Named Entity Recognition for Clinical Portuguese Corpus with Conditional Random Fields and Semantic Groups

João Vitor Andrioli de Souza[1], Yohan Bonescki Gumiel[1], Lucas Emanuel Silva e Oliveira[1], Claudia Maria Cabral Moro[1]

[1]Programa de Pós-graduação em Tecnologia em Saúde – Pontifícia Universidade Católica do Paraná (PUCPR) - Curitiba, PR - Brazil

{joao.souza,yohan.bonescki}@pucpr.edu.br,{lucas.oliveira,claudia.moro}@pucpr.br

Abstract. *Considering the difficulties of extracting entities from Electronic Health Records (EHR) texts in Portuguese, we explore the Conditional Random Fields (CRF) algorithm to build a Named Entity Recognition (NER) system based on a corpus of clinical Portuguese data annotated by experts. We acquaint the challenges and methods to classify Abbreviations, Disorders, Procedures and Chemicals within the texts. By selecting a meaningful set of features, and parameters with the best performance the results demonstrate that the method is promising and may support other biomedical tasks, nonetheless, further experiments with more features, different architectures and sophisticated preprocessing steps are needed.*

1. Introduction

The Electronic Health Records (EHR) digitally store the patient's data, from personal information to care history, designed to improve the operational efficiency of the health services. Among the data found in the EHR, the clinical narratives are one of the most important ones, due to its rich source of information, supporting several application tasks and clinical computer research, such as: extraction of medical concepts, mapping of terminologies, decision support, ontologies construction and text summarization [Shickel et al. 2017]. The narratives are open texts written by healthcare professionals to describe details about the patients and their treatment.

The manipulation of clinical narratives by computational algorithms is a challenging task due to its lack of formal structure and organization, widespread use of acronyms and clinical jargons, grammatical errors and data redundancy. Thus, to automatically extract and identify entities in the middle of these texts, it is necessary to use Natural Language Processing (NLP) techniques, more specifically Named Entity Recognition (NER) algorithms. The NER can be classified into two categories, rule-based and machine-learning (ML) based. Rule-based approaches depend on a set of rules carefully created for a specific entity by a specialist on the subject, it is costly and not flexible in adapting to new entities, while ML approaches are easy to adapt and less costly to develop [Saha et al. 2015][Sebastiani 2002].

Most ML-based biomedical NER (Bio-NER) algorithms are supervised and require texts previously annotated by specialists with the entity to be extracted, an entity denotes an object found within a word or a set of words. Clinical narratives are an important source of entities and despite the dependence on laborious manual annotation

and manual resources [Oliveira et al. 2017], NLP and ML techniques are widely used to extract, identify and summarize EHR data.

In the literature, most of the NER studies are based on the English language, with rare studies related to the Portuguese language, especially in the health area, where to the best of our knowledge a Bio-NER system doesn't exist. This work aims to explore conditional random fields to fulfill the gap of a Bio-NER algorithm for the Portuguese language.

2. Materials and Methods

2.1 Semantically Annotated Clinical Corpus

A set of EHR data with 1.000 texts from three hospitals was used to generate a clinical corpus with semantic annotations, aiming to create a gold standard for algorithms to extract and identify clinical concepts in the clinical narratives. This set is composed of different types of clinical narratives (i.e., discharge summaries, ambulatory records and nursing notes) and multiple clinical specialties, mainly in Nephrology, Cardiology and Endocrinology areas.

The UMLS [Lindberg et al. 1993] is known to contain several clinical terminologies and coding standards, in addition, it provides interoperability between biomedical information systems, due to this, its semantic types[1] were used in the annotation of the texts, along with the "Negation" and "Abbreviation" entities. Each clinical concept (or entity) can be annotated with more than one semantic type at the same time. Figure 1 shows an excerpt of a text and their respective annotations.

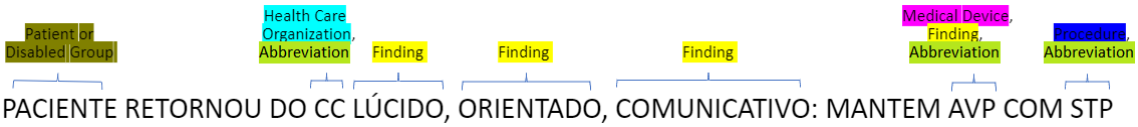

Figure 1. Excerpt of text from a clinical narrative and their entity annotations.

The annotation process took place with 8 annotators and 2 adjudicators. All annotators and adjudicators have experience in writing and interpreting clinical narratives within the hospital settings. The texts underwent a double-annotation process (each text was annotated by two different annotators), and then by the adjudication process in which the adjudicator resolved all divergences between the two annotators. More details on the tool used for annotation and the process itself can be found in [Oliveira et al. 2017]. The annotation process resulted in a gold standard composed of 64.549 entities, 12.955 unique tokens with 89 semantic types and 16 semantic groups[2].

2.2 Machine Learning Algorithms for Named Entity Recognition

A survey was conducted to check the state of the art in Bio-NER. Among the studies found, two ML algorithms stood out for their high adoption rate and good results in their evaluations: the Support Vector Machine (SVM) [Cortes and Vapnik 1995] and Conditional Random Fields (CRF) [Lafferty et al. 2001]. Recently, studies have applied

[1] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[2] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

Neural Networks approaches to process EHR data [Miotto, et al. 2016] [Jagannatha and Yu 2016], and in several cases, obtained better results than the SVM and CRF.

The NER algorithms, for the most part, adopt the IOB2 labeling model as input, where "B" represents the beginning of an entity, "I" represents the interior and "O" represents words outside the entity.

2.3 Experimental setup

The 1.000 annotated texts were divided into 9.624 sentences distributed in different amounts for training and testing, according to Figure 2.

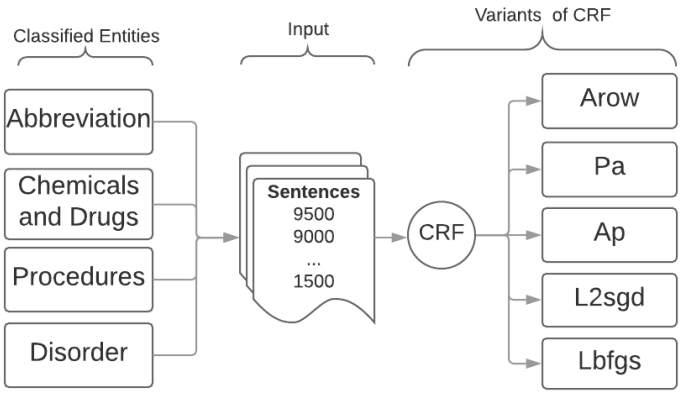

Figure 2. Training and testing steps with the entities, the different number of sentences and the variations of the CRF algorithm.

Due to the contrasting number of annotations of certain semantic types, it was chosen to group the semantic types to obtain a gold standard with lower granularity. Table 1 lists the semantic types and groups that participated in the experiments, as well as the respective number of annotations in the corpus for each of them.

Table 1. The semantic groups / types and the number of their annotations.

Semantic Group	Nº annotations	Semantic Type	Nº annotations
Disorder	27111	Finding	11.093
		Sign or Symptom	9.624
		Disease or Syndrome	4.203
		Others	2.191
Procedures	16338	Therapeutic or Prev. Procedure	9.624
		Health Care Activity	3.189
		Diagnostic Procedure	2.854
		Others	671
Chemicals and Drugs	5666	Pharmacologic Substance	3.008
		Organic Chemical	1.941
		Hormone	208
		Others	509
Abbreviation	7477	Abbreviation	7.477

Features used in NER algorithms, as in [Al-Hegami et al. 2017] were passed to the classifier, described in Table 2.

We have chosen to use the CRF algorithm because it currently has more adoptions in Bio-NER, and maintain various of the best results [Dingcheng et al. 2008]. Experiments were done with 5 variations of CRF (see Table 3 for results).

The performance of the methods was evaluated with a cross-validation model with 10 partitions, which infers the generalization capacity of the models. The values of Precision, Recall and F1-scores were measured.

Table 2. Features extracted from texts.

Nº	Meaning	Nº	Meaning
1	Word is lowercase	9	Word is uppercase
2	Word has more than 2 consecutive consonants	10	Maximum number of consecutive consonants
3	Word begins with a capital letter	11	Word is at the beginning of the sentence
4	All words in sentence are uppercase	12	All words in sentence are lowercase
5	Number of letters in the word	13	Word has Accent
6	Number of word vowels	14	Word only has numbers.
7	Word has no vowels	15	Part-of-speech Tag
8	Maximum number of consecutive vowels	16	Word is at the end of the sentence

Besides utilizing the IOB2 model for labeling, due to the discrepant numbers of words annotated with O the F1 scores shown in results was calculated with B and I only.

3. Partial Results

Previous experiments have shown a gradual worsening in the results and computational cost according to the increase of sentences used. Thus, subsequent experiments were performed using the optimal number of 3000 sentences, as it achieved the best results.

Table 3. CRF algorithms with Begin and Inside in abbreviation.

Algorithm	IOB2	Precision	Recall	F1	Average F1
Lbfgs	B_Abbreviation	0.73	0.61	0.67	0.52
	I_Abbreviation	0.39	0.14	0.21	
L2sgd	B_Abbreviation	0.75	0.62	0.68	0.53
	I_Abbreviation	0.42	0.14	0.21	
Ap	B_Abbreviation	0.75	0.59	0.66	0.49
	I_Abbreviation	0.52	0.07	0.12	
Pa	B_Abbreviation	0.72	0.66	0.69	0.56
	I_Abbreviation	0.44	0.22	0.29	
Arow	B_Abbreviation	0.60	0.55	0.57	0.48
	I_Abbreviation	0.30	0.24	0.27	

Table 4. Result of semantic groups using CRF.

Semantic Group	IOB2	Precision	Recall	F1	Average F1
Disorders	B_Disorders	0.73	0.68	0.70	0.65
	I_Disorders	0.66	0.53	0.59	
Procedures	B_Procedures	0.68	0.59	0.63	0.60
	I_Procedures	0.65	0.48	0.55	
Chemicals & Drugs	B_Chemicals & Drugs	0.86	0.39	0.54	0.42
	I_Chemicals & Drugs	0.25	0.03	0.05	

Table 3 presents the results of the different CRF algorithms classifying abbreviations with 3000 sentences, the best results for each IOB2 tag are in blue. The results for the semantic groups of Table 4 were obtained using passive aggressive CRF with 3000 sentences.

4. Discussion and future work

During the experiments with CRF algorithm variants the best results in F1 and Recall were always seen with the passive aggressive method, however, the average perceptron method had superior Precision results.

Disorder was the group with the best result. Although Chemicals and Drugs achieved the best Accuracy with Begin, this group had the worst results. Among the factors that may have led to a decreased performance are: annotation quality, granularity and specificity of selected semantic types and groups, features that did not sufficiently cover the entity characteristics, the classifiers used, specificities of Portuguese texts, or even for using texts from different institutions, types and medical specialties. In spite of the use of features and parameters similar to other Bio-NER's works that deal with English texts, we have some differences in the results, comparing to works as [Dingcheng et al. 2008] and [Abacha and Zweigenbaum 2011] that obtained 0.86 and 0.77 respectively of F1 in Disorder and [Denecke 2014] that obtained 0.59 and 0.69 of F1 in Disease and Procedure.

The difficulties seen during this work were mainly on finding a set of features that could represent the Portuguese language and the semantic groups. Features used on other languages and types of texts do not guarantee to have the same results on clinical Portuguese texts, an important aspect of the semantic group is that they have more domain knowledge grouped, making words with different characteristics being in the same entity.

For future work, we will focus on different pre-processing techniques, update algorithm's parameters and features, and use Genetic Algorithms to find a good set of features to train the CRF for each entity. Furthermore, we will perform hybrid experiments with rules and unsupervised approaches, use different ML architectures (such as the use of Neural Networks), and incorporate different types of key features from different models and architectures [Yadav and Bethard 2018].

5. Conclusion

In this work, we explore the CRF algorithm for Bio-NER in Portuguese texts. Preliminary results demonstrate that by selecting a set of features that covers and represents the characteristics of the entities, and the right CRF parameters, the method may support a variety of biomedical application tasks. Nonetheless, additional studies are needed to achieve results comparable to other state-of-the-art methods, especially in the refinement of features, improvement in preprocessing, and use of other ML classifiers and architectures such as Neural Networks.

References

Abacha A.B., Zweigenbaum P. (2011). Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. Workshop on Biomedical Natural Language Processing, 56-64.

Al-Hegami A. S., Othman A.M.F., Bagash F.T.. (2017). A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set. IJCSNS International Journal of Computer Science and Network Security, Vol.17.1. 170-176.

Cortes C., Vapnik V. (1995). Support-Vector Networks. Kluwer Academic Publishers, Boston. Manufactured in the Netherlands. Machine Learning, 20, 273-297.

Denecke K. (2014). Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study.

Dingcheng Li, Kipper-Schuler K., Savova G.. (2008). Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. Current Trends in Biomedical Natural Language Processing. 94–95.

Jagannatha, A.N., Yu, H. (2016). Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc. Conf. Assoc. Comput. Linguist. North Am. Chapter. Meet. 473–482.

Lafferty J. D., McCallum A., Perreira F.C.N. (2001). Conditional Random Fields: Boston. Probabilistic Models for Segmenting and Labeling Sequence Data. Department of Computer & Information Science. 282-289.

Lindberg D.A.B., Humphreys B.L., McCray A.T.. (1993). The Unified Medical Language System. Methods Inf Med. 32:281–91.

Miotto, R., Li, L., Kidd, B.A., Dudley, J.T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci. Rep. 6, 1–10.

Oliveira, L. E. S., Gebeluca, C. P., Silva, A. M. P., Moro, C. M. C., Hasan, S. A., Farri, O. (2017). A Statistics and UMLS-based Tool for Assisted Semantic Annotation of Brazilian Clinical Documents. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 1072-1078.

Saha, S., Ekbal, A., Sikdar, U. K. (2015). Named entity recognition and classification in biomedical text using classifier ensemble. International Journal of Data Mining and Bioinformatics, 11(4), 365. doi:10.1504/ijdmb.2015.067954.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol.34.1. 1–47.

Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR. (2017). A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J. Biomed. Heal. Informatics. 1–1.

Yadav V., Bethard S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. Proceedings of the 27th International Conference on Computational Linguistics. 2145-2158.