J. Health Inform. 2020 Número Especial SBIS - Dezembro: 366-72

Artigo Original

JHI JOURNAL OF HEALTH INFORMATICS

CBIS'20
7 a 11 de Dezembro
Foz do Iguaçú - Paraná
Evento Virtual

A Multilabel Approach to Portuguese Clinical Named Entity Recognition

Uma Abordagem Multirrótulo para Reconhecimento de Entidade Nomeada Clínica em Português

Enfoque Multi-Etiqueta para el Reconocimiento de Entidad Nombrada Clínica en Portugués

João Vitor Andrioli de Souza[1], Elisa Terumi Rubel Schneider[2], Josilaine Oliveira Cezar[3], Lucas Emanuel Silva e Oliveira[4], Yohan Bonescki Gumiel[5], Emerson Cabrera Paraiso[6], Douglas Teodoro[7], Claudia Maria Cabral Moro Barra[8]

Keywords: Clinical Named Entity Recognition; Label Powerset; BERT

ABSTRACT

Objectives: Clinical Named Entity Recognition is a critical Natural Language Processing task, as it could support biomedical research and healthcare systems. While most extracted clinical entities are based on single-label concepts, it is very common in the clinical domain entities with more than one semantic category simultaneously. This work proposes BERT-based models to support multilabel clinical named entity recognition in the Portuguese language. **Methods:** For the experiment, we used the Label Powerset method applied to the multilabel corpus SemClinBr. **Results:** We compare our results with a Conditional Random Fields baseline, reaching +2.1 in precision, +11.2 in recall, and +7.4 in F1 with a clinical-biomedical BERT model (BioBERTpt). **Conclusion:** We achieved higher results for both exact and partial metrics, contributing to the multilabel semantic processing of clinical narratives in Portuguese.

Descritores: Reconhecimento de Entidade Nomeada Clínica; Label Powerset; BERT

RESUMO

Objetivos: O Reconhecimento de Entidade Nomeada Clínico é uma tarefa crítica do Processamento de Linguagem Natural, uma vez que apoia a pesquisa biomédica e os sistemas de saúde. Embora a maioria das entidades clínicas extraídas seja baseada em conceitos de rótulo único, é muito comum no domínio clínico entidades com mais de uma categoria semântica simultaneamente. Neste trabalho, propomos modelos baseados em BERT para suportar o reconhecimento de entidade nomeada clínico multirrótulo na língua portuguesa. **Métodos:** Para o experimento, utilizamos o corpus multirrótulo SemClinBr com o método Label Powerset. **Resultados:** Comparamos nossos resultados com o baseline Campos Aleatórios Condicionais, atingindo +2,1 em precisão, +11,2 em recall e +7,4 em F1 com um modelo clínico-biomédico de BERT (BioBERTpt). **Conclusão:** Obtivemos resultados superiores para as métricas exatas e parciais, o que contribui para o processamento semântico multirrótulo de narrativas clínicas em português.

Descriptores: Reconocimiento de Entidad Nombrada Clínica; Label Powerset; BERT

RESUMEN

Metas: Reconocimiento de Entidades Nombradas Clínico es una tarea fundamental del procesamiento del lenguaje natural, ya que apoya la investigación biomédica y los sistemas de salud. Aunque la mayoría de las entidades clínicas extraídas se basan en conceptos de etiqueta única, es muy común en el dominio clínico tener entidades con más de una categoría semántica simultáneamente. En este trabajo, proponemos modelos basados en BERT para apoyar el reconocimiento de entidad clínica multi-etiqueta en lengua portuguesa. **Métodos:** Para el experimento, usamos el corpus de múltiples etiquetas SemClinBr con el método Label Powerset. **Resultados:** Comparamos nuestros resultados con la línea de base de los Campos Aleatorios Condicionales, alcanzando +2,1 en precisión, +11,2 en recuerdo y +7,4 en F1 con un modelo BERT clínico-biomédico (BioBERTpt). **Conclusión:** Obtuvimos resultados superiores para las métricas exactas y parciales, lo que contribuye al procesamiento semántico de múltiples etiquetas de las narrativas clínicas en portugués.

[1] *Master's student in the Graduate Program in Health Technology - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*
[2] *Master in Bioinformatics - Federal University of Paraná (UFPR) - Curitiba (PR), Brasil.*
[3] *Librarian. Master's student in the Graduate Program in Health Technology - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*
[4] *Full Professor at Polytechnic School - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*
[5] *Doctor of Health Technology - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*
[6] *Full Professor at Graduate Program of Informatic - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*
[7] *Assistant Professor at Haute École de Gestion de Genève - Haute École Spécialisée de Suisse Occidentale (HES-SO) - Genève, Switzerland.*
[8] *Full Professor at Health Technology Graduate Program - Pontifical Catholic University of Paraná (PUCPR) - Curitiba (PR), Brasil.*

Autor Coorespondente: **João Vitor Andrioli de Souza**
e-mail: **joao.vitor.andrioli@gmail.com**

INTRODUCTION

Access to information is essential to offer quality healthcare assistance and to face the challenges of urgent health issues released by the World Health Organization (WHO). The COVID-19 pandemic has evidenced the importance of identifying information and relations between medical concepts to share among international researchers. i.e., to establish the best treatment.

Most of the valuable information in the Electronic Health Records (EHR) is available as free text, susceptible to grammatical errors, lack of structure, use of many acronyms, and jargon. These characteristics make the automatic extraction of clinical concepts a challenging task. Natural Language Processing (NLP) and Machine Learning (ML) tasks, such as Named Entity Recognition (NER), are widely used to process EHR data in both clinical practice and biomedical research (such as clinical trials and pharmacovigilance).

Traditionally, the NER task is defined as the function of, given a sequence of words, return for every word or term an entity type from a predefined category list. This leads to the premise that an entity has only one semantic type[1]. However, in the clinical domain, a named entity (or clinical concept) often has more than one semantic type, resulting in ambiguity or loss of information when labeling the entity with just one class. For example, in the sentence *"The patient received insulin"*, the term "*insulin"* can both belong to a hormone-like entity and protein-like entity simultaneously. Therefore, works with multilabel classification methods in NER can minimize ambiguity and increase the prediction performance since the classifier will be able to assign more than one semantic category within all possible labels for each concept extracted[2]. The multilabel classification is performed by transforming the problem into one or more single-label problems or adapting a classifier to handle multilabel data.

The adaptation of algorithms to handle multilabel data are costly to build and may require extensive parameter tuning, while the transformation-based approaches are easier to manipulate and reproduce, beneficial as a first approach towards multilabel problems.

The commonly used transformation-based approaches for multilabel classification are methods that use multiple classifiers, such as Binary Relevance (BR) and Classifier Chains (CC), and methods that transform the labels by grouping, eliminating, filtering, or converting multilabel instances. An example is Label Powerset (LP), which groups multiple labels to single-label strings.

The BR and CC methods use an ensemble of classifiers, where each classifier predicts only a specific class. In the BR method, all the classifiers' output is joined, resulting in a multilabel output, similar to the one-against-all in a multiclass problem. The CC method uses a chain of classifiers in which the output of a classifier is an input attribute for the next classifier. The main limitation of these methods is the large number of classifiers produced when the dataset has many labels.

The LP method assumes that the labels are dependent, combining each set of labels into a new single label of a multiclass problem with *k* classes, where *k* is the number of possible combinations of labels. As we can see in Figure 1, for each instance X of the dataset, it is created a new unique class for each combination of the instance labels Y so that the task can be executed as a single-label classification. Although it considers correlations between classes, this method has two main limitations[3]: the number of labels generated could lead to a combinatorial explosion, and some label combinations can have very few positive examples, resulting in data imbalance problems. It can cause low results for these specific classes, however, as they do not have high weight, the overall results may not be affected. There are other methods of adaptation to a single-label problem, such as the random k-labelsets (RAKEL) algorithm[4], a set of LP multiclass classifiers, but it also has the complexity of creating many classifiers and for this reason was not selected in our work.

Although multi-types entities are common in the clinical domain, there is still little research on this topic, compared to traditional single-label NER. Furthermore, there is a lack of fundamental computational tools and resources for textual information extraction on Portuguese clinical narratives. The creation of models for Portuguese NER can support numerous important tasks and consequently advance biomedical research, evolve the clinical practice, and fill the gap of semantic algorithms for processing clinical narratives in Portuguese.

This paper presents our contribution to recognizing named entities in the clinical domain using multilabel transformation methods and contextual word embeddings, aiming to reduce the ambiguity and loss of information in the labeling process. We performed a novel experiment with the LP method in a multilabel NER clinical corpus.

Related Work

There are many studies using models based on Neural Networks for NER tasks applied to clinical texts. Some of them consider entities from standard terminologies, such as SNOMED-CT and UMLS (Unified Medical Language System).

Supervised learning is the main approach for the NER task, containing resource-based machine learning algorithms and deep learning approaches that have reached state-of-the-art in various corpora for NLP tasks[5].

X	Y1	Y2	Y3	Yk	New class
x1	1	0	0	1	1001
x2	1	1	0	0	1100
x3	0	1	0	0	0100
x4	0	1	1	0	0110
xn	1	0	1	1	1011

Figure 1 - LP method: each set of labels Yk of an instance Xn is transformed into a new class.

The Conditional Random Fields (CRF) is a resource-based supervised learning algorithm for sequence labeling, usually uses resources from different linguistic levels, including orthographic information (e.g., capital letters, prefix, suffix), syntactic information, n-grams words, and also information semantics such as UMLS unique concept identifiers. Among deep learning architectures, Transformer[6] has proven to capture better the global dependencies of input texts and to have greater power of word representation[7]. Models based on this architecture achieved better performance in several learning tasks such as automatic translation and text generation[8]. An example of a Transformer-based model is Bidirectional Encoder Representations from Transformer (BERT). BERT achieved state-of-the-art in various NLP tasks, and several models are based on its architecture[9]. In Table 1, we present some characteristics of clinical NER studies.

The characteristics of the clinical domain often influence the performance of the contextual word embeddings. Models as BioBERT[17], ClinicalBERT[7], EhrBERT[21], and BioBERTpt[14], when trained with data from the clinical and biomedical domain, surpasses the results of the general BERT models. The authors of BioBERT state that the system achieves next-generation scores in the NER task.

METHODS

In this section, we explain the corpus used in our experiment, the LP settings as the entities filtering and grouping performed to adapt the original corpus, the BERT-based models used, and the evaluation criteria.

Corpus

For our multilabel NER experiment, we used the SemClinBr[22], a semantically annotated corpus for the Brazilian Portuguese clinical NLP tasks. This corpus has the characteristic of being multi-type, which means that each entity can have more than one associated label. SemClinBr contains 1,000 clinical notes from Nephrology, Cardiology, and Endocrinology areas, with 10,000 sentences and 147,164 tokens, of which 16,315 are unique tokens. The corpus was manually labeled by a group of healthcare specialists with 89 UMLS semantic types, which can be automatically mapped and grouped using the UMLS semantic concept tree. The UMLS semantic network, as a metathesaurus, categorizes the types and semantic relations of health and biomedical concepts[23].

The NER corpora are usually annotated using labeling schemes such as IO, IOB2, and IOBES[1]. The need for using these schemes is to characterize the taxonomy of the words and separate occurrences of two different objects with the same semantics. In our work, we adopted the IO scheme, once it performs better with the LP method, by reducing the number of new classes generated.

Label Powerset settings

We decided to use the LP method since applying BR, CC, or RAKEL would significantly increase the computational complexity, creating many partitions. BERT-based models are one of the most computationally expensive components and although models trained as LP deal with more classes, the change in complexity is negligible compared to training multiple BERT-based models.

Given the increased number of classes and the class unbalance created by the LP method, we adopted strategies to deal with low occurrence entity-tags and reduce the number of entity-tags using filters and grouping entities during LP transformation. Firstly, we grouped the entities using the UMLS semantic concepts, where each specific concept was mapped to a more generic one. Thus, the granularity is decreased and the entities, well generalized.

Table 1 - Summary of the methods, datasets, and languages for clinical NER.

Author(s)	Date	Method(s)	Dataset(s) Used	Language(s)
Lopes, Teixeira e Oliveira[10]	2020	BiLSTM+CRF	3678 clinical texts	Portuguese (Portugal)
Santos et al.[11]	2019	BiLSTM+CRF	HAREM	Portuguese (Brazil)
Souza, Nogueira e Lotufo[12]	2020	BERT+CRF	HAREM I and brWaC	Portuguese (Brazil)
Souza et al.[13]	2019	CRF	SemClinBr	Portuguese (Brazil)
Schneider et al.[14]	2020	BioBERTpt	SemClinBr	Portuguese (Brazil and Portugal)
Huang, Altosaar e Ranganath[15]	2019	ClinicalBERT	MIMIC III	English
Alsentzer et al.[7]	2019	ClinicalBERT and BioBERT	MIMIC III, MedNLI and i2b2 2010 & 2012	English
Miftahutdinov, Alimova e Tutubalina[16]	2020	BERT and LSTM+CRF	From 4 clinical datasets	English and Russian
Lee et al.[17]	2020	BERT base and BioBERT	From PubMed, PMC, Wikipedia and BookCorpus	English
Sun, Yang[18]	2019	Multilingual BERT and BioBERT	PharmaCoNER	Spanish
Ji, Wei e Xu[19]	2020	BERT and BioBERT and ClinicalBERT	298 clinical notes	English
Wei et al.[20]	2020	BERT and FT-BERT and FC-BERT	n2c2 and i2b2	English
Li et al.[21]	2019	EhrBERT	MADE, NCBI and CDR	English

[1] https://repositorio-aberto.up.pt/bitstream/10216/114087/2/277689.pdf

The initial grouping resulted in 13 labels (from the original 89 labels) and 94 labels after the LP transformation method (from 424, as we will see below). Secondly, we transformed the multilabel to single-label using the LP method, i.e., we generated new classes by combining the set of labels from each instance. We obtained a number of entity-tags using IOBES, IOB2, and IO tagging formats 4~6 times greater than the average label size: the transformation with the IOBES tagging format created 1,073 entity tags; with the IOB2, 704; and with the IO, 424 entity tags. To minimize the problem of data unbalance, we reduced the annotation scheme to the IO (inside-outside) tagging format since it is the simplest format with the lowest number of entities. Thirdly, we filtered entity sets with few occurrences and replaced them with the closest subset with the highest occurrence (e.g. {Disorder, Phenomena} -> {Disorder}). If the set does not have a valid subset, then the label set was not used. This action decreased the number of entity-tags while preserving the main entities of the instance and assured that the tags have a minimum occurrence, helping to deal with the class unbalance.

Figure 2 shows a summary of the adopted process. Firstly, the corpus was grouped using the UMLS concepts, remaining 13 groups of entities from the original 89. Secondly, we applied the LP transformation, generating new classes from a combination of instance labels, resulting in 424 entities and 94 groups of entities, and then we filtered the labels with lower occurrence. These steps resulted in three different corpora for our experiments, henceforth named: Groups, Filtered Groups, and Filtered Entities, with 94, 30, and 50 classes, respectively. The NER task was executed with these new single-label corpora, using BERT-based models for the classification.

We analyzed the new labels using the cardinality and density metrics proposed by[24]. The cardinality metric represents the mean number of labels for all words. According to Table 2, every word in the original corpus was annotated on average with 1.168 labels. The density metric is similar to the cardinality metric, however, it takes into account the number of classes. As it is only 0.013 for the SemClinBr corpus, this means it has many classes.

Models

Models based on the Transformer architecture[6], such as BERT, have become a new paradigm for NLP tasks, by providing contextualized word representation using the attention mechanism. BERT also provides the fine-tuning process, where the model can be re-trained for a specific task[9].

In our experiment, we used the following contextual models: BERT multilingual uncased, BERT multilingual cased, BioBERTpt(all), BioBERTpt(clin), BioBERTpt(bio), Portuguese BERT large, and Portuguese BERT base[II]. The BioBERTpt models are fresh clinical-biomedical BERT-based models for Portuguese[14], fine-tuned from BERT-multilingual-cased and free available in the HAILab repository[III]. BioBERTpt(clin) was trained with Portuguese clinical notes, BioBERTpt(bio) was trained with scientific-biomedical abstracts, and BioBERTpt(all) represents the full version. The Portuguese BERT models[12] are BERT-models pre-trained on a large Portuguese corpus (BrWaC, Brazilian Web as Corpus), out of the clinical domain.

As a baseline, we selected the CRF with the same

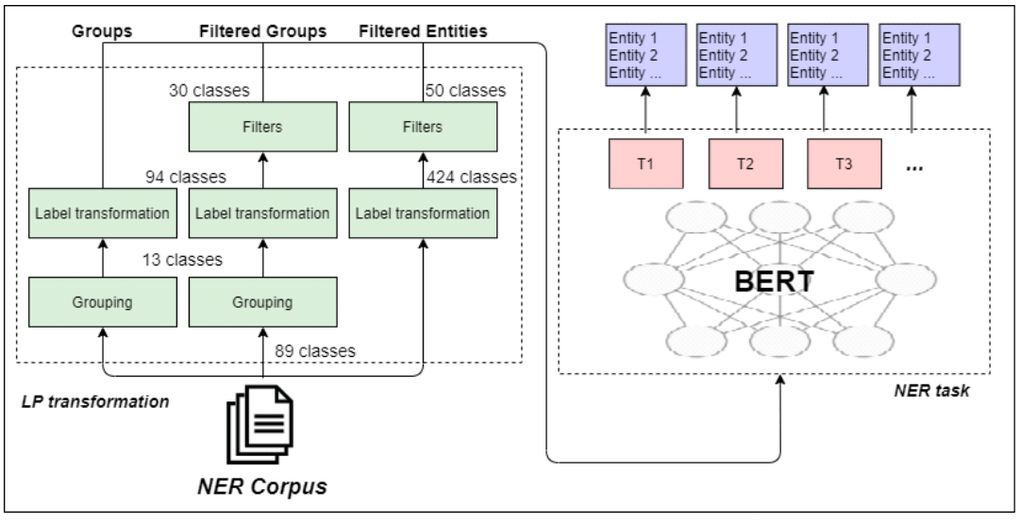

Figure 2 - Summary of the process with label transformation and entity classification.

Table 2 - Characteristics of the original and the generated corpora.

	Original Corpus	Filtered Entities	Groups	Filtered Groups
Entity Quantity	89	50	94	30
Annotated Tokens	66,422	64,715	66,422	64,727
Annotated Unique Tokens	12,025	11,594	12,025	11,970
Label Cardinality	1.168	1.0	1.0	1.0
Label Density	0.013	0.019	0.010	0.032

[II] https://huggingface.co/models
[III] https://github.com/HAILab-PUCPR/BioBERTpt

parameters selected in our previous work[13], which used the same corpus of this study, with the BR method.

We used the PyTorch implementation of Hugging Face API, adding at the top of the BERT models a token level classifier. This linear layer uses the last hidden state of a sequence, performing the fine-tuning process for the NER task. The NER experiments were processed with this configuration: AdamW optimizer, 0.01 as weight decay, 4 as batch size, 256 as maximum length, 3e-5 as learning rate, 10 as maximum epoch, and 0.1 for warmup proportion.

Evaluation Setup

We performed our experiment with holdout using a corpus split of 60% for training, 20% for validation, and 20% for test. We decided to use all the BERT models and the CRF on the group selection and select only the best approach - in this case, BioBERTpt (all) - to compare with our three selections (groups, filtered groups, and filtered entities).

Since the LP method transforms multilabel to single-label problems, the evaluation metrics can be the traditional single-label ones for precision, recall, and micro F1-score. As we can change the single-label back to multilabel and handle the evaluation as a multilabel problem, we also evaluated the models with the multilabel metrics of precision, recall and micro F1-score which accounts for partial correctness[25].

RESULTS

Table 3 shows the results for both single-label (exact match) and multilabel precision, recall, and F1-score for the CRF baseline and all BERT-based models, using the groups selection.

Figure 3 shows the F1-score by the number of occurrences that a class had in the test set of Groups selection (with Pearson correlation of 0.53 and Spearman correlation of 0.84).

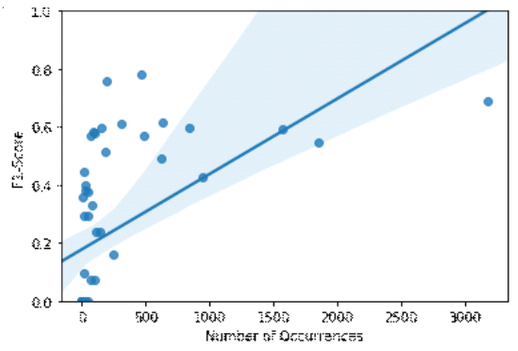

Figure 3 - F1-score by the number of occurrences that a class had (Groups selection).

DISCUSSION

As expected, the results of the in-domain models (BioBERTpt) were higher since we are working with a clinical corpus. Figure 3 corroborates our hypothesis that entities with lower occurrences have lower results, justifying why we have filtered and mapped entities with lower occurrences.

Training a model with LP transformation leads to a situation where class sets with the same entities (e.g., Abbreviation+Disorder and Abbreviation) are trained as different instances. Thus, even if the model predicts partially correct, the loss function will fully penalize the model, forcing it to search in the vector space for better distinctions between these classes in order to predict the exact class set. It creates a tendency to separate the classes

Table 3 - Evaluation for all methods on the Groups selection using the exact and partial match versions of precision, recall and F1-Score.

Method	Precision (%)	Recall (%)	F1-Score (%)	Multilabel Precision (%)	Multilabel Recall (%)	Multilabel F1-Score (%)
Conditional Random Fields (CRF) *	54.5	45.0	48.7	71.3	71.3	71.3
BERT base multilingual uncased	53.1 (-1.4)	55.8 (+10,8)	54.4 (+5.70)	73.7 (+2.4)	73.9 (+2.6)	73.8 (+2.5)
BERT base multilingual cased	51.8 (-2.7)	53.8 (+8,8)	52.7 (+4.00)	74.0 (+2.7)	74.1 (+2.8)	74.1 (+2.8)
Portuguese BERT base	54.9 (+0,4)	53.6 (+8.6)	54.2 (+5.5)	74.7 (+3.4)	74.6 (+3.3)	74.7 (+3.4)
Portuguese BERT large	50.0 (-4.50)	54.5 (+9.5)	52.1 (+3.4)	74.1 (+2.8)	74.1 (+2.8)	74.1 (+2.8)
BioBERTpt(bio)	51.8 (-2.7)	**56.2** (+11.2)	53.9 (+5.2)	74.5 (+3.2)	74.4 (+3.1)	74.5 (+3.3)
BioBERTpt(clin)	55.7 (+1,20)	54.6 (+9.6)	55.1 (+6.4)	**75.2** (+3.9)	**75.3** (+4.0)	**75.2** (+3.9)
BioBERTpt(all)	**56.6** (+2.1)	55.7 (+10.7)	**56.1** (+7.4)	74.6 (+3.3)	75.0 (+3.7)	74.8 (+3.5)

Table 4 - Results for all generated corpora using BioBERTpt(all).

Method	Precision (%)	Recall (%)	F1-Score (%)	Multilabel Precision (%)	Multilabel Recall (%)	Multilabel F1-Score (%)
Groups	**56.6**	55.7	56.1	74.6	75.0	74.8
Filtered Groups	55.2	**58.2**	**56.7**	**75.4**	**75.2**	**75.3**
Filtered Entities	50.7	54.1	52.3	72.1	71.4	71.7

more than necessary, which may lead the model to lose generalization.

As shown in Table 4, the BioBERTpt(all) had the best normal F1-score, and BioBERTpt(clin) had the best multilabel F1-score, suggesting that BioBERTpt(clin) has better generalization capability than BioBERTpt(all). Even though BioBERTpt(clin) predicted the exact class set wrongly more often than BioBERTpt(all), it maintained more representative feature weights, inducing the model to predict the overall subset of classes with higher accuracy than BioBERTpt(all).

Grouping the entities was beneficial since it decreased substantially the number of classes created after applying the LP method. Grouping also increased the performance, although it is not recommended if the task needs granular and specific classes.

Filtering and mapping the entities also increased the performance. Still, this technique is only preferable if the entities with lower occurrence are not significant to the task, which applies to our case. If the corpus has a high label density and fewer classes, then filtering might not be necessary.

The NER tagging schemes (as IOBES, IOB2, and IO) affected the number of classes created after applying the LP transformation. The IO scheme loses minimal information on cases where two subsequent words are different objects of the same entity type, making both words a single object. This IO problem is only observable on specific entities. We hypothesize that one possible way to remediate it is by using IOB2 on fewer manually selected entities and using the remaining ones as IO, however, this was not explored. As this is uncommon, especially on datasets with low label density, such as SemClinBr, and high label cardinality, this makes IO a feasible tagging scheme for LP.

The use of BERT-based models in this work positively affected our results, taking advantage of the Transformer architecture and the BERT fine-tuning process. Although the contextual word embeddings require a minimum memory size and GPU to be used, the LP method creates only one model for prediction to all classes, unlike other methods such as BR, CC, and RAKEL that create many models, making it a feasible multilabel method for the clinical area.

CONCLUSION

In this work, we performed a novelty multilabel approach in a Portuguese clinical NER corpus, using BERT-based models and LP techniques. Our results showed that the LP method benefits multilabel NER problems, which, combined with BERT-based models, leads to interesting results. All BERT models, even the out-of-domain, improved the results comparing to the CRF baseline, with BioBERTpt achieving the highest results. In recall and F1 metrics, all models trained in this work had better results than the previous work. We analyzed the results with single and multilabel metrics (precision, recall, and F1) and achieved better results for both scenarios. We expect to contribute to the clinical NER for the Portuguese language. Moreover, our experiments can be replicated in other languages and domains as well. In the future, we would like to adapt a single-label algorithm to deal with multilabel problems, explore data augmentation techniques and experiment with other contextual models based on Transformer.

ACKNOWLEDGMENT

The authors 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.

REFERÊNCIAS

1. DAI X. Recognizing complex entity mentions: a review and future directions. Proceedings of ACL 2018, Student Research Workshop. 2018 Jul 37-44. Melbourne, Australia: Association for Computational Linguistics (ACL); 2018. Available from: https://www.aclweb.org/anthology/P18-3006/
2. Lin W, Ji D, Lu Y. Disorder recognition in clinical texts using multilabel structured SVM [internet]. BMC Bioinformatics. 2017 Jan 31;18(1):75. doi: 10.1186/s12859-017-1476-4.
3. Zhang M, Zhou Z. A Review on Multi-Label learning algorithms [internet]. BMC Bioinformatics. 2014 Aug; vol. 26, no. 8, pp. 1819-1837, doi: 10.1109/TKDE.2013.39.
4. Tsoumakas G, Katakis I, Vlahavas I. Random k-Labelsets for Multilabel classification [internet]. IEEE Transactions on Knowledge and Data Engineering. 2011 Jul;23(7):1079-1089. doi: 10.1109/TKDE.2010.164.
5. Li J, Sun A, Han J, Li C. A survey on Deep Learning for Named Entity Recognition [internet]. IEEE Transactions on Knowledge and Data Engineering. 2020 Mar 17;1-1. doi: 10.1109/TKDE.2020.2981314.
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A et al. Attention is all you need [internet]. Polosukhin. 2017 Dec 6;5. Available from: https://arxiv.org/abs/1706.03762
7. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T et al. Publicly available clinical BERT embeddings [internet]. 2019 Jun 20. Available from: https://arxiv.org/pdf/1904.03323.pdf
8. Liu Q, Kusner M, Blunson P. A survey on contextual embeddings [internet]. 2020 Apr 13;2. Available from: https://arxiv.org/abs/2003.07278
9. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. Available from: https://arxiv.org/abs/1810.04805
10. Lopes F, Teixeira C, Gonçalo O H. Comparing different methods for Named Entity Recognition in Portuguese Neurology text [internet]. J Med Syst. 2020 Feb 28;44(4):77. doi: 10.1007/s10916-020-1542-8.
11. Santos J, Terra, J, Consoli, B S, Vieira, R. Multidomain contextual embeddings for Named Entity Recognition. Proceedings of the Iberian Languages Evaluation Forum (IberLEF); 2019. Bilbao, Espanha: CEUR-WS; 2019 Set:2421. 434-441 Available from:https://www.semanticscholar.org/paper/Multidomain-Contextual-Embeddings-for-Named-Entity-Santos-Terra/c655cb5b49b6afdd72c8c72c096f223c412edff5
12. Souza F, Nogueira R, Lotufo R. Portuguese Named Entity Recognition using BERT-CRF [internet]. 2020 Feb 27. Available from: https://arxiv.org/pdf/1909.10649.pdf
13. Souza JVA, Gumiel YB, Oliveira LES, Moro CMC.Named

Entity Recognition for clinical portuguese corpus with Conditional Random Fields and Semantic Groups. Anais do 19 Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), 2019, Niterói. Porto Alegre: Sociedade Brasileira de Computação, 2019 Abr 18:318-323. DOI: https://doi.org/10.5753/sbcas.2019.6269

14. Schneider ETR, Souza JVA, Knafou J, Oliveira LES, Copara J, Gumiel YB. BioBERTpt: a portuguese neural language model for clinical Named Entity Recognition. Proceedings EMNLP. 2020. p65-72. https://www.aclweb.org/anthology/2020.clinicalnlp-1.7
15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. 2019 Apr 11;2. Available from: https://arxiv.org/abs/1904.05342
16. Miftahutdinov Z, Alimova I, Tutubalina E. On biomedical Named Entity Recognition: experiments in interlingual transfer for clinical and social media texts. In: Jose J et al. editors. Advances in information retrieval: ECIR 2020. Lecture Notes in Computer Science: Springer; 2020. p. 12036. Available from: https://doi.org/10.1007/978-3-030-45442-5_35
17. Lee J et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [internet]. Bioinformatics. 2020;36(4):1234–1240. Available from: https://doi.org/10.1093/bioinformatics/btz682
18. Sun C, Yang Z. Transfer learning in biomedical Named Entity Recognition: an evaluation of BERT in the PharmaCoNER task. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks; 2019; Hong Kong, China: Association for Computational Linguistics; 2019 Nov:100-104. doi: 10.18653/v1/D19-5715
19. Ji Z, Wei Q, Xu H. BERT-based ranking for biomedical Entity Normalization. AMIA Jt Summits Transl Sci Proc. 2020 May 30:269-277. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233044/
20. Wei Q, Ji Z, Si Y, Du J, Wang J, Tiryaki F. Relation extraction from clinical narratives using pre-trained language models. AMIA Annu Symp Proc. 2020 Mar 4; 2019:1236-1245. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153059/
21. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based models on Large-Scale Electronic Health Record Notes: an empirical study. JMIR Med Inform. 2019; 7(3):e14830. doi: 10.2196/14830.
22. Oliveira LES, Peters AC, Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM. SemClinBr: a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks [internet]. 2020 Jan 27. Available from: https://arxiv.org/abs/2001.10071
23. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. doi: 10.1093/nar/gkh061.
24. Tsoumakas G, Katakis I. Multi-Label classification: an overview. International Journal of Data Warehousing and Mining. 2007;3(3):1-13. Available from: https://econpapers.repec.org/article/iggjdwm00/v_3a3_3ay_3a2007_3ai_3a3_3ap_3a1-13.htm
25. Godbole S, Sarawagi G. Discriminative Methods for Multi-labeled Classification. In: Dai H, Srikant R, Zhang C editors. Advances in Knowledge Discovery and Data Mining: PAKDD 2004. p.22-30. Available from: https://link.springer.com/chapter/10.1007/978-3-540-24775-3_5