Published in Proceedings of the 3rd Clinical Natural Language Processing Workshop, 2020, November 19, online, which should be cited to refer to this work.

BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition

Elisa Terumi Rubel Schneider[1], **João Vitor Andrioli de Souza**[1], **Julien Knafou**[2], **Jenny Copara**[2], **Lucas E. S. e Oliveira**[1], **Yohan B. Gumiel**[1], **Lucas F. A. de Oliveira**[1], **Douglas Teodoro**[2], **Emerson Cabrera Paraiso**[1] and **Claudia Moro**[1]
[1]Pontifícia Universidade Católica do Paraná, Brazil
[2]University of Applied Sciences and Arts of Western Switzerland
{elisa.rubel, joao.souza}@pucpr.edu.br, paraiso@ppgia.pucpr.br, c.moro@pucpr.br

Abstract

With the growing number of electronic health record data, clinical NLP tasks have become increasingly relevant to unlock valuable information from unstructured clinical text. Although the performance of downstream NLP tasks, such as named-entity recognition (NER), in English corpus has recently improved by contextualised language models, less research is available for clinical texts in low resource languages. Our goal is to assess a deep contextual embedding model for Portuguese, so called BioBERTpt, to support clinical and biomedical NER. We transfer learned information encoded in a multilingual-BERT model to a corpora of clinical narratives and biomedical-scientific papers in Brazilian Portuguese. To evaluate the performance of BioBERTpt, we ran NER experiments on two annotated corpora containing clinical narratives and compared the results with existing BERT models. Our in-domain model outperformed the baseline model in F1-score by 2.72%, achieving higher performance in 11 out of 13 assessed entities. We demonstrate that enriching contextual embedding models with domain literature can play an important role in improving performance for specific NLP tasks. The transfer learning process enhanced the Portuguese biomedical NER model by reducing the necessity of labeled data and the demand for retraining a whole new model.

1 Introduction

Despite recent increases in the availability of machine learning methods, extracting structured information from large amounts of unstructured and noisy clinical documents, as available in electronic health record (EHR) systems, is still a challenging task. Patient's EHR are filled with clinical concepts, often misspelled, abbreviated and represented by a variety of synonyms. Nevertheless, they contain valuable and detailed patient information (Lopes et al., 2019). Natural language processing (NLP) tasks, such as Named Entity Recognition (NER), are used for acquiring knowledge from unstructured texts, by recognizing meaningful entities in text passages. In the clinical domain, NER can be used to identify clinical concepts, such as diseases, signs, procedures and drugs, supporting other data analysis as prediction of future clinical events, summarization, and relation extraction between entities (e.g., drug-to-drug interaction).

Rule-based NER approaches, supported by dictionary resources, perform well in simple contexts (Eftimov et al., 2017). However, they are limited to work with the complexity of clinical texts. For complex corpora, machine learning approaches, such as conditional random fields (CRF) (Lafferty et al., 2001) and, lately, a combination with Bidirectional Long Short-Term Memory (BiLSTM) models, have been proposed (Lample et al., 2016). These supervised approaches have a considerable performance gain when trained on huge amounts of labeled data. Neural network language models introduced the idea of deep learning into language modeling by learning a distributed representation of words. These distributed word representations, trained on massive amounts of unannotated textual data, have been proved to provide good lower dimension feature representations in a wide range of NLP tasks (Wang et al., 2020). The Continuous Bag-of-Words and Skip-gram models

proposed to reduce the computational complexity were considered as a milestone in the development of the so-called word embeddings (Mikolov et al., 2013), followed by the Global Vector (GloVe) (Pennington et al., 2014) and the fastText (Bojanowski et al., 2016) models.

While these approaches work with a single global representation for each word, several context-dependent representations models have been recently proposed, such as embeddings from language models (ELMo) (Peters et al., 2018), flair embeddings (Akbik et al., 2018), the Universal Language Model Fine-tuning (ULMFit) (Howard and Ruder, 2018) and bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018). Contextual embedding models pre-trained on large-scale unlabelled corpora, particularly those supported by the transformer architecture (Vaswani et al., 2017), reached the state-of-the-art performance on many NLP tasks (Liu et al., 2020). Nevertheless, when applying the general word representation models in healthcare text mining, the characteristics of clinical texts are not considered, known to be noisy, with a different vocabulary, expressions, and word distribution (Knake et al., 2016). Therefore, contextual word embedding models, like BERT, can be fine-tuned, i.e., have their last layers updated to adapt to a specific domain, like clinical and biomedical, using domain-specific training data. These transfer learning process allows the training of a general domain model with medical domain corpus, proving to be a viable technique to medical NLP tasks (Ranti et al., 2020).

Despite the low availability of clinical narratives, given the sensitive nature of health data and privacy concerns (Berman, 2002), several models were trained on clinical and biomedical corpora. In 2013, the word2vec model was trained on biomedical corpora (Pyysalo et al., 2013), creating a language model with high-quality vector space representations. BioBERT (Lee et al., 2019) is a BERT model trained from scratch using PubMed and PubMed Central (PMC) scientific texts, reaching the state-of-the-art results on some biomedical NLP tasks. Clinical BERT (Alsentzer et al., 2019) demonstrated that the pre-trained model with clinical data improved performance in three common clinical NLP tasks. Li et al. (2019) reached state-of-the-art for biomedical and clinical entity normalization with a model trained using EHR data. Despite the essential contributions of contextual word embeddings on clinical NER, all these studies used English corpora. Indeed, there are few studies in lower resources languages for the clinical domain. In Portuguese, Lopes et al. (2019) proposed a fastText model trained with clinical texts, which achieved higher results when compared to out-of-domain embeddings. In a recent work, de Souza et al. (2019) explored the CRF algorithm for the NER task on SemClinBr (Oliveira et al., 2020), the same annotated corpus we used in this work. They classified three clinical entities (*Disorders*, *Procedures* and *Chemicals and Drugs*) and some medical text abbreviations, achieving promising results. A Portuguese clinical word embedding model were trained using Skip-gram with negative sampling and evaluated on a downstream biomedical NLP task for Urinary Tract Infection disease identification (Oliveira et al., 2019). Their results showed that larger, coarse-grained models achieve a slightly better outcome when compared with small, fine-grained models in the proposed task.

Although these previous works achieved relevant results, we have not found studies for clinical Portuguese using attention-based architectures, such as BERT, which have been achieving the state-of-the-art for most of English NLP tasks. Even with the existence of multilingual models, like BERT-multilingual, it is important to investigate what can be the contribution in creating a domain fine-tuned model for a lower-resource language. As demonstrated in the work of Peng et al. (2019), pre-trained BERT models with biomedical and clinical data achieves better results in the BLUE benchmark for English. This leads us to believe that the same is valid for Portuguese. Thus, the objective of this work is to assess the performance of a domain specific attention-based model, BioBERTpt, to support NER tasks in Portuguese clinical narratives. We intend to investigate how an in-domain model can influence the performance of BERT-based models for NER in clinical data. Also, as knowledge encoded in transformer-based language models can be leveraged to several downstream NLP tasks, we release publicly the first BERT-based model trained on clinical data for Portuguese [1].

2 Methods

In this section, we first describe how BioBERTpt was developed using clinical notes and scientific

[1] https://github.com/HAILab-PUCPR/BioBERTpt

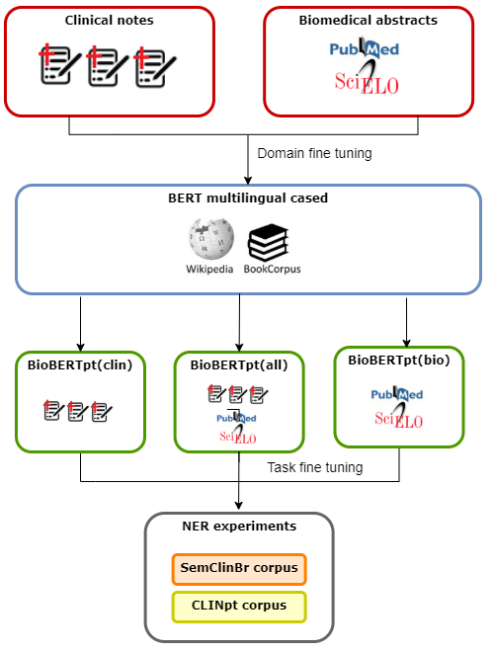

Figure 1: Clinical notes and scientific biomedical abstracts are fed to a pre-trained BERT multilingual model to create BioBERTpt(clin), BioBERTpt(bio) and BioBERTpt(all). These models are then used to extract information from Portuguese clinical notes, evaluated in the clinical NER corpora SemClinBr and CLINpt.

abstracts. Next, we introduce the corpora used for the NER tasks and the evaluation metrics used in our experiments.

2.1 Development of BioBERTpt

In this paper, we fine-tuned three BERT-based models on Portuguese clinical and biomedical corpora, initialized with multilingual BERT weights provided by Devlin et al. (2018).

With the approval from the PUCPR Research Ethics Committee with CAAE (Certificate of presentation for Ethical Appreciation), number 51376015.4.0000.0020, we collected 2,100,546 clinical notes from Brazilian hospitals, from 2002 to 2018. All the clinical text have been properly de-identified, to respect patient's privacy. This corpus contains multi specialty information, including cardiology, nephrology and endocrinology, from different types of clinical texts (narratives), such as discharge summaries, nurse notes and ambulatory notes. In total, the clinical notes contain 3.8 million sentences with 27.7 million words. Our clinical model was trained with this corpus, benefiting from the weights already trained in the multilingual BERT model.

We also trained a biomedical model, using titles and abstracts from Portuguese scientific papers published in Pubmed and in the Scielo (Scientific Electronic Library Online)[2], an integrated database that contains Brazilian's scientific journal publications in multidisciplinary areas such as health. These texts were obtained from the Biomedical Translation Task in the First Conference on Machine Translation (WMT16), which evaluated the translation of scientific abstracts between English, French, Spanish and Portuguese (Bojar et al., 2016). In this work, we used only the Portuguese part, composed by documents from Scielo and Pubmed databases about biological and health, resulting in 16.4 million words. The text corpora used for training our models are listed in Table 1.

In the preprocessing step, we split the notes and abstracts into sentences and tokenize them with the default BERT wordpiece tokenizer (Devlin et al., 2018). All models were trained for 5 epochs on a GPU GTX2080Ti Titan 12 GB, with the hyperparameters: batch size as 4, learning rate as 2e-5 and block size as 512. We used the PyTorch implementation of Bert proposed by Hugging Face[3].

To investigate how the domain can influence the task performance, we trained: a) a model with the clinical data, from the narratives of Brazilian hospitals, b) a model with the biomedical data, from the scientific papers abstracts, and c) a full version, i.e., using both clinical and biomedical data. Throughout this paper, we will refer to these corresponding models as BioBERTpt(clin), BioBERTpt(bio) and BioBERTpt(all), respectively.

2.2 NER experiments

Corpora: In our first NER experiment, we use SemClinBr (Oliveira et al., 2020), a semantically annotated corpus for Portuguese clinical NER, containing 1,000 labeled clinical notes. This corpus comprehended 100 UMLS semantic types, summarized in 13 groups of entities: *Disorders*, *Chemicals and Drugs*, *Medical Procedure*, *Diagnostic Procedure*, *Disease Or Syndrome*, *Findings*, *Health Care Activity*, *Laboratory or Test Result*, *Medical Device*, *Pharmacologic Substance*, *Quantitative Concept*, *Sign or Symptom* and *Therapeutic or Preventive Procedure*. Although SemClinBr supports IOB2 (aka BIO) and IOBES (aka BILOU) tagging schemes, we report our experiment in IOB2, widely

[2]https://scielo.org/
[3]https://github.com/huggingface/transformers

Table 1: List of text corpora used for BioBERTpt

Corpus	Source	Nº of sentences	Nº of words	Domain
Clinical notes	EHR from Brazilian hospitals	3.8 million	27.7 million	Clinical
Scielo: Health area	Literature titles and abstracts	532,920	12.4 million	Biomedical
Scielo: Biological area	Literature titles and abstracts	130,098	3.2 million	Biomedical
Pubmed	Literature titles	74,451	812,711	Biomedical

used in the literature.

For the second NER experiment, we run our models in a small dataset with IOBES format, proposed by Lopes et al. (2019). This corpus is a collection of 281 Neurology clinical case descriptions, with manually-annotated named entities, from now on called CLINpt. These cases were collected from a clinical journal published by the Portuguese Society of Neurology.

Execution: Our experiments were performed with holdout using a corpus split of 60% for training, 20% for validation and 20% for test. We used the Hugging Face API, which provides the BertForTokenClassification class. This class adds a token-level classifier, a linear layer that uses the last hidden state of the sequence. For both NER tasks we used this configuration: AdamW optimizer, weight decay as 0.01, batch size as 4, maximum length as 256, learning rate as 3e-5, maximum epoch as 10, and the linear schedule that decreases the learning rate throughout the epochs with warmup as 0.1.

Evaluation criteria: We evaluate the results using precision, recall and F1-score metrics. As in SemClinBr each entity can have more than one semantic type associated (similar to a multi-label classification), we used the label-based metrics, an adaptation of existing single-label problem metrics, to measure the model general performance. We calculated the micro-average metric, when the score is computed globally over all instances and then over all class labels (Sorower, 2010).

In addition, we also analyzed statistical significance between the F1-score of the models for all entities in SemClinBr. We defined seven samples, where each one corresponds to a set of the F1-score values of all entities in the corpus, calculated for each respective model. As the Friedman test only indicates if there is a difference between the means of the samples, without identifying which sample(s) is(are) different from the set, we applied a Wilcoxon signed-ranks pair-wise as post-test. The Wilcoxon signed-rank test was calculated between pairs of samples, in order to show which pairs of samples have different means. The results are considered statistically significant for P value $<.05$.

We compare BioBERTpt with the already existing contextual models: BERT multilingual uncased, BERT multilingual cased, Portuguese BERT base and Portuguese BERT large. Both BERT multilingual are large versions and provide Portuguese language support, called in this work BERT multi(u) for the uncased version and BERT multi(c) for the cased version. The Portuguese BERT models, proposed by Souza et al. (2019), are BERT-models trained on the BrWaC (Brazilian Web as Corpus), a large Portuguese corpus, with whole-word mask. We used both base and large versions, called here BERT PT(b) and BERT PT(l), respectively. All these word embeddings are out-of-domain, i.e., trained in general context corpora, like Wikipedia and books.

3 Results

Table 2 shows the average precision, recall and F1-score values for all BERT models on SemClinBr and CLINpt corpora, where our in-domain models outperformed in the average scores.

In the SemClinBr corpus, BioBERTpt(bio) improved 0.1 in precision, BioBERTpt(all), 2.0 in recall and 1.6 in F1-score, over the out-of-domain model with better performance. Full F1-score values for each entity are provided on our repository. Analyzing the performance by entity, the in-domain models in general were better at recall and F1-score. Our models obtained better results in precision for 4 entities, recall for 8 and F1-score for 11. The out-of-domain models obtained better results for 9 entities in precision, 5 in recall and 2 in F1-score. The results of the Friedman test evidenced that there is a difference between some models. The post-test Wilcoxon signed-ranks pair-wise showed the statistical relevance between models over all entities, as shown in Figure 2.

Table 2: The average scores of the NER tasks, for each model evaluated. In bold, the best results

Corpus / model	Precision	Recall	F1
SemClinBr			
BERT multi (u)[a]	0.623	0.566	0.588
BERT multi (c)[b]	0.604	0.567	0.582
BERT PT(b)[c]	0.595	0.587	0.585
BERT PT(l)[d]	0.563	0.531	0.541
BioBERTpt(bio)	**0.624**	0.586	0.602
BioBERTpt(clin)	0.609	0.603	0.602
BioBERTpt(all)	0.608	**0.607**	**0.604**
CLINpt			
BiLSTM-CRF [e]	0.753	0.745	0.749
BERT multi (u)	0.903	0.921	0.912
BERT multi (c)	0.912	0.931	0.921
BERT PT(b)	0.910	0.922	0.916
BERT PT(l)	0.898	0.927	0.912
BioBERTpt(bio)	**0.917**	0.925	0.921
BioBERTpt(clin)	**0.917**	**0.935**	**0.926**
BioBERTpt(all)	0.912	0.929	0.920

[a]BERT multilingual uncased
[b]BERT multilingual cased
[c]Portuguese BERT base
[d]Portuguese BERT large
[e]Baseline from previous work Lopes et al. (2019), where the authors used Fastext as word embeddings.

BioBERTpt(all) had statistically higher results on F1-score than BERT multilingual uncased (*P* value as 0.04640), Portuguese BERT large (*P* value as 0.00298) and Portuguese BERT base (*P* value as 0.01750). BioBERTpt(clin) had its performance statistically higher in relation to Portuguese BERT large (0.00713) and Portuguese BERT base (*P* value as 0.01075), and BioBERTpt(bio), in relation to Portuguese BERT large (*P* value as 0.01750). Also, BERT multilingual uncased had a significant higher performance in relation to Portuguese BERT large (*P* value as 0.03305).

The results on the CLINpt corpus, also presented in table 4, shows that BioBERTpt(clin) improved precision in 0.5, recall in 0.4 and F1-score in 0.5.

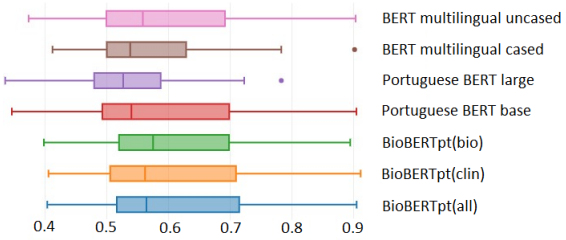

Figure 2: F1-scores of all entities from SemClinBr for evaluation of the models (Wilcoxon signed-ranks pairwise post-test).

Despite CLINpt cases are not representative of the usual clinical notes and narratives found in EHRs, our clinical model presented the best results. Although with little improvement compared to BERT multilingual cased, BioBERTpt(clin) reached the state-of-the-art on this corpus for these three metrics.

4 Discussion

4.1 Effect of domain

Our results show that the in-domain models outperform the general models in average precision, recall and F1-score on the two Portuguese corpora. These results are aligned with previous experiments in English, where domain-specific models outperform generic models (Lee et al., 2019; Alsentzer et al., 2019; Li et al., 2019; Pyysalo et al., 2013). BioBERTpt trained on clinical narratives had overall better performance when compared with the model trained only on biomedical texts, reaching higher results for entities with more clinical-domain-specific vocabulary, such as *Laboratory*, *Pharmacologic Substance* and *Chemical and Drugs*. The better performance of BioBERTpt(clin) over BioBERTpt(bio) was expected, since the NER evaluation set only contains clinical narratives. Although we evaluated both schemes, IOBES and IOB2, we report only IOB2 as there was no significant difference between them.

The F1-score performance of the *Chemical and Drugs* entities was the most assertive for all models, reaching 0.911 with BioBERTpt(clin). Due to specific characteristics of each entity, such as granularity, specificity and different vocabulary across institutions, some entities achieved low performance, like *Laboratory*, which reached only 0.453 as its highest F1-score with BioBERTpt(clin). The use of imbalanced data can also affect the results, since the entities with lower frequency have fewer and

Table 3: F1-score values for three SemClinBr group of entities, for comparison with baseline. In bold, the highest values.

Entity / Model	Disorder	Proced.[a]	Drug
CRF [b]	0.65	0.60	0.42
BioBERTpt(bio)	**0.79**	0.69	0.89
BioBERTpt(clin)	0.78	0.69	**0.91**
BioBERTpt(all)	**0.79**	**0.70**	0.90

[a]Procedure
[b]Baseline from previous work (de Souza et al., 2019)

selected vocabulary, leading the models to achieve lower results or overfit the vocabulary vectors.

By evaluating BioBERTpt, we found that the domain can influence the performance of BERT-based models, particularly for domains with unique characteristics such as medical. Our in-domain models achieved higher results for the average metrics. As shown in the statistical tests, the results were significant in relation to the BERT uncased model and the Portuguese BERT versions.

4.2 Effect of the contextualized language model

By providing a contextualized word representation and taking advantage of the transformer architecture, BERT-based language models have become a new paradigm for NLP tasks (Liu et al., 2020). The use of BERT-base models in our work had a positive impact on the results when compared to previous works with traditional machine learning algorithms and word embeddings for NER in Portuguese clinical text (de Souza et al., 2019; Lopes et al., 2019). For examples, de Souza et al. (2019) evaluated three groups of entities from the SemClinBr corpus using CRF, without any word embedding. AS shown in Table 3, they obtained for *Disorder* 0.65 of F1-score, compared to our 0.79; for *Procedure*, they achieved 0.60 compared to our 0.70 and for *Drug*, they achieved 0.42 compared to our 0.91. In the work of Lopes et al. (2019), where the authors used BiLSTM-CRF plus fastText on the CLINpt corpus, they achieved 0.759 with their in-domain model for micro F1-score, compared with 0.926 with BioBERTpt(clin), as we can see in Table 2. In general, all BERT-based models performed better in both corpora compared to the results of previous works. Indeed, the generic BERT models performed reasonably well on clinical NER tasks, probably because they were trained with a considerable amount of data, which embraced most of the semantics and syntax of the medical context.

4.3 Effect of language

Although the in-domain models performed better than out-of-domain models, the generic Portuguese BERT models (Souza et al., 2019) were outperformed by the BERT multilingual versions. The statistical analyses showed that the Portuguese BERT large version was significantly outperformed not only by the in-domain models, but also by the BERT multilingual uncased. This may be due to a local minima problem or the catastrophic forgetting. As shown by Xu *et al.*, catastrophic forgetting can happen during fine-tuning step, by overwriting previous knowledge of the model with new distinct knowledge, leading to a loss of information on lower layers (Xu et al., 2019). This may have occurred since the linguistic characteristics of clinical texts are very different from the Portuguese corpus used during pre-training phase of Portuguese BERT. As they were trained from a Web Corpus, collected using a search engine with random pairs of content words from 120,000 different Brazilian websites, maybe the new data in the fine-tuning process did not adequately represented the knowledge included in the original training data. The catastrophic forgetting probably occurred because the pre-trained model had to learn new input patterns, or needed to be adapted to a very distinct environment. On the other hand, for the multilingual model, this effect is less noticeable due to the larger and more generic corpus used for training.

4.4 Clinical relevance

The World Health Organization (WHO) recently released a list of 13 urgent health challenges the world will face over next decade, which highlights a range of issues, including health care equity and topping infectious diseases (WHO). To face these challenges, access to quality health information is essential, specially considering the information provided only in EHR's clinical narratives.

The BERT-based models proposed in this study and publicly released will support clinical NLP tasks for Portuguese, a language with relative lower resources, in particular in the health domain. Extracting structured information from a large amount of available clinical documents can provide health care assistance and help in the clinical decision-

making process, supporting other biomedical tasks and contributing to the urgent health challenges for the next decade [4].

5 Conclusions and future work

We proposed a new publicly available Portuguese BERT-based model to support clinical and biomedical NLP tasks. Our NER experiments showed that, compared to out-of-domain contextual word embeddings, BioBERTpt reaches the state-of-the-art on the CLINpt corpus. Additionally, it has better performance for most entities analyzed on the SemClinBR corpus. Our preliminary results are aligned with previous results in other languages, evidencing that domain transfer learning can benefit clinical tasks, in a statistically significant way. In the future, we would like to explore larger transformers-based models in the clinical Portuguese domain and evaluate our model in different clinical NLP tasks, such as negation detection, summarization and de-identification.

Acknowledgments

This work is related to a project supported by the Leading House for the Latin American Region - Seed Money Grant (No.1922) - of the Centro Latinoamericano-Suizo de la Universidad de San Gallen CLS-HSG. The authors also would like to thank Fundação Araucária, CAPES (Brazilian Coordination for the Improvement of Higher Education Personnel) and CNPq (Brazilian National Council of Scientific and Technologic Development) for their support in this research.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[4] https://www.who.int/news-room/photo-story/photo-story-detail/urgent-health-challenges-for-the-next-decade

Jules Berman. 2002. Confidentiality issues for medical data miners. *Artificial intelligence in medicine*, 26:25–36.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLOS ONE*, 12(6):1–32.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. pages 328–339.

Lindsey Knake, Monika Ahuja, Erin McDonald, Kelli Ryckman, Nancy Weathers, Todd Burstain, John Dagle, Jeffrey Murray, and Prakash Nadkarni. 2016. Quality of ehr data extractions for studies of preterm birth in a tertiary care center: Guidelines for obtaining reliable data. *BMC Pediatrics*, 16.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. pages 260–270.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. *JMIR Medical Informatics*, 7.

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *ArXiv*, abs/2003.07278.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in Portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Lucas Oliveira, Yohan Gumiel, Lilian Cintho, Sadid Hasan, Deborah Carvalho, Claudia Moro, and Arnon Santos. 2019. Learning portuguese clinical word embeddings: a multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task.

Lucas Oliveira, Ana Peters, Adalniza Silva, Caroline Gebeluca, Yohan Gumiel, Lilian Cintho, Deborah Carvalho, Sadid Hasan, and Claudia Moro. 2020. Semclinbr – a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Jeffrey Pennington, Richard Socher, and Christoper Manning. 2014. Glove: Global vectors for word representation. volume 14, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Sampo Pyysalo, F Ginter, Hans Moen, T Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine*.

Daniel Ranti, Katie Hanss, Shan Zhao, Varun Arvind, Joseph Titano, Anthony Costa, and Eric Oermann. 2020. The utility of general domain transfer learning for medical language tasks.

Mohammad S. Sorower. 2010. A literature survey on algorithms for multi-label learning.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf.

João Vitor de Souza, Yohan Gumiel, Lucas Emanuel Oliveira, and Claudia Maria Moro. 2019. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In *Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 318–323, Porto Alegre, RS, Brasil. SBC.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*.

WHO. World health organization.

Y. Xu, X. Zhong, A. Yepes, and J. Lau. 2019. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension.