Recognizing pharmacovigilance named entities in Brazilian Portuguese with CoreNLP

Alexandre M. da Cunha[1], **Kele T. Belloze**[1], **Gustavo P. Guedes**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ
Rio de Janeiro – RJ – Brasil

alexandre.cunha@eic.cefet-rj.br,
{kele.belloze, gustavo.guedes}@cefet-rj.br

Abstract. *Textual data sources may assist in the detection of adverse events not predicted for a particular drug. However, given the amount of information available in several sources, it is reasonable to adopt a computational approach to analyze these sources to search for adverse events. In this scenario, we created an extension of CoreNLP to process Brazilian Portuguese texts from pharmacovigilance area. We trained three natural language models: a Part-of-speech tagger, a parser and a Named Entity Recognizer. Preliminary results indicate success in generating a dependency tree for phrases in the pharmacovigilance area and in identifying pharmacovigilance named entities.*

1. Introduction

Textual data sources are an important resource for extracting useful information about a given domain. In the area of pharmacovigilance, much of this information exists in textual sources such as notifications from the National Agency for Sanitary Surveillance (ANVISA), scientific articles, package leaflets and social media publications. Such sources are important because they may assist in the detection of signs of adverse events not predicted for a particular drug. However, to extract the information automatically and try to establish correlations between them, it is necessary to identify the entities of the different domains present in the texts (e.g., medicine, symptom, adverse event). In this way, a computational approach, such as natural language processing, must be adopted.

The Natural Language Processing (NLP) is an area of computer science directed to the interaction between computers and human language, in particular, refers to how to program computers to understand and manipulate natural language [Chowdhury 2003]. Some techniques may be used to extract information from texts in natural language, among them, tokenization, sentence splitting and identification of named entities. There are some tools to help the NLP tasks, among them, the literature highlights the CoreNLP which is highly cited (more than 2,000 citations). CoreNLP [Manning et al. 2014] is a tool capable of extracting the word radical, marking sentence structure, and identifying some entities such as person, location, organization, date, and numerical data.

In this scenario, the objective of the present work is to extend CoreNLP to process Brazilian Portuguese texts in the pharmacovigilance area, identifying the named entities and producing the dependency tree of the sentences. To achieve this goal, we produced three processing models: a Part-of-speech tagger, a Parser and a Named Entity Recog-

nizer. The models were trained based on data sets from Universal Dependencies[1](UD), CID-10[2] and BULARIO[3].

2. Related Work

In the last years, the NLP for text analysis from different pharmacovigilance sources has been gained the focus of some works in the computing area. The work of [Benton et al. 2011] analyze clinical narrative in electronic medical records and adopt association statistics to raise potential adverse drug events. Medical case reports are analyzed by [Gurulingappa et al. 2012] and data extracted from Twitter by [Nikfarjam et al. 2015]. Both works make use of a machine learning-based system for identifying the relations between drugs and adverse events. The text analysis from social media like Twitter are also observed in the works of [O'Connor et al. 2014] and [Cocos et al. 2017]. The first one uses Cohen's kappa to calculate the inter-annotator agreement between the drug and the adverse event. The second one adopts a deep learning approach to detect adverse drug reaction. All of these works employ controlled vocabularies such as Unified Medical Language System (UMLS)[4] and its sources for corpus annotation and the named entity recognition system. To the best of our knowledge, there are no works for the same purpose as the above-mentioned works in which the adopt language is Brazilian Portuguese. However, controlled vocabularies such as CID-10 and data sets of domain-specific terms and concepts such as BULARIO are presented in Brazilian Portuguese and can be exploited for the analysis of texts in the Brazilian pharmacovigilance area, in order to support pharmacovigilance actions in the country.

3. CoreNLP

CoreNLP [Manning et al. 2014] is a set of human language technology tools, structured in a Java annotation pipeline format, that provides most common processes for language processing in natural format. Figure 1 shows an overview of CoreNLP architecture.

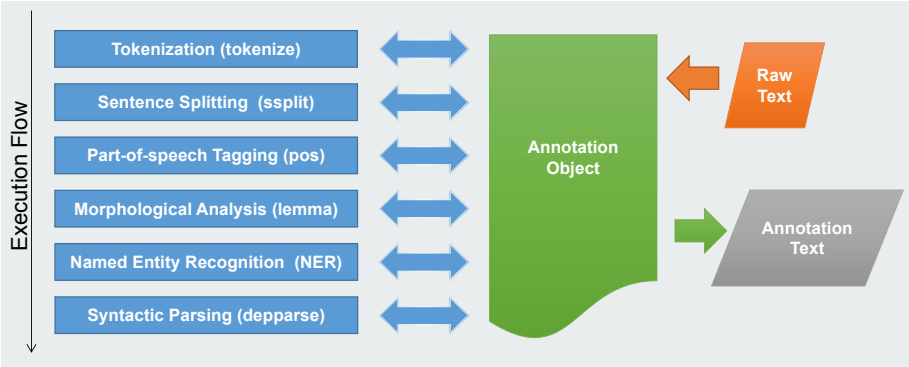

Figure 1. Overview of CoreNLP architecture, adapted from [Manning et al. 2014]

The CoreNLP architecture has six initial phases: `Tokenization`, splits the sentence into the tokens (words); `Sentence Splitting`, transform a sequence of tokens

[1]http://universaldependencies.org/
[2]http://www.datasus.gov.br/cid10/V2008/cid10.htm
[3]http://www.bulario.com
[4]https://www.nlm.nih.gov/research/umls/

in sentences; `Part-of-speech (POS) Tagging`, assigns a part of speech to each word based on a manually annotated lexicon; `Morphological Analysis`, identify the stem of a word; `Named Entity Recognition`, consists of identifying tokens that correspond to named entities; `syntactic Parsing`, generates the syntactic tree of a sentence [Manning et al. 2014]. Unfortunately, there is no native support for the Brazilian Portuguese language in CoreNLP. However we can train our models for execution.

4. Methodology

In this work we created an extension of CoreNLP to process Brazilian Portuguese. So that, we trained three natural language models: a POS Tagger (*pos*), a Parser (*depparse*) and a NER model (*NER*).

POS tagger training (*pos*). In the first step to create an extension of CoreNLP for Brazilian Portuguese, we trained a Portuguese POS tagger based on CoreNLP MaxentTagger class. Also, to accomplish this task, we used the Brazilian Portuguese Universal Dependencies[5]. The result is a Brazilian Portuguese tagger model.

Parser training (*depparse*). The first step to create a Brazilian Portuguese parser, is to configure it with a POS tagger. Then, we set it with the tagger built above. In the training phase, it is strongly recommended to use word embeddings to improve results. Thus, we used a 50 dimensions pretrained word embeddings distributed by NILC [Hartmann et al. 2017]. After that, we configure a DependencyParser to create our model.

Pharmacovigilance NER training (*ner*). To create our Portuguese NER model, we trained a linear chain conditional random field (CRF) to recognize pharmacovigilance entities. The CRF (provided by CoreNLP) was trained with three datasets created by our research group: (i) the first one contains $6,701$ commercial drug names and $1,778$ names of active substances (drugs), extracted from package leaflets provided by the website bulario.com; (ii) the third is a public dataset, named CID-10, which contains several disease names. Since CID-10 contains terms that are not diseases, we select a subset of CID-10, related to infectious or parasitic diseases, neoplasms (tumors), mental and behavioral disorders and diseases of the digestive system. We named this dataset CID-10-Diseases. All the three datasets are available in https://github.com/LaCAfe/AEventPT-br.

5. Preliminary results

We tested our trained CoreNLP with the phrase: "Ela teve cefaleia tensional e tomou Maxalt, mas teve ataxia.". The result is shown in Figure 2. It is possible to note that all pharmacovigilance entities were detected. Furthermore, the dependency tree was successfully generated for Brazilian Portuguese.

6. Final Remarks

The test carried out in the trained Brazilian Portuguese CoreNLP model showed that the entities were adequately recognized and the dependency tree were correctly produced, indicating the feasibility of the methodology. As next steps of this work, we can highlight the adoption of a corpus to improve the named entity recognizer model. To achieve this

[5]`https://github.com/UniversalDependencies/UD_Portuguese-GSD/tree/dev`

```
Print: Word: [Ela] Tagger: [PRON] Entity: [O]
Print: Word: [teve] Tagger: [VERB] Entity: [O]
Print: Word: [cefaleia] Tagger: [NOUN] Entity: [DISEASE]
Print: Word: [tensional] Tagger: [ADJ] Entity: [DISEASE]
Print: Word: [e] Tagger: [CONJ] Entity: [O]
Print: Word: [tomou] Tagger: [VERB] Entity: [O]
Print: Word: [Maxalt] Tagger: [PNOUN] Entity: [DRUG]
Print: Word: [,] Tagger: [.] Entity: [O]
Print: Word: [mas] Tagger: [CONJ] Entity: [O]
Print: Word: [teve] Tagger: [VERB] Entity: [O]
Print: Word: [ataxia] Tagger: [NOUN] Entity: [DISEASE]
Print: Word: [.] Tagger: [.] Entity: [O]
```

Figure 2. Named Entity Recognition

goal, we intend to use specific adverse event data sets. Also, we will conduct a survey on controlled vocabularies in the area of pharmacovigilance in Brazilian Portuguese.

References

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C. E., and Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Cocos, A., Fiks, A. G., and Masino, A. J. (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Gurulingappa, H., Mateen-Rajpu, A., and Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., and Gonzalez, G. (2014). Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.