An Ontology Based Natural Language Processing Pipeline for Brazilian COVID-19 EMR

Raquel A. J. Gritz[1], Rafael S. Pereira[1], Henrique Matheus F. da Silva[1], Henrique G. Zatti[2], Laura E. A. Viana[2], Karol C. S. F. Navarro[5], Thalita R. Dias[3], Viviane S. B. Oliveira[4], Ricardo A. Souza[2], Vinícius A. Oliveira[4], Manoel Barral Netto[4], Fabio Porto[1]

[1]DEXL – Laboratório Nacional de Computação Científica (LNCC)
Av. Getúlio Vargas, 333, Petrópolis, RJ, Brasil

[2]Faculdade de Medicina da Universidade Federal de Minas Gerais (UFMG)
Av. Prof. Alfredo Balena, 190, Belo Horizonte, MG, Brasil

[3]Hospital Universitário de Brasília (UNB) - SGAN 605, Asa Norte, Brasília, Brasil

[4]Fundação Oswaldo Cruz (Fiocruz) – Instituto Gonçalo Moniz (IGM)
Rua Waldemar Falcão, 121, Salvador, BA, Brasil

[5]Universidade Federal de Minas Gerais (EEFFTO - UFMG) - Av. Pres. Antônio Carlos, 6627, Belo Horizonte, MG, Brasil

{rgritz,rpereira,matheusf,fporto}@lncc.br,

{hgzatti,lauraelisaviana,to.salomaonavarr, thalitadyas,ric.alex}@gmail.com,

{hviviane.boaventura,vinicius.oliveira manoel.barral}@fiocruz.br

Abstract. *COVID-19 became a pandemic infecting more than 100 million people across the world and has been going on for over a year. A huge amount of data has been produced as electronic medical records in the form of textual data because of patient visits. Extracting this information may be very useful in better understanding the COVID-19 disease. However, challenges exist in interpreting the medical records typed as free text as doctors may use different terms to type in their observations. In order to deal with the latter, we created an ontology in Portuguese to describe the terms used in COVID-19 medical records in Brazil. In this paper, we present a brief overview of the ontology and how we are using it as the first step of a more complex NLP task.*

1. Introduction

COVID-19 become a pandemic infecting more than 100 million people, at the time of writing. The creation of health care strategies in the face of this pandemic is a major challenge. When considering the local limitations and the lack of global systematization for the care of patients with different presentations of the disease, both in light and moderate profile, this challenge is even greater.

As a result of the interactions between doctors and patients in the context of COVID-19 treatment, an increasing amount of rich textual data has become available in the form of electronic medical records (EMR). However, automatically processing this type of data is hard, as each doctor, health institution, or medical specialization records differently in text the log of patients visits.

Nevertheless, exploring this kind of data can be very useful in different aspects related to the fight against the pandemic, such as: in helping to extract knowledge about the disease; in identifying conditions that lead to disease progression; in defining the best strategies for treatment, among many more interesting queries.

Such data can also be integrated into machine learning systems helping to answer these questions by highlighting terms of interest to help a healthcare professional, or even by classifying the EMR text into whether the patient would progress to a more severe state or if she will have mild complications.

However, dealing with such data brings many challenges because of the complexity of the text, as well as its high variability. Thus, the creation of instruments that assist in the description and systematization of updated knowledge about the disease, such as an ontology, was essential in the context of the project, provide different tools to face this pandemic. From this complex characterization of COVID-19, it will be possible to create alerts for patients with the potential to present forms of the disease with unfavorable evolution, which, in the scenario of Brazil, may contribute to the reduction of morbidity and mortality.

Given this in this paper we present an approach to process EMR, using Natural Language Processing (NLP) and an ontology in order to extract meaningful information. With these techniques, we can therefore identify characteristics that may or may not define the severity of a disease based on information contained in the EMR and assist in medical decisions based on the incidence of terms, using machine learning techniques.

The rest of this paper is organized as follows: in section 2, we discuss the use of NLP in EMR. Section 3.1 presents the proposed ontology and how it can be used to improve EMR text processing. Section 4 presents the methodology of our experiment and the obtained results. Finally, we conclude on section 5.

2. Preliminaries: Natural Language Processing in medical data

NLP is a field where text data is processed via algorithms in order to extract information. Medical data can be expressed in textual format in EMR which contains information about the patient. While there are many advances in NLP [Hirschberg and Manning 2015], most require significant amounts of annotated data.

This becomes harder in the medical field given privacy concerns which limits the availability of larger datasets. Because of this many works that apply NLP in medical data build their own datasets which are not shared [Chen et al. 2018] [Cai et al. 2019].

Also, most works in the medical field limit their data to a single hospital or even a single department in order to minimize variance on how the same expressions can be reflected in the text.

In contrast, in our work, we aim to process and integrate data that come from multiple sources like outpatient clinics, health centers, general hospitals, intensive care unit (ICU), among others, since the data we have comes from several health units spread over some cities in Brazil, containing information that differs in the writing pattern of the EMRs among themselves. Given this, in this paper, we explore NLP techniques on a sample of medical records and present how through the use of our ontology we can extract meaningful results from the data even facing all the aforementioned challenges.

3. Methodology

In this section, we describe a Portuguese language ontology-based NLP pipeline for COVID-19 EMR. The Data Source is the Electronic Health Record of the Brazilian Company of Hospital Services[1] [EBSERH]. Data was provided to the research team anonymized, thus not implying risks of leakage of personal data.

3.1. Brazilian COVID-19 ontology for NLP

Given the complexity in processing the textual data appearing in EMR, we propose to integrate into the NLP pipeline an ontology for COVID-19 in the Portuguese language. The ontology shall help in dealing with the ambiguities produced during text interpretation, as it can lemmatize all different forms of a concept to a base form using synonyms, and we can leverage the relationships between different terms. To the best of our knowledge, we are the first group to design a Portuguese medical ontology for COVID-19 that will be available in the public domain.

3.1.1. A Portuguese COVID-19 Ontology

The terminology used in health is technical, however it considers subjectivity, different and various clinical information, and complementary exams. Beyond that, it is broad for various conditions of illness and conditions, helping forming hypotheses and diagnoses of probability. To represent this medical language, we use an ontology.

Ontology is a branch of philosophy that aims to study and understand the concepts of being, existence, and reality. As [Gruber 1993] originally defines, "An ontology is an explicit specification of a conceptualization. For knowledge-based systems, what "exists" is exactly that which can be represented". Thus, through ontology, if something exists it is possible to portray its nature and existence, its structure, its principles, its components, and its relations.

In this context, an ontology in the medical field seeks to describe the information concerning the problem of interest in the medical domain, thus structuring the knowledge about this domain.

In this work, we aimed to build, through a joint effort between computer scientists and a group of medical studentes and doctors in medicine in different areas of expertise, an ontology that sought to formally describe relevant information contained in COVID-19 EMR. The creation of COVID-19 ontology was a challenging exercise, requiring a work of reflexive thought and collective construction. Thus, we seek to understand the concepts for the creation of an ontology, as well as concepts and characteristics of COVID-19, and implement the techniques used for the creation of ontologies[Noy and McGuinness] [Farinelli and Almeida 2019].

In this context, we started studyng for medical ontologies from other domains [Smith et al. 2007] [Schriml et al. 2019]. Then, we studied the characteristics and aspects that are part of disease [McIntosh et al. 2020] [de Andrade 2013] we analyzed the EMRs, and identified the classes that structured the main concepts in the domain: defining its

[1] Available in http://www2.ebserh.gov.br/web/portal-ebserh/inicio

terms, hierarchies, and relations. We use *Protégé*[2] [Musen 2015], a tool developed by Stanford University School of Medicine for the creation and visualization of ontologies. Figure 1 shows a structure of part of the ontology developed in this project.

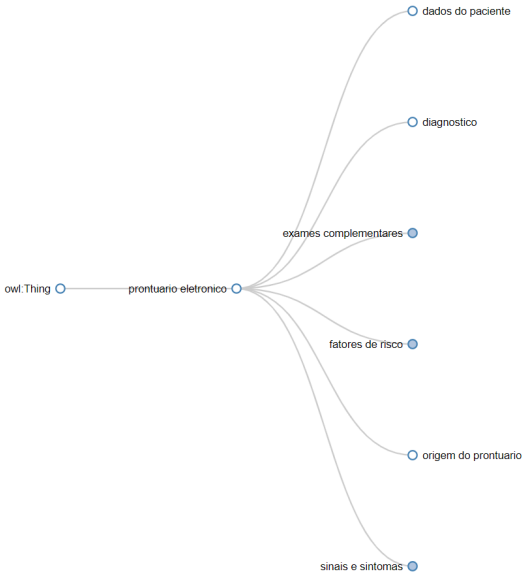

Figure 1. Example for a snippet of the Portuguese COVID-19 ontology

When creating the ontology, one of the challenges faced was to describe a disease almost unknown in the current moment of social and scientific commotion, in which science focuses on the complexity of COVID-19 to generate new learning. Changes in their understanding required us to keep up to date at all times while working on deciphering this issue.

3.1.2. Ontology processing: Owlready

To use the Portuguese COVID-19 ontology as part of the NLP pipeline, we needed a tool with a programmatic API to reason on its classes and answer some inference queries. We adopted the Owlready tool [Lamy 2017]. With Owlready, we access the ontology created in *Protégé* to check whether a term, extracted from the EMR, exists as a class, returning its target synonym and the superclass, or NULL in case the term is not a class in the ontology. We assume that each term should be a specialization of one of the main concepts in the ontology. For example: *cefaleia* is a specialization of the main concept *sintoma*. In case a term shares different synonymous, the ontology point to the one that is considered the *target* synonyms and this is the one assumed for the text interpretation.

Figure 2 presents an example of searching three medical terms: *cefaleia*, *RT_PCR_nao_reagente*, and *bisturi*. The *cefaleia* query returned as its target synonym, *cefaleia* and its superclass was *sinais_e_sintomas*.When querying *RT_PCR_nao_reagente* we obtained *RT_PCR_negativo* as the synonym target and its superclass was *exame_complementar*. In the last query the term *bisturi* does not exist in the COVID-19 ontology, in this case we return NULL.

[2]Available in `https://protege.stanford.edu/`

```
a=consulta_de_termo_na_ontologia("cefaleia",covid)
b=consulta_de_termo_na_ontologia("RT_PCR_nao_reagente",covid)
c=consulta_de_termo_na_ontologia("bisturi",covid)
print(a)
print(b)
print(c)
```

```
['cefaleia', 'sinais_e_sintomas']
['RT_PCR_negativo', 'exame_complementar']
['NULL']
```

Figure 2. Query existing terms in COVID-19 ontology

At the current moment our ontology is being used via owlready in order tô extract from the EMRS only the terms contained in our ontology in order to apply nlp techniques.

3.2. Global project overview

We aim to later extend our ontology by training a named entity recognition model [Li et al. 2020] to classify the words that we predefined into our ontology into their main class, for example, 'symptoms', 'complementary exams', etc. Then this model will be later used on new texts to try to extract other possible terms that should be contained in our ontology. This is but the beginning of our workflow for this project, which is seen in figure 3.

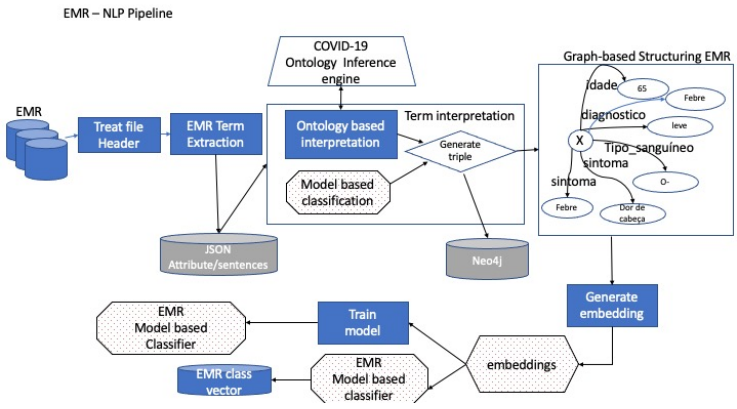

Figure 3. Complete NLP pipeline of the project. This paper presents the EMR term extraction module including the ontology inference engine

4. Experiment

As an initial evaluation of the outcome of the proposed NLP pipeline supported by the Portuguese COVID-19 ontology, we explore how the elements of the ontology interact with one another in actual medical text data. To do so we tokenize the text into sentences. Then, for each sentence, we filter the terms that appear in the ontology. Next, from the set of filtered terms at each sentence, we create a term-document matrix.

To build the adjacency matrix, we take the cross product of the term-document matrix. This is summarized in the algorithm 1.

Algorithm 1 Obtains the graph's adjacency matrix

$Text \leftarrow PreProcess(Text)$
$Text \leftarrow Tokenize(Text,' sentences')$
$Text \leftarrow Filter(Text, Ontology)$
$TDM \leftarrow TermDocumentMatrix(Text)$
$Text \leftarrow AdjacencyMatrix = CrossProduct(TDM)$
$Return\ AdjacencyMatrix$

4.1. Results

This work sought to explore how the classes defined in the ontology of COVID-19 are related to each other based on the words that appear in textual data sets. The data set used in this work, seeking to verify the functionality of the ontology in NLP, was the Tele Coronavirus, which has descriptive information of signs, symptoms, risk classification reported by people who sought telephone assistance for presenting symptoms they identified as related to COVID-19. As a result of this experiment, we obtained a graph composed of 514 nodes.

To find the graph we calculated a co-occurrence matrix of words according to the methodology already described to define the graph adjacency matrix. A co-occurrence matrix shows that if two words are highly connected in a graph it means they usually appear together according to how we tokenized the text (via sentences), and groups in this graph can be explored to find some common theme.

Figure 4 represents a subgraph, part of the graph of 514 vertices, containing the classes with the greatest relations between them. When observing the graph nodes, a strong relationship can be identified between the terms of the ontology that define classes of signs and symptoms, classes of exams requested by doctors, described in the medical records, and also classes that make up the group of risk factors for COVID-19.

In this way, we were able to demonstrate that the information contained in the ontology, determining whether the information in the EMRs can relate to the characteristics of COVID-19 through the use of ontology. where, we can identify a greater chance or not of disease severity and make decisions about the diagnosis of a particular patient.

5. Conclusion

In this paper, we present the importance and functionality of an ontology in helping solutions that integrate NLP pipelines, as the medical and computational areas are able to converse using ontology to identify the relevance or not of certain related terms in an EMR. Our first results, present how we could find groups using a concurrency matrix of terms in free text when using our ontology.

Which will be useful later when building the final project denoted in section 3.2 to be able to process EMR from all over the country and help with the COVID-19 pandemic. In the next steps, we will conclude the pipeline shown in the figure 3, applying machine learning algorithms to the preprocessing described in this article, in order to assist medical decisions.

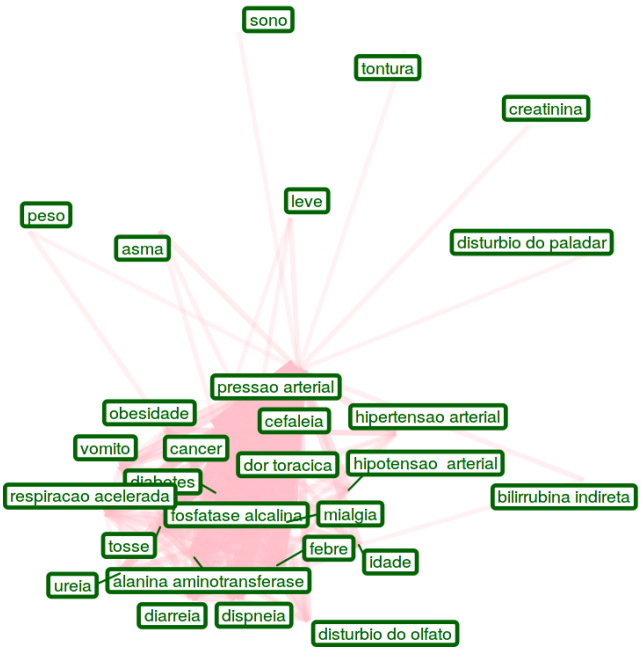

Figure 4. Subgraph visualization of terms of our ontology extracted from medical texts obtained from the project containing only the top 30% strongest links.

6. Acknowledgements

We thank JBL SA Program "Fazer o bem faz o bem" for sponsoring this research. Fiocruz for the institutional support and coordination of the parent research project. The Dataset was provided by Ebserh as a institutional participant of the project, we specially thank for the support of the Information Technology Director, Simone Scholze. We also thank Tales Mota, who anonimized the data but didn't participate on this branch of the research.

References

Cai, X., Dong, S., and Hu, J. (2019). A deep learning model incorporating part of speech and self-matching attention for named entity recognition of chinese electronic medical records. *BMC medical informatics and decision making*, 19(2):101–109.

Chen, Q., Du, J., Kim, S., Wilbur, W. J., and Lu, Z. (2018). Combining rich features and deep learning for finding similar sentences in electronic medical records. *Proceedings of the BioCreative/OHNLP Challenge*, pages 5–8.

de Andrade, A. Q. (2013). A linguagem médica utilizada em prontuários e suas representações em sistemas de informação: as ontologias e os modelos de informação.

EBSERH. Brasil. Ministério da Educação. Empresa Brasileira de Serviços Hospitalares.

Farinelli, F. and Almeida, M. B. (2019). Ontologias biomédicas: teoria e prática. *Sociedade Brasileira de Computação*.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Lamy, J.-B. (2017). Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, 80:11–28.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge Data Engineering*, (01):1–1.

McIntosh, K., Hirsch, M., and Bloom, A. (2020). Coronavirus disease 2019 (covid-19): Epidemiology, virology, and prevention. *Lancet. Infect. Dis*, 1:2019–2020.

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12.

Noy, N. F. and McGuinness, D. L. A guide to creating your first ontology 2001. *URL:[http://www. protege. stanford. edu/publications/ontologydevelopment/ontology101. html2001]*.

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2019). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.