Developing a Transformer-based Clinical Part-of-Speech Tagger for Brazilian Portuguese

Desenvolvimento de um etiquetador morfossintático baseado em Transformer para textos clínicos brasileiros

Desarrollo de un etiquetador morfosintáctico basado en Transformer para textos clínicos brasileños

Elisa Terumi Rubel Schneider[1], Yohan Bonescki Gumiel[2,3], Lucas Ferro Antunes de Oliveira[1,2], Carolina de Oliveira Montenegro[1], Laura Rubel Barzotto[1], Claudia Moro[1], Adriana Pagano[2], Emerson Cabrera Paraiso[1]

1 Pontifícia Universidade Católica do Paraná - PUCPR
2 Universidade Federal de Minas Gerais - UFMG
3 Laboratório de Informática Biomédica - Instituto do Coração - HC FMUSP

Autor correspondente: Elisa Terumi Rubel Schneider
E-mail: elisa.rubel@pucpr.edu.br

Resumo

O Prontuário Eletrônico do Paciente contém informações valiosas, passíveis de serem extraídas por meio de tarefas de processamento de língua natural (PLN), como o etiquetamento morfossintático de palavras. Embora grandes avanços em PLN no domínio da saúde tenham sido observados, como a arquitetura Transformer, línguas como o português ainda estão subrepresentadas. Neste artigo, apresentamos etiquetadores desenvolvidos para textos em português, refinados a partir dos modelos BioBERTpt (clínico/biomédico) e BERTimbau (genérico) em um corpus com anotações morfossintáticas. Atingimos 0.9826 em acurácia, estado-da-arte para o corpus utilizado. Além disso, realizamos uma avaliação por humanos dos modelos treinados e outros da literatura, utilizando narrativas clínicas autênticas. Nosso modelo clínico atingiu 0.8145 em acurácia comparado com 0.7656 do modelo genérico. Também apresentou resultados competitivos em relação a modelos treinados especificamente com textos clínicos, evidenciando o impacto do domínio no modelo de base em tarefas de PLN.

Descritores: Processamento de Linguagem Natural; Registros Eletrônicos de Saúde; Aprendizado Profundo

Abstract

Electronic Health Records are a valuable source of information to be extracted by means of natural language processing (NLP) tasks, such as morphosyntactic word tagging. Although there have been significant advances in health NLP, such as the Transformer architecture, languages such as Portuguese are still underrepresented. This paper presents taggers developed for Portuguese texts, fine-tuned using BioBERTpt (clinical/biomedical) and BERTimbau (generic) models on a POS-tagged corpus. We achieved an accuracy of 0.9826, state-of-the-art for the corpus used. In addition, we performed a human-based evaluation of the trained models and others in the literature, using authentic clinical narratives. Our clinical model achieved 0.8145 in accuracy compared to 0.7656 for the generic model. It also showed competitive results compared to models trained specifically with clinical texts, evidencing domain impact on the base model in NLP tasks.

Keywords: Natural language processing; Electronic Health Records; Deep Learning

Resumen

La historia clínica electrónica contiene información valiosa que puede extraerse mediante tareas de procesamiento del lenguaje natural (PLN), como el etiquetado morfosintáctico. Aunque se han observado grandes avances en el PNL para la salud, como la arquitectura Transformer, lenguas como el portugués continúan subrepresentadas. En este trabajo, presentamos etiquetadores desarrollados para textos en portugués, refinados usando los modelos BioBERtpt (clínico/biomédico) y BERTimbau (genérico) en un corpus con anotaciones morfosintácticas. Alcanzamos una exactitud de 0,9826, estado del arte para el corpus utilizado. Además, realizamos una evaluación por humanos de los modelos entrenados y de otros en la literatura, utilizando narrativas clínicas auténticas. Nuestro modelo clínico alcanzó una exactitud de 0,8145 en comparación con 0,7656 del modelo genérico. También mostró resultados competitivos frente a modelos entrenados específicamente con textos clínicos, lo que pone de manifiesto el impacto del dominio sobre el modelo base en tareas de PLN.

Descriptores: Procesamiento de Lenguaje Natural; Historia Clínica Electrónica; Aprendizaje Profundo

Introduction

Electronic health records (EHRs) contain patient-related information comprising both structured and unstructured data. As these clinical data are essential for health decisions, Natural Language Processing (NLP) techniques have been researched to extract relevant information and acquire a wide variety of knowledge from unstructured health documents. An important NLP task is part-of-speech (POS) tagging, which categorizes the morphosyntactic value of words in a text, labeling them with a grammatical class (e.g., NOUN, PRONOUN, NUMERAL, etc.) [1]. POS tagging can support other tasks, such as named entity recognition (NER), a powerful task to assist in healthcare decision-making.

Despite the recent advances in NLP, there are still few studies in low-resource languages, such as Portuguese. Regarding POS tagging in Portuguese in particular, we can mention the development of the Mac-Morpho corpus [2] [3] [4]; a neural network trained with the Mac-Morpho corpus [5]; a POS-tagger structure learning framework [6]; and a Bi-LSTM architecture trained with two corpora [7]. However, when it comes to the clinical domain, we have found few research initiatives targeting Portuguese. Worth highlighting are the study by [8], which trained a POS-tagger on clinical texts written in Portuguese; the study by [9], which trained a clinical POS-tagger model with Flair; and the study by [10], which created a Brazilian Portuguese corpus containing clinical texts annotated with morphological and syntactic information. To the best of our knowledge, no research has addressed Portuguese clinical POS-tagging using the Transformer architecture [11], which is currently responsible for the state-of-the-art for several NLP tasks.

Processing information in EHRs still poses many challenges, since they are noisy clinical texts and contain acronyms and shortened forms, a wide range of synonyms, flexible formatting, and very often typos. Moreover, domain-specific vocabulary and assumptions are widely used, demanding particular solutions for clinical texts. Hence, our objective was to develop a POS-tagger model for Portuguese medical texts, using contextualized pre-trained language models based on the Transformer architecture as BERT-based models [12]. As the language model can benefit several downstream NLP tasks, we have released it in a public repository. To the best of our knowledge, this is the first POS-tagger Transformer-based model released for the Portuguese clinical and biomedical domains. Additionally, for evaluation purposes, we also trained a POS-tagger model for generic Portuguese, which was also made available to other researchers.

This paper is organized as follows: the next section provides details about our method, such as the corpus used for training the models with implementation and evaluation details; subsequently, we present our results and their discussion; finally, we present the conclusions drawn from the observed results.

Methods

We first present Mac-Morpho, the labeled corpus for POS-tagger that we used in fine-tuning; then describe the BERT-based models used as checkpoints, the training steps, and the metrics used. Figure 1 shows an overview of the experimental setup, comprising our model training process, clinical pos-tagging annotation process, and our two evaluation processes (general and clinical domain).

Mac-Morpho corpus

Mac-Morpho is a POS-annotated corpus for Portuguese, with approximately one million words of news reports from the Brazilian newspaper Folha de São Paulo [1]. It was first released in 2003 [2], and two revisions have been made to improve its quality. The first revision consisted in cleaning noise in the data and changing the tagset to include preposition-article contractions [3]. A second revision included corrections of problematic

[1] https://www.folha.uol.com.br/

sentences and a change in the tagset, removing tags that needed knowledge above the morpho-syntactic level to be correctly detected [4]. We used the latest version of the corpus[2], containing almost 50,000 sentences with nearly 950,000 tokens in total, following the hold-out split encouraged by the authors in order to make consistent comparisons possible (76%, 4%, and 20% for train, dev, and test, respectively). Although Mac-Morpho is a collection of news texts, far removed from clinical texts, we chose this corpus, since it is currently the most extensive, manually POS-tagged corpus of modern Brazilian Portuguese [1], with 26 different tags for word labeling.

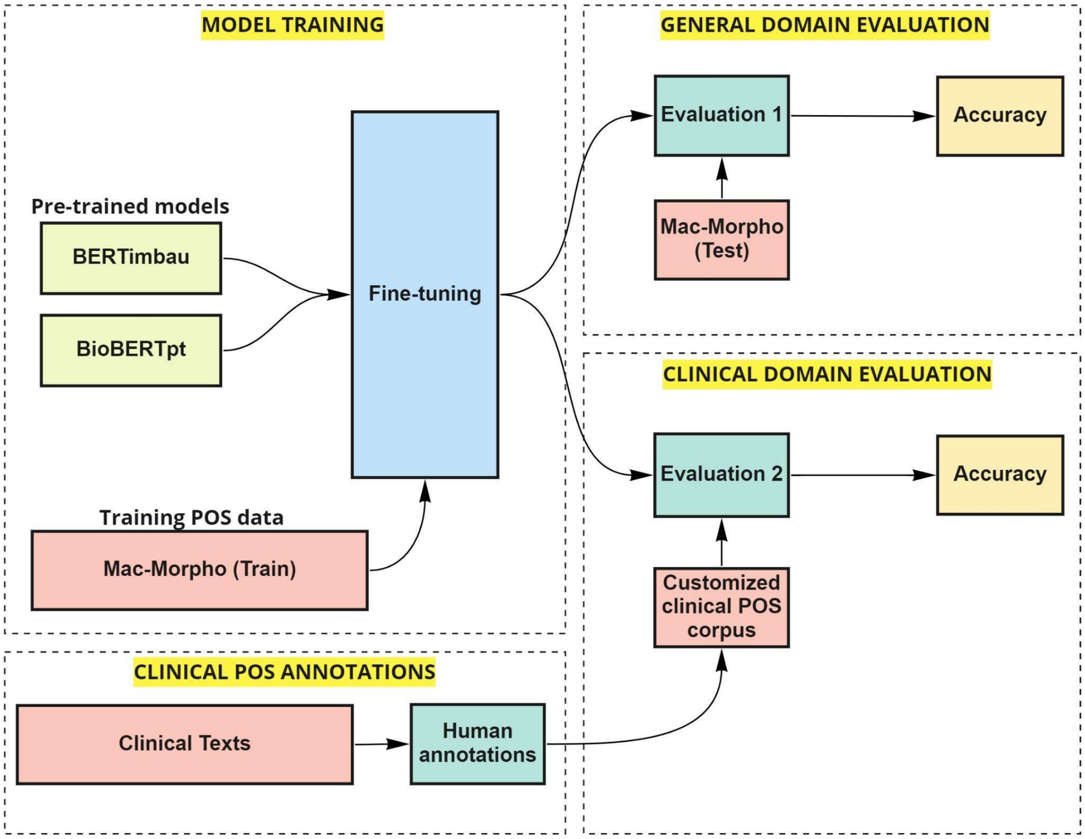

Figure 1 – Overview of the experimental setup.

[2] http://nilc.icmc.usp.br/macmorpho/

BERT-based models

Contextual language models pre-trained on large-scale unstructured data, especially those supported by the Transformer architecture [11], have reached state-of-the-art performance in several NLP tasks. From this training on unlabeled data, these models can be fine-tuned for a specific task through labeled data, updating its last layers to adapt for a downstream task. Bidirectional Encoder Representations from Transformers (BERT) [12] is a transformer-based representation model, pre-trained on two generic tasks: a) masked language modeling, where 15% of tokens are masked, and BERT is trained to predict them from context, and b) next-sentence-prediction, in which BERT is trained to predict whether a selected sentence is a probable sentence or not to follow the previous one. As a result of this training process, the model learns contextual embeddings for words.

In our fine-tuning step, we used the weights of the BioBERTpt(all) model [13] as checkpoints, trained with a large clinical and biomedical dataset in the Portuguese language. BioBERTpt(all) was trained on multilingual BERT and reached the state-of-the-art in two NER corpora.

In order to investigate the importance of domain in the base model, we also trained a generic POS-tagger model for Portuguese. This out-of-domain model was trained using BERTimbau [14] as a checkpoint, a generic Portuguese BERT-based model trained using a whole-word mask on a large Brazilian corpus.

Training steps

A preprocessing step in the corpus was performed to replace NOUN tags with NUM tags in the case of numbers, since some numerals are tagged as NOUN (instead of NUM) in Mac-Morpho. We split the texts into sentences, arranged them in iob2 format, and tokenized them with the AutoTokenizer from the Hugging Face library [16]. We trained our POS-tagger models for ten epochs, with a learning rate of 1e-5, batch size of 32, AdamW as an optimizer, and a maximum length of 200.

This paper will refer to the in-domain and generic domain models as POStagger-BioBERTpt and POStagger-BERTimbau, respectively.

Evaluation

We first analyzed the F1 score and accuracy calculated by the test set of the Mac-Morpho corpus to verify how accurate the model performed in texts from the same corpus of the training.

Also, we evaluated the trained models on a set of clinical notes taken from SemClinBr [15], a corpus containing clinical narratives from Brazilian hospitals. We randomly selected 50 sentences containing between 6 and 15 tokens, which were manually POS-annotated by a human linguist, referred to in this paper as human annotation. For comparison purposes, we also evaluated two clinical POS-tagger models, trained with Flair [9] and Spacy [10], referred to in this paper as Flair and Spacy. For evaluation, we separated words pertaining to distinct functions in contracted forms in Portuguese, such as the contraction of a preposition with an article (eg "NA" was separated into "EM + A"). We also removed whitespaces and tokenized the sentences by line breaks and tabs. All the sentences were submitted to the models and, based on their output, we calculated the number of hits and errors of each model (accuracy), both as a general score and a score by category. In a post-processing stage, a) we tagged all the symbols in the models' outputs as SYM, since this tag is common in clinical texts but was not learned by the models, and b) we normalized tags using the tagset proposed by Universal Dependencies[3], as each model has its own tagset (we used the same structure of the work of [9]). In addition, we performed the following experiments: 1) Sentence evaluation with the text in their actual typed form; 2) Sentence evaluation in lowercase; and 3) Automatic tagging of all numbers as NUM and sentences in lowercase.

Ethical aspects

For our study, no data with human participants were collected. The clinical corpus we used to train our models was previously published in [15], for which authors report to have obtained Ethics approval.

[3] https://universaldependencies.org/

Results and discussion

Our results for the Mac-Morpho corpus are shown in Table 1. We selected some studies for comparison, reporting their results for the MacMorpho corpus, favoring results for the latest version when several were available. We achieved an F1 score and accuracy of 0.9826 for the POStagger-BERTimbau model and slightly lower results for the POStagger-BioBERTpt model, with an F1 score and accuracy of 0.9818. Even though it was trained on clinical and biomedical texts, POStagger-BioBERTpt yielded similar results (below by only 0.0008) to POStagger-BERTimbau, trained over general domain corpora and expected to have superior results in a journalistic corpus such as the Mac-Morpho rather than in a clinical one. Furthermore, the POStagger-BERTimbau model achieved state-of-the-art for the Mac-Morpho corpus when accuracy is considered. The De Oliveira et al. [9] (ALL) model, a FLAIR model concatenating journalistic, clinical, and biomedical corpora, achieved slightly superior results when the F1 score was considered. The De Oliveira et al. [9] (MAC) model, trained on the Mac-Morpho corpus solely, achieved slightly worse results than De Oliveira et al. [9] (ALL), suggesting that concatenating several annotated corpora for training can yield better results.

Table 1 – Model results for the Mac-Morpho corpus

Model	Accuracy	F1 score
Our models		
POStagger-BERTimbau	**0.9826**	0.9826
POStagger-BioBERTpt	0.9818	0.9818
Baseline		
De Oliveira et al. [9] (MAC)	0.9612	0.9803
De Oliveira et al. [9] (ALL)	0.9730	**0.9863**
dos Santos and Zadrozny [17]	0.9731	*not reported*
Oleynik et al. [8]	0.753	*not reported*
Fonseca, Rosa and Aluisio [2]	0.9773	*not reported*

We also achieved superior results for the Mac-Morphos corpus using BERT-based models compared to models trained over other architectures reported in the literature, evidencing the impact of models based on the Transformer architecture, as BERT-based models, in NLP tasks. Regarding architectures: [9] used FLAIR; [17] used Convolutional Neural Networks (CNNs); [8] used openNLP[4] training with only 595 sentences from Mac-Morpho; [2] used a multilayer perceptron (MLP) neural network.

To evaluate the performance of our POS-tagger models in the clinical domain, we evaluated their performance on a set of clinical notes taken from the SemClinBr corpus, generating gold standard POS-tagging annotations to validate our models. The accuracy values are shown in Table 2, reporting the results for (1) Sentences in their actual form (Normal); (2) Sentences in lowercase (Lower); and (3) Automatic tagging of all numbers as NUM and sentences in lowercase (LowerNUM).

Table 2 – Accuracy for the models evaluated on clinical texts.

Model	Normal	Lower	LowerNUM
Our models			
POStagger-BERTimbau	0.7656	0.9109	0.9147
POStagger-BioBERTpt	0.8145	0.9012	0.9167
Baseline			
De Oliveira et al. [10] (SPACY)	**0.9238**	**0.9225**	**0.9225**
De Oliveira et al. [9] (FLAIR)	0.9141	0.9089	0.9089

We found that the lack of standardization of clinical texts directly impacted the POStagger-BERTimbau and POStagger-BioBERTpt models' results. Both models were trained with Mac-Morpho, a news texts corpora, and may not be robust enough to deal

[4] http://opennlp.apache.org/

with some aspects particular to the clinical domain, such as extensive use of abbreviations and acronyms, uppercase terms, domain-specific vocabulary, flexible formatting, and atypical grammatical constructions [18]. Setting the whole text to lowercase impacted the results, as some clinical texts are fully written in uppercase; this step improved the performance for both models. Further, as the Mac-Morpho corpus had instances where numbers were tagged as NOUN or with other labels, our strategy of normalizing the text and tagging all numbers with the NUM tag slightly improved the results. In clinical texts, numbers usually refer to values of diagnostic exams, drug dosage, and results from a physical examination.

Figure 2 shows an example where our two models - POStagger-BioBERTpt and POStagger-BERTimbau - outperformed all other models and had an output closely matching human annotation. The human annotations are at the top of the sentence (our gold standard), and below are the words incorrectly tagged by the models when compared to human annotation. POStagger-BioBERTpt got only one incorrect tag while POStagger-BERTimbau got three incorrect ones. Spacy and Flair get over four incorrect tags.

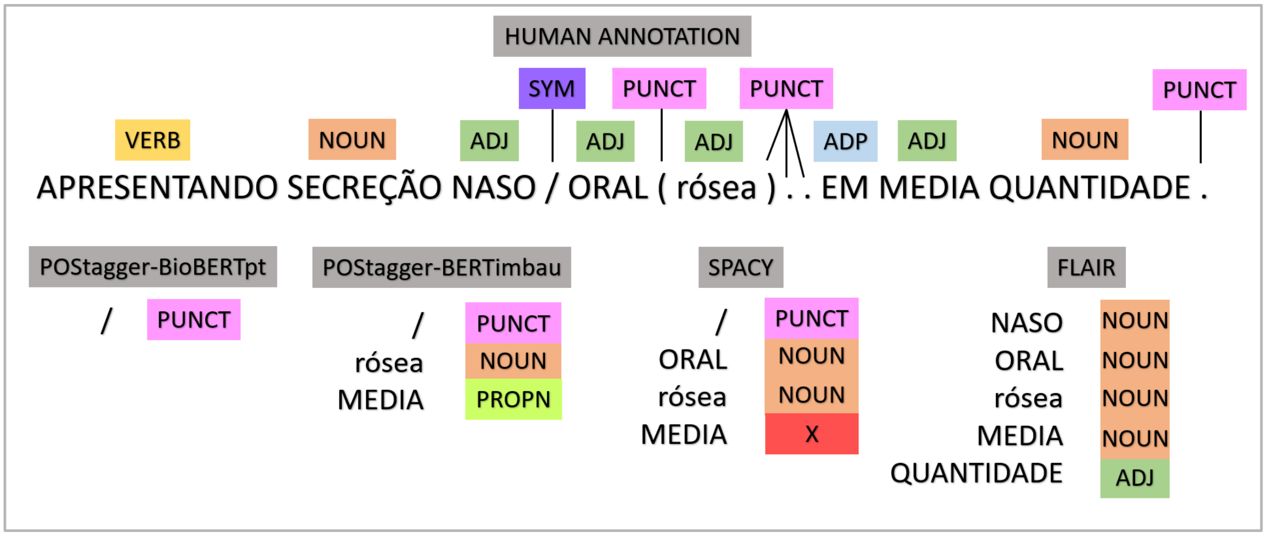

Figure 2 - Human annotation and model output for a sample POS-tagged sentence

Overall, results point to the model proposed by [10], trained with SPACY and 2,900 sentences from SemClinBr, as achieving the best results for our 50 sample sentences. This can be accounted for by the fact that those sentences were sampled from SemClinBr and were similar to the ones the model was trained with. However, considering the small number of sentences used in training our model, our results are promising in that augmenting the training sample will certainly improve the model's performance. This corroborates the fact that an in-domain corpus, no matter how small it is, has a positive impact on model performance [10]. For the purpose of our research, this implies that our POS-tagger can be fine-tuned with a clinical corpus so as to further enhance its performance.

Conclusion

We proposed a new publicly available Portuguese POS-tagger model to support clinical and biomedical tasks, using Transformer architecture, state-of-the-art for NLP. Our experiment showed that even using a journalistic corpus for training, the in-domain knowledge coded in BioBERTpt can benefit clinical tasks, showing promising results. The POStagger-BioBERTpt presented competitive results compared to other models trained with clinical texts, ranking second in lowercase texts (common in clinical notes). Our in-domain model contributes to NLP in Portuguese, a low-resource language with scarce resources in the clinical domain. Also, as the generic Portuguese POS-tagger can assist research in other domains, we released the POStagger-BERTimbau, which achieves the state-of-the-art for Mac-Morpho corpus. Since the BERT-based models demonstrated promising results, in future work, we intend to train a BERT-based model for POS-tagger with the same clinical corpora used in [9] and [10] to compare with the current Flair and Spacy models, and further investigate the role of the Transformer architecture in POS-tagger tasks.

Acknowledgment

This study was funded in part by the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, and the National Council for Scientific and Technological Development (CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico), Grant 313103/2021-6.

References

1. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second. Prentice Hall; 2008.

2. Fonseca ER, Rosa JLG, Aluísio SM. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. Journal of the Brazilian Computer Society. 2014;21:1-14.

3. Aluísio S, Pelizzoni J, Marchi AR, de Oliveira L, Manenti R, Marquiafável V. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: Mamede NJ, Trancoso I, Baptista J, das Graças Volpe Nunes M, editors. Computational Processing of the Portuguese Language. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 110-7.

4. Fonseca ER, Rosa JLG. Mac-Morpho revisited: towards robust part-of-speech tagging [Internet]. Proceedings. 2013 ;[citado 2022 ago. 09] Available from: http://www.lbd.dcc.ufmg.br/colecoes/stil/2013/0011.pdf

5. Dos Santos CN, Zadrozny B. Learning character-level representations for part-of-speech tagging. ICML'14 Proc. 31st Int. Conf. Int. Conf. Mach Learn. 2014;32:1818–26.

6. Fernandes ER, Rodrigues IM, Milidiu RL. Portuguese Part-of-Speech Tagging with Large Margin Structure Learning. 2014 Brazilian Conf. Intell. Syst., IEEE; 2014, p. 25–30. doi: https://doi.org/10. 1109/BRACIS.2014.16.

7. De Sousa RCC, Lopes H. Portuguese POS Tagging Using BLSTM Without Handcrafted Features. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11896 LNCS, 2019, p. 120–30. doi: https://doi.org/10.1007/978-3- 030-33904-3_11.

8. Oleynik M, Nohama P, Cancian PS, Schulz S. Performance analysis of a POS tagger applied to discharge summaries in portuguese. Stud Health Technol Inform. 2010;160:959–63. https://doi.org/10.3233/ 978-1-60750-588-4-959.

9. Ferro Antunes de Oliveira L, Oliveira L, Gumiel Y, Carvalho D, Moro C. Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. Research on Biomedical Engineering. 2020 06;36.

10. De Oliveira LFA, Pagano A, e Oliveira LES, Moro C. Challenges in Annotating a Treebank of Clinical Narratives in Brazilian Portuguese. In: Pinheiro V, Gamallo P, Amaro R, Scarton C, Batista F, Silva D, et al., editors. Computational Processing of the Portuguese Language. Cham: Springer International Publishing; 2022. p. 90-100.

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R,

Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017.

12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86.

13. Schneider ETR, de Souza JVA, Knafou J, Oliveira LESe, Copara J, Gumiel YB, et al. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 65-72.

14. Souza F, Nogueira R, Lotufo R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri R, Prati RC, editors. Intelligent Systems. Cham: Springer International Publishing; 2020. p. 403-17.

15. E Oliveira LES, 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;13(1).

16. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 38-45.

17. Santos CND, Zadrozny B. Training state-of-the-art portuguese POS taggers without handcrafted features. In: International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2014. p. 82-93.

18. Gumiel YB, Oliveira LES, Claveau V, Grabar N, Paraiso EC, Moro C, et al. Temporal Relation Extraction in Clinical Texts: A Systematic Review. ACM Computing Surveys (CSUR), v. 54, n. 7, p. 1-36, 2021.