Desafios e Dificuldades na Extração de Entidades Nomeadas de Notas Clínicas de Oncologia

Challenges and Issues on Extracting Named Entities from Oncology Clinical Notes

Desafíos y Problemas en la Extracción de Entidades Nombradas de las Notas Clínicas de Oncología

Luiz Henrique Pereira Niero[1,2], João Vitor Andrioli de Souza[1,3], Luciana Martins Gomes da Silva[1], Yohan Bonescki Gumiel[5], Nícolas Henrique Borges[1], Gustavo Henrique Munhoz Piotto[1,4,] Gustavo Giavarini[1], Lucas Emanuel Silva e Oliveira[1,5]

[1] Comsentimento NLP Lab, São Paulo, Brasil
[2] Paulista State University "Júlio de Mesquita Filho" - UNESP, Rio Claro (SP), Brasil.
[3] Pontifical Catholic University of Paraná - PUCPR, Curitiba (PR), Brasil.
[4] DASA oncology, Brasil.
[5] Biomedical Informatics Laboratory - Instituto do Coração - HC FMUSP

Autor correspondente: Luiz Henrique Pereira Niero
E-mail: luiz.niero@unesp.br

Resumo

Este artigo tem como objetivo descrever o processo de anotação de um corpus multi institucional de textos clínicos da especialidade de oncologia e treinar modelos para o Reconhecimento de Entidade Nomeadas. Utilizamos o corpus anotado para treinar modelos com diferentes quantidades de dados e comparar o resultado do modelo com a quantidade de dados utilizados no treinamento. O treinamento dos modelos foi feito a partir do *fine-tuning* do *Bidirectional Encoder Representations from Transformers* adaptado ao domínio médico-biológico da língua portuguesa (BioBERTpt). Para comparar o comportamento do modelo com o aumento dos dados de treinamento, os modelos foram treinados com quantidades incrementais de dados. Como resultado, obtivemos que os modelos treinados com conjuntos de dados menores porém totalmente revisados tiveram melhor resultado que modelos treinados com conjuntos de dados maiores com pouca revisão.

Descritores: Processamento de Linguagem Natural; Registros Eletrônicos de Saúde; Oncologia.

Abstract

This article aims to describe the annotation process of a multi-institutional corpus of clinical texts in the oncology specialty and to train models for the Recognition of Named Entities. We use the annotated corpus to train models with different amounts of data and compare the model result with the amount of data used in training. The training of the models was done from the fine-tuning of the Bidirectional Encoder Representations from Transformers adapted to the medical-biological domain of the Portuguese language (BioBERTpt). To compare model behavior with increasing training data, models were trained with incremental amounts of data. As a result, we found that models trained with smaller but fully revised datasets performed better than models trained with larger datasets with little revision.

Keywords: Natural Language Processing; Electronic Health Records; Medical Oncology.

Resumen

Este artículo tiene como objetivo describir el proceso de anotación de un corpus multiinstitucional de textos clínicos en la especialidad de oncología y entrenar modelos para el Reconocimiento de Entidades Nombradas. Usamos el corpus anotado para entrenar modelos con diferentes cantidades de datos y comparamos el resultado del modelo con la cantidad de datos utilizados en el entrenamiento. El entrenamiento de los modelos se hizo a partir de la puesta a punto de las Representaciones de Codificadores Bidireccionales de Transformadores adaptados al dominio médico-biológico de la lengua portuguesa (BioBERTpt). Para comparar el comportamiento del modelo con el aumento de los datos de entrenamiento, los modelos se entrenaron con cantidades incrementales de datos. Como resultado, encontramos que los modelos entrenados con conjuntos de datos más pequeños pero completamente revisados funcionaron mejor que los modelos entrenados con conjuntos de datos más grandes con poca revisión..

Descriptores: Procesamiento de Lenguaje Natural; Registros Electrónicos de Salud; Oncología Médica.

Introduction

Electronic health records (EHR) are an important source of data for patient care and for countless other applications that involve data analysis, decision support, or even the creation of predictive models. Much of this data is stored in an unstructured way, as free texts (i.e., discharge summaries, ambulatory notes), thus, incomprehensible to machines without some processing.[1,2] Hence, using natural language processing (NLP) is essential to transform this unstructured data into a structured format that can be understood by machines and used in other pipelines.[3]

Named entity recognition (NER) is an important NLP task, as it automatically identifies named entities of interest in the text. For the clinical domain, named entities can be drugs, adverse events, disorders, and risk factors. Further, NER is often the first step in several more complex NLP applications, such as timeline creation and data summarization.

The oncology domain is a topic of interest among the possible applications of NER in the clinical field.[1] For instance, according to the World Health Organization (WHO)[4], cancer is a leading cause of death worldwide, responsible for almost 10 million deaths in 2020, representing nearly one in every six deaths. For that reason, it is frequently subject of scientific research, drug development and discovery, and innovation.

In Oncology, for each type of cancer, different criteria are used for staging and clinical management. As in all technical fields, specific terminologies are used in biomedical sciences, and among those, different terms and abbreviations are used in the variety of existing medical subspecialties. Scientists and healthcare professionals use specific terms related to diagnostics, procedures, clinical conditions, substances, medications, among others. It is also worth noting that in different settings, cultural, regional, geographical, and even institutional differences can determine how the same term is referred to in different ways. For example, sometimes in one hospital, a disease is referred to using an acronym in English, and in another hospital, the same condition is known by its acronym in Portuguese. Therefore, developing tools that identify and correlate those variations in terminology could greatly enhance the ability of pooling and correlating data from different settings, which in turn can have great impact in improving the general quality of care or the development of new treatments.

Objectives

This work contained four main objectives: (1) to annotate the first corpus of clinical notes of oncology in Portuguese; (2) to describe the process of annotation and training NER models; (3) to point out some of the challenges and difficulties encountered during the annotation process; (4) to create and compare the behavior of trained token prediction models with multi-institucional corpus of different sizes, generated incrementally.

Contributions

It is a novel study regarding all the techniques involving the extraction of named entities from clinical texts in oncology written in Portuguese. Some studies also addressed clinical texts written in Portuguese: three groups worked with the extraction of Unified Medical Language System® (UMLS) entities from a multispecialty corpus[5-7]; dos Santos et al.[8] addressed adverse events to detect fall events; Lopes et al.[9] addressed the extraction of entities from neurology reports. However, none of those dealt especially with oncology and Portuguese. We also generate a multi-institutional corpus of oncology. Lastly, to our knowledge, there is no published work that has compared the f1-score of NER models trained with clinical NER corpus of different sizes, incrementally generated, which can give us evidence about the volume of data needed to train this type of model.

Methods

In order to carry out this work, it was necessary to follow two major steps: The first major step was to obtain, annotate and sub-divide the corpus, a process that involved a great deal of human work and, as it is specialized annotation, it is humanly costly. The second step was the training of 40 NER models from the corpus generated in the previous step. The processes performed are well described in the sections of this chapter.

Data Source

In this study, we considered clinical records from two hospitals specialized in cancer in Brazil. Those records contain information about the clinical evolution, procedures performed, medication administered, routine care, and all information pertaining to the clinical management of the patients. The first hospital contributed with

350 clinical records of patients in treatment for breast cancer, and the second contributed with 1,150 clinical records of patients in treatment for prostate cancer, two of most recurrent cancer types in the world[10].

All records were de-identified by removing any Personally Identifiable Information (PII). Names of people or institutions, birth dates, places, phone, record, or medical board numbers, professions, those were all replaced with a coded representation such as '_PPPPPP_'.

Corpus Annotation

The majority of NER models require data that have been manually annotated to train models, like the transformer based models.[11] These annotations, in the medical domain, due to their peculiar linguistic characteristics, need to be taken by domain specialists, such as physicians or other healthcare professionals who are familiar with the clinical context.[12]

In order to establish a gold standard during the training and annotation process, the annotators underwent an intense process of adaptation to the annotation environment (presented in a future section) with continuous supervision by a data scientist and a medical oncologist. The training sessions aimed to explain the concepts used in the artificial intelligence algorithm, the entities and attributes used, and to provide examples of past experiences to guide the process and avoid errors and missing data. In order to document all the important information for the annotation process, guidelines were created. Table 1 summarizes the main labels of the guidelines used to train the models in this article. The labels were built based on Elile's guideline, created to extract data from eligibility criteria from clinical studies.[13]

Table 1 - Entity and Attribute utilized in the Guideline

Item	Description
Condition	An illness or medical condition determined by the clinical staff or reported by the patient.

Medication/Drug Substance	Records about the inferred use of a biochemical substance with a physiological therapeutic effect when ingested or otherwise introduced into the body. Medications include prescription and over-the-counter drugs, vaccines, and large molecule biologic therapies. Drug exposure is inferred from clinical events associated with ordering, written prescriptions, pharmacy dispensing, procedural administration, and other patient-reported information.
Measurement	A structured value (focus on numerical values) obtained through the systematic examination of a person or sample. It captures measurement orders and measurement results. The measurement domain may contain laboratory results, vital signs, or quantitative findings from pathology reports.
Anatomic Location	Location of the body associated with the ENTITIES.

Annotation process

An initial batch of 5 documents was sent to each annotator who returned the batches properly annotated to an adjudicator, who solved the problems and sent feedback to respective annotators. This process was repeated until the annotators arrived at a certain level of agreement, when the annotators were finally considered able to start the annotation of the real corpus. It took 8 rounds of annotation, resulting in 40 documents each annotator to reach an agreement considered good for the clinical domain, above 0.61.[14]

After all annotators were considered suitable, a joint annotation round was carried out. In this round, a batch of 150 documents was annotated by a team of three annotators. An adjudicator reviewed and gave feedback to the annotators, with the objective of standardizing the form of annotation of the team. These 150 documents were considered gold plus standard, as in addition to being double annotated, they were searched.

After this batch, each one of the tree annotators began to annotate their own batches, with 450 documents each, which were considered a gold standard, until completing the corpus of 1,500 documents. During the annotation of gold standard documents, an adjudicator chose random documents to verify that the annotators were following the defined rules.

Annotation tool

In order to carry out the annotation process, the use of computational tools was necessary to help the work.[15] To annotate our corpus, the tool chosen was the Multi-purpose Annotation Environment (MAE).[4] MAE was chosen because it is easy to use and often flexible for many annotation tasks.[16]

POS-tagging

After annotated with clinical entity types, the corpus was automatically annotated with POS-tagging entities, using a state-of-the-art model for Portuguese entity texts.[17] The annotation with POS-tagging is necessary to standardize the corpus according to the CoNLL-2003 standard[18], commonly used as input data in libraries of NLP.

Model development

The NER task can be considered a task of classification, where the model needs to assign a named entity label to every word in a sentence. A single named entity can span one or many tokens. For the representation of each token in a named entity, it is necessary to use some type of representation at token level. The chosen representation for our tokens was the Inside Outside Begin 2 (IOB2). IOB2 is one of the most common and simple representations for tokens, where the first token of an entity will be represented by 'B-*label*', the other tokens of the same entity by 'I-*label*', and tokens that are not part of any entity are represented by 'O'.[11]

The most powerful techniques to NER are the transformers based architecture models like BERT.[11] Models like BERT contain two phases of training: the 'pre processing', a time-consuming self-supervised phase where the model learns to 'understand' the language, processing large amounts of data, and the 'fine-tuning', a typically very fast phase where model can be adapted to linguistic sub domains, to perform downstream tasks, like token classification, or both.

In this study, we used the BioBERTpt[6], a pre-trained BERT model for the Portuguese language, fine-tuned by domain adaptation for clinical and biological text domains. Moreover, we performed an additional fine-tuning for a specific task (i.e., NER), to generate all of our NER prediction models.

The corpus is multilabel, with possible overlap of entities in a same token, thus it

is necessary to use a multilabel approach, we used the binary relevance[19], in this approach each semantic type is used in separate, training one model for each semantic type.

Sub-corpus Generation

Although our guidelines have 8 labels of semantic entities, we chose 4 of them to be part of this study: Condition, Anatomic Location, Drug/Substance, and Measurement. We chose those categories due to the time required to perform the processing necessary to train the models in the proposed way. Thus, the entire corpus containing all the labels was divided into 4 new corpora with the same clinical documents, where each one contained only the labels related to its semantic type. This training technique is known as binary relevance, where each semantic type is trained separately, thus, creating one model for each semantic type.

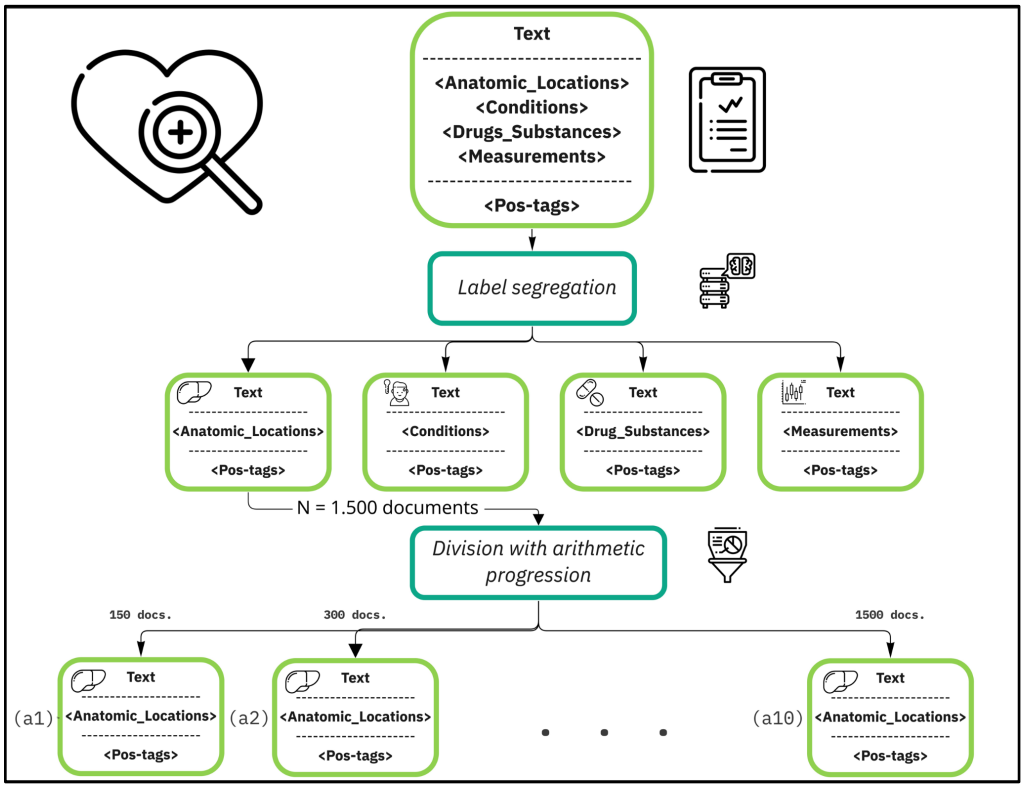

Figure 1 - The sub-corpus generation

Each corpus generated in the previous step, containing 1,500 documents each, was then split again in order to create 10 new sub-corpus, with 150 documents each, for each label. The final 40 corpora, used in the experimental setup, were generated by the

agglutination of the subcorpuses of 150 documents, creating the corpus according to an arithmetic progression of n=1 corpus or n=150 documents, so that a corpus called 'C ' contained documents from corpus 'C-1' plus 150, and corpus 'C+1' contained documents from corpus 'C' plus 150. Figure 1 illustrates the divisions that created the corpus.

Experimental Setup

A total of 40 models for NER were trained, one for each of the generated sub-corpus. For training, we performed the BioBERTpt fine-tuning step for each of our 40 sub-corpus. The division between training, validation and testing data was done randomly. To aid training, the Simpletransformers library[20] was used and all the processing was done in the Google Collab[21] environment. The parameters of all training can be seen at Table 2.

Table 2 - Parameters of model training

Parameter	Value
Batch Size	14
Pre-seed	True
Learning Rate	4e-5
Training data	70% of the execution corpus
Validation data	15% of the execution corpus
Test data	15% of the execution corpus

Results and Discussion

All the 40 trained models were measured according to the f1-score. Figure 2 contains the f-score of each model and shows also the variation of f1-score of the models of the same label according to the increase of documents in the corpus that trained the model. In Figure 2 it is possible to observe variations in the f1-score of the models at

certain times, indicated by vertical gray lines. The exact f1-score values of each of the models are summarized at Table 3.

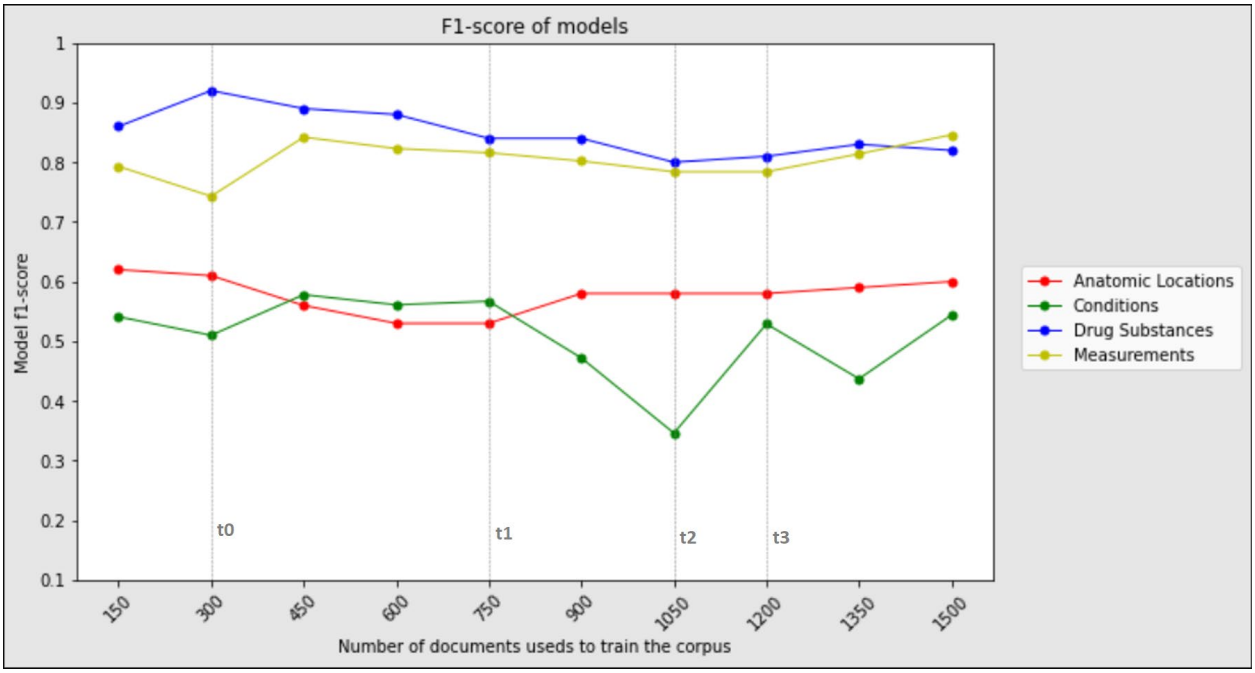

Figure 2 - F1-score of all models

Table 3 - F1-score of models

f1-score of models	**Number of documents of models**									
Type of label of models	150	300	450	600	750	900	1050	1200	1350	1500
Anatomic Location	0,44	0,36	0,35	0,43	0,50	0,38	0,24	0,26	0,36	0,36
Condition	0,54	0,51	0,58	0,56	0,57	0,47	0,35	0,53	0,44	0,54
Drug Substance	0,86	0,92	0,89	0,88	0,84	0,84	0,80	0,81	0,83	0,82
Measurements	0,79	0,74	0,84	0,82	0,82	0,80	0,78	0,78	0,81	0,85

At t_0 three of all four models showed a decrease of their f1-score compared to their respective immediately previous models. This decrease could be explained due the fact that in the previous respective training, with 150 documents each, all documents were triple annotated and became part of the corpus only after a complete review by an expert. The multiple annotation and expert review made the corpus generated by these documents have a higher agreement pattern than the other corpus, which could possibly

result in a good f-score of the models. However, at this point (t_0) the 150 additional documents that were part of this model were not double-annotated and revised as were the first 150, resulting in a corpus with a slightly lower quality than the initial one, reflecting directly on the f1-score of the model. We should also note an exception: the model of Drug Substances labels, where an increase in the f1-score was observed, possibly due to the ease of annotating such entities.

In the models that are represented within the area between the lines that indicate the moments t_1 and t_2, there is a relative decrease in the f1-score of all models. This could be possibly because at t_1 new annotators, still with little experience, were added to the team.

Finally, after t_2, the f1-score of the models rises again, probably due to the fact that the annotators have improved their expertise in annotation or just because of the increase in documents in the corpus.

The score behavior of the Anatomic Locations and Conditions models was very similar. Surprisingly, the results for Anatomic Locations were much lower than the other entities, and we expected Anatomic Location to have a much better performance than Conditions, since it is a simpler label to annotate and identify. To pinpoint what could have happened, an analysis was carried out that identified the main annotation errors in the Anatomic Locations labels. The errors were corrected by post-processing and the Anatomic Locations score improved. Table 4 shows the f1-score between Anatomic Locations models scores after the post-processing.

Table 4 - F1-score comparison of Anatomical Location models before and after post-processing

F1-score of Anatomic Locations after post-processing	**Number of documents of models**									
	150	300	450	600	750	900	1050	1200	1350	1500
Anatomic Locations	0,62	0,61	0,56	0,53	0,53	0,58	0,58	0,58	0,59	0,6

Conclusion

In this article, we built a clinical corpus with oncology records annotated for the NER task. We reported the human labeling process and pointed out challenges regarding the clinical annotation. Also, we incrementally trained and evaluated the NER models with

distinct dimensions. Even using fine-tuning of pre-trained models, we noticed that there is still a need for a greater number of annotated texts for the models to generalize, so we understand that the results could still be improved, either through an annotation effort achieving a larger volume of data, or even the use of data augmentation techniques.(22)

We also verified that while the models do not have a sufficient number of documents to generalize, the quality of the annotation significantly affects the performance of the model, since models with fewer documents but higher quality of annotation performed better than models with more documents and lower annotation quality, reinforcing the trade-off between the number of annotators and the quality of the annotation process.

It is also important to note that this analysis was done for a limited quantity of documents and in a specific and complex sub-domain, and may behave differently for larger numbers of documents or different domains. For future work, we intend to annotate more texts, carry out experiments with the other labels that we have in the corpus and carry out an extensive error analysis and annotation harmonization process.

Acknowledgements

We thank the annotators who contributed to the annotation of the corpus.

References

1. Jensen, PB; Jensen, LJ, Brunak, S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 2012;13(6): 395-405.
2. Jian F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: Past, present and future, Stroke Vasc. Neurol. 2 2017;230–243.
3. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics 2019;21:1-18.
4. World Health Organization. Cancer. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer

5. Oliveira LES e, Peters AC, da Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. Journal of Biomedical Semantics. 2022 May 8;13(1).

6. Schneider ETR, de Souza JVA, Knafou J, Oliveira LES e, Copara J, Gumiel YB, et al. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. Proceedings of the 3rd Clinical Natural Language Processing Workshop. 2020;

7. de Souza JVA, Gumiel YB, Silva EL, Moro CM. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic group.Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde; 2019, 318-323.

8. dos Santos HDP, Silva AP, Maciel MCO, Burin HMV, Urbanetto JS, Vieira R. Fall Detection in EHR using Word Embeddings and Deep Learning. Proceedings of the 19th International Conference on Bioinformatics and Bioengineering (BIBE); 2019, 265-268.

9. Lopes F, Teixeira C, Oliveira HG. Contributions to Clinical Named Entity Recognition in Portuguese. Proceedings of the 18th BioNLP Workshop and Shared Task; 2019, 223–233.

10. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians [Internet]. 2021 Feb 4;71(3):209–49. Available from: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21660

11. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition [Internet]. arXiv.org. 2016. Available from: https://arxiv.org/abs/1603.01360

12. Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annual Symposium Proceedings AMIA Symposium [Internet]. 2007 [cited 2022 Aug 27];2007:625–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655900/

13. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association [Internet]. 2017 Apr 1 [cited 2021 Dec 1];24(6):1062–71. Available from: https://academic.oup.com/jamia/article/24/6/1062/3098256?login=true

14. Richard LJ, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics [Internet]. 1977;33(1):159–74. Available from: https://www.jstor.org/stable/2529310?seq=7#metadata_info_tab_contents

15. Stubbs A. MAE and MAI: Lightweight Annotation and Adjudication Tools [Internet]. Association for Computational Linguistics; 2011 [cited 2022 Aug 29] p. 23–4. Available from: https://aclanthology.org/W11-0416.pdf

16. de Oliveira LFA, Oliveira LES, Gumiel YB, Carvalho DR, Moro CMC. Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. Research on Biomedical Engineering. 2020 Jun 19;36(3):267–76.

17. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv.org. 2018. Available from: https://arxiv.org/abs/1810.04805

18. CoNLL-2003 standard https://aclanthology.org/W03-0419.pdf

19. Souza JVA, Schneider ETR, Oliveira LES, Gumiel YB, Paraiso EC, Teodoro D, Barra CMCM. A Multilabel approach to Portuguese clinical named entity recognition. Journal of Health Informatics. 2020 Dez; (special number SBIS): 366-72.

20. Simpletransformers library. Available from: https://simpletransformers.ai/

21. Google Collab. Available from: https://colab.research.google.com/

22. Issifu AM, Ganiz MC. A simple data augmentation method to improve the performance of named entity recognition models in medical domain. 2021. 6th International Conference on Computer Science and Engineering (UBMK): 763-768.