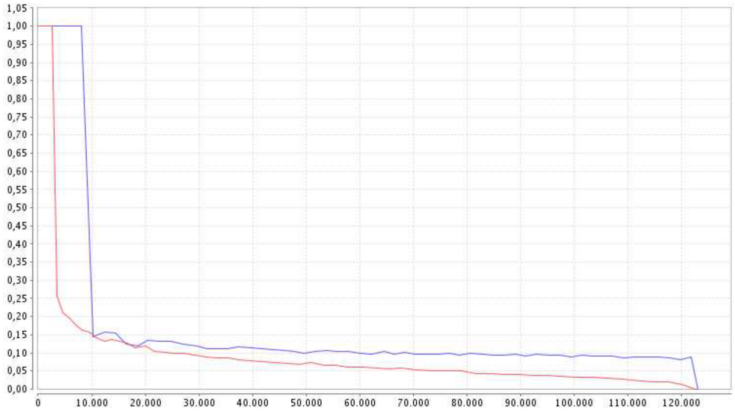

Figure 3 - Sentence disagreement index for active learning-based selection (red line) and random selection (blue line).

Discussion

When testing different approaches for active learning as proposed above (see "Materials and Methods"), we found similar results to those presented before. We then realized that active learning strategy should be adapted for POS tagging task in order to significantly reduce the manual tagging effort.

However, our results lack some statistical rigor. For example, a more precise result would have been found if we had run the random selection process at least three times. Nevertheless, our experience shows that data would have been similar and would not have affected conclusions presented here.

Despite that, a rigorous study would require the same initial set for random and active learning selections. The initial set influence was tested as described in the third approach presented above. We found that even an optimal initial set did not produce better active learning results.

Nonetheless, we could have tested other methodological approaches. The first one would be plotting learning curves as a function of number of sentences, not tokens. The second one would use 3/4 of the data available per iteration for the training process of each tagger, as proposed by [9]. Anyway, based on our experiments, we are certain that those variations would not considerably affect the conclusions of our work.

Conclusion

Our work confirm Campbell and Johnson [4] results: taggers trained on newspaper data cannot be readily used to tag medical data. At least in Portuguese, an *OpenNLP* model trained on a collection of 4,533 sentences (123,019 tokens), randomly chosen from the *MAC-Morpho* corpus of Brazilian newspaper, tags correctly only 75,3% of tokens from a collection of 595 sentences (12,451 tokens) arbitrarily chosen from a corpus of discharge summaries in Portuguese.

Comparing *OpenNLP* learning curves in the newspaper and in the medical corpus, we noted that newspaper data is more heterogeneous and more grammatically complex than medical data. Not only is the learning process in newspaper data slower and more sinuous than on medical data, but also the sentences are longer on that domain.

Considering that accuracy rates around 96.4% are expected [10], we encourage a domain-specific corpus to be build. However, our implementation of the active learning strategy for corpus construction [6] reached different results. We could not reduce manual correction effort by means of active learning when compared to a simple strategy like random selection.

Settles and Craven [11] reported similar results on a larger study with eight corpora from different domains. They realized that *vote entropy* is biased toward querying shorter sentences, which in our work showed as little or no value.

Our work showed additionally that manual tagging of around 30,000 tokens is enough to build a probabilistic POS tagger that accounts for an accuracy of 90% in the medical domain in Portuguese. Using a four times bigger sample, the tagger reached an accuracy index of 91.5%, which assures *OpenNLP*'s feasibility to annotate Brazilian hospital discharge summaries with POS tags.

Acknowledgments

We would like to thank Ana Carolina Peters, Claudia Seiko Yokoyama and Priscilla Koppe for their effort in corpus construction. We also thank Edson José Pacheco for his collaboration.

Our work is funded by the International Bureau of the German Ministry of Education and Research (BRA 07/022) and the Brazilian National Research Council (CNPq).

References

[1] Stenzhorn H, Pacheco EJ, Nohama, P, Schulz, S. Automatic mapping of clinical documentation to Snomed CT. In: Adlassnig KP, Blobel B, Mantas J, Masic I, editors. Proceedings of 22nd International Congress of the European Federation for Medical Informatics; 2009; Sarajevo, Bosnia and Herzegovina. Amsterdam: IOS Press; 2009. p. 228-232.

[2] Interinstitutional Center for Research and Development in Computational Linguistics, Lácio-Web Project [Online]. 2004 Jun 28 [cited 2009 Oct 14]; Available from: URL:http://www.nilc.icmc.usp.br/lacioweb.

[3] Hahn U, Wermter J. High-performance tagging on medical texts. In: Proceedings of the 20th International Conference on Computational Linguistics; 2004 Aug 23-27, Genebra, Switzerland. Morristown: Association for Computational Linguistics, 2004. p. 973.

[4] Campbell AD, Johnson SB. Comparing syntactic complexity in medical and non-medical corpora. In: Proceedings of the 5th American Medical Informatics Association Symposium; 2001; New York. p. 90-94.

[5] Lezius W, Rapp R, Wettler M. A morphology-system and part-of-speech tagger for German. In: Results of 3rd Konvens Conference; 1996; Berlim. Hawthorne: Mouton de Gruyter, 1996.

[6] Tomanek K; Wermter J, Hahn U. An approach to text corpus construction which cuts annotation costs and maintain reusability of annotated data. In: Proceedings of the Joint Conference on EMNLP-CoNLL; 2007; Praga. p. 486-495.

[7] Dagan I, Engelson, S. Committee-based sampling for training probabilistic classifiers. In: Proceedings of the International Conference on Machine Learning; 1995; Tahoe City. p. 150-157.

[8] OpenNLP [Online]. 2008 Nov 28 [cited 2009 Oct 14]. Available from: URL:http://opennlp.sourceforge.net.

[9] Reichart R, Tomanek K, Hahn U, Rappoport A. Multi-task active learning for linguistic annotations. In: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics; 2008. Columbus: HLT, 2008. p. 861-869.

[10]Morton T. Using semantic relations to improve information retrieval. Pennsylvania: University of Pennsylvania, 2005.

[11]Settles M, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2008, Honolulu. Morristown: ACL Press, 2008. p. 1069-1078

Address for correspondence

Michel Oleynik
Pontifical Catholic University of Parana
+55 41 3271-2446
michelole+medinfo2010@gmail.com