Pontifícia Universidade Católica do Paraná
Programa de Pós-Graduação em Informática

BioNestedNER: A Two-Phase Method For Recognizing Nested, Discontinuous, And Multi-Type Named Entities Using Transformers And Multi-Label CRF

Elisa Terumi Rubel Schneider

Supervisor

Prof. Dr. Emerson Cabrera Paraiso

Co-Supervisor

Prof. Dr. Cláudia Maria Cabral Moro Barra

Curitiba
2023

Pontifícia Universidade Católica do Paraná
Programa de Pós-Graduação em Informática

BioNestedNER: A Two-Phase Method For Recognizing Nested, Discontinuous, And Multi-Type Named Entities Using Transformers And Multi-Label CRF

Elisa Terumi Rubel Schneider

Thesis Project presented to the *Programa de Pós-Graduação em Informática* as a partial requirement for the degree of Doctor in Informatics.

Major Field: Computer Science

Supervisor: Prof. Dr. Emerson Cabrera Paraiso

Co-supervisor: Prof. Dr. Cláudia Maria Cabral Moro Barra

Curitiba
2023

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Sônia Maria Magalhães da Silva – CRB 9/1191

Schneider, Elisa Terumi Rubel
S358b
2023
BioNestedNer : a two-plase method for recognizing nested, discontinuous, and multi-type named entities using transformers and multi-label CRF / Elisa Terumi Rubel Schneider ; supervisor: Emerson Cabrera Paraiso ; co-supervisor: Cláudia Maria Cabral Moro Barra. – 2023
xvi, 154 f. ; il. : 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2023
Bibliografia: f. 139-154

1. Processamento de linguagem natural (Computação). 2. Processamento de eventos (Ciência da computação). 3. Aprendizado do computador. 4. Informática. I. Paraiso, Emerson Cabrera. II. Barra, Cláudia Maria Cabral Moro, 1969-. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD. 20. ed. – 004

SCIENTIA, VITA ET FIDES
PUCPR
GRUPO MARISTA

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 01 de setembro de 2023.

74-2023

DECLARAÇÃO

Declaro para os devidos fins, que **ELISA TERUMI RUBEL SCHNEIDER** defendeu a tese intitulada **“BioNestedNER: A Two-Phase Method For Recognizing Nested, Discontinuous, And Multi-Type Named Entities Using Transformers And Multi-Label CRF”**, na área de concentração Ciência da Computação no dia 24 de julho de 2023, a qual foi aprovada.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade firmo a presente declaração.

gov.br
Documento assinado digitalmente
EMERSON CABRERA PARAISO
Data: 04/09/2023 15:17:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática

A mente que se abre a uma nova ideia jamais volta ao seu tamanho original. - Albert Einstein

Abstract

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that allows relevant information to be extracted from unstructured data. NER can serve as a foundation for other NLP tasks, such as relation and event extraction, and has applications in many fields. Although there are different methods and techniques for entity recognition, the traditional NER approach assumes that entities are continuous and non-overlapping, which is not always true in real-world scenarios. Nested and discontinuous entities are common in texts, such as in clinical and biomedical domains, where multiple entities can be nested within each other, and mentions can have gaps between them. Moreover, entities can have multiple types, making NER a multi-label classification problem. This thesis proposes BioNestedNER, a two-phase method for nested, discontinuous, and multi-type entity recognition in clinical and biomedical texts. Our method is formed by a) a Transformer-based model utilizing a Machine Reading Comprehension NER approach, where the NER task is formulated into a Question-Answering similar task (the mention of the entity is the answer to a question and the sentence, the paragraph), and b) a Conditional Random Field trained to address multi-label sequence labeling, particularly useful as nested entities can be handled as multi-type entities. In nine NER experiments using six corpora (in English and Portuguese), we evaluate our method in the clinical and biomedical domains, obtaining state-of-the-art results in the micro F1 score in six experiments. We also found that our method was more effective in identifying these complex entities than similar methods. We are also releasing a new clinical corpus in the Brazilian Portuguese language annotated with nested and discontinuous entities. This corpus is a new resource for developing and evaluating models that can handle the complexity of these entities and facilitate the advancement of tools and language models for clinical NER, which can significantly impact healthcare applications. Our proposed method provides a flexible and efficient NER solution that can handle nested, discontinuous, and multi-type entities. It can benefit many applications, including drug discovery, biomedical information retrieval, and clinical decision-making.

Key-words: natural language processing, named entity recognition, complex entities, Transformer architecture, machine learning, conditional random field, language models

Resumo

O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa fundamental no Processamento de Linguagem Natural (PLN) que permite extrair informações relevantes de dados não estruturados. O REN pode servir como base para outras tarefas de PLN, como extração de relações e eventos, e tem aplicações em muitos campos. Embora existam diferentes métodos e técnicas para o reconhecimento de entidades, a abordagem tradicional de REN assume que as entidades são contínuas e não se sobrepõem, o que nem sempre ocorre no mundo real. Entidades aninhadas e descontínuas são comuns em textos, como nos domínios clínico e biomédico, onde múltiplas entidades podem estar aninhadas umas nas outras, e menções podem ter lacunas. Além disso, as entidades podem pertencer a vários tipos, tornando o REN um problema de classificação multi-rótulo. Nesta tese, propomos BioNestedNER, um método de duas fases para o reconhecimento de entidades aninhadas, descontínuas e multi-tipo em textos clínicos e biomédicos. Nosso método é formado por a) um modelo baseado em *Transformer* que utiliza uma abordagem de Compreensão de Leitura de Máquina, no qual a tarefa de REN é formulada como uma tarefa semelhante a *Question-Answering* (a menção da entidade é a resposta a uma pergunta e a sentença, o parágrafo), e b) um modelo *Conditional Random Field* treinado para lidar com rotulação de sequência multirrótulo, particularmente útil uma vez que entidades aninhadas podem ser tratadas como entidades multi-tipo. Avaliamos nosso método nos domínios clínico e biomédico, em nove experimentos de REN usando seis *corpora* (em inglês e português), obtendo resultados estado-da-arte na métrica F1 micro em seis experimentos. Nosso método também foi mais eficaz em identificar essas entidades complexas em comparação com métodos semelhantes. Além disso, estamos disponibilizando um novo *corpus* clínico em língua portuguesa (do Brasil) anotado com entidades aninhadas e descontínuas. Esse *corpus* é um novo recurso para desenvolver e avaliar modelos que possam lidar com a complexidade dessas entidades, e facilitar o avanço de ferramentas e modelos de linguagem para REN clínico, o que pode impactar significativamente aplicações de saúde. Nosso método proposto oferece uma solução flexível e eficiente de REN que pode lidar com entidades aninhadas, descontínuas e multi-tipo. O método tem o potencial de beneficiar muitas aplicações, incluindo descoberta de medicamentos, recuperação de informações biomédicas e tomada de decisões clínicas.

Palavras-chave: processamento de linguagem natural, reconhecimento de entidades nomeadas, entidades complexas, arquitetura Transformer, aprendizagem de máquina, *conditional random field*, modelos de linguagem, textos clínicos.

Acknowledgements

I would like to thank my advisors, Emerson Cabrera Paraiso and Claudia Moro, who, during these years, have always placed enormous trust in my work. It was a great privilege to have you by my side on this journey.

I also thank my husband, Fernando Henrique Schneider, who has always supported me, and my children Nicole, Erick, and Kevin, who even without knowing it, gave me the strength to always move forward. Love you all.

I am very grateful to my mom, dad (in memorial), and my sisters, Fernanda and Paula, for all the care, learning, love, and friendship in all these years. I am also grateful for the support of all my family and my husband's family.

To my research colleagues and friends from HaiLab and PPGTS (PUCPR), I would like to thank for all support, in special Yohan, Lucas Oliveira, João Vitor, Lucas Ferro, Giovanni, Josilaine, Lilian, and Carol. And I also thank my colleagues from PPGIA (PUCPR), in special João Pedro, Eduardo Tieppo, Bruno, and Ricardo for the partnership.

I would like to thank Professor Paloma Martinez and colleagues from Universidad Carlos III de Madrid (UC3M), for allowing me to do research together in such a prestigious research group, where much knowledge was shared; as well as Professor Douglas Teodoro and colleagues, especially Julien Knafou and Jenny Copara, from the University of Geneva and HES-SO, for the interesting work we did together.

A special thanks to all professors at PUCPR, who contributed immensely to my learning, and to Celepar, the company I've been working for 15 years so far.

And finally, I want to thank CAPES for funding this research with a tax exemption and a "sandwich" doctoral scholarship, and PUCPR for being an important part of my professional qualification.

Contents

List of Figures

List of Tables

List of Charts

List of Algorithms

List of Acronyms

ACC Accuracy

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bidirectional Long Short-Term Memory

BPE Byte-Pair Encoding

BR Binary Relevance

CE Cross Entropy

CNN Convolutional Neural Network

CUDA Compute Unified Device Architecture

CRF Conditional Random Fields

DE Discontinuous Entity

EHR Electronic Health Records

ELMo Embeddings From Language Models

FN False Negatives

FP False Positives

GPT Generative Pre-trained Transformer

GPU Graphics Processing Unit

HMMs Hidden Markov Models

IAA Inter Annotator Agreement

ICD International Classification of Diseases

LP Label Powerset

ME Multi-type Entity

ML Machine Learning

MRC Machine Reading Comprehension

NE Nested Entity

NER Named Entity Recognition

NLP Natural Language Processing

POS Part-of-speech

QA Question Answering

RoBERTa Robustly Optimized BERT Pretraining Approach

SOTA State-Of-The-Art

SVM Support Vector Machines

RNN Recurrent Neural Networks

TP True Positives

TPE Time Per Epoch

UMLS Unified Medical Language System

WHO World Health Organization

1

Introduction

Natural Language Processing (NLP) is a field of study that focuses on the interaction between computers and humans using natural language, i.e., the written or spoken language that humans use to communicate. NLP tasks deal with the automatic processing and analysis of human language, enabling computers to understand, interpret, extract, generate, and utilize human language in a useful way. NLP techniques include language modeling, information extraction (as named entity recognition), semantic analysis, text classification, machine translation, and others. With the increasing use of electronic health record (EHR) systems, NLP techniques are widely used in the medical domain, extracting patients' valuable information and supporting decision-making related to the health area.

Named Entity Recognition (NER) is one of the most used NLP tasks, which allows the machine to acquire knowledge from unstructured texts by recognizing and identifying meaningful entities in text passages, such as persons, organizations, and locations. Usually, NER is used as support for other NLP tasks, like document summarization, question answering, and relationship extraction, among others. In the clinical domain, NER can identify medical concepts, such as diseases, symptoms, and drugs, providing a basis for other data analysis like predicting future clinical events and relation extraction between entities. The application of NER in the clinical domain can support clinical research, pharmacovigilance, diagnostic support, biomedical research, treatment customization, and others.

Traditionally, the NER task is defined as: given a sequence of tokens [1], return a list of tuples <*Is, Ie, t*>, each mentioning an entity. In this work, the "mention of the entity" is treated as a synonym for the word "entity", and the semantic groups that define the entity are called "entity types". *Is* and *Ie* represent the start and end indexes of the entity mentioned, respectively, and *t* is the entity type from a predefined category. Therefore, we have two premises: 1) a mention of an entity consists in a continuous sequence, whose words belong to the interval between *Is* and *Ie*, and 2) mentions do not overlap (DAI, 2018). This traditional entity is called a “flat entity”.

Besides flat entities, the authors of (DAI, 2018) define "complex entities" as entities that are nested, incorporated, overlapped, discontinuous, and/or multi-type. These entities do not adhere to the previously mentioned assumptions for the NER task. Such entities often contain valuable information for downstream tasks and are very common in clinical and biomedical texts (FINKEL; MANNING, 2009). For instance, various biological entities of interest are frequently composed of one another, such as proteins, genes, and chemical substances, forming nested or overlapping entities (WANG; LU, 2018), (ALEX; HADDOW; GROVER, 2007), (LU; ROTH, 2015). Discontinuous entities occur when entity mentions consist of non-sequential words in the text. An example from the GENIA corpus (KIM et al., 2003) is the expression "alpha and beta-globin", where "alpha-globin" is considered a discontinuous entity. Another scenario involves entities having multiple types simultaneously, referred to as "multi-type" entities. In this case, each mention can belong to more than one entity type, resembling a multi-label scenario. As an example, in the fictional sentence "The patient received insulin", the term "insulin" can be associated with hormone, pharmacologic substance, and protein-like entity types concurrently, a common situation in the SemClinBr corpus (OLIVEIRA et al., 2022)

Although the term “complex entities” has been recently used to reference an entity composed of multiple components, formed by any linguistic constituent (as titles of creative works), semantically ambiguous, or in some way, difficult to recognize (as in SemEval-2022 task 11 (MALMASI et al., 2022)), in this work, we refer to complex entities following the definition of (DAI, 2018), where these

[1] "Token" refers to a sequence of contiguous characters that represent a semantic unit in a text. To facilitate understanding, in this work, we consider “word” synonymous with “token”, although they are different in some situations. A comma, for example, is a token but not a word.

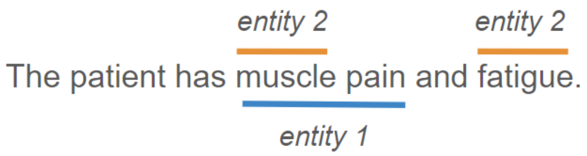


Figure 1.1: Fictitious examples of complex entities in a clinical text.

entities are formed by: a) nested and overlapping, b) discontinuous, and c) multi-type mentions.

Figure 1.1 presents examples of complex entities that can occur in clinical text. Entity 1 refers to "muscle pain", and entity 2, "muscle fatigue". The word "muscular" in both entities represents an example of mentions with overlapping, and the expression "muscle fatigue" is an example of a discontinuous entity.

As presented by (LI et al., 2022), four approaches are used to execute the traditional NER task: rule-based, unsupervised learning, resource-based supervised learning, and deep learning. Deep learning approaches have reached the state-of-the-art in various corpora for NLP tasks such as NER, discovering data representations with various levels of abstraction necessary for entity classification.

Among deep learning architectures, Transformer (VASWANI et al., 2017) proved to better capture the global dependencies of input texts in relation to other architectures, such as those based on recurrent neural networks (RNN), performing better in several learning tasks (LI et al., 2022). BERT (Bidirectional Encoder Representations from Transformers) is an example of a Transformer-based model, which reached state-of-the-art in 11 NLP tasks and inspired several models based on its architecture (DEVLIN et al., 2019).

1.1 Motivation

Although nested, discontinuous, and multi-type entities are common in several domains, such as clinical and biomedical, the traditional NER methods are not naturally prepared to deal with the characteristics of these complex entities, typically requiring the use of advanced NLP techniques and/or sophisticated applications (WANG et al., 2022).

Approaches to the recognition of complex entities involving incorporated and discontinuous entities generally present, according to (DAI, 2018):

- Lack of expressivity, e.g., in token-level approaches, usually the tagging schema presents some intrinsic restrictions; and/or
- Computational complexity, as in sentence-level approaches since the exhaustive enumeration of possible entities is exponential and relative to the length of the sentence.

Besides, most proposed methods do not consider recognizing all complex entities, generally focusing on nested entities and leaving the discontinuous and multi-type ones aside. Some NER methods have been adapted to recognize nested entities, violating the first premise of the NER task presented above, guaranteeing the identification of embedded entities. However, the second premise is still little studied and requires handling with discontinuous mentions, which are particularly challenging (DAI, 2018). Moreover, few works focus on recognizing multi-type entities, contributing to their high ambiguity in assigning the correct label to a mentioned span. Recognizing discontinuous and multi-type entities represents gaps in recognizing complex entities, a common situation in clinical and biomedical texts. The failure to identify complex entities can lead to the loss of relevant information. For instance, within the biomedical corpus GENIA (KIM et al., 2003), around 31.64% of entities are nested (CHEN et al., 2020). Neglecting these entities translates to disregarding a significant portion of potentially valuable information. Furthermore, in the context of the Rare Disease corpus (MARTíNEZ-DEMIGUEL et al., 2022), which includes clinical concepts, nested entities are very common in sign, disease, and rare disease entity types, often with overlapping mentions.

Another gap we identified is the small number of corpora annotated with complex entities in health sciences. To the best of our knowledge, the Portuguese language has no corpus with nested and discontinuous entities in the clinical domain.

This work proposes a two-phase method for recognizing nested, discontinuous, and multi-type entities. The first phase employs an MRC-based NER approach, also called QA-NER, where the NER task is framed as a question-answer machine reading comprehension task. This approach leverages the power of machine comprehension techniques to extract entities in a more accurate and context-aware manner. In the second phase, a Conditional Random Field (CRF) model is trained specifically to handle multi-label sequence labeling, handling the complexities of identifying entities that can have multiple labels or be part of nested structures. The final results were obtained by combining the outputs of the two different models.

In the MRC-based NER approach, proposed by (LI et al., 2020) and explored in the works of (ZHANG et al., 2020), (SHEN et al., 2022) and (BANERJEE et al., 2021), a model is trained similarly to a Question-Answering task, extracting the entities as answers spans to the question. In the same way as (BANERJEE et al., 2021), our method differs from (LI et al., 2020) and (SHEN et al., 2022) in the way the model is trained to return responses. While in the traditional QA task, the output refers to the index(es) of the entity(ies) found in the sentence, our method appropriates the process in which the NER task works, returning the output in a token-level way. We consider a hybrid task between QA and NER, presenting computational simplicity, flexibility, and reaching the state-of-the-art in some situations.

The proposed method addresses the two challenges highlighted by (DAI, 2018) in the recognition of complex entities. Firstly, it allows the utilization of the NER tagging scheme, such as IOBES or IOB2, since each entity found is a response to a specific query. Secondly, the method does not demand computational complexity like methods that exhaustively enumerate all regions in the text.

We further simplified the work of (BANERJEE et al., 2021) by removing the CNN module, since Transformer models already worked with word context, and added a layer to handle discontinuous and nested entities of the same type, in an end-to-end model. Additionally, we improved their method by adding a treatment to improve class imbalance, as this approach generates more "O" (non-entity) type tokens than normal NER. In our method, we also trained CRF models adapted to find multi-type entities, using both syntactic and semantic characteristics of the words and their context as input features. The final results are the union of the outputs generated by the QA-NER-based method with the results from the CRF, helping to increase recall by adding results from two different methods. Combining methods can improve coverage by leveraging the complementarity and redundancy of different approaches.

We also use the Transformer architecture (VASWANI et al., 2017), state-of-the-art to NLP tasks, and the fine-tuning technique to take advantage of the weights of a generic pre-trained base model on a massive amount of text data, fine-tuning on a specific downstream task with relatively few labels. As we focus on finding the complex entities in clinical and biomedical texts (but not restricted to), we call our method "BioNestedNER".

Since complex entities are common in health sciences texts, as verified by (FINKEL; MANNING, 2009) and (WANG; LU, 2018), we implement some experiments with our method (both with complex and flat entities), in clinical and biomedical domains, and in English and Portuguese languages. Also, to further validate our method and promote research in this domain, we annotated a small corpus called "NestedClinBr", containing clinical notes labeled with complex entities in Brazilian Portuguese.

1.2 Objectives

The main objective of this research is to develop a named entity recognition method that also considers nested, discontinuous, and multi-type entities, using state-of-the-art architecture for NLP such as Transformer architecture and deep learning.

The specific objectives of this project are presented as follows:

- To study existing approaches that address the recognizing of complex named entities to compare with the proposed method (Chapter 3 - Related Works);
- To search available NER corpora containing complex entities in English and Portuguese languages to perform experiments (Chapter 4 - Methodological Procedures, item 4.2 - Exploratory Phase);
- To develop a two-phase method that combines the QA-NER approach with CRF to recognize nested, discontinuous, and multi-type entities, while achieving competitive results without the high computational demands associated with exhaustive methods (Chapter 5 - Method);
- To develop a guideline for human annotations of nested and discontinuous entities in clinical texts in Portuguese (Chapter 6 - A New Portuguese-language Clinical Corpus, item 6.0.4 - Annotation Guidelines);
- To build a corpus with clinical texts in Brazilian Portuguese containing nested and discontinuous entities (Chapter 6 - A New Portuguese-Language Clinical Corpus);
- To train clinical and biomedical Transformer-based models for Portuguese language (Chapter 7 - Portuguese-language Models for Clinical and Biomedical Domains);
- To evaluate the proposed method in both English and Portuguese (a low-resource language), in clinical and biomedical domains (Chapter 8 - Experiments, Results, and Discussion).

1.3 Hypotheses

This research presents three hypotheses, each one being discussed and evaluated throughout this work:

H1 A new NLP task, which combines aspects of both NER and QA, allows the successful recognition of nested, multi-type, and discontinuous entities, yielding competitive results with literature methods;

H2 By incorporating a multi-label CRF model into the Transformer-based model, the method improves the coverage of nested and multi-type entities;

H3 We hypothesize that our method achieves state-of-the-art performance in NER task, when performed in corpora containing complex entities.

1.4 Contributions

The scientific contributions of this thesis are:

- A enhanced QA-NER task, with a treatment for class imbalance and also adapted for recognizing discontinuous entities and nested entities of the same type;
- A new publicly available Brazilian-Portuguese clinical corpus manually annotated with nested and discontinuous entities.

The technological contributions of this thesis are:

- The implementation of multi-label CRF models;
- A Wod2vec model trained with Brazilian clinical texts;
- A publicly available clinical Transformer-based POS-tagger model;
- Publicly available pre-trained models for Portuguese in the clinical and biomedical domains.

The recognition of nested, multi-type, and discontinuous entities is highly challenging and has received limited attention. Our method addresses this gap by proposing a new approach to identify and extract these intricate entity structures. By combining techniques from both QA and NER domains, our method opens new avenues for more accurate and comprehensive entity recognition in natural language texts.

All source code to execute the method is being made publicly available. As a contribution to the health domain, we can mention that extracting information from clinical notes and biomedical texts (unstructured data) can contribute to the clinical practice as an improvement of medical processes and decision support systems and in biomedical research (e.g., clinical trials, pharmacovigilance), being integrated into many pipelines and supporting other NLP tasks.

1.5 Scope

The scope of this research is limited to the recognition of flat and complex named entities using CRF, deep learning, and Transformer architecture, being evaluated by publicly available datasets for the research community. Although we focus on clinical and biomedical domains and for English and Portuguese languages, the method can be helpful in other domains and languages.

1.6 Financial Support

This thesis was financed in part by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil" (CAPES), a research agency from the Ministry of Education from Brazil (Ministério da Educação – MEC), with Finance Code 001. Also, between September 2021 and January 2022, the Ph.D. candidate stayed at the Universidad Carlos III de Madrid (UC3M), Spain, under the supervision of Professor Paloma Martinez, financed by CAPES under the "Programa de Doutorado-sanduíche no Exterior" (PDSE) program. Finally, for five days in November 2021, the development of this thesis took place at HES-SO, in Geneva, Switzerland, under the supervision of Professor Douglas Teodoro, with the travel expenses financed by the program "Leading House for the Latin American Region - Seed Money Grants 2019 (SMG nº 1922), Centro Latinoamericano-Suizo de la Universidad de San Gallen (CLS-HSG)", in the project "Exploring deep language models to leverage Portuguese and French biomedical semantic resources".

1.7 Overview

The research project document is structured as follows:

- Chapter 1, the current chapter, offers an overview of the context in which this research work is inserted, identifying the objective, hypotheses, contributions, and other relevant information;
- Chapter 2 introduces all the theoretical foundation, with the main topics, challenges, and concepts related to the research to understand the work and support the study development;
- Chapter 3 presents the state-of-the-art and related work on complex entity recognition, yielding a survey of existing works;
- Chapter 4 presents the methodological procedures adopted in work;
- Chapter 5 presents the proposed method with its assumptions and limitations;
- Chapter 6 exposes the results of experiment E1, which refers to the development of a new corpus, NestedClinBr;
- Chapter 7 explains the new resources developed, results of the experiment E2, i.e., the training of Portuguese clinical and biomedical language models;
- Chapter 8 shows the results of the experiments E3 to E9 with BioNestedNER, evaluating and comparing with baseline and literature methods, presenting a discussion about the results, and revisiting the research goals and hypothesis;
- Finally, in Chapter 9, the conclusion of this thesis is presented, with the research contributions and future work.

2
Background

In this chapter, we first present some characteristics, challenges, and contributions to the use of NLP in the clinical and biomedical domains. Next, the basic concepts about the named entity recognition task, complex entities, tagging schemas, NER approaches, and class imbalance are presented. We address the Conditional Random Field model, the contextual language models of word representation, and the Transformer architecture. We present some pre-trained language models, as well as the concept of fine-tuning.

2.1 NLP on Clinical and Biomedical Domains

Research on Artificial Intelligence (AI) with a focus on the healthcare sector has been enabling a wide variety of potential applications that are particularly valuable in the medical context (DUDCHENKO; GANZINGER; KOPANITSA, 2020). The EHR systems store a large volume of clinical patient data, such as demographic information, care histories, medical developments, clinical narratives, and hospital discharge summaries, which can support hospital processes, health services, and clinical research. Access to information is essential to offer quality healthcare assistance, which has improved the NLP techniques applied in clinical narratives.

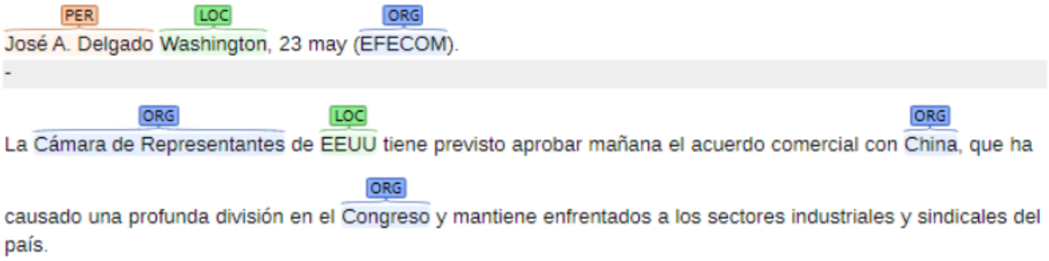
However, besides the clinical texts being written in free text, they have some particular characteristics that make the extraction of medical information a real challenge. Usually, these texts do not have a defined formal structure and are likely to contain grammatical errors, typos, high use of medical acronyms, medical jargon, lexical and semantic sparse, noises, and complex dependencies between variables (DALIANIS, 2018). Also, another challenge working with clinical narratives is its low availability, given the sensitive nature of health data and privacy concerns (SCHNEIDER et al., 2020).

Among the information extraction subtasks, NER is one of the most used in the clinical domain since it automatically extracts and encodes clinical concepts, entities (such as drugs, diseases, procedures, infections, comorbidities), and events, as pointed out by (DALIANIS, 2018) and (ZHANG; ELHADAD, 2013), as well as in biomedical research, providing valuable clinical and biomedical information.

2.2 Named Entity Recognition

Named entity recognition is an NLP subtask that aims to locate and classify named entities present in a text into predefined categories such as person names, organizations, locations, medical codes, and time expressions. The goal of NER is to extract structured information from unstructured data, according to their meaning in the text. NER is a token classification task, like the Part-of-Speech (POS) tagging, in which each token in the sentence has an output, representing its type of entity (or "O" for non-entity). One of the most important NLP tasks, as it serves as a basis for other tasks, NER can be divided into two subtasks: a) identifying mentions of relevant entities in the text, and b) classifying them into predefined categories of interest. Figure 2.1 presents an example of named entities identified in a random Spanish text from the NER Conll 2002 corpus (SANG, 2002), highlighting entities of type Person (in orange), Location (in green) and Organization (in blue).

Although it seems simple for humans, entity recognition is a challenge for computational models, since each entity can present ambiguity or different meanings depending on its context. Relevant entity extraction is useful in many problems such as machine translation, information retrieval, question-and-answer systems, and summarization. In biology, for example, the NER task can extract predefined concepts from raw texts, such as protein names. In clinical text processing, it can be useful to identify adverse drug events in patients or extract key information from patients’ electronic medical records. These ap-



José A. Delgado Washington, 23 may (EFECOM).

-

La Cámara de Representantes de EEUU tiene previsto aprobar mañana el acuerdo comercial con China, que ha causado una profunda división en el Congreso y mantiene enfrentados a los sectores industriales y sindicales del país.

Figure 2.1: Examples of entities found in a Spanish text present in the Conll2002 corpus (SANG, 2002)

plications require the identification of specific entities from the biomedical and clinical domains, respectively (DAI, 2018).

With the adoption of electronic health records (or electronic patient records) and the growing number of publications in the health domain, a large amount of clinical and biomedical texts are becoming available. At the same time, the academic community has devoted significant efforts to the creation of standardized terminologies and knowledge bases, facilitating the extraction of information from raw data (ZHANG; ELHADAD, 2013). The barrier in clinical information processing, therefore, is no longer collecting data, but using available data through scalable models to process large amounts of text. The quality of entity recognition in the health area (clinical and biomedical) strongly impacts the performance of other tasks, fundamental in clinical language processing, identifying and mapping terms into semantic categories (ZHANG; ELHADAD, 2013).

Complex Entities

While flat entities are mentions in the text formed by one or more continuous terms that do not overlap, in this work we consider complex entities [1] the entities that involve nested, overlapping, discontinuous mentions and/or belonging to more than one semantic type (e.g. caused by polysemy). As an example, in the sentence "Bill and Hillary Clinton went to Canada", we have the following references to the entity type "Person": "Bill Clinton" and "Hillary Clinton". The first mention is an example of a discontinuous entity, formed by non-sequential words in the text, and the word "Clinton" is an example of an overlapping

[1] We do not follow the definition adopted by SemEval-2022 task 11, as explained earlier.

mention, used in two separate mentions. In the text "Bank of China", the word "China" represents an entity of type "Local" and "Bank of China", "Organization". This is an example of nested or embedded entities.

Following the definition of (DAI, 2018), in this work, we consider complex entities the entities formed by mentions:

- Nested or embedded: when a mention of an entity is completely incorporated by another. We call the involved mentions "nested mentions". In Figure 2.2(a) we can see nested entities, a Protein "TNF-alpha" and a DNA "human TNF-alpha promoter".

- Overlapping: when two mentions overlap but are not completely contained by each other. Besides other situations, this can occur with discontinuous entities, as an example in Figure 2.2(b), where the "c-jun early response genes" and "c-fos early response genes" entities share tokens in common ("early response genes").

- Discontinuous: when the mention consists of discontinuous tokens, i.e. the mention contains at least one gap or interval between the terms (tokens) that compose it. An example can be seen in Figure 2.2(b), where the expression "c-jun early response genes" is an entity formed by discontinuous terms;

- Multi-type, where a mention may belong to more than one type of entity (also called a "multi-category" entity in (WANG et al., 2022)). This may be due to polysemy, when a mention can be classified into different semantic classes. An example is presented in figure 2.2(c), where "nitric oxide synthase" is both a Protein and a DNA entity, according to the annotation in the GENIA corpus.

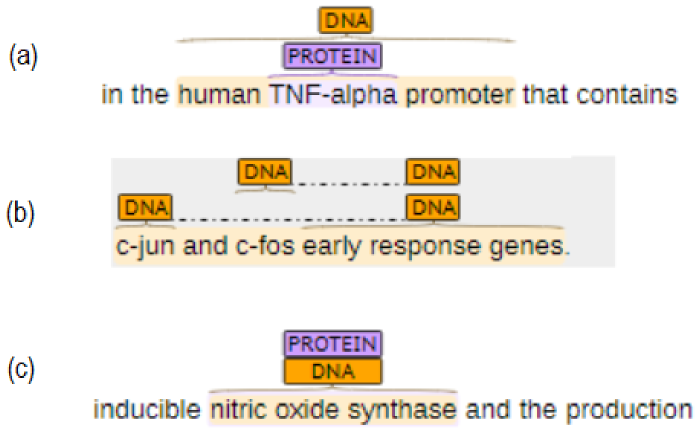

(a)
DNA
PROTEIN
in the human TNF-alpha promoter that contains
(b)
DNA
DNA
DNA
DNA
c-jun and c-fos early response genes.
(c)
PROTEIN
DNA
inducible nitric oxide synthase and the production

Figure 2.2: Examples of complex entities, present in the GENIA corpus (KIM et al., 2003)

These complex entities can contain useful information for other tasks, for example, the nested and overlapping structure itself can be a good indicator of the relationship between the different entities involved. Also, the recognition of complex mentions can contribute to building a knowledge base, and identifying discontinuous mentions can improve the performance of a machine translation system. In addition, complex structures can also exist for other NLP tasks, such as recognition of multiword phrases (LI et al., 2022). The proposed methods for recognizing complex entities can be applied to face similar difficulties in tasks other than NER.

Labeling schemes

As NER is a token-level task, there are some tagging schemes applied in the labeling of instances to identify the limits of the mention. The most common in the literature are:

- IO, where *I* represents an internal token of an entity ("inside") formed by one or several tokens, and *O*, a token that does not represent any entity ("outside").
- IOB: proposed by (RAMSHAW; MARCUS, 1995), has a similar format to the IO with the inclusion of the *B* that indicates the first token of an entity ("begin") but just when followed by another, delimiting the entities.
- IOB2 (or BIO): a variant of the IOB format, proposed by (SANG; VEENSTRA, 1999), where the *B* tag is used at the beginning of all mentions.
- IOE: in this format, the *E* indicates the last token ("end") of a mention involving several tokens (SANG; VEENSTRA, 1999).
- IOE2: similar to IOE, but the *E* tag is used at the end of all mentions (SANG; VEENSTRA, 1999).
- IOBES (or BILOU): the *B*, *I*, *O* tags have the same usage as the IOB2 format, the *E* tag has the same usage as the IOE format, and *S* indicates an entity formed by just one token ("single") (DAI et al., 2015).

Table 2.1 presents an example of the different tagging formats for representing local (LOC) and person (PER) entities.

Table 2.1: Example of tagging for NER in the different formats available.

Format	The	artist	was	born	in	Rio	de	Janeiro	,	RJ	.
IO	O	I-PER	O	O	O	I-LOC	I-LOC	I-LOC	O	I-LOC	O
IOB	O	I-PER	O	O	O	B-LOC	I-LOC	I-LOC	O	I-LOC	O
IOB2	O	B-PER	O	O	O	B-LOC	I-LOC	I-LOC	O	B-LOC	O
IOE	O	I-PER	O	O	O	I-LOC	I-LOC	E-LOC	O	I-LOC	O
IOE2	O	E-PER	O	O	O	I-LOC	I-LOC	E-LOC	O	E-LOC	O
IOBES	O	S-PER	O	O	O	B-LOC	I-LOC	E-LOC	O	S-LOC	O

Although these labeling schemes are widely used in the literature, they cannot represent nested and discontinuous entities.

Approaches

There are four main approaches used in named entity recognition, according to (LI et al., 2022):

1. Rule-based: Early NER research used rules created by humans to extract entities, usually with a set of grammatical patterns, linguistic analyses, and dictionaries. They present reasonable results in restricted domains but with little portability and robustness and high maintenance cost of the rules (MANSOURI; AFFENDEY; MAMAT, 2008).

2. Based on unsupervised learning: In the unsupervised approach, NER systems are cluster-based, extracting named entities from groups based on context similarity. The lexical features, patterns, and statistics computed in a large corpus are used to infer mentions of named entities, as in the work of (ZHANG; ELHADAD, 2013).

3. Based on supervised learning: In the supervised approach, NER is used as a token classification, or labeling task. Through the data samples, features are carefully designed to represent each training example, allowing machine learning algorithms to learn a model to recognize similar patterns. Added to the machine learning algorithm, the vector representation of words was widely used in supervised NER systems.

4. Based on deep learning: In recent years, deep learning approaches to NLP tasks, including NER, have become dominant and have reached the state-of-the-art in various corpora. Deep learning allows the discovery of representations of data with various levels of abstraction, with deep artificial neural networks, discovering necessary representations for the classification or detection of entities. A traditional use for the NER task was the combination of BiLSTM (a deep learning neural network, being a type of recurrent neural network) with Conditional Random Fields, as

> in the work proposed by (MA; HOVY, 2016). The CRF algorithm is a statistical modeling method often applied in NLP, as it takes into account the context of the instance to be classified, implementing dependencies between predictions. There are three strengths of the deep learning application: a) the NER benefits from the non-linear transformation, being able to learn complex data patterns via non-linear activation functions, b) the deep learning saves significant effort in the generation of labeled data, not requiring a considerable amount of domain skill and experience; c) NER models with deep learning can be trained in the end-to-end paradigm, by gradient descent, allowing to design more complex systems (LI et al., 2022).

Allied with deep learning, the recent NER methods are being developed with contextual word embeddings and Transformer architecture, which will be seen next.

Class Imbalance

Class imbalance is a common problem in NLP tasks where one or more classes are significantly underrepresented in the data, which usually occurs in classes rare or infrequent. Class imbalance can negatively impact the performance of NLP models, leading to poor recall and precision for the minority class(es).

The problem of class imbalance is particularly severe in NER tasks, where the class "O" (corresponding to non-entities) is typically much more frequent than the other classes. This imbalance can significantly impact the performance of NER systems, especially in terms of recall, as they tend to prioritize the accuracy of the majority class at the expense of the minority classes. For example, the ratio of the class "O" is up to 5 times more common in CoNLL03 corpus and 8 times in Ontonotes5.0 (AMOR; GRANITZER; MITROVIć, 2023).

In Figure 2.3, we have examples of class balancing in GENIA (a), Nested-ClinBr (b), and Rare Disease (c) corpora. The number of tokens with annotation of entities vs non-entities (class O) is shown, where we notice that the number of non-entities is naturally higher.

Several techniques have been proposed in the literature to mitigate the impact of class imbalance in NLP, including sampling techniques (e.g. oversampling, undersampling, and position bias), modified loss functions, and ensemble methods. The work of (GRANCHAROVA; BERG; DALIANIS, 2020) show that both undersampling the negative class and oversampling the minority positive classes can improve recall in NER for class imbalanced data, however with a

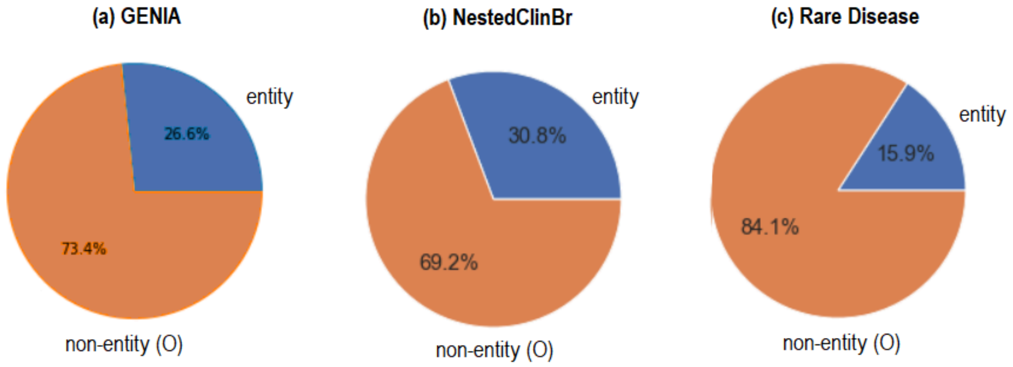


Figure 2.3: Class imbalance (entity vs non-entity) in GENIA, NestedClinBr, and Rare Disease corpora, in train dataset.

negative impact on precision, since the models have been trained on higher ratios of positive samples than are present in the test data, causing them to tend towards labeling more samples as positive. A recent work has proposed two methods to deal with the class imbalance and position bias of positive examples in token classification tasks, with Random Position Shifting and Context Perturbation techniques (AMOR; GRANITZER; MITROVIć, 2023). For the NER task, the authors obtained improvement in results on the CoNLL03 corpus, however, on OntoNotes5.0 no improvements were achieved, indicating that further investigations should be performed.

Another approach is to modify the loss function to give more weight to the minority classes in training, penalizing incorrect classification of samples from minority classes more than that of samples from majority ones. The loss function is a key component of training a machine learning model, as Cross Entropy (CE), a commonly used loss function in classification tasks. It measures the difference between the predicted probability distribution and the actual probability distribution of the classes.

In cases where the classes are imbalanced, it can be useful to define weights for each class to ensure that the model does not prioritize the majority class and instead learns to predict all classes accurately. The class weights can be assigned based on the frequency of each class in the training data or based on the importance of each class to the specific task. By incorporating class weights, the model can learn to give equal importance to all classes, leading to better overall performance (PANCHAPAGESAN et al., 2016). Using class weights can be advantageous in managing the positive-to-negative token ratio, as it preserves the text structure, unlike undersampling. Oversampling can negatively impact precision, possibly due to mismatches between the structure of training and test data. This implies that utilizing class weights rather than undersampling can be a potential solution for improvement, as pointed out by (GRANCHAROVA; BERG; DALIANIS, 2020).

2.3 Conditional Random Field

CRF is a probabilistic model very used for sequence labeling tasks such as NER. The input sequence is typically a sentence, and the output is a sequence of entity labels, one for each token in the input. The CRF model learns to assign a probability to each possible output sequence given the input sequence, selecting the most likely output sequence as the predicted labels (LAFFERTY; MCCALLUM; PEREIRA, 2001). Let $X = (x_1, x_2, ..., x_n)$ a sequence of words in a sentence, to determine the best sequence of labels $Y = (y_1, y_2, ..., y_n)$ for these words (corresponding to the categories of entities or "O" to non-entity), the CRF models a conditional distribution $p(y \mid x)$ that represents the probability of obtaining the output y given the input x. A linear-chain CRF can be formulated as Equation (1), where the normalization constant $Z(x)$ is a sum of all possible state sequences such that the total becomes 1.

$$p(y|x) = \frac{1}{Z(x)} \sum_{t=1}^{T} exp \left\{ \sum_{k=1}^{K} \theta_{\mathrm{k}} f_{\mathrm{k}}(y_{\mathrm{t}}, y_{\mathrm{t-1}}, x_{\mathrm{t}}) \right\} \tag{1}$$

This method can model dependencies between neighboring labels, which is important in NER since the label of a token is often dependent on the labels of its neighboring tokens. For example, if the label of the previous token is "B-DNA" (beginning of an entity), the current token is more likely to be labeled as "I-DNA" (inside an entity) than "O" (outside).

CRFs are usually trained on labeled data using a maximum likelihood estimation approach, maximizing the likelihood of the training data given the model parameters. The model parameters include feature weights (which capture the importance of different features in predicting the labels), and transition weights (which capture the importance of transitioning from one label to another). In NER tasks, the features typically include information about the token itself, such as its POS tag and its word shape, as well as contextual information such as the labels of neighboring tokens. The transition features capture information about the likelihood of transitioning from one label to another, based on patterns observed in the training data.

Despite the recent advances in contextual models such as the Transformer architecture, CRFs remain an important tool in NLP, providing a way to model sequential data taking into account the dependencies between labels of adjacent tokens. This is especially useful for tasks such as NER, where the context of a token can be critical in determining its label. A comparative study between CRF and Support Vector Machines (SVM) for clinical NER was presented by (LI; SAVOVA; KIPPER-SCHULER, 2008). It was evaluated in a set of gold standard NER and demonstrated that CRFs outperformed SVMs in terms of F-score. Other studies in biomedical NER task were made by (PONOMAREVA et al., 2007) and (MADY; AFIFY; BADR, 2022), where the first one compared CRF with Hidden Markov Models (HMMs), demonstrating that CRF-based models were superior to HMM in F-score, and the second, achieved satisfactory results in GENIA corpus with SVM and CRF. Outside the biomedical domain, other research with CRF has also been conducted, such as the study by (PATIL; PATIL; PAWAR, 2020) on electronic newspapers in the Marathi language. For the Portuguese language, we highlight the work of (PIROVANI; OLIVEIRA, 2018), (SOUZA et al., 2019), and (SOUZA; NOGUEIRA; LOTUFO, 2019). While contextualized deep learning models have achieved state-of-the-art performance on many NLP tasks, they still have limitations in dealing with long-range dependencies and handling noisy or incomplete data. CRFs can complement these models by providing a structured output that is more interpretable, helping improve the overall performance by enforcing consistency and coherence in the labeling of the sequence. Additionally, CRFs can be trained efficiently with small amounts of labeled data, making them particularly useful in low-resource settings.

2.4 Word Embeddings and Contextual Models

Word representation, a small semantic element in natural language, has always been a relevant research topic in NLP. In recent years, several vector representations of low-dimensional words have been trained, with huge amounts of unannotated textual data. These vectors, known as word embeddings, have proven to be effective in various NLP tasks, such as syntactic analysis, entity recognition, and machine translation (WANG et al., 2020a).

The development of these traditional word representations can be divided into two stages. In the first phase, the vectors used to represent words are sparse and high-dimensional, not indicating the semantic distance between words and being difficult to use them. In the second phase, the vectors are trained with large textual data, being dense and of low dimension. The Neural Network Language Model is a pioneering work that uses deep learning in language modeling, through a model that predicts the next word (BENGIO et al., 2003). Later, two prominent Word2Vec architectures were proposed, the Continuous Bag-of-Words and the Skip-gram models, reducing computational complexity and being considered a milestone in the development of distributed representation (MIKOLOV et al., 2013). Several other models have emerged, such as the Global Vector (GloVe) that uses the statistics of the co-occurrence of words in a corpus (PENNINGTON; SOCHER; MANNING, 2014) and fastText, an extension of the Skip-gram model that considers n-gram characters (BOJANOWSKI et al., 2016).

Although these distributed representations have achieved great success in NLP tasks, each word is represented by a single vector that does not take into account its context. For example, the word "guard" could mean a person who guards, like a sentinel, or a solid protective shield, distinguished by its context in the sentence. With traditional word embeddings, this word is represented by the same vector representation, although it has different semantic meanings in the sentences. This static representation does not capture the contextualized semantic meaning of the words. To deal with the problem of polysemy, dynamic representations (contextual word embeddings) emerged, such as ELMo, ULMFit, BERT, and XLNet.

ELMo (Embeddings from Language Models), released by (PETERS et al., 2018), uses language modeling to explore unlabeled data, reaching state-of-the-art in several tasks at the time of its release. ELMo extracts context-dependent representations of the word using a bidirectional language model, in which two LSTM are applied to encode the left and right contexts of the word. In each layer, the contextualized representations are generated by concatenating the left-to-right and right-to-left representations.

In the same period, Universal Language Model Fine tuning (ULMFit) was proposed by (HOWARD; RUDER, 2018), which also employs language modeling based on LSTM to explore large unlabeled data. ULMFiT enables "transfer learning" through the general pre-training of the model, and fine-tuning the classifier to the target task. ULMFiT uses a simple 3-layer LSTM, and this unique architecture is used throughout the entire process, from pre-training to fine-tuning.

BERT (Bidirectional Encoder Representations from Transformers) uses a bidirectional transformer (the Transformer encoder, which will be seen later) to pre-train richer contextualized representations. In training, BERT generates the representation of each word within the context of the sentence, using attention mechanisms to store the level of relevance of each word in the sentence. With a greater ability to capture contextual information from both sides, BERT reached the state-of-the-art in 11 NLP tasks (DEVLIN et al., 2019) and inspired several models based on its architecture.

RoBERTa (Robustly Optimized BERT Pretraining Approach) is a language model developed by (ZHUANG et al., 2021). Similar to BERT, it is based on a Transformer architecture that utilizes self-attention mechanisms to model the relationships between words in a sentence. RoBERTa differs from BERT in using a larger pretraining corpus, over 160GB of text, which allows the model to learn a more comprehensive representation of the language. Additionally, RoBERTa uses a different pretraining objective, masking out more words in each instance, which leads to a robust model. In terms of fine-tuning, it uses dynamic control over the learning rate schedule and removes the next-sentence prediction objective used in BERT, resulting in a model less sensitive to the choice of hyperparameters.

GPT (Generative Pretrained Transformer) is a language model developed by OpenAI (RADFORD et al., 2019) known to use deep learning techniques to generate natural language text. The model is also based on the Transformer architecture, pre-trained on a large corpus of text data using unsupervised learning, specifically masked language modeling, where the model is trained to predict missing words in a sentence given the context. GPT-3 is the latest version of OpenAI's language model[2] that has been trained on a large corpus of text data using a variant of the Transformer architecture, specifically the Transformer decoder architecture, which has been shown to be highly effective in a number of language processing tasks. The model was trained on a massive amount of diverse text data from the internet, including websites, books, and articles, which allowed it to learn the patterns and structures of human language. This

[2] February 2023 information

massive amount of training data, combined with the highly effective Transformer architecture, has enabled GPT-3 to generate human-like text with accuracy and fluency, leading to its widespread use in a variety of applications, including language translation, text summarization, and conversation modeling.

2.5 Transformer Architecture

The Transformer architecture, proposed by (VASWANI et al., 2017), uses stacked and connected layers of an individual attention mechanism, building basic building blocks for the encoder and decoder. The attention mechanism is a central component of recent deep learning models used to model relationships between words in an input sequence, assigning a weight to each element in the input sequence, which reflects its relative importance to the current task. These weights are computed dynamically for each input sequence and are used to compute a weighted sum of the elements, which serves as an effective representation of the input and the relationships between words in a sentence. The Transformer architecture consists of multi-head self-attention mechanisms, feed-forward neural networks, and positional encoding, which allows the model to capture long-range dependencies in the input sequence. The basic neuron of the architecture is shown in Figure 2.4. Each neuron is formed by an encoder on the left side and a decoder on the right side. The encoder is composed of a self-attention layer, which applies the attention mechanism to the received text, and a feed-forward layer, which converts the result into a shorter-length vector. The decoder, in addition to having these units, also has an encoder-decoder attention layer, which maps the result of its self-attention layer with the vectors generated by the encoder.

As seen previously, several works have been proposed using the Transformer architecture, such as GPT and BERT.

BERT

According to (LI et al., 2022), language models using the BERT architecture are becoming a new paradigm for the NER task. As seen in the "Word Representation and Contextual Models" section, the word representations generated by BERT are contextualized and can be used to replace traditional representations, such as Word2vec and GloVe. BERT makes use of the Transformer architecture,

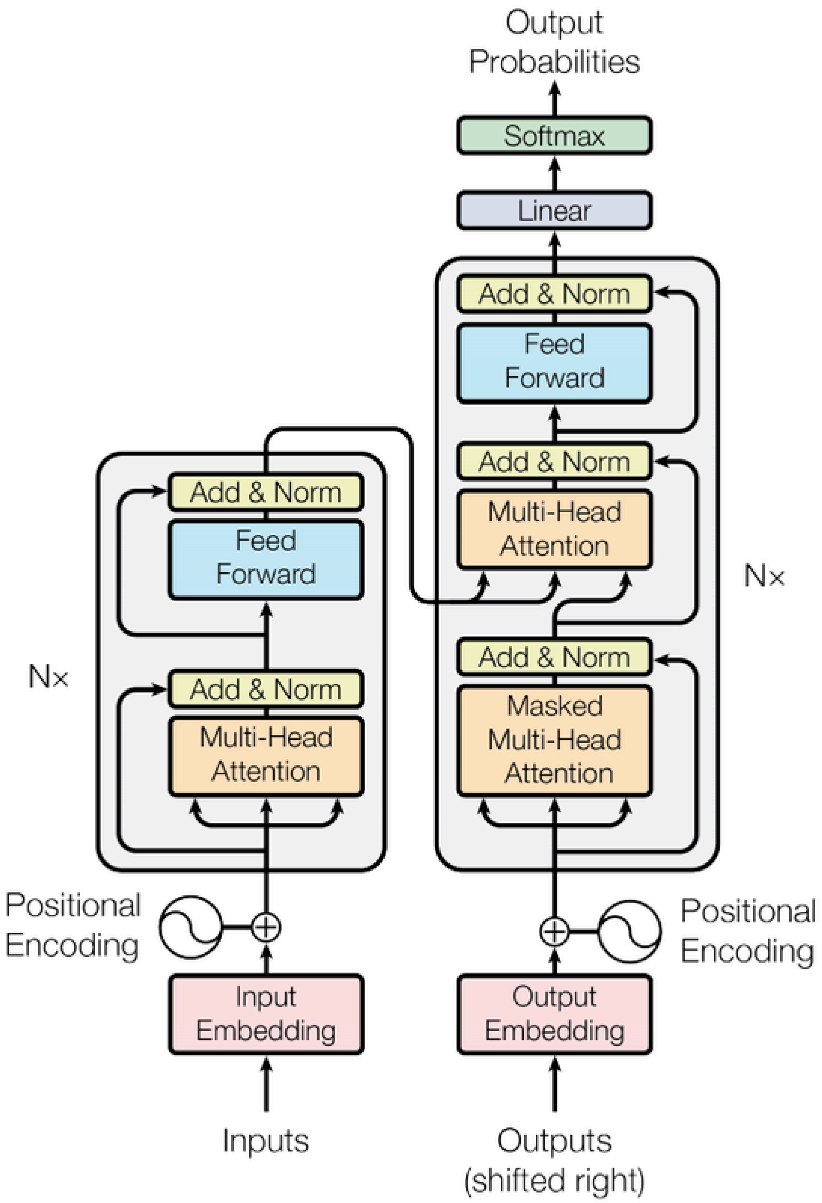


Figure 2.4: Transformer architecture (VASWANI et al., 2017).

which uses an attention engine that learns contextual relationships between words in a text, as seen earlier. Whereas the Transformer architecture includes two separate engines, an encoder that reads the input text and a decoder that performs the task prediction, BERT uses only the encoder engine to generate the language model. The Transformer encoder reads the entire sequence of words at once in a bidirectional way. This allows the model to learn the context of a word based on the words to the left and right of the word.

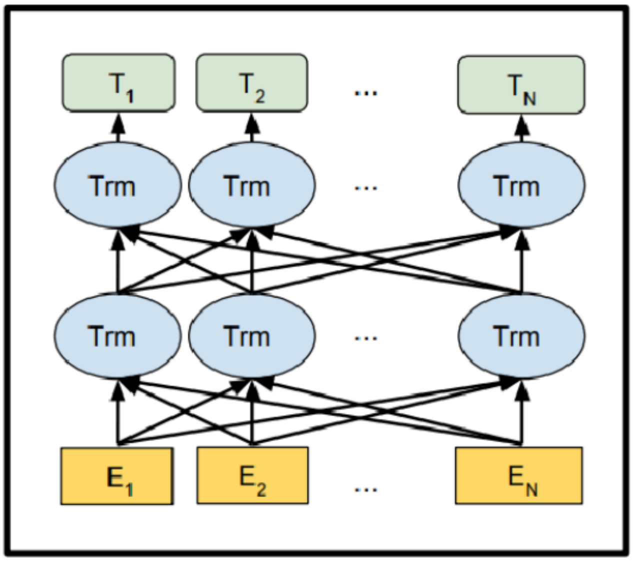


Figure 2.5: BERT architecture for contextual representation of words (DEVLIN et al., 2019).

In Figure 2.5, it is possible to visualize the architecture used for generating contextual word representations, as described by the BERT authors (DEVLIN et al., 2019). In E1, the word is transformed into its embedding representation. Each layer of the architecture performs a multi-headed attention calculation on the previous layer's word representation to create a new intermediate representation (*Trm*), generating the final output (*T1*). In a 12 layers BERT model, a token will have 12 intermediate representations, being both the last one and the combination of the last three generally used.

The BERT authors (DEVLIN et al., 2019) provided some pre-trained models, which can be used and adjusted (via fine-tuning) for some NLP tasks, as:

- BERT-base, with 12 layers (or transformer blocks), 12 attention heads, and 110 million parameters;
- BERT-large, with 24 layers, 16 attention heads, and 340 million parameters;
- BERT-multilingual-cased, a model that takes uppercase and lowercase words into account, with support for 104 languages including Portuguese, with 12 layers, 12 attention heads, and 110 million parameters (called here mBERT);
- BERT-multilingual-uncased, a model that does not consider uppercase or lowercase words, which also supports several languages such as Portuguese, with 12 layers, 12 attention heads, and 110 million parameters.

Also, a wide variety of BERT-based models were trained and released by several researchers, for various languages and domains. Among them, we can highlight:

- BERT-base-portuguese-cased, a language model trained in a Brazilian Portuguese massive corpus, called here BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020);
- BERT-large-portuguese-cased, a large version of BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020);
- BioBERT, BERT-based models trained in English biomedical language, which have achieved state-of-the-art in various biomedical tasks (LEE et al., 2019);
- ClinicalBERT, BERT models trained with generic clinical texts and discharge summaries, in English (ALSENTZER et al., 2019);
- PubMedBERT, a pre-trained model trained using abstracts and full-text articles from PubMedCentral (GU et al., 2021);
- Bio-ClinicalBERT, a language model initialized from BioBERT and trained on all MIMIC notes (ALSENTZER et al., 2019);
- Bio-Discharge-Summary_BERT, a language model initialized from BioBERT and trained on discharge summaries from MIMIC (ALSENTZER et al., 2019).

Fine-tuning to downstream NLP tasks

Transfer learning in NLP is a technique where a model pre-trained on a large dataset in one domain is fine-tuned on a smaller dataset in a different but related domain. Therefore, the knowledge learned from the large pre-training dataset can improve the performance of the smaller fine-tuning dataset, as the latter may not have enough data to train a model from scratch. This approach can be especially useful for NLP tasks with limited annotated data, such as named entity recognition or sentiment analysis in a specific language or domain. This process can be used to specialize a language model for another domain or for a specific task.

Fine-tuning can be considered a type of transfer learning that involves training all the layers of the network on the new task, not just the top layers. As the entire network is updated to minimize the loss on the new task, it makes it more specialized to the new task when compared to regular transfer learning, where only the top layer(s) of the network are pre-trained on the new data.

Hence, the pre-trained models as BERT can be specialized for a wide variety of NLP tasks, with the addition of an activation layer (e.g. softmax) that calculates the probabilities of the labels. With this fine-tuning process, the language model's architecture also allows the execution of an NLP task trained at the same time that the word embeddings. As seen, no layer is frozen, as all pre-trained layers along with task-specific parameters are trained simultaneously. The final hidden states, i.e. the Transformer output of each token, are fed into the classification layer to obtain the prediction for each token.

Figure 2.6 shows the BERT fine-tuning architecture for some NLP tasks. In token-classification tasks, such as NER, a tag (or label) must be provided for each input word, as seen in 2.6 (d). The final hidden states, i.e. the transformer output of each token, are fed into the classification layer to obtain the prediction for each token. The input is a sequence of tokens, which are embedded in arrays and then processed in a neural network. A classification layer is fed with the sequence of vectors, in this case, to predict the label of each token.

Moreover, fine-tuning is also widely used to train generic models in a particular domain, such as clinical and biomedical (where data is scarce), so that it can better perform in this domain while still leveraging the previous generic knowledge.

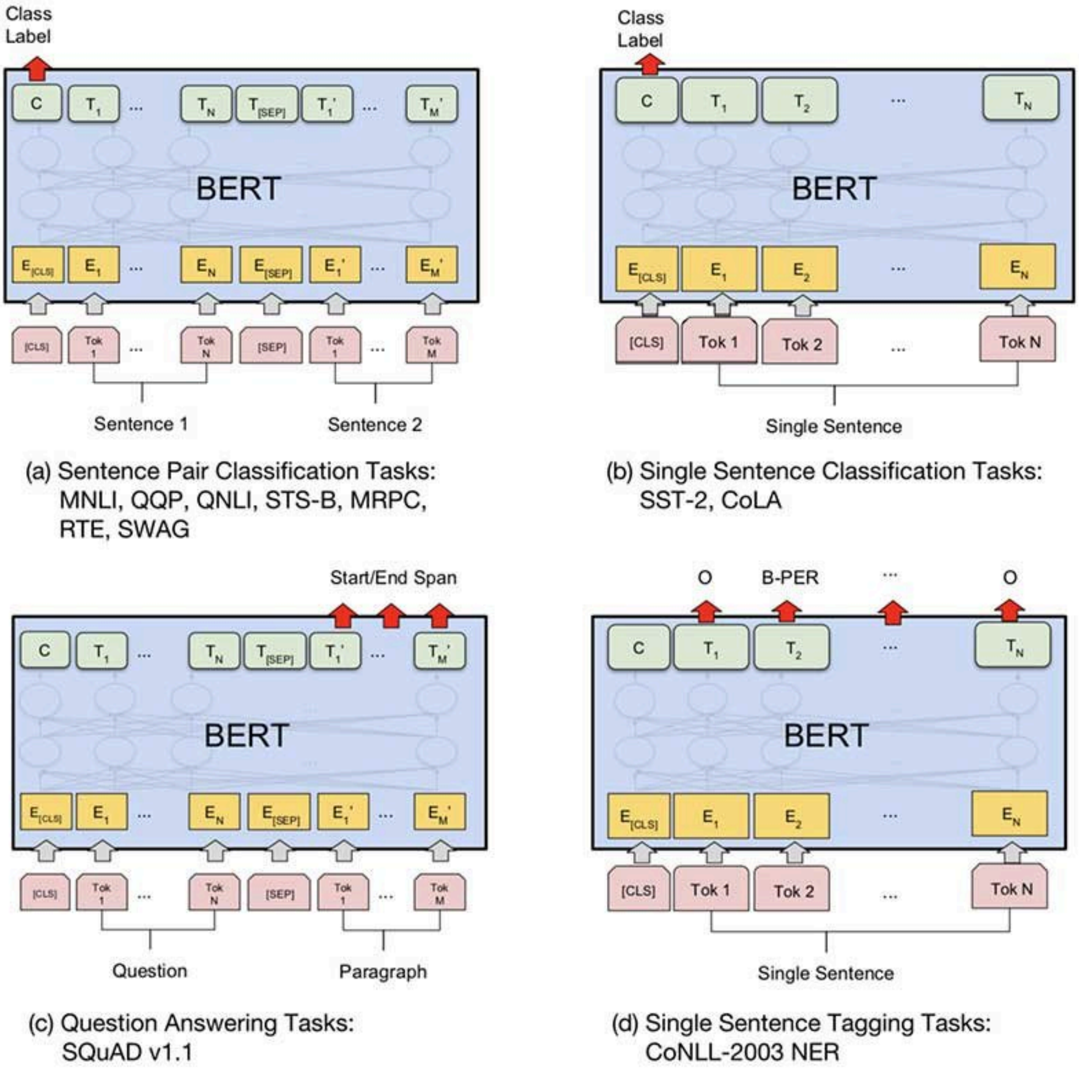

Class Label
C
T1
TN
T[SEP]
T1'
TM'
BERT
E[CLS]
E1
EN
E[SEP]
E1'
EM'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
Sentence 1
Sentence 2
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
Class Label
T2
E2
Tok 2
Single Sentence
(b) Single Sentence Classification Tasks: SST-2, CoLA
Start/End Span
Question
Paragraph
(c) Question Answering Tasks: SQuAD v1.1
O
B-PER
O
Single Sentence
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Figure 2.6: BERT architecture for fine-tuning in the tasks: sentence-pairs classification (a), text classification (b), question answering (c), and token-level classification (d). Source: (DEVLIN et al., 2019).

3

Related Work

This chapter presents the works related to the scope of this research and the state-of-the-art (SOTA) methods. To collect the papers, we used the scientific databases: ACM Digital Library (Full text-collection) [1], ScienceDirect[2], Pubmed[3], SpringerLink[4], IEEE Xplorer[5], and ACL Anthology[6]. Research selected using the "snowballing" technique was also included.

The search string considers the words *"nested entity"*, *"complex entity"*, *"entity overlap"*, *"entity discontinuous"*, *"multi type entity"*, *"irregular entity"*, *"structured entity"* and *"cascaded entity"*, in addition to *"entity recognition"* and its variations and synonyms. To verify similar works in Portuguese, a search was also performed in Portuguese in the same databases, but no relevant results were found. The search was implemented initially between 05/26/2020 to 06/14/2020 and performed again on March 9, 2023, to complement with new research published since then. Only works written in English or Portuguese that contained the search terms in the title or abstract were selected. Works that do not involve the NER task or that do not involve complex entities (nested, discontinuous, and/or multi-type) were discarded. Table 3.1 displays the number of works found in each indexed base, excluding duplicate articles. The search queries used in each

[1] https://dl.acm.org/

[2] https://www.sciencedirect.com/

[3] https://www.ncbi.nlm.nih.gov/pubmed/

[4] https://link.springer.com/advanced-search

[5] http://ieeexplore.ieee.org/

[6] https://www.aclweb.org/anthology

base can be found in Appendix 10.1.

Table 3.1: Number of papers searched, discarded, and selected per scientific database.

Scientific database	Total papers	Discarded	Selected
Pubmed	55	38	17
ACM DL	62	39	23
ScienceDirect	49	35	14
Springer Link	316	287	29
IEEE Explorer	199	177	22
ACL Anthology	146	96	50
ACL Anthology - Portuguese	9	9	0
Snowballing	7	-	7
Total	**843**	**681**	**162**

After discarding unrelated works, in total, **162** scientific papers related to the recognition of complex named entities were selected.

The nested NER approaches will be introduced according to their model architectures, similar to the division proposed by (WANG et al., 2022)[7]: early rule-based, layered-based, region-based, hypergraph-based, QA-based, transition-based, and others approaches, which will be detailed next.

Of the selected papers, 15 use the rule-based approach, 26 use the layered-based approach, 57 use the region-based approach, 8 use the hypergraph-based approach, 8 use the QA-based approach, 15 use the transition-based approach, and 33 use others or hybrid approaches.

In Table 3.2 we summarize these approaches according to the model architecture, listing some examples of each approach[8].

[7] We added the "QA-based" category

[8] Due to space constraints and the scope of this work, it was necessary to make a selection of the most relevant works of each approach, to provide a comprehensive but focused review

Table 3.2: Selected related works found in the literature review.

Year	Title	Reference
Rule-based		
2003	Effective Adaptation of Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain	(SHEN et al., 2003)
2004	Recognizing Names in Biomedical Texts: a Machine Learning Approach	(ZHOU et al., 2004)
2006	Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid	(ZHOU, 2006)
Layered-based		
2004	Enhancing HMM-based biomedical named entity recognition by studying special phenomena	(ZHANG et al., 2004)
2007	Recognising nested named entities in biomedical text	(ALEX; HADDOW; GROVER, 2007)
2018	A neural layered model for nested named entity recognition	(JU; MIWA; ANANIADOU, 2018)
2019	Merge and label: A novel neural network architecture for nested NER	(FISHER; VLACHOS, 2019)
2020	Dispatched attention with multi-task learning for nested mention recognition	(FEI; REN; JI, 2020)
2020	Pyramid: A layered model for nested named entity recognition	(WANG et al., 2020)
2020	Nested named entity recognition via second-best sequence learning and decoding	(SHIBUYA; HOVY, 2020)
Region-based		
2007	Nested named entity recognition in historical archive text	(BYRNE, 2007)
2017	A local detection approach for named entity recognition and mention detection	(XU; JIANG; WATCHARAWITTAYAKUL, 2017)
2018	Deep exhaustive model for nested named entity recognition	(SOHRAB; MIWA, 2018)
2019	Gazetteer-enhanced attentive neural networks for named entity recognition	(LIN et al., 2019a)
2019	A boundary-aware neural model for nested named entity recognition	(ZHENG et al., 2019a)

Table 3.2: Selected related works found in the literature review.

Year	Title	Reference
2020	Instance-based learning of span representations: A case study through named entity recognition	(OUCHI et al., 2020)
2020	Joint learning of token context and span feature for span-based nested NER	(SUN et al., 2020)
2020	Hierarchical region learning for nested named entity recognition	(LONG; NIU; LI, 2020)
2020	Boundary enhanced neural span classification for nested named entity recognition	(TAN et al., 2020)
2020	HIT: Nested named entity recognition via head-tail pair and token interaction	(WANG et al., 2020b)
Hypergraph-based		
2015	Joint mention extraction and classification with mention hypergraphs	(LU; ROTH, 2015)
2018	Neural segmental hypergraphs for overlapping mention recognition	(WANG; LU, 2018)
2018	Nested named entity recognition revisited	(KATIYAR; CARDIE, 2018)
Transition-based		
2009	Nested named entity recognition	(FINKEL; MANNING, 2009)
2018	A neural transition-based model for nested mention recognition	(WANG et al., 2018)
2019	Hierarchical nested named entity recognition	(MARINHO et al., 2019)
2020	Named entity recognition as dependency parsing	(YU; BOHNET; POESIO, 2020)
QA-based		
2020	A unified MRC framework for named entity recognition	(LI et al., 2020)
2020	A Question Answering-Based Framework for One-Step Event Argument Extraction	(ZHANG et al., 2020)
2021	Bridge Inspection Named Entity Recognition via BERT and Lexicon Augmented Machine Reading Comprehension Neural Model	(LI et al., 2021)

Table 3.2: Selected related works found in the literature review.

Year	Title	Reference
2021	Biomedical Named Entity Recognition via Knowledge Guidance and Question Answering	(BANERJEE et al., 2021)
2022	Parallel Instance Query Network for Named Entity Recognition	(SHEN et al., 2022)
2022	MRC-based Medical NER with Multi-task Learning and Multi-strategies	(DU; YUXIANG; HONGYING, 2022)
2023	Judicial nested named entity recognition method with MRC framework	(ZHANG et al., 2023)
2023	A novel MRC framework for evidence extracts in judgment documents	(ZHOU et al., 2023)
Other approaches		
2006	Recognizing nested named entities in GENIA corpus	(GU, 2006)
2017	Labeling gaps between words: Recognizing overlapping mentions with mention separators	(MUIS; LU, 2017)
2018	Neural architectures for nested NER through linearization	(STRAKOVÁ; STRAKA; HAJIC, 2019)
2019	Sequence-to-nuggets: Nested entity mention detection via anchor-region networks	(LIN et al., 2019b)
2020	Bipartite flat-graph network for nested named entity recognition	(LUO; ZHAO, 2020)

3.1 Early Rule-based

Early approaches for nested NER mainly rely on hand-craft rules and rule-based post-processing. In the work of (SHEN et al., 2003), the authors defined four basic patterns corresponding to types of nested entities, leveraging the Hidden Markov Models and integrating deterministic, morphological, POS, and semantic trigger features. Similar approaches were proposed by (ZHOU et al., 2004) and (ZHOU, 2006), which enhanced the rule-based post-processing to automatically extract rules from training data for the nested NER task.

3.2 Layered-based

The Layered-based approach treats nested NER task as multiple flat NER task, in a cascade structure connected in series. In this solution, the models generally contain multiple layers according to the hierarchical nature of nested entities, where each level can identify a group of entities.

An HMM-based approach with a layered structure for recognizing nested entities was introduced by (ZHANG et al., 2004). The training involved two models: the first model was used to recognize short embedded entities, and the second model focused on the extended short entities.

In the work of (ALEX; HADDOW; GROVER, 2007), three CRF models were used, reducing the nested NER problem into sequence tagging problems, using inside-out and outside-in layered CRFs. Their work demonstrated the superiority of CRF models over traditional Hidden Markov Models for nested NER, with the cascaded CRF model achieving the best performance.

A neural layered model for identifying nested entities was proposed by (JU; MIWA; ANANIADOU, 2018). This model employs the dynamic stacking of flat NER layers in an inside-out manner, utilizing BiLSTM and a CRF decoder. Each flat NER layer is composed of a BiLSTM encoder and a CRF decoder. Within the model, the encoder outputs from the current layer are merged to generate new entity representations, which are subsequently passed to the subsequent layer. This approach facilitates the identification of outer entities by capitalizing on information from corresponding inner entities.

A dispatched attention model with multitask learning for nested NER was introduced by (FEI; REN; JI, 2020). In this model, each task is responsible for recognizing entities at a specific level, utilizing BiLSTM and a CRF decoder. Each layer module contains a position- and syntax-aware attention-based encoder. Furthermore, a dispatched attention mechanism was introduced to facilitate the transfer of knowledge inside-out, sequentially capturing information across layers.

A neural model was developed by (FISHER; VLACHOS, 2019). This model first merges tokens and/or entities into entities, forming nested structures, and subsequently classifies them independently. It begins by identifying inner entities and then proceeds to recognize outer entities.

Another approach was presented by (WANG et al., 2020), involving a neural layered model for identifying nested entities. This model follows an inside-out approach and consists of a stack of interconnected layers, each of which predicts whether a region is an entity.

In the work of (SHIBUYA; HOVY, 2020), a CRF-based decoding approach is used, recognizing entities in an outside-in manner. First, they encoded the input sentence with BiLSTM, then a CRF for each entity category decodes and extracts outermost entities and inner entities without re-encoding, and finally, their model recursively extracts inner entities, called the 2nd best path, until no multi-token entities are detected in each region.

3.3 Region-based

Region-based nested NER approaches treat the nested NER task as a multi-class classification task, classifying each potential region (entity candidate) into one of the classes (types of entity).

The work of (MCDONALD; CRAMMER; PEREIRA, 2005) proposed a new approach to NER as a structured multi-label classification to represent overlapping segments in a sentence. This region-based model is flexible, as it allows finding mentions made up of discontinuous tokens and overlapping or nested mentions. The method was developed using CRF. A disadvantage of this model is the high computational complexity, as the number of labels for classification is exponential and depends on the number of words in the sentence.

A waterfall approach is employed in the work of (CHAN; LAM; YU, 2008), dividing the NER task into segmentation and classification. The segmentation task resembles the work of (MCDONALD; CRAMMER; PEREIRA, 2005), involving the segmentation of phrases and identification of possible segments containing biomedical named entities. Subsequently, the identified segments are classified into potential named or rejected entity types using a passive-aggressive algorithm and CRF.

An enumeration-based nested NER model was developed by (BYRNE, 2007). This model initially enumerates all regions from the input sentence and subsequently learns their corresponding region representations. Likewise, the work of (XU; JIANG; WATCHARAWITTAYAKUL, 2017) introduced a neural nested NER approach through local detection. They enumerate all regions of a certain length in a sentence, similar to the approach of (BYRNE, 2007).

The work of (SOHRAB; MIWA, 2018) is also based on regions, where mentions are detected by identifying subsequences of a sentence. They proposed a simple deep neural model for recognizing nested named entities, looking for possible regions in the text as potential entities to be classified. However, this exhaustive method suffers from too many irrelevant regions, in addition to classifying types and regions individually, not considering contextual information. In the same direction, a neural network was proposed by (LIN et al., 2019a), which models the candidate region along with its corresponding contextual information. This information is then fed into a Multi-Layer Perceptron (MLP) classifier to achieve entity prediction.

An instance-based nested NER approach was developed by (OUCHI et al., 2020), treating NER as a multiclass classification problem. All region representations are enumerated, and a category label is assigned to each of them. Similarly, the end-to-end region-based model in the work of (SUN et al., 2020) jointly learns the token context and region feature in sentences.

A multi-grained model was presented by (XIA et al., 2019), which first detects all possible regions before categorizing them. This model includes a detector and a classifier. The work of (LONG; NIU; LI, 2020) proposed a hierarchical region learning framework that generates a tree hierarchy of candidate regions and incorporates structure information into region representations for improved classification.

A boundary-based strategy was introduced by (ZHENG et al., 2019b), where the candidate boundary of the entity is detected instead of enumerating all regions in the sentence. This approach locates entities by identifying candidate boundaries using sequence tagging models. The boundary-relevant regions are then utilized to predict entity category labels. Furthermore, a region-based neural classification method that incorporates an additional boundary detection task for predicting words (entity boundaries) was proposed by (TAN et al., 2020). This method mitigates the computational complexity associated with the region-based approach by allowing the model to learn from better representations with boundary supervision. Bidirectional LSTMs and BERT are employed for word representation.

The use of graph-based dependency parsing for recognizing named entities was proposed by (YU; BOHNET; POESIO, 2020). This approach leverages BiLSTM to acquire contextual representation and subsequently employs two MLPs to classify the start/end representations.

A biaffine-based head-tail detector was developed by (WANG et al., 2020b) to classify each pair of tokens at the boundary of an entity. A token interaction tagger characterizes the internal token connection within the head-tail pair, and a region classifier is used for entity recognition.

3.4 Hypergraph-based

These approaches use the hypergraph to represent the nested structure of the entities in the text, where the hyperarcs naturally express that tokens belong to several entities.

A directed hypergraph for boundary detection and category prediction was presented by (LU; ROTH, 2015), consisting of five types of nodes representing entities of different semantic categories and boundaries.

In the work of (WANG; LU, 2018), a neural segmental hypergraph model is employed, which utilizes neural networks to acquire distributed feature representations. This model addresses the structural ambiguity issue identified in the study by (LU; ROTH, 2015).

A hypergraph structure based on the IOBES tag scheme was introduced by (KATIYAR; CARDIE, 2018), along with an LSTM-based model that learns a hypergraph representation for nested entities within the input sentence.

3.5 Transition-based

These approaches parse a sentence from left to right, building a tree using greedy decoding one action at a time, like in transition-based parsers.

A discriminative constituent parser for recognizing nested entities was developed by (FINKEL; MANNING, 2009). They extracted a constituency-based parsing tree from a sentence to represent its nested structure. This approach resembled the chart-based PCFG parser but distinguished itself by employing clique potentials for local subtrees instead of probabilities over rules.

In the work of (WANG et al., 2018), a neural transition-based approach was proposed to model the nested structure of entities. The sentence with nested entities was transformed into a forest structure, with each entity serving as a constituent within the forest. The system utilized three types of transition actions (SHIFT, REDUCE, UNARY) and employed a stack to temporarily store processed nested elements. At each step, one of the three action types was applied to modify the system's state.

Similarly, the work of (MARINHO et al., 2019) introduced a neural transition-based approach named Hierarchical and Nested Named Entity Recognition (HNNER), model to address various levels of nested entities. Their transition system integrated a word stack, word buffer, mention stack, and output buffer, incorporating four system actions (OUT, SHIFT, TRANSITION, and REDUCE). They also introduced a set of modifier classes introducing specific concepts that altered the entity's meaning, such as absence or uncertainty about a given entity.

3.6 Question-Answering-based

In the QA-based approach, the NER task is formulated as a question-answering problem, where the extraction of the nested entities occurs in response to specific questions related to the text, tackling naturally the nested structure problem.

The approach was initially proposed by (LI et al., 2020) as a machine reading comprehension task, where the tagging-style annotated NER dataset was transformed into a set of QUESTION, ANSWER, CONTEXT tuples. In their work, for each question (entity type), all possible starts and ends of the answer were identified by the model. Subsequently, all potential combinations of start and end in candidate answers were classified, considering the possibility of multiple entities of the same category. The machine reading comprehension framework was introduced into judicial named entity recognition by (ZHANG et al., 2023), which also involves nested entities. In a similar way, (LI et al., 2021) proposed a lexicon-augmented MRC-based NER neural model for identifying both flat and nested entities from Chinese bridge inspection text, following the same approach as (LI et al., 2020).

The Parallel Instance Query Network (PIQN) was introduced by (SHEN et al., 2022), which established global and learnable instance queries to extract entities from a sentence, similar to the approach in (LI et al., 2020), but utilizing a parallel approach. Each instance query predicts one entity, and by feeding all instance queries simultaneously, all entities are recognized in parallel.

An evidence extraction architecture was presented by (ZHOU et al., 2023), formalized as a QA problem, where all evidence spans were screened as potential correct answers. To address the data imbalance problem in the judgment documents, the authors revised the loss function.

In the work by (ZHANG et al., 2020), a one-step question-answering-based framework was proposed for simultaneous argument candidate extraction and argument role classification. Since conventional QA task cannot be directly applied, the authors designed QA-based Sequence Labeling, referred to as QA-NER, for this purpose. Although the authors focused on event argument extraction, to identify arguments of specific events and label their roles, their method can be applied to nested entities as well. Unlike methods like (LI et al., 2020) and others based on it, this work merged the NER task with QA, creating a hybrid task where the model receives the question (entity type) and the paragraph (sentence input), and provides outputs for each token instead of the start and end indexes of the response.

The work of (BANERJEE et al., 2021) also formulated the NER task as a multi-answer knowledge-guided QA task, specifically focusing on NER. The authors trained a large biomedical model using a combined dataset of 18 different datasets, achieving state-of-the-art results for 11 of the 18 biomedical NER datasets. Since this approach naturally allows the extraction of nested entities without increasing computational complexity, our method follows this approach, introducing enhancements like handling discontinuous entities and addressing the class imbalance.

Finally, (DU; YUXIANG; HONGYING, 2022) proposed an MRC-based approach with multi-task learning and multi-strategies, integrating an MRC-CRF model for sequence labeling with an MRC-Biaffine model for span boundary detection, ultimately selecting the more efficient MRC-CRF as the final decoder.

3.7 Other Approaches

Apart from the above mentioned approaches, there are particular works that use other specific approaches. The work of (GU, 2006) used SVM to classify nested entities, treating the NER task as a binary classification problem, using outmost and inner labeling.

A gap-based tag schema was proposed by (MUIS; LU, 2017) to capture nested entities, with the mention separators representing nested named entities.

Two neural network architectures for nested NER were developed by (STRAKOVÁ; STRAKA; HAJIC, 2019), where the nested entity multiple labels were concatenated into a single multi-label. An LSTM-CRF was used to predict the label of each token, followed by a sequence-to-sequence task.

A sequence-to-nuggets architecture (Anchor-Region Networks or ARNs) for recognizing nested entities was introduced by (LIN et al., 2019b). In this approach, the anchor words of an entity were first identified, followed by the determination of entity boundaries for each anchor word.

In the work of (LUO; ZHAO, 2020), a novel bipartite flat-graph (BiFlaG) network for nested NER was proposed, containing a flat NER module and a graph module. This system can jointly learn flat entities and recognize the outermost entities and construct entity graphs.

3.8 Shared Tasks

In addition to the research performed in the scientific databases, we also searched for shared tasks that encompass nested, multi-type, and/or discontinuous NER, as there are relevant works in these venues. Although there have been numerous shared tasks on related subjects such as entity recognition and entity relations extraction, most focus on flat entities. In recognition of complex entities, we find three related shared tasks:

- GermEval 2014 Named Entity Recognition Shared Task: Companion Paper (BENIKOVA CHRIS BIEMANN; PADO., 2014): This event makes available German data with NER annotation with the goal of advancing the state of the art in German NER and nested representations of named entities. Of the 11 participating teams, 5 used handcraft rules, one was based on a CRF model, and the others used machine learning approaches.

- RuNNE-2022 Shared Task: Recognizing Nested Named Entities (ARTEMOVA et al., 2022): This shared task addresses nested named entity recognition, using the Russian NEREL dataset. It has received 156 submissions, with systems based on MRC, region, layered models, and other approaches. Although the MRC-based models showed better accuracy than the mean accuracy, the best result was obtained by a rule-based system.

- NER Shared Task 2023 - Subtask 2: Nested NER (TALAFHA, 2023): This shared task aims to mitigate the lack of resources in Arabic NER by introducing a comprehensive Arabic NER corpus with 21 annotated entity types, with a nested version. This shared task is currently in progress as of the writing of this thesis.

Although the shared task "MultiCoNER: SemEval-2022 Task 11" (MALMASI et al., 2022) deals with the recognition of complex entities, we do not consider it here because it uses a concept different from ours about complex entities.

The shared task is about the challenge of detecting semantically ambiguous or formed by any linguistic constituent as titles of creative works, difficult to recognize.

3.9 Final Considerations

In summary, we found 162 related works in the researched scientific bases, categorized into 6 categories (plus others).

Methods based on early rule-based, which relies on hand-craft rules and rule-based postprocessing, work well in specific situations, but can be labor-intensive and less adaptable to complex language patterns. The layered-based methods provide an intuitive solution for nested NER, with multiple layers that work in a cascade. A disadvantage of this method is the need to train more than one model, one for each layer. The region-based method transforms NER into a multiclass classification problem. It naturally addresses the recognition of nested and discontinuous entities, but can suffer from an imbalance between entity vs non-entity and computational complexity in the exhaustive enumeration of all regions. The hypergraph-based method uses a hypergraph structure to represent nested entities, but it can suffer from structural ambiguity and also computational complexity. The QA-based formulates the NER task as a Question-Answering task, naturally addressing the nested structure problem. However, it needs a treatment to find more than one entity per question and to recognize discontinuous entities, and can also suffer from an imbalance between entity vs non-entity. The transition-based method uses discriminative dependency parsers to represent nested entities, but can also suffer from computational complexity.

Among these approaches, we selected the QA-based for our work, which provides a simple solution without requiring great computational complexity and without the need to train a model for each layer. Our method was based on the (BANERJEE et al., 2021) method, which already has the ability to return more than one answer per question. We implemented certain adaptations to enhance class balancing and enable the model to identify discontinuous entities.

4

Methodological Procedures

This chapter presents the methodological approach used in the development of this work. The research structure is divided into four phases: Initial Planning, Exploratory Phase, Development, and Evaluation. Figure 4.1 presents an overview of these phases and their respective tasks.

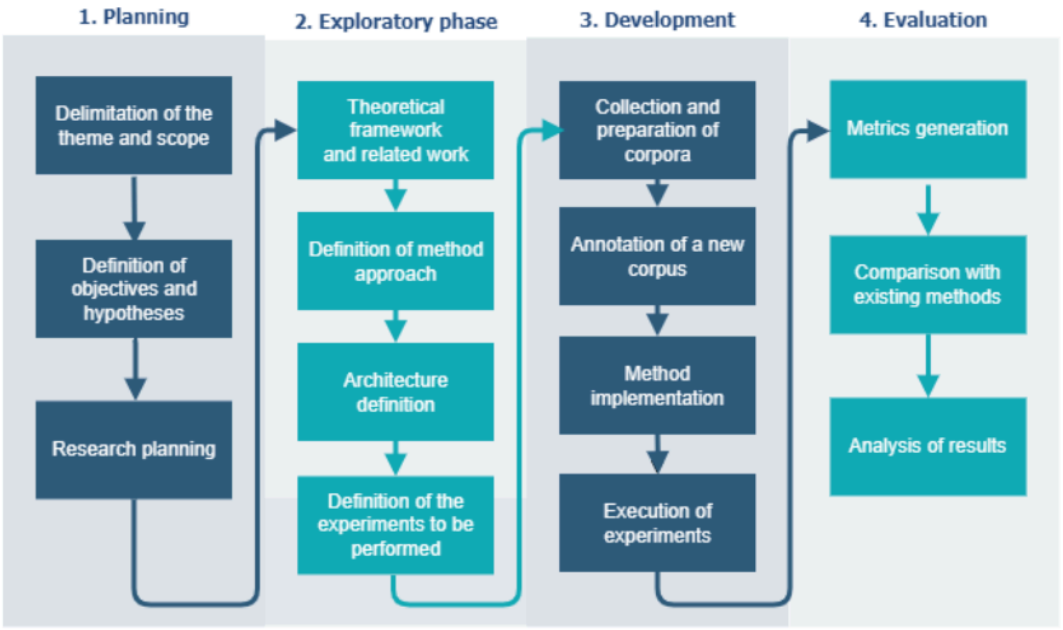

1. Planning
Delimitation of the theme and scope
Definition of objectives and hypotheses
Research planning
2. Exploratory phase
Theoretical framework and related work
Definition of method approach
Architecture definition
Definition of the experiments to be performed
3. Development
Collection and preparation of corpora
Annotation of a new corpus
Method implementation
Execution of experiments
4. Evaluation
Metrics generation
Comparison with existing methods
Analysis of results

Figure 4.1: Summary of research steps. Although they follow this order, some steps may have been conducted concurrently.

The following sections present the phases and tasks of the research method used.

4.1 Planning

The planning stage refers to all research preparation, involving scope delimitation, definition of work objectives and hypotheses, and research planning.

A literature review allowed a better understanding of the problem and the identification of gaps in the state-of-the-art NER research, identified in this work as methods capable of working with complex entities. Hence, our general objective was defined: the development of a method for the recognition of nested, discontinuous, and multi-type entities. Through literature reviews on this theme, the potential of Transformer-based models, such as BERT, proved efficient in several NLP tasks, leading us to follow this path.

In this phase, a challenge faced during our studies was the absence of a clinical corpus in the Portuguese language with nested and discontinuous entities, motivating us to annotate a new corpus for experiments.

We design a plan of all necessary activities, defining a schedule to be followed. One planned stage was the academic exchange lasting almost five months in Madrid, Spain. During this period, research would be performed with researchers from the UC3M (Universidad Carlos III de Madrid[1]) to share experiences and knowledge about state-of-the-art models in NLP.

4.2 Exploratory Phase

This phase concerns the development of a theoretical framework to support the research process, achieved by exploring publications in scientific databases. Exploratory research was executed on related work to verify current gaps and how the proposed method could contribute. Chapter 3 presents the research and the related works found.

In the exploratory phase, we also encompass the definition of our approach, the architecture, the resources, and the experiments necessary to lead this thesis. We verified that recent works obtained interesting results using the Question Answering technique, training models that identify entities through machine reading comprehension. Through several studies, we defined that our approach would be to take advantage of some concepts of the QA task while maintaining

[1] https://www.uc3m.es/Home

the way NER works, as in the work of (BANERJEE et al., 2021). We define our approach to recognizing nested, discontinuous, and multi-type entities as a new NLP task, the merge between the QA and NER. We also verified that an adapted to multi-label version of the CRF could be useful to find the nested and multi-type entities since the CRF is a machine learning technique that allows modeling the dependency between the labels of each position in the sequence, which is particularly useful for tasks like NER.

In this phase, we also research and define the NER corpora that were used in the work, in addition to the new corpus proposed:

- GENIA (KIM et al., 2003), a collection of biomedical literature compiled and annotated, containing 2,000 Medline abstracts (from PubMed), with discontinuous and nested entity annotations;
- SemClinBr (OLIVEIRA et al., 2022), a semantically annotated corpus for the Portuguese language, containing 1,000 clinical notes humanly labeled with UMLS- compatible concepts, with multi-type entities;
- RareDisease (MARTíNEZ-DEMIGUEL et al., 2022), a clinical corpus annotated with rare diseases, their signs and symptoms, containing nested and discontinuous entities [2].
- PortugueseClinicalNER (LOPES; TEIXEIRA; OLIVEIRA, 2019), a collection of 281 clinical texts in Portuguese, with manually-annotated named entities, which, although it does not have complex entities, was used to verify the performance of the method in a biomedical corpus in Portuguese with flat entities;
- JNLPBA (COLLIER; KIM, 2004), a biomedical dataset created from the GENIA corpus, by removing all nested and discontinuous entities. As it is a copy of GENIA that contains only flat entities, it will be interesting to test this corpus to compare how complex entities interfere with the model's efficiency.

The DDI corpus (HERRERO-ZAZO et al., 2013), an annotated corpus with pharmacological substances and drug-drug interactions, was not considered in our experiments since it has a low index of nested and discontinuous entities (0.2% and 0.3%, respectively, according to our analysis).

The Medical Case Report Corpus (SCHULZ et al., 2020) is a very interesting corpus as it contains nested, discontinuous, and multi-type entities. Still, until

[2] The experiments executed in this dataset cannot be compared with the baseline as they have not been evaluated in the original test set, which was not available until the writing date of this document.

the date of writing this thesis, we did not have access to the corpus, by a technical issue.

Despite ACE 2004 (MITCHELL ALEXIS, 2005), ACE 2005 (WALKER CHRISTOPHER, 2006), and NNE (RINGLAND et al., 2019) having related entities, they were not used in our experiments for not being in clinical or biomedical domains, and also they do not have free public access.

The NLPMedTerm (CAMPILLOS-LLANOS L., 2021) and CWLCE (Chilean Waiting List Corpus) (BÁEZ et al., 2020) are corpora available containing entities annotated in the health domain in Spanish texts. The first one contains 1,200 clinical trials labeled with entities from the UMLS, with 13.98% of nested entities. The second one consists of 900 referrals for several specialty consultations, with 9,029 entities, of which 32.2% are nested. Although publicly available, these two corpora were not selected in this research since they are not in English or Portuguese, being left for future work. The same goes for NEREL-BIO (LOUKACHEVITCH et al., 2023), a corpus of PubMed abstracts in Russian and English, containing nested entities.

Also, although the second version of HAREM (FREITAS et al., 2010), an initiative to evaluate the identification of proper names in the Portuguese language, has discontinuous entities (tagged with the "ALT" attribute), it is outside the scope of this work for not being in the clinical or biomedical domains.

We did not find any clinical Portuguese corpus containing nested and/or discontinuous entities.

4.3 Development

This phase can be divided into four tasks: 1) obtaining and preparing the corpora for the experiments, 2) creating a guideline and annotation of the new corpus, 3) language models training and implementation of the method, and 4) execution of experiments.

In this phase, the corpora to be used in the experiments and the annotation of the new corpus proposed in the work, NestedClinBr, were collected and prepared. The new corpus was annotated in the BRAT rapid annotation tool (STENETORP et al., 2012), a web-based text annotation tool designed for structured annotation.

We also detected a lack of trained language models in the clinical and biomedical domains in the Portuguese language. Therefore, we have developed some contextual BERT-based models trained with clinical narratives and biomedical data in Portuguese.

With the method defined, we have implemented our code using Python, the PyTorch version of the "transformers" Hugging Faces library (WOLF et al., 2020), and the package "Crfsuite" from "Sklearn" (PEDREGOSA et al., 2011).

We also have executed several empirical tests to define the best settings and the algorithm used. The developed resources and experiments performed to validate the method are:

E1 Construction of a corpus with nested and discontinuous entities for the Portuguese language;

E2 The training of clinical and biomedical language models for the Portuguese language;

E3 NER experiments in the NestedClinBr corpus, with nested entities (E3.1) and nested and discontinuous entities (E3.2);

E4 NER experiments in the SemClinBr corpus;

E5 NER experiments in the GENIA corpus, with nested entities (E5.1), nested and discontinuous entities (E5.2), and a few-shot experiment (E5.3);

E6 NER experiments in the Rare Disease corpus;

E7 NER experiments in the PortugueseClinicalNER.

E8 NER experiments in the JNLPBA corpus.

4.4 Evaluation

This phase consists of evaluating the proposed method, analyzing the results, and extracting conclusions. After implementing the method, several tests were performed to define the best configuration and parameters. In addition to these tests, experiments were conducted with a baseline (using binary relevance) and some similar methods (employing the same approach) to facilitate comparisons with the proposed method.

In this section, we present the metrics used to measure the performance of the method, in order to compare it with other methods, the statistical approach used to measure whether or not the results are significant, and the agreement metrics used to measure the level of our corpus annotation.

4.4.1 Method Performance

In token-level classification tasks, such as NER, when comparing the golden standard annotations with the output of a system, different scenarios might occur:

I. Surface string and entity type match;

II. The system hypothesized an entity;

III. The system misses an entity;

IV. The system assigns the wrong entity type;

V. The system gets the boundaries of the surface string wrong;

VI. The system gets the boundaries and entity type wrong.

In the literature, there are several ways to evaluate a NER system, such as the way used by Conll (SANG; MEULDER, 2003) that considers only scenarios I, II, and III, discarding the others; the Message Understanding Conference (MUC) that defines the number of correct, incorrect, partial, missing and spurious entities (CHINCHOR; SUNDHEIM, 1993); and the one introduced in SemEval (SEGURA-BEDMAR; MARTÍNEZ; HERRERO-ZAZO, 2013), where it measures the performance accounts for correct, incorrect, partial, missed and spurious in different ways.

While traditional named entity recognition systems utilize these NER metrics, our study evaluates complex NER as a standard classification task, in alignment with other comparable studies such as (LI et al., 2020), (SHEN et al., 2022), and (SOHRAB; MIWA, 2018). In the context of complex entity recognition, assessment typically adopts an entity-based approach (as opposed to a token-based one), requiring accurate prediction of both the span and entity type, while disregarding partial matches. The metrics employed in our experiments encompass Precision (P), Recall (R), and F1-score (F1).

Precision is a measure of the correct named entities identified by the models, which is defined by Equation (2). Recall, also known as sensitivity, calculates the ratio of true positives to the total actual positives in the dataset, as defined by Equation (3). The F1-measure metric represents the harmonic mean between precision and recall, defined by Equation (4).

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

$$F1 = 2 * \frac{Precison * Recall}{Precison + Recall} \qquad (4)$$

In these equations:

- True positives (TP) refer to the count of accurately predicted entities. This means the model correctly identifies an entity that exists in the gold standard.
- False positives (FP) correspond to the count of incorrect predictions classified as positive. This involves instances where the model incorrectly identifies an entity that is not present in the gold standard or inaccurately identifies an entity from the gold standard.
- False negatives (FN) represent the count of incorrect predictions classified as negative. This occurs when the model fails to identify an entity that is present in the gold standard.

There are two approaches for evaluating system performance: the micro-average and the macro-average. In the micro-average, first, we sum all error types of all documents and then make the average of each metric (Equations (5) and (6)). In the macro-average, we first calculate the Precision, Recall, and F1-score for each document and then make the average for all instances (Equations (7) and (8)). Since the micro-average weighs each instance separately, it can capture the imbalance between the documents, being more suitable for our work.

$$Precision_{\mathrm{micro}} = \frac{TP_1 + TP_2 + ... + TP_{\mathrm{n}}}{TP_1 + TP_2 + ... + TP_{\mathrm{n}} + FP_1 + FP_2 + ... + FP_{\mathrm{n}}} \qquad (5)$$

$$Recall_{\mathrm{micro}} = \frac{TP_1 + TP_2 + ... + TP_{\mathrm{n}}}{TP_1 + TP_2 + ... + TP_{\mathrm{n}} + FN_1 + FN_2 + ... + FN_{\mathrm{n}}} \qquad (6)$$

$$Precision_{\mathrm{macro}} = \frac{P_1 + P_2 + ... + P_{\mathrm{n}}}{n} \qquad (7)$$

$$Recall_{\mathrm{macro}} = \frac{R_1 + R_2 + ... + R_{\mathrm{n}}}{n} \qquad (8)$$

4.4.2 Statistical Evaluation

There are some statistical methods in order to verify if the results were statistically significant, such as the non-parametric tests, which do not make assumptions about the underlying distribution of the data, as normality. These tests are usually used when the sample size is small and the distribution of the data is unknown. Some examples of non-parametric tests include the Wilcoxon rank-sum test, the Kruskal-Wallis test, the Mann-Whitney U test, the Friedman test, and Spearman's rank correlation coefficient, all used for a variety of purposes.

The Friedman test assesses the differences between repeated-measures data, where the same set of subjects is measured under different conditions or treatments (SHESKIN, 2007). It provides a useful alternative when the assumptions of the repeated-measures ANOVA are not satisfied. After applying the Friedman test, the Nemenyi post-hoc test can determine which pairs of groups are significantly different from one another. Being a rank-based test, it uses the concept of the Nemenyi distance (the difference between two groups based on their ranks in the original data) to determine the significance of the differences between pairs of groups.

4.4.3 Evaluation Corpus Annotation

There are also specific metrics to assess the consistency and quality of a new corpus, such as the Inter Annotator Agreement (IAA), to find possible disagreements between annotators and avoid inconsistencies in the annotation guidelines. Cohen's Kappa is a common measure for IAA, however, for NER annotation the F1-measure has become the standard metric, as the "O" token has a higher frequency and the Kappa score would be misguidedly too high (DELEGER et al., 2012). For NER annotation, the F1-measure is used to evaluate the agreement between two annotators, in a token-based way, checking the consistency under the exact match criteria of both annotators.

5

Method

This chapter presents the proposed two-phase method for recognizing nested, discontinuous, and multi-type named entities. The first phase of the method is a QA-Based NER task, in which the NER is formulated as a machine reading comprehension task. Therefore, extracting Protein-like entities, for example, is formalized as extracting answer spans to the question "Which proteins are mentioned in the text?". Since the query encodes informative prior knowledge, this strategy facilitates the entity extraction process. Naturally, it tackles the overlapping entity issue in nested NER, answering two (or more) independent questions. Our approach consists of a token-level classification, as in NER, but sending a query to the model along with the sentence tokens, receiving as output all the entities related to that query. Therefore, the method can deal with nested and multi-type entities naturally, since the outputs are independent of the class for each input query. The second phase concerns the training of a CRF model adapted to deal with multi-label outputs through a predefined threshold. Finally, the results of the QA-NER model are added to those of the multi-label CRF, similar to an ensemble technique, generating the final result.

Hence, to address the recognition of complex named entities in the clinical and biomedical domain, we propose BioNestedNER, an ensemble-based approach formed by a two-phase method, as illustrated in Figure 5.1.

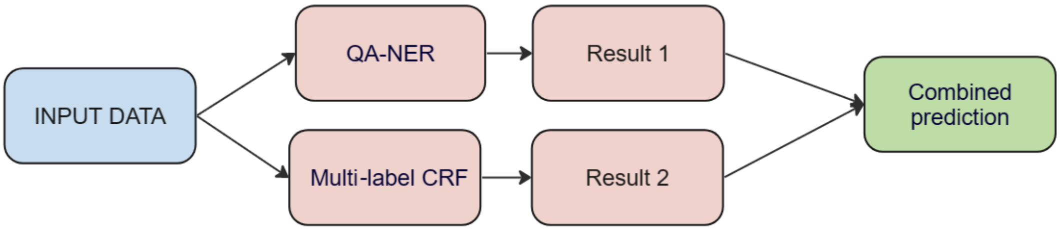


Figure 5.1: A two-phase method for nested, discontinuous, and multi-type entity recognition.

5.0.1 Assumptions

Based on the objective and motivation defined at the beginning of this research, the method has the following assumptions:

- It must recognize nested, multi-type, and discontinuous entities in addition to flat entities;
- It should be flexible and consume less computational resources than exhaustive models;
- It must classify the entity type at the same moment of extraction, in a single step, in a real multi-label approach, without the necessity of training one model for each class as in binary relevance;
- Should be possible to use it in different domains and languages.

Next, we present the two phases of the method separately.

5.0.2 Phase 1- QA-NER Approach

In the first phase, we create models to recognize named entities using the QA-NER approach. With this approach, the model can identify nested entities in the text eliminating the necessity of training multiple models as required by the layered-based approaches, or enumerate all possible regions in the text as seen in exhaustive region-based methods. The QA-NER approach enables an efficient identification of nested entities without the need for computationally intensive procedures.

Some similar works also apply a QA-NER approach to find nested and flat entities such as (LI et al., 2020), (SHEN et al., 2022), and(BANERJEE et al., 2021), but in the same way that (BANERJEE et al., 2021), our model is trained to return the results in NER format (e.g., IOBES or IOB2), instead of the index of the beginning and end of the entity, as occurs in QA tasks. In this way, with a single query, the model can return all the entities related to the query, token by token, at once, without the need to call the same query several times to find more entities of the same type in the sentence, nor to perform post-processing with word indexes. We have improved the QA-NER model proposed by (BANERJEE et al., 2021) by implementing some modifications, such as the removal of the CNN layer (for a simpler architecture, since the Transformer architecture already allows the contextualization of words), application of treatment for class imbalance, adjust segments embeddings, and add some specific treatments for discontinuous and nested entities of the same type.

In this work, we explore BERT-based language models to generate word representations and make the final classification for each token, as BERT architecture has proven efficient for NLP [1].

An overview of the method is presented in Figure 5.2.

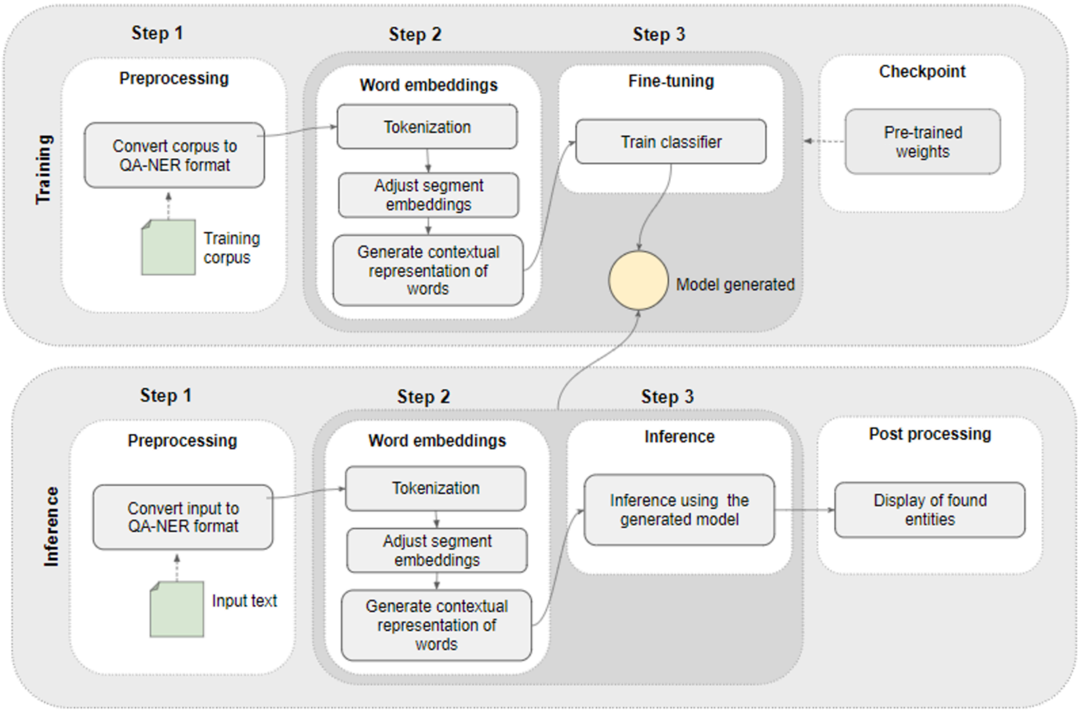

Training
Step 1
Preprocessing
Convert corpus to QA-NER format
Training corpus
Step 2
Word embeddings
Tokenization
Adjust segment embeddings
Generate contextual representation of words
Step 3
Fine-tuning
Train classifier
Checkpoint
Pre-trained weights
Model generated
Inference
Step 1
Preprocessing
Convert input to QA-NER format
Input text
Step 2
Word embeddings
Tokenization
Adjust segment embeddings
Generate contextual representation of words
Step 3
Inference
Inference using the generated model
Post processing
Display of found entities

Figure 5.2: Overview of phase 1 of the method.

[1] While we selected BERT for our method, it's worth noting that other language models can also be employed.

Task Formalization

Given an input sequence $X = x_1, x_2, ..., x_n$ where n denotes the length of the sequence in tokens, the model is trained to assign a label $y \in Y$ to each token x, where Y corresponds to entity boundary tagging label in the sentence. In IOBES tagging scheme, Y could be defined as *Y = B-ENT, I-ENT, E-ENT, S-ENT, O*. For IOB2, *Y = B-ENT, I-ENT, O*. The output of the model (y_n) only indicates the start, continuation, and end of the entity, without needing to provide the type of entity, since the tag label is sent in the input text, concatenated with the sentence as a query (or question).

A query q_t, where $t \in T$, is defined as a natural language identifier, in which T is the predefined list of all possible entity types (e.g. person, location, organization). Therefore, to an input (q_t, X), the output is $y_1, y_2, ..., y_n$, indicating that for the t tag, the entities found are all y that are different from "O". The query must be represented by a word or a set of words whose semantic meaning is really close to the label, in most cases, the label itself can be used as a query. Hence, for each query (i.e. for each t), the same sentence must be sent to the model, concatenated (with a [SEP] token) with the query corresponding to t (q_t).

Training

The method has three steps in the model training stage: a preprocessing step, the generation of the contextual representation of tokens, and the fine-tuning for the task. Next, each step of the method will be discussed in detail.

Step 1 Preprocessing the input text

Initially, the training corpus must be converted to NER QA-based format, in order to be processed in the following steps, as can be seen in Figure 5.3. As we used BERT-based models, the first token is a special [CLS] token, indicating the beginning of the sentence. Next, we add the query, i.e., a word (or a small set of words) that describes the entity type we are looking for (e.g. Protein). We add a separator token, in this case [SEP], and concatenate the entire sentence, indicating the entities of this specific type in NER format (as IOB2). This process is repeated for each type of entity, replicating the same sentence and concatenating with the new query. In Appendix 10.2, we have examples of inputs in the JSON format, expected by the system.

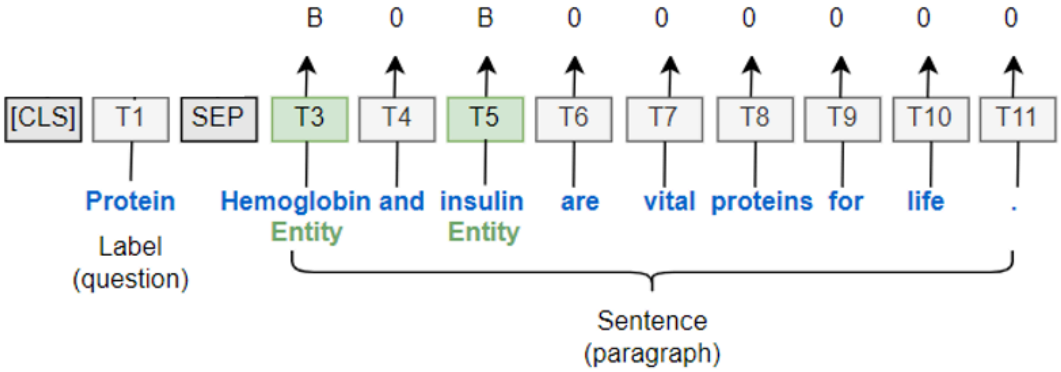


Figure 5.3: Representation of QA-NER input format, using IOB2 schema.

Step 2 Generation of Word Embeddings

In this step, the contextual representation of each word of the sentence is generated, using the weights of pre-trained BERT models and their derivatives as checkpoints.

Firstly, the sentence is tokenized, using tokenizers such as SentencePiece or Byte-Pair Encoding, for example. The process of tokenization is essential in NLP tasks, where the stream of text is split into separate, smaller "tokens" (any meaningful unit that the tokenizer has been programmed to identify). We made an adjustment so that the special tokens are added in the right places, especially the sentence separation token ([SEP]).

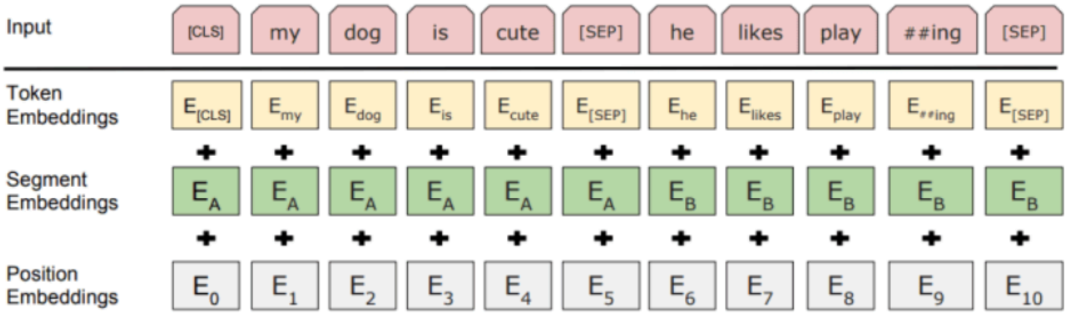


Figure 5.4: BERT word embedding layer. Source: (DEVLIN et al., 2019)

Next, we adjust the segment embeddings, to indicate to the model that there has been a sentence break in the input text (E_A to tokens of the first sentence, i.e. the query, and E_B to tokens of the second sentence). This is required in our method, as we use the sentence separator token [SEP] to separate the input query from the input sentence. The other layers that the encoder receives, the token embeddings (matrix of embeddings) and position embeddings (numbers that give importance to the order of words) were not changed. Figure 5.4 shows the embedding layers that are summed and sent to the BERT architecture.

Finally, the model generates the word embeddings, vector representations for each word in a text segment, where each dimension of the vector encodes a different aspect of the meaning of the word. These word embeddings capture the semantic relationships between words, in a way that words with similar meanings have similar embeddings. Transformer-based models can capture the context-dependent meaning of words, important where the meaning of a word can change based on the context in which it is used.

Step 3 Fine-tuning for the Task

In this step, the fine-tuning to the new task is performed.

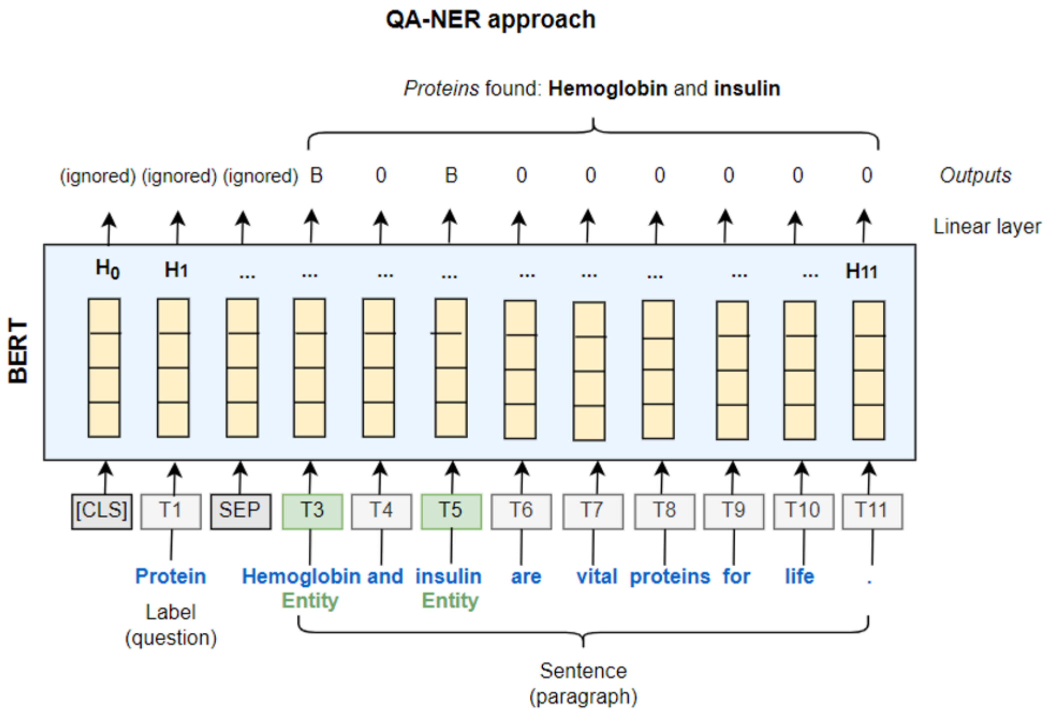

QA-NER approach
Proteins found: Hemoglobin and insulin
(ignored) (ignored) (ignored) B 0 B 0 0 0 0 0 0
Outputs
Linear layer
H0 H1 ... ... H11
BERT
[CLS] T1 SEP T3 T4 T5 T6 T7 T8 T9 T10 T11
Protein
Label
(question)
Hemoglobin and insulin are vital proteins for life .
Entity
Entity
Sentence
(paragraph)

Figure 5.5: Fine-tuning architecture for the hybrid task combining NER and QA.

Fine-tuning refers to the process of adapting a pre-trained machine learning model to a specific task by training it on a smaller dataset for that task. A linear layer for token-level classification is added on top of the model, and for each input token, an output is generated. The linear layer, also known as dense layer or fully connected layer, performs a linear operation on its inputs (word embeddings) and produces a vector of outputs through a matrix multiplication and an optional bias term.

In our method, as in the traditional QA task, the outputs of the [CLS] token and of the tokens from the first sentence (the query) were ignored. We leverage the outputs from the first token of the second sentence, which returns the outputs in NER format, indicating only the entity's delimitation (and not the entity's type, since it is in the query). The architecture of this step can be seen in Figure 5.5.

Adaptation to Find Discontinuous Entities

As one of the main gaps in the recognition of complex entities is recognizing discontinuous entities, common in clinical and biomedical texts, we adapted the QA-NER method in order to also recognize discontinuous entities.

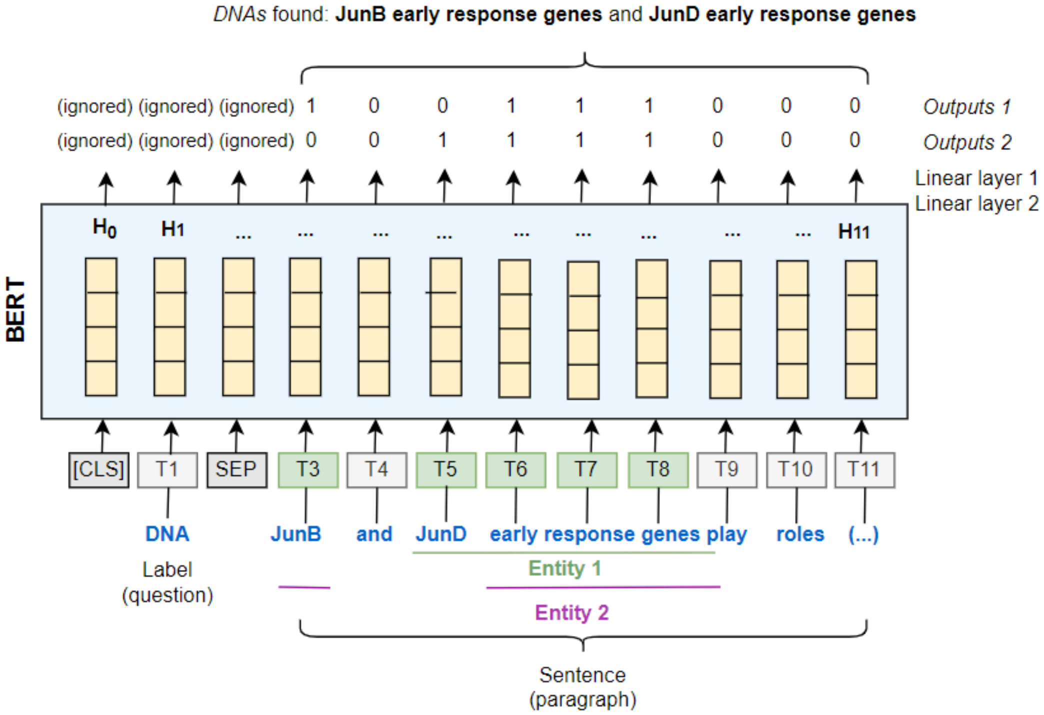

DNAs found: JunB early response genes and JunD early response genes
(ignored) (ignored) (ignored) 1 0 0 1 1 1 0 0 0
(ignored) (ignored) (ignored) 0 0 1 1 1 1 0 0 0
Outputs 1
Outputs 2
Linear layer 1
Linear layer 2
BERT
H0 H1 ... ... H11
[CLS] T1 SEP T3 T4 T5 T6 T7 T8 T9 T10 T11
DNA
Label
(question)
JunB and JunD early response genes play roles (...)
Entity 1
Entity 2
Sentence
(paragraph)

Figure 5.6: Adaptation to find discontinuous entities.

To detect discontinuous entities, separated into non-sequential parts, we transformed the model into an end-to-end model trained to identify both regular entities and discontinuous ones simultaneously, sharing the same embeddings and loss during training and at the end, the same model weights. We have experimented with this strategy on GENIA and NestedClinBr corpora.

In the pre-processing step, we need to send two labels for each token, one for each classifier, in both the training and inference stages. In the training step, we train two classifiers simultaneously, each one specialized for a sub-task, i.e. find regular entities and discontinuous entities, as shown in Figure 5.6. At the end, each classifier will learn to classify its specific sub-task, and during inference, both types of entities are identified. For extracting the discontinuous entities, we used the IOB tagging scheme, which marks the beginning (B) and the continuation (I) of the entity, and with this, it is possible to reconstruct the entity at the end.

The same applies to nested entities of the same type, as occurs in the GENIA corpus. With an end-to-end model, it is possible to recognize entities with this characteristic, as the example shown in Figure 5.7.

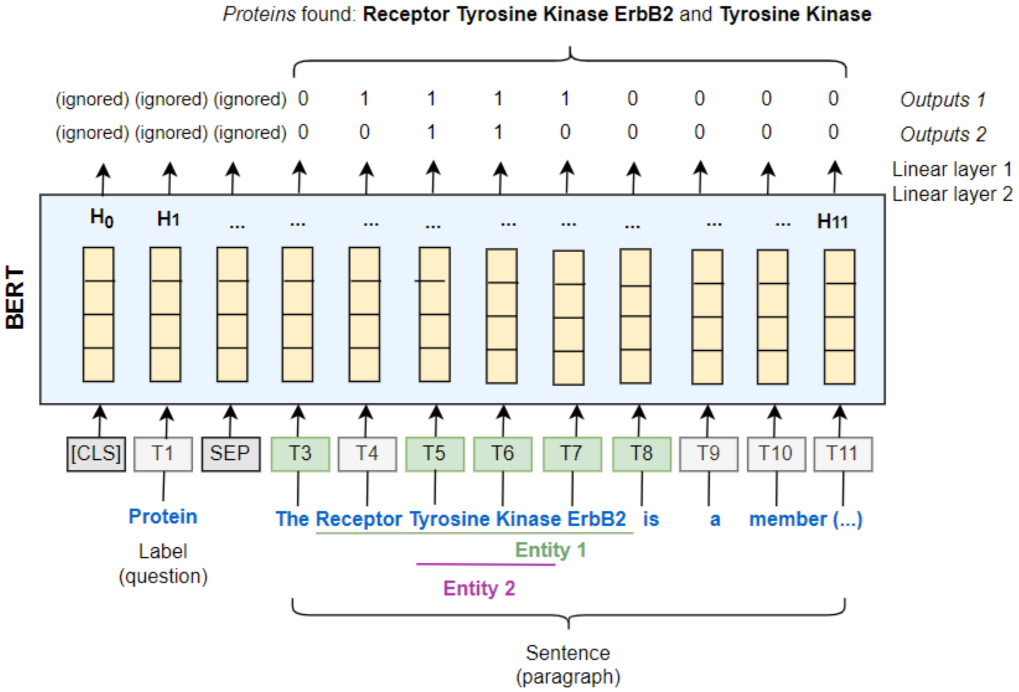

Proteins found: Receptor Tyrosine Kinase ErbB2 and Tyrosine Kinase
(ignored) (ignored) (ignored) 0 1 1 1 1 0 0 0 0
(ignored) (ignored) (ignored) 0 0 1 1 0 0 0 0 0
Outputs 1
Outputs 2
Linear layer 1
Linear layer 2
BERT
H0
H1
...
H11
[CLS]
T1
SEP
T3
T4
T5
T6
T7
T8
T9
T10
T11
Protein
Label
(question)
The Receptor Tyrosine Kinase ErbB2 is a member (...)
Entity 1
Entity 2
Sentence
(paragraph)

Figure 5.7: Adaptation to find nested entities of the same type.

Treatment for Class Imbalance

As in this technique there is an increase in the number of non-entity ("O") classes (more than in a traditional NER), which can result in class imbalance, we have adapted the method by applying class weights to the underrepresented classes. This approach helps to improve the overall performance of the model by reducing the impact of the class imbalance on the learning process and enhancing the model's ability to correctly identify entities of all types. Given a set of class labels and the corresponding frequency of each class in the training data, it was calculated a weight for each class that can be used to adjust the contribution of each class during the training process. The weights are inversely proportional to the frequency of each class, which means that the less frequent classes are assigned higher weights to increase their impact on the learning process. This helps to improve the overall performance of the model by reducing the impact of class imbalance on the learning process and improving the model's ability to correctly classify minority classes. The formula used can be seen in Equation (9), where $n_{samples}$ is the total number of samples in the training data, $n_{classes}$ is the total number of unique classes, and *np.bincount(y)* counts the frequency of each class label (y) in the training data.

$$class_{\text{weight}} = \frac{n_{\text{samples}}}{(n_{\text{classes}} * np.bincount(y))} \qquad (9)$$

We have used the $class_{weight}$ lib from Sklearn[2] to compute the class weights, using the "balanced" parameter. Besides the traditional method for calculating a weight for each individual class, based on its frequency in the training data (*ClassWeights*), we also proposed a new way to calculate class weights for QA-NER approaches. This involves computing binary weights that consider weights for just class and non-class (e.g. "entity" vs "O"), in a smoother way, prioritizing the search for entities, regardless of their type (*BinaryWeights*).

In Figure 5.8, we observe the percentage of each entity type present in the NestedClinBr corpus. In the ClassWeights variant, we calculate the weights for each class using their individual percentages (e.g., 3.9% for the Anatomy class), while in the BinaryWeights variant, we calculate the weights for all entities (using the percentage of 33.2%).

[2] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

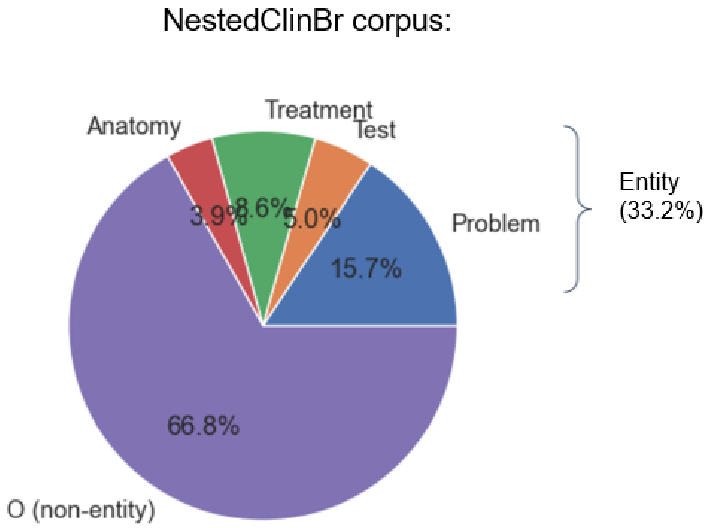


Figure 5.8: Percentage of each type of entity present in the NestedClinBr corpus.

In Equation (10), we can see the formula of the Cross-entropy loss function:

$$l(x,y) = L = l_1, ..., l_{\mathrm{n}} = -\sum_{c=1}^{C} w_c log \frac{exp(x_{\mathrm{n,c}})}{\sum_{i=1}^{C} exp(x_{\mathrm{n,i}})} y_{\mathrm{n,c}} \qquad (10)$$

, where x is the input, y is the target, w is the weight, C is the number of classes, and N spans the minibatch dimension.

We also test some different reduction parameters in the CE loss function (how the individual losses are aggregated or averaged across the batch). "Mean" is the default value, where the individual losses are averaged over all samples in the batch, minimizing the average loss across the entire training data. We have tested with "sum" as well, where the individual losses are summed over all samples in the batch, computing the total sum of the losses (Equation 11 shows the adapted formulas[3]). This option is useful to minimize the total loss across the entire training data, more appropriate to account for the skewed distribution of samples across the classes.

$$l(x,y) = \begin{cases} \frac{\sum_{n=1}^{N} l_{\mathrm{n}}}{N} \quad if \ \ reduction = "mean"; \\ \hline \sum_{n=1}^{N} l_{\mathrm{n}}, \quad if \ \ reduction = "sum". \end{cases} \qquad (11)$$

[3] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

Inference and Test

During inference and testing, the method will also have the same preprocessing and word embeddings generation steps, as shown in Figure 5.2. The only difference is in step 3, where the previously trained model is used to generate the outputs. After inference by the model, a post-processing step is necessary to display the entities found in a user-friendly format. The overview of the inference process is shown in Algorithm 1.

Algorithm 1 Overview of the inference process

```
Require: $X$: input tokens ( $X = \{x_1, x_2, ..., x_n\}$); $T$: entity types ( $T = \{t_1, t_2, ..., t_n\}$).
queries = [t for t in T]
listAllEntities = [ ] # all entities extracted from the sentence
for query q in queries do
    # Concatenates the query with the input sentence and adds the separator token
    $X'$ = concatenate(q, X)
    # Tokenizes and adjusts the segment embeddings
    $X'$ = tokenize($X'$)
    # Gets the labels from $X'$
    Y = NER($X'$)
    # Gets the entities of type t from Y
    entities = processEntities(Y)
    if entities != null then
        listAllEntities.append(entities)
    end if
end for
# Return the list with all entities found in the input sentence
return listAllEntities
```

Parameters

The proposed method includes the following parameters:

- *max_length*: maximum number of tokens per sentence;
- *t*: number of entity types (determined from the corpus);
- *q*: number of queries (determined from *t*)
- *s*: number of sentences in the corpus;
- *i*: number of instances for training.

The real number of instances for training is $i = s * q$, since each sentence will be replicated as many times as the number of queries. Both number of queries (q) and number of training sentences (s) determine the training velocity, where the smaller s and q, the faster the model will be trained. The same applies for *max_length*, making it challenging to determine the ideal value since if set too high, can result in increased memory usage and slower performance, but if it is set too low, it can lead to loss of important information and degraded performance. The number of queries (q) can interfere with performance, usually when the fewer classes for the model to learn, the better its efficiency.

The selection of the pre-trained model as a checkpoint to the fine-tuning process also is an important adoption that could have a large impact on the results. Usually, applying domain-specific models can generate better results for tasks in this domain (or similar domain), as well as language-specific models. Hence, the pre-trained language model to be used as a checkpoint depends on the experiment to be performed. In addition to the method parameters, there are some generic hyper-parameters of the architecture, which have been defined during our experiments (empirically):

- Dropout rate, a regularization technique for reducing overfitting in neural networks with many layers, where a random number of neurons is temporarily excluded during training;
- Learning rate, which determines the step size at which the optimizer makes updates to the model parameters. A high learning rate can lead to fast convergence, but with the risk of getting stuck in suboptimal minimum, while a low learning rate can increase the stability of the optimization process, but slowly;
- Warmup, a technique applied at the start of the training process where the learning rate is gradually increased from a low value to a higher value over a few training iterations, to allow the model parameters to gradually adjust to the optimization process;
- Optimizer, a method for finding the minimum or maximum of the loss function, finding the best parameters for a machine learning model;
- Weight decay, a regularization technique that adds a penalty term to the loss function during training to discourage the model from having large weights, aiming to reduce overfitting and improve the generalization performance of the model;
- Batch size, which defines the number of samples used in one iteration of training, to update the model's parameters in one forward and backward pass, impacting on the model's convergence speed, memory usage, and computational cost;

- Number of epochs, that defines the number of times the entire training dataset is passed through the model during training, determining how many times the model will be exposed to the training data;
- Early stop, to prevent overfitting and improve the generalization performance of the model, interrupting the training process before the model has completed all the specified number of epochs, based on a monitoring metric such as the validation loss or accuracy.

Adjusts to Other Architectures

Although the method is focused on the BERT architecture, it can be adapted to work with other architectures, such as RoBERTa, Flair, GPT-3, etc. In these cases, if the tokenizer is not the WordPiece, as the Byte-Pair Encoding used by RoBERTa and GPT-3, one must change the special tokens (e.g. [SEP] to </s>) and also the alignment function.

5.0.3 Phase 2- Multi-label CRF Approach

Combining the outputs of a CRF layer with our Transformer-based model can further improve the general coverage, by combining the strengths of both models.

In phase 2 of the method, we train CRF models, using as features the Morphological, Orthographical, Context, POS, and Semantic information, frequently used in Biomedical NER, following the works of (MADY; AFIFY; BADR, 2022), (ZHANG et al., 2004), and (CAMPOS; MATOS; OLIVEIRA, 2012).

The morphological features analyze the constituents of words and their interactions, examining the structural similarities between words. Orthographic features group words with similar forms and are often used to capture information about word formation. Contextual features consider the words before and after a token to determine its class label. In our work, we considered a window of 4 tokens left and right. Part-of-speech features identify named entities based on POS information, e. g. nouns are typically strong candidates for named entities, whereas verbs and prepositions often indicate named entity boundaries. In our work, we trained a Transformer-based clinical POS tagger for Brazilian Portuguese (SCHNEIDER et al., 2022) to extract part-of-speech information from texts in Portuguese. Semantic features focus on the meaning of words and their relationships with each other in a sentence or text. We also included a clustering-based feature, as in (MADY; AFIFY; BADR, 2022), where features are extracted from clustering algorithms applied to text data. To create the clusters, for English we used the Word2vec models provided by (CHIU et al., 2016), using the K-means algorithm for grouping similar data points and then creating features for each data point that indicate the distance from the center of the cluster. For the experiments in the Portuguese language, we trained a clinical Word2vec using data from a Brazilian Hospital, containing 157,929 de-identified clinical narratives[4]. For both English and Portuguese, we have created five worksheets containing clustered words, with cluster numbers defined as 5, 10, 50, 100, and 300. Table 5.1 shows more details of features utilized in the proposed approach.

In a multi-label NER task, each token can be associated with multiple labels, rather than a single label. Nested entities can also be treated as multi-label, since each token can have more than one label type, as in the example "Bank of Brazil", where the token Brazil can be a "local" and an "organization" at the same time.

We adapted the CRF training and inference steps, transforming the single-label problem into a multi-label problem, without the need to train a binary model for each class. We used the Problem Transformation Method 5 (PT5) strategy presented by (TSOUMAKAS; KATAKIS, 2009), which decomposes each example *(x, Y)* into $|Y|$ examples *(x, l)* for all $l \in Y$ and learns a single-label coverage-based classifier from the transformed dataset. In other words, for each existing multi-label token in the input sentence, we replicate the same information, each with a different label, as can be seen in Figure 5.9.

For each input token, the CRF model returns a probability distribution over each label, indicating the probability that the token belongs to each label type. To adapt CRF to a multi-label NER task, we defined a threshold value to classify the model's output into binary categories (positive or negative) for each class. We defined several threshold values (e.g. values between 0.15 to 0.8) and empirically tested in each corpus to define its ideal value. For example, if the threshold is set to 0.5, any predicted probability above 0.5 is classified as positive and any predicted probability below 0.5 is classified as negative. If the threshold is increased, the model becomes more conservative in its predictions and may

[4] Approved by the Institutional Review Board with the ethical approval n. 5944847.

Table 5.1: Features extract to the CRF model training.

Feature type	Feature name	Description	Example
Morphological	Prefix	Set of characters that are taken from the leftmost location of the words, with length 3.	"cyc" for the word "cycloheximide"
	Sufix	Set of characters that are taken from the rightmost location of the words, with length 3.	"ide" for the word "cycloheximide"
Orthographical	Is Upper	Check if the word is uppercase.	"True" for the word "HB24"
	Is title	Check if the first letter is uppercase	"True" for the word "The"
	Has digit	Check if all the characters are digits.	"False" for the word "HB24"
Semantic	UMLS concept	A unique identifier assigned to a concept in the UMLS Metathesaurus, a comprehensive biomedical terminology database.	"chem" for the word "cycloheximide"
	Clustering-based feature	Cluster number where the word is found. Similar words tend to group together in the same cluster.	"5" for the word "disorder"
Part of speech	POS	The POS of each token.	"verb" for the word "was"
Context	Context feature	Refer to tokens and their information that appear within a 4-word window size.	Four tokens to the right and four tokens to the left of the token

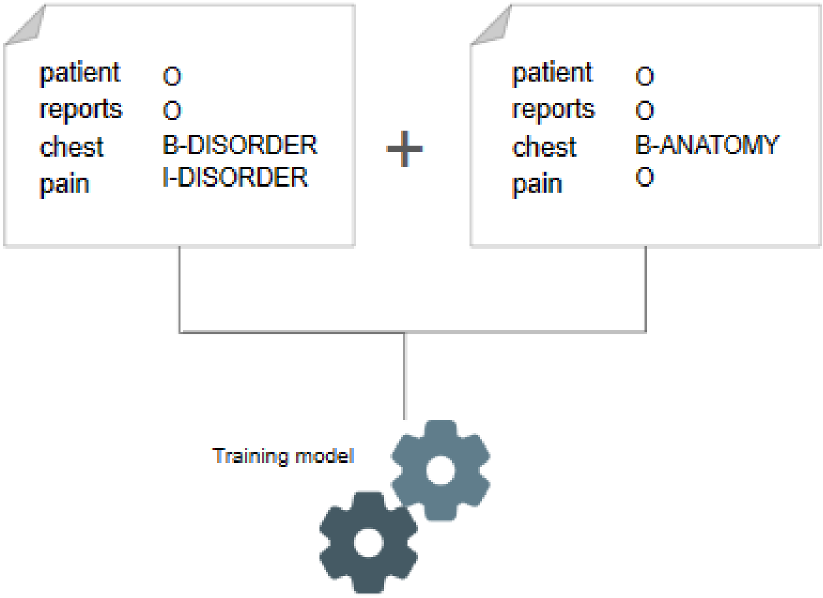


Figure 5.9: Input example for CRF model with nested entities.

achieve higher precision but lower recall, however, if the threshold is decreased, the model becomes more liberal in its predictions and may achieve higher recall but lower precision.

In the training of the CRF models, we employed the following settings: a maximum of 100 iterations, utilizing the "lbfgs" optimization algorithm, and the Viterbi algorithm for calculating the energy function, from the "Crfsuite" package within Scikit-Learn (PEDREGOSA et al., 2011).

5.0.4 Limitations

Although the method does not require as much computational power as the exhaustive methods, the training time of the Transformer-based model may vary according to the number of entity types, where the more types, the longer the time. This occurs because each sentence is sent *t* times to the model during training, where *t* is the number of entity types. For a corpus with many sentences, the training time can be impactful. The adapted model (end-to-end) to find discontinuous and nested entities of the same type can also be slower than a simple QA-NER model. The model also has limitations in finding nested entities of the same type with multiple levels, limited to two levels of nesting. The same goes for discontinuous entities. Also, the treatment to deal with class imbalance only concerns "entity" vs. "non-entity" and does not consider the imbalance that can occur between different classes in the corpus, although the *ClassWeights* variant we proposed alleviate this situation.

The training of CRF models, in turn, is extremely fast and does not require GPUs. However, to find the best configuration of the number of clusters, the number of words per window, and the threshold, several CRF models need to be trained and evaluated, which can demand more work. Also, the CRF model has limitations in finding discontinuous entities, only working to find multi-type and nested entities. The CRF models alone show inferior results compared to deep learning models trained with Transformer architecture, serving only as a complement to the method.

5.0.5 Final Considerations

In summary, our method consists of two phases that can be assessed separately or combined for greater coverage. In Phase 1, the method uses the QA-NER approach, proposed by (BANERJEE et al., 2021), with some improvements:

- A treatment for class imbalance, assigning different weights to classes to compensate for imbalances in the data;
- Adjust segments embeddings to separate sentence one (question) from sentence two (context) during training;
- An adaptation to find discontinuous entities, using an end-to-end model with two classifiers, one for normal entities and one for discontinuous entities;
- An adaptation to find nested entities of the same type, a situation that occurs in some corpora such as GENIA. Since the original method does not deal with this situation, we also adapted the method providing an end-to-end model with two classifiers, in which one searches for the outside entities and the other, the inside ones.

In Phase 2, we propose a multi-label CRF adapted to work in a multi-label way, using thresholds to define the labels.

The proposed method provides a flexible and efficient NER solution that can handle nested, discontinuous, and multi-type entities, which can benefit many clinical applications.

6

A new Portuguese-language clinical corpus

In this section, we detail the development of NestedClinBr, a new corpus containing nested and discontinuous entities in Brazilian Portuguese clinical narratives.

The main goal of NestedClinBr is to provide a human-annotated corpus that can be used for learning and evaluating different machine learning models to extract valuable medical information in the Portuguese language, in special nested and discontinuous entities, an important but less explored task.

In the context of clinical NLP, the recognition of entities is commonly used for the identification of diseases, body parts, medications, and other relevant information, facilitating, for example, the detection of risk factors and medical decision-making (DALIANIS, 2018). As NestedClinBr, although small, can contribute to the healthcare domain, it will be freely available to the research community.

6.0.1 Data Acquisition

The data for the construction of the corpus originates from TempClinBr (GUMIEL et al., 2023), a corpus containing clinical notes on the cardiology domain in Portuguese, annotated for entity recognition and temporal relations. Formed by 126 clinical notes from hospitals in Brazil, containing both structured and unstructured data, the corpus comprises 2,347 sentences and 20,907 tokens.

These texts were selected according to some criteria, such as being from a single specialty (in this case, cardiology), which contains mentions of specific problems, treatments, and diagnostic procedures. This selection approach enhances the coverage of the guideline with a high degree of entity representativeness. All the clinical texts have been properly de-identified, to respect patient privacy and the Brazilian General Data Protection Law (LGPD) [1]. The research was approved by the Ethical Committee databases (Certificate of Presentation for Ethical Appreciation number 51376015.4.0000.0020).

TempClinBr is originally annotated with entities of the types: "Problem", "Treatment", "Test", "Evidence", "Occurrence", and "Clinical Department", in addition to other annotations such as polarity and temporal relations. To generate NestedClinBr, we have discarded the tag of entities "Evidence", "Occurrence", and "Clinical Department", as well as the other tags, keeping "Problem", "Treatment", and "Tests". Following the example of (CAMPILLOS-LLANOS L., 2021) and (BÁEZ et al., 2020) corpora, we added the entity "Anatomy", which refers to the location of the human body, or "Bodypart" as it is called in (BÁEZ et al., 2020). Besides its medical relevance, which allows the extraction of relevant medical information, this entity is important for the study of nested entities as usually a problem or treatment occurs or is linked to a body part. In addition to the inclusion of the new entity, the maintained entities ("Problem", "Treatment" and "Test") were also revised in order to label the discontinuous mentions.

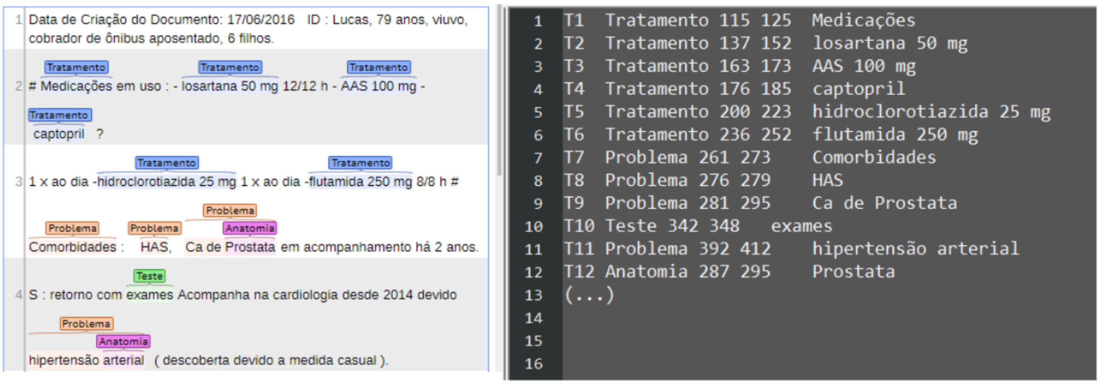
6.0.2 Annotation Tool

To perform the manual annotations, the BRAT rapid annotation tool[2] (STENETORP et al., 2012) was used, a web-based tool for text annotation, i.e. the addition of notes to existing text documents. Designed for structured annotation, where the notes have a fixed form that can be automatically processed and interpreted by a computer, allows the visual identification of marked mentions and the relation between them. We have selected this tool since it allows annotating discontinuous, multi-type, and nested entities in an intuitive and user-friendly way, besides being one of the most comprehensive tools and most popular, regarding the number of citations according to (NEVES; ŠEVA, 2019).

[1] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[2] https://brat.nlplab.org/

For each text file, BRAT creates an ANN file containing the corresponding annotations for that text, becoming a standard of corpora annotations for NLP tasks. Figure 6.1 shows an example of text annotated with the BRAT tool. In (a), we can see the visual interface, and in (b), its corresponding annotation file is displayed, which stores the type of entity, span indexes, and strings for each mention.



1 Data de Criação do Documento: 17/06/2016 ID : Lucas, 79 anos, viuvo, cobrador de ônibus aposentado, 6 filhos.

2 # Medicações em uso : - losartana 50 mg 12/12 h - AAS 100 mg - captopril ?

3 1 x ao dia -hidroclorotiazida 25 mg 1 x ao dia -flutamida 250 mg 8/8 h # Comorbidades : HAS, Ca de Prostata em acompanhamento há 2 anos.

4 S : retorno com exames Acompanha na cardiologia desde 2014 devido hipertensão arterial (descoberta devido a medida casual).

```
T1  Tratamento 115 125  Medicações
T2  Tratamento 137 152  losartana 50 mg
T3  Tratamento 163 173  AAS 100 mg
T4  Tratamento 176 185  captopril
T5  Tratamento 200 223  hidroclorotiazida 25 mg
T6  Tratamento 236 252  flutamida 250 mg
T7  Problema 261 273    Comorbidades
T8  Problema 276 279    HAS
T9  Problema 281 295    Ca de Prostata
T10 Teste 342 348   exames
T11 Problema 392 412    hipertensão arterial
T12 Anatomia 287 295    Prostata
(...)
```

Figure 6.1: Example of an annotated text with BRAT, from NestedClinBr.

6.0.3 Annotation Process

The NestedClinBr was an end-to-end double annotation project, where all the texts were annotated by two different annotators. The differences were resolved by a third annotator (i.e., the adjudicator), who cannot remove annotations made by both annotators, and neither create new annotations. Performing a double annotation of a document prevents bias and makes it possible to check the annotation quality, by measuring the agreement between both annotators. This process resulted in the creation of a gold standard corpus.

As verified by (GURULINGAPPA et al., 2012), clinically trained annotators are better than linguists and computer scientists at annotating clinical text. We have selected annotators with expertise in the health domain: two students [3] from the Medicine course at the Pontifícia Universidade Católica do Paraná (PUCPR), and a master [4] in Bioinformatics at Universidade Federal do Paraná (UFPR) for adjudication. We also had the support of two doctors [5] in Informa-

[3] Carolina de Oliveira Montenegro and Laura Rubel Barzotto

[4] Elisa Terumi Rubel Schneider

[5] Yohan Bonescki Gumiel and Lilian Mie Mukai

tion Technology from PUCPR and a doctor [6] in Biomedical Informatics from the Università degli Studi di Pavia (UNIPV), who helped with specific knowledge during the creation of the guidelines and to answer questions from the annotators.

A training phase was provided so the annotators could familiarize themselves with the annotation tool and the process. The annotation process started with the release of one-third of the texts for annotation in this phase, performed by two annotators separately. As the corpus is small, and already pre-annotated with "Problems", "Treatment" and "Tests", only two round was performed to resolve doubts and disagreements, check the consistency of annotations, and improve the guideline. During each iteration, a small set of documents was subjected to double annotation. After addressing uncertainties and ensuring consistency, the annotation guidelines were refined. Following the last round, with all documents double annotated, we had an adjudication step to generate the gold standard, and the Inter Annotator Agreement was measured using F1-measure.

6.0.4 Annotation Guidelines

The annotation guidelines provide details on how to annotate each concept, listing a set of useful examples and serving as a guide to annotators. They are essential to maintain homogeneity during the annotation process and ensure the gold standard quality.

Following the works of (MARTíNEZ-DEMIGUEL et al., 2022), (OLIVEIRA et al., 2022), (GUMIEL et al., 2023), and (DOGAN; LEAMAN; LU, 2014), our guidelines provide accurate descriptions of entities, as well as illustrative examples to help during the annotation phase. Table 6.1 provides the definitions and some examples of the entity included in NestedClinBr corpus. We chose to maintain the same descriptions of TempClinBr, for entities of type "Problem", "Treatment", and "Test", but with the addition of examples of nested and discontinuous mentions. The importance of nested and discontinuous mention annotations was emphasized, showing clear examples, since this would be a differential of the corpus.

The entities to be marked refer to important events or mentions of the pa-

[6] Claudia Maria Cabral Moro Barra

Table 6.1: Event types with their respective description and examples.

Entity type	Definition	UMLS group	Examples (free translation)
Problem	Mentions that differ from normal expected conditions, including the location (body part), characterization, and severity, when available in the text.	Disorders	Injury, chest pain, SAH, severe dyspnea on exertion.
Treatment	Mentions relating to any procedure or intervention used to treat problems, including the dosage, in the case of drugs, and the location (body part), when available in the text.	Chemicals & Drugs, Devices, Procedures	Pacemaker, angioplasty, Enalapril 10 mg, mitral valve repair.
Test	Used to detect and evaluate problems (such as diagnostic procedures and physical examination), also including the location (body part), when available in the text.	Phenomena, Physiology	HDL, potassium, cardiac catheterization, myocardial scintigraphy.
Anatomy	Refers to body location, region, organ or organ component.	Anatomy	Heart valves, left hemithorax, mitral.

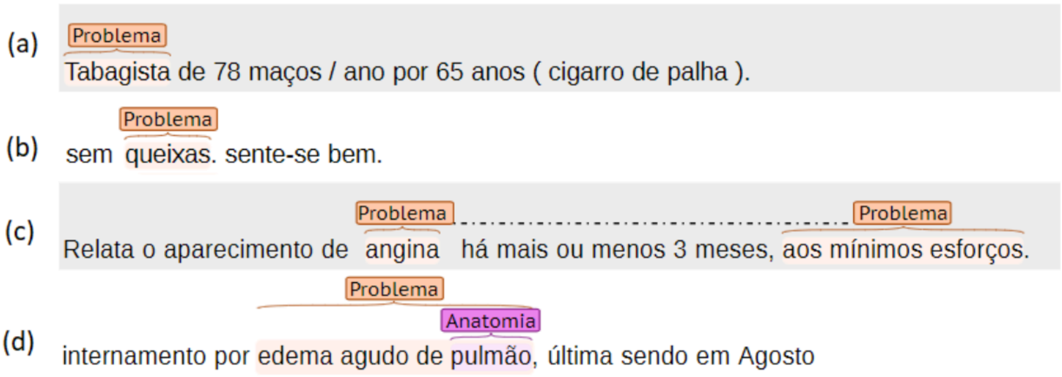
tient, related to the Unified Medical Language System semantic categories. In addition, some specific guidelines for each type were defined:

Problem Following the guideline proposed by (GUMIEL et al., 2023), any patient situation that differs from the normal and/or expected situation should be marked as a "Problem", including diseases, disorders, syndromes, clinical findings, signs, symptoms, disorders, injuries, or poisoning.

When labeling Problem-type entities, all the tokens that formed the mention must be selected, even if they are not sequential in the text. "Problems" should include, when available: a) location, which refers to the site of the disease or signal and irradiation pattern (localized or diffused), b) characterization, the description of what the patient understands by the symptom and its characterization as acute, constant, etc., c) severity, which describes the degree of severity or intensity of the signal or symptom.

Normal conditions such as "blushed", "lucid" and test results that indicate normality should not be marked. Laboratory test results such as “creatinine

2 mg/dl” should not be marked as "Problem", yet as "Test", as well as “Blood pressure 145/95”. Some marking examples are shown in Figure 6.2. In example (a), we have the expression “*Tabagista*” (smoker) which, despite being subjective, is related to the patient’s social history and has codes in the International Classification of Diseases (ICD), and for this reason in our guideline it is considered a "Problem". In example (b) we have the identification of “*queixas*” (complaints) as a "Problem", even though it is denied in the sentence. In example (c), we have an example of a discontinuous entity, where the mention would be “*angina aos mínimos esforços*” (angina on minimal exertion), i.e., angina plus its characterization. In example (d) we have a mention of a problem “*edema agudo*” (acute edema) with its location, “*pulmão*” (lung), being an example of a nested entity with an anatomy part.



Tabagista de 78 maços / ano por 65 anos (cigarro de palha).

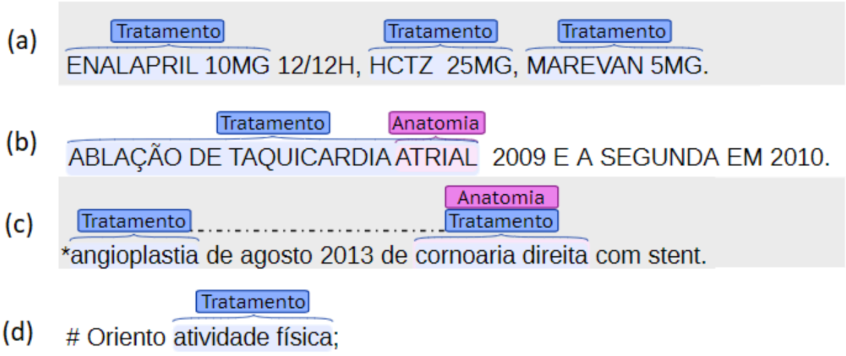
sem queixas. sente-se bem.

Relata o aparecimento de angina há mais ou menos 3 meses, aos mínimos esforços.

internamento por edema agudo de pulmão, última sendo em Agosto

Figure 6.2: Examples of labeled entities of the "Problem" type (in orange), from NestedClinBr corpus.

Treatment Still following the guideline by (GUMIEL et al., 2023), the treatment labeling should maintain specific markings, as the location in the case of procedures (e.g. "stent in the right coronary artery" instead of just "stent") and the dosage of medicines such as "simvastatin 20 mg". Treatment-like entities include the measures proposed by the health professional from the diagnosis, involving pharmacological substances, devices, procedures, and interventions performed. Figure 6.3 presents some examples of marking "Treatment" entities. In example (a), we have mentions containing the drug along with the dosage, very common in the corpus. In (b), we have an example of a treatment, "*ablação de taquicardia atrial*" (atrial tachycardia ablation), that contains the location of the treatment, creating a nested entity with "Anatomy". Example (c) shows a treatment with its location, "*angioplastia coronária direita*" (right coronary angioplasty), in non-sequential words in the text, originating an entity that is both nested and discontinuous. In (d), a more subjective example, where the expression "*atividade física*" (physical activity) is a medical recommendation, and for this reason, it is considered a Treatment.



(a) ENALAPRIL 10MG 12/12H, HCTZ 25MG, MAREVAN 5MG.

(b) ABLAÇÃO DE TAQUICARDIA ATRIAL 2009 E A SEGUNDA EM 2010.

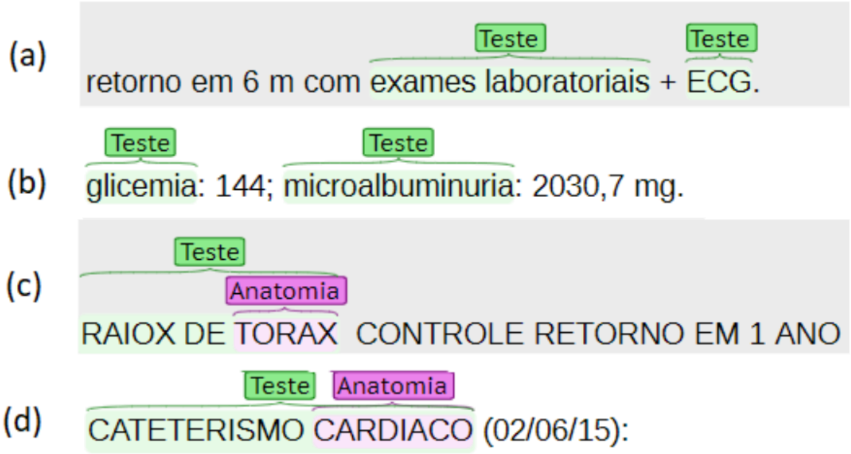
(c) *angioplastia de agosto 2013 de cornoaria direita com stent.

(d) # Oriento atividade física;

Figure 6.3: Examples of labeled entities of the "Treatment" type (in blue), from NestedClinBr corpus.

Test "Test" mentions involve physical, visual, and laboratory examinations, as well as diagnostic tests, and should also include the most complete term possible, such as "transesophageal echocardiogram", rather than just "echocardiogram", as proposed by (GUMIEL et al., 2023). Results should not be marked to the tests nor mention of verbs related to its performance. Figure 6.4 presents some examples of test labeling. In (a), we have the generic term "*exames laboratoriais*" (laboratory exams) as a "Test", as well as ECG (an acronym for electrocardiogram). In example (b), we have the tests identified, "glicemia" and "microalbuminuria", without considering their results. In (c) and (d), we have examples of test identification with its location, "*raio x de torax*" (chest x-ray) and "*cateterismo cardíaco*" (cardiac catheterization), configuring nested entities with the "Anatomy" type. Although granted in our guideline, "Test" entities formed by discontinuous tokens were not found.

Anatomy Anatomy-like entities are a differential of the new corpus and thence, they must be labeled from scratch. "Anatomy" is an important concept in medical texts, presented in various health corpora such as (CAMPILLOS-LLANOS L., 2021), (BÁEZ et al., 2020), (OLIVEIRA et al., 2022), and (LOPES; TEIXEIRA;



retorno em 6 m com exames laboratoriais + ECG.

glicemia: 144; microalbuminuria: 2030,7 mg.

RAIOX DE TORAX CONTROLE RETORNO EM 1 ANO

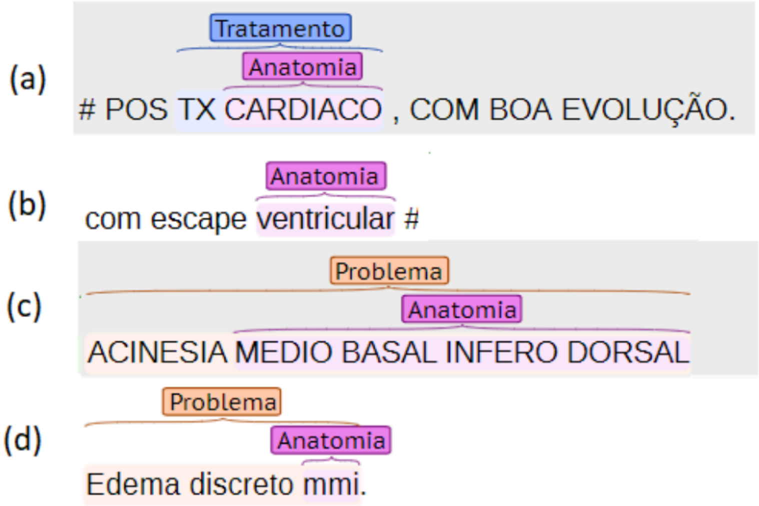
CATETERISMO CARDIACO (02/06/15):

Figure 6.4: Examples of labeled entities of the "Test" type (in green), from NestedClinBr corpus.

OLIVEIRA, 2019). "Anatomy" means any region or location of the human body, which may be organs, components, substances, systems, cell components, tissue, or other anatomical structures. Also, any expression that refers to the anatomical location must be labeled, for example, “abdominal” which refers to the abdomen region. This entity also accepts discontinuous mentions, however only 3% of the training set contained discontinuous anatomy entities. When the mention is associated with a "Problem", "Treatment", or "Test", it will usually be nested within this entity. Figure 6.5 shows some examples of "Anatomy" mentions. In (a), we have an example where the word *“cardiaco”* (cardiac) refers to the location of the treatment, and for this reason, it should be marked as "Anatomy", as well as in example (b), where *“ventricular”* goes to the ventricle. In (c) an example of anatomy formed by several words, *“medio basal infero dorsal”* (mid-basal inferodorsal), and in (d) a very common example in texts, involving edema in the lower limbs (*“mmii”*).

Generic Guidelines

- Discontinuous entities must always be formed by the same semantic type and be in the same sentence;
- Nested entities of the same type must not be marked, (e.g. “ventricle” is a mention of "Anatomy" type, but if in a larger expression as "left ventricle", the entire expression (more specific) must be marked, with no need to select only "ventricle");
- "Problems", "Treatments", and "Tests" must be associated with at most one



POS TX CARDIACO , COM BOA EVOLUÇÃO.

com escape ventricular #

ACINESIA MEDIO BASAL INFERO DORSAL

Edema discreto mmi.

Figure 6.5: Examples of labeled entities of the "Anatomy" type (in pink), from NestedClinBr corpus.

anatomy-type entity. If there's more than one body location involved, then two mentions must be labeled, one for each region;

- Abbreviated concepts must also be labeled (e.g. *"VE"*, *"mmii"*), as well as concepts with typos or grammatical errors.

6.0.5 Inter-Annotator Agreement

Given that the inter-annotator agreement assesses the consistency and quality of the corpus, we calculated the IAA of all the data, using the F-1 measure in a token-level way. As explained in the methodology section, for named entities annotation tasks, Cohen's Kappa may not be suitable due to the lack of a fixed number of negative cases, being the F-measure (not reliant on negative case count) more appropriate for named entity annotation (DELEGER et al., 2012).

As dealing with a small corpus, this metric was calculated only at the end of the annotation process, as follows: 1) we considered the annotations made by the first annotator our gold standard, 2) we calculated the precision, recall, and F1-measure between the gold standard and the annotations of the second annotator, under exact match criteria (i.e., the annotations should exactly coincide by entity type and the mention boundaries). As the work of (MARTíNEZ-DEMIGUEL et al., 2022), the "Bratiaa" library [7] was used to compute the F1-measure, a Python library that computes the agreement under exact match (type and mention) for entities.

The final IAA value for entity types was 94.08%, a high F1-measure which represents a substantial agreement between annotators. Figure 6.6 shows the IAA values per entity type.

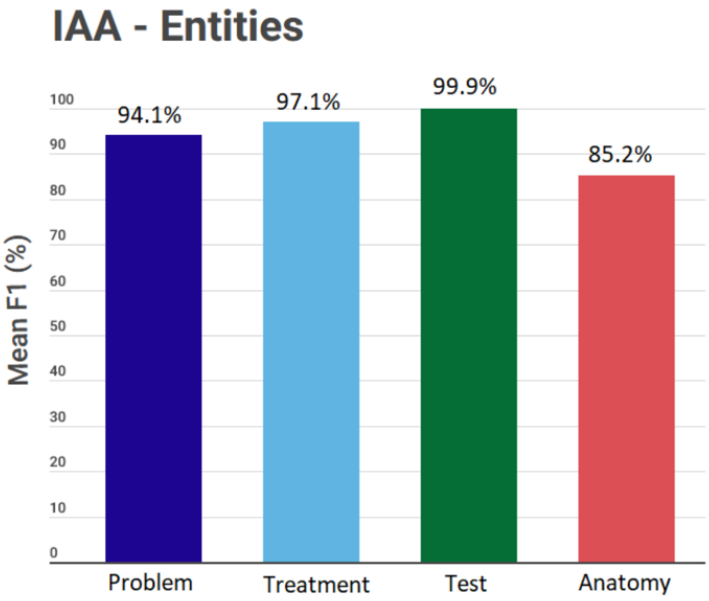


Figure 6.6: IAA scores for entities in NestedClinBr.

6.0.6 Corpus Statistics

As the annotations present in our corpus can be very useful to train models to detect medical information from unannotated texts, we split it into training and test datasets in the ratio 80:20. Table 6.2 shows some basic statistics about the number of tokens, sentences, and documents in NestedClinBr.

Table 6.3 shows the numbers of the annotated entities, with some statistics per entity type.

Figure 6.7 displays in (a) the percentage of entities of each type in the training corpus, and in (b) the number of tokens per entity, where the minimum is 1 and

[7] https://pypi.org/project/bratiaa/

Table 6.2: Number of documents, sentences, and tokens present in NestedClinBr.

	Training	Test	Total
Documents	100	26	126
Sentences	2,273	693	2,966
Tokens	17,154	5,310	22,464

Table 6.3: Statistics of the NestedClinBr corpus.

Item	Training	Test	Total
Problem			
Nested	315 (72.92%)	117 (27.08%)	432
Discontinuous	80 (65.04%)	43 (34.96%)	123
Total	1,110 (77.19%)	328 (22.81%)	1,438
Entity avg. length	2.6	2.5	-
Treatment			
Nested	50 (76.92%)	15 (23.08%)	65
Discontinuous	4 (100%)	0 (0%)	4
Total	761 (78.05%)	214 (21.95%)	975
Entity avg. length	2.1	2.1	-
Test			
Nested	15 (71.43%)	6 (28.57%)	21
Discontinuous	0 (0%)	0 (0%)	0
Total	772 (75.98%)	244 (24.02%)	1,016
Entity avg. length	1.2	1.3	-
Anatomy			
Nested	395 (75.38%)	129 (24.62%)	524
Discontinuous	5 (83.33%)	1 (16.66%)	6
Total	543 (73.58%)	195 (26.42%)	738
Entity avg. length	1.4	1.3	-
Overall			
Nested	778 (74.45%)	267 (25.55%)	1,045
Discontinuous	89 (66.92%)	44 (33.08%)	133
Total	3,186 (76.46%)	981 (23.54%)	4,167
Percentage of entities vs ’O’ (balancing)	30.8%	30.1%	-
Entity avg. length	1.9	1.9	-
Max. tokens per sentence	192	146	-

the maximum is 27.

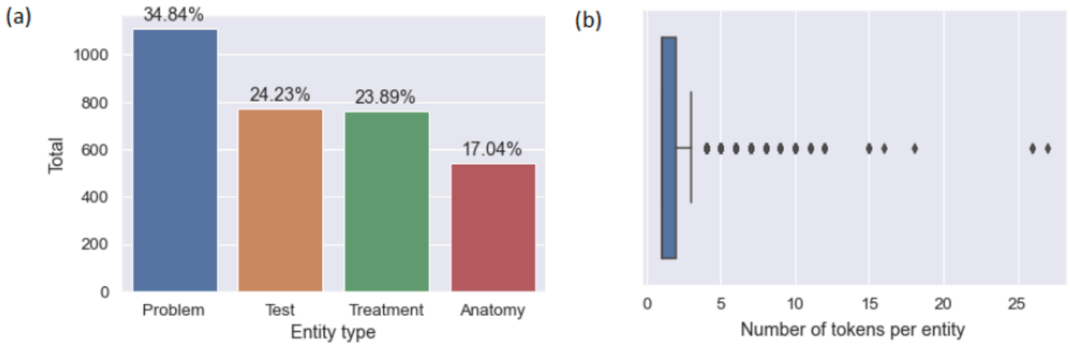


Figure 6.7: Percentage of entities and the number of tokens per entity in NestedClinBr.

The code used to generate corpus statistics will also be available in a publicly accessible repository, serving as a resource for generating corpus statistics annotated in the BRAT standard.

6.0.7 Discussion

We proposed NestedClinBr, a Brazilian-Portuguese corpus that includes the annotation of clinical concepts, in flat, nested, and discontinuous format, built from the TempClinBr (GUMIEL et al., 2023) corpus. To the best of our knowledge, this is the first resource for clinical natural language processing containing nested and discontinuous entities, in Portuguese language. The nested and discontinuous entities address some NER challenges that have hardly been addressed.

In our analysis of the development of NestedClinBr corpus, the IAA score was calculated, to ensure its quality and consistency as well as provide insights into the quality of the guidelines created. Our IAA values indicate the high quality of the corpus, showing a high agreement for "Problem" (94.1%), "Treatment" (97.1%), "Test" (99.9%) and "Anatomy" (85.2%) entities. The high value for the "Problem", "Treatment" and "Test" entities was expected, since the mentions were already pre-marked, coming from the TempClinBr corpus. It is worth mentioning that one of the NestedClinBr annotators participated in the TempClinBr annotation, as well as two of the researchers who provided support, already bringing their previously acquired knowledge. However, new markings were performed to signalize nested and discontinuous entities, something that did not exist in the original corpus.

The entity of type "Anatomy", annotated from scratch, had the lowest agreement value between annotators. One of the reasons for this might be the labeling of words that are not anatomy per see but refer to the location of the "Problem", "Treatment", or "Test". For example, in the words "cardiac" and "abdominal", as can be seen in Fig 6.8.Example 1, just the second annotator (b) correctly labeled it. Some acronyms also went unnoticed, as in 6.8. Example 2, when the mention of CD, abbreviation for "*coronária direita*" (right coronary) was missing by the first annotator (a). Also, mentions of "Anatomy", although representing 17% of all mentions, represent 51.2% of all nested mentions in the corpus, which may bring some complexity when labeling it.

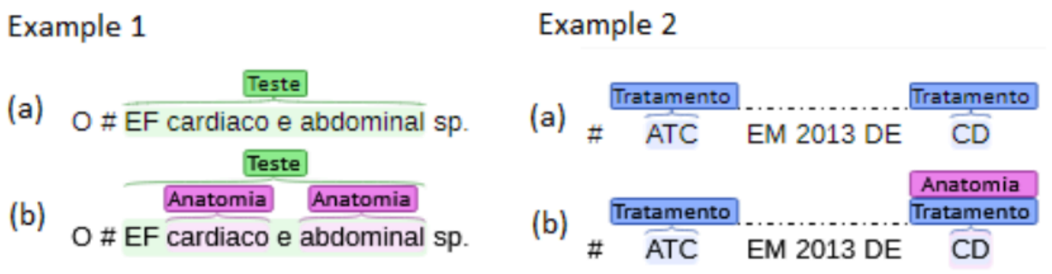
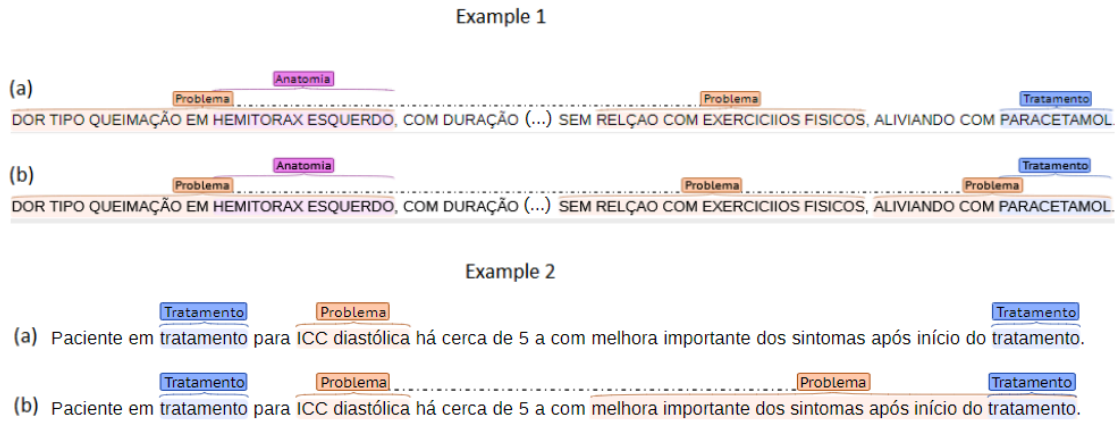


Figure 6.8: Comparison between labeled entities of the two annotators.

One of the most important challenges is the accurate annotation of discontinuous entities, as different annotators might produce very similar annotations but with some small differences. In Figure 6.9.Example 1 we can see that while the second annotator (b) correctly identified the expression "*dor tipo queimação em hemitorax esquerdo sem relção com exerciciios fisicos*" (sic) (burning pain in the left hemithorax without relation to physical exercises), the first annotator did not mark the word "*sem*" (without) in the annotation of this discontinuous entity. Although both annotators correctly detected a problem and its location and characterization, these small disagreements penalized the global IAA. Another disagreement example of a discontinuous entity can be seen in Figure 6.9. Example 2, where the first annotator did not label the severity of the "Problem", generating inconsistency ("*ICC diastólica melhora importante dos sintomas após início do tratamento*", in English "*Diastolic CHF significantly improves symptoms after starting treatment*"). In our guideline, all location, characterization, and severity must be marked united with the problem, when available, even if in non-sequential words in the text.



Example 1

(a) DOR TIPO QUEIMAÇÃO EM HEMITORAX ESQUERDO, COM DURAÇÃO (...) SEM RELÇAO COM EXERCICIIOS FISICOS, ALIVIANDO COM PARACETAMOL.

(b) DOR TIPO QUEIMAÇÃO EM HEMITORAX ESQUERDO, COM DURAÇÃO (...) SEM RELÇAO COM EXERCICIIOS FISICOS, ALIVIANDO COM PARACETAMOL.

Example 2

(a) Paciente em tratamento para ICC diastólica há cerca de 5 a com melhora importante dos sintomas após início do tratamento.

(b) Paciente em tratamento para ICC diastólica há cerca de 5 a com melhora importante dos sintomas após início do tratamento.

Figure 6.9: Comparison between labeled entities of the two annotators.

Annotating nested and discontinuous entities is a challenging task that involves identifying and marking entities that can occur at different levels of nesting and can also be interrupted by other entities, causing disagreements between annotators, as pointed out by (MARTíNEZ-DEMIGUEL et al., 2022). To overcome these challenges, it is important to have clear guidelines, adequate training for annotators, and iteration between annotators.

One of the limitations of the corpus is its small size, formed by 126 clinical notes from Brazilian hospitals. However, as seen in the experiments conducted with NestedClinBr, it was possible to train machine learning models to recognize these medical entities with a reasonable level of performance. Also, our corpus could be used to develop semi-supervised approaches, providing gold-standard seeds to augment the training data.

NestedClinBr can be considered a gold-standard corpus since it was manually annotated and its quality was confirmed by the IAA measurement between different annotators.

7

Portuguese-language models for clinical and biomedical domains

In this chapter, we will present the clinical models for the Portuguese language developed during the research. All language models trained were publicly released [1]. More details can be seen in the published paper (SCHNEIDER et al., 2020).

7.1 Methods

We fine-tuned three BERT-based models on Portuguese clinical and biomedical corpora: a) BioBERTpt(clin), a clinical model, b) BioBERTpt(bio), a biomedical model, and c) BioBERTpt(all), a clinical + biomedical model, using as checkpoints the weights provided by multilingual BERT (DEVLIN et al., 2019).

We have used 2,100,546 clinical notes from Brazilian hospitals [2], properly de-identified, containing multi-specialty information, including cardiology, nephrology, and endocrinology. These clinical notes, formed by 3.8 million sentences and 27.7 million words, were used to train a Portuguese clinical BERT-base model (BioBERTpt(clin)). For training the biomedical model (BioBERTpt(bio)), we collect titles and abstracts from Portuguese scientific papers published in Pubmed and in the Scielo (Scientific Electronic Library Online), obtained from

[1] https://huggingface.co/pucpr and https://github.com/HAILab-PUCPR/

[2] Certificate of presentation for Ethical Appreciation number 51376015.4.0000.0020

Table 7.1: List of text corpora used for BioBERTpt.

Texts	Source	Sentences	Words	Domain
Clinical narratives	EHR from Brazilian Hospitals	3.8 million	27.7 million	Clinical
Scielo (heath and biological areas)	Literature titles abstracts	663,018	15.6 million	Biomedical
Pubmed	Literature titles	74,451	812,711	Biomedical

the Biomedical Translation Task in the First Conference on Machine Translation (BOJAR et al., 2016); resulting in 16.4 million words. We also have trained a model with all corpora, BioBERTpt(all). All documents used for training BioBERTpt models are listed in Table 7.1.

We split the notes and abstracts into sentences and tokenize them with the default BERT Wordpiece tokenizer (DEVLIN et al., 2019). We trained the models 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 (WOLF et al., 2020).

In order to validate the in-domain encoded information in the models, we performed some experiments for the NER task [3] with the following corpora: SemClinBr (OLIVEIRA et al., 2022) and PortugueseClinicalNER (LOPES; TEIXEIRA; OLIVEIRA, 2019). To evaluate the performance, we used as metrics the micro precision, recall, and F1-measure, and compared them with other BERT-based models. For both experiments, we used 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 linear schedule with warm up as 0.1. Figure 7.1 shows an overview of the method.

7.2 NER Results

Table 7.2 shows the models results on SemClinBr corpus, and Table 7.3, on PortugueseClinicalNER. BioBERTpt has achieved the highest results for both corpora, compared to existing models (BERT-multilingual (DEVLIN et al., 2019) and BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020)).

[3] We evaluated the BioBERTpt models in the traditional NER task, i.e. with flat entities.

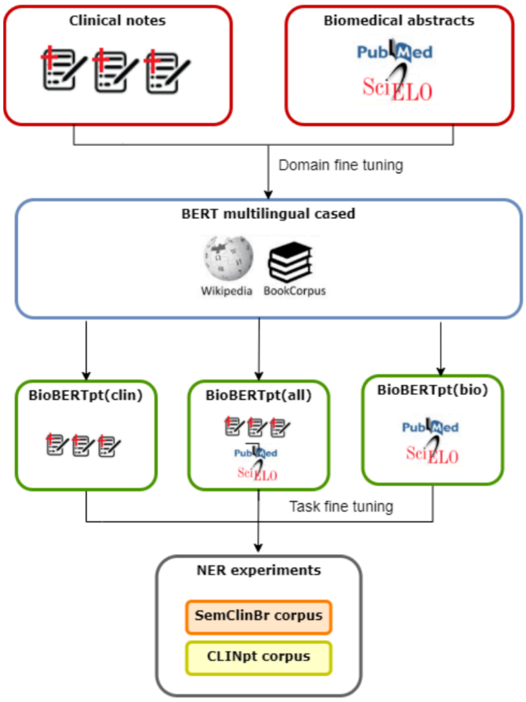


Figure 7.1: Overview of BioBERTpt training.

Table 7.2: The average scores of the NER task in SemClinBr corpus, for each model evaluated.

Model	Precision	Recall	F1
BERT-based models			
mBERT-uncase	0.623	0.566	0.588
mBERT-cased	0.604	0.567	0.582
BERTimbau-base	0.595	0.587	0.585
BERTimbau-large	0.563	0.531	0.541
Ours			
BioBERTpt(bio)	**0.624**	0.586	0.602
BioBERTpt(clin)	0.609	0.603	0.602
BioBERTpt(all)	0.608	**0.607**	**0.604**

7.3 Discussion

We developed three new language models for clinical texts in Portuguese: a) a biomedical, BioBERTpt(bio), b) a clinical, BioBERtpt(clin), and c) a clinical and biomedical model, BioBERTpt(all). To evaluate the models and assess if they

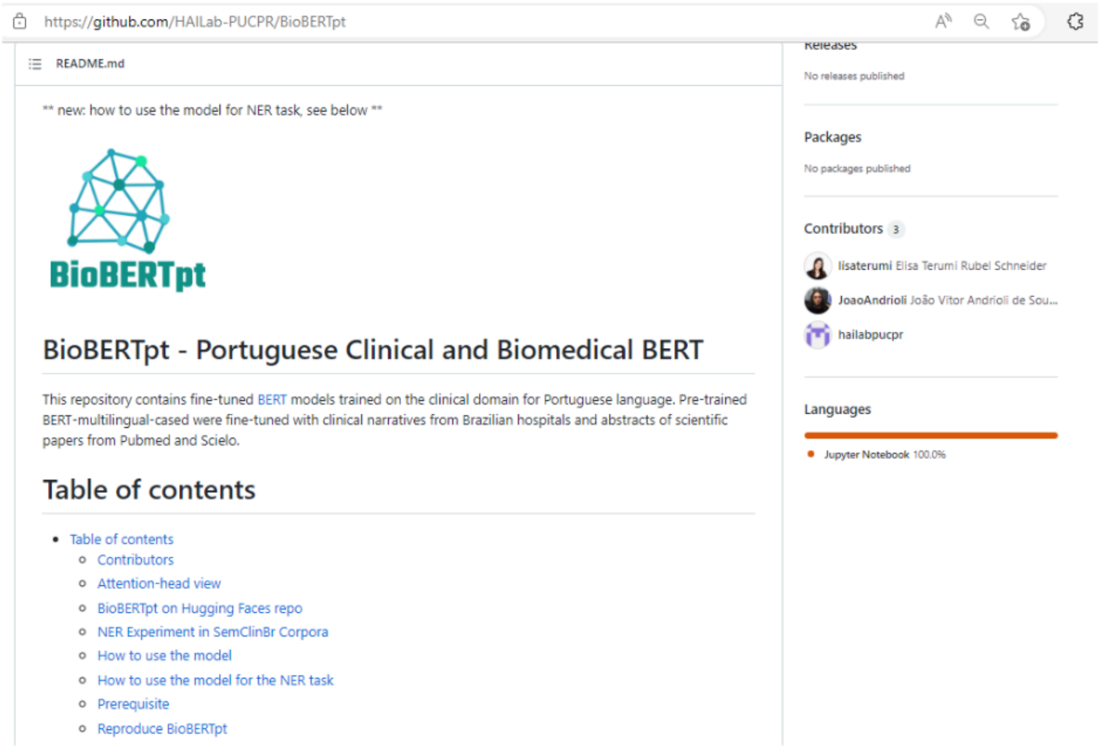
Table 7.3: The average scores of the NER task in PortugueseClinicalNER corpus, for each model evaluated.

Model	Precision	Recall	F1
Baseline			
BiLSTM-CRF	0.753	0.745	0.749
BERT-based models			
mBERT-uncase	0.903	0.921	0.912
mBERT-cased	0.912	0.931	0.921
BERTimbau-base	0.910	0.922	0.916
BERTimbau-large	0.898	0.927	0.912
Ours			
BioBERTpt(bio)	**0.917**	0.925	0.921
BioBERTpt(clin)	**0.917**	**0.935**	**0.926**
BioBERTpt(all)	0.912	0.929	0.920

are relevant to the medical area, we did two NER experiments using the corpora SemClinBr and PortugueseClinicalNER. Our models reached state-of-the-art in both corpora for recall, precision, and F1, including significantly superior results in terms of F1 to the mBERT, BERTimbau-large, and BERTimbau-base, on SemClinBr corpus.

These results showed that the in-domain models outperform the general models in the evaluated metrics, particularly for domains with unique characteristics such as medical, corroborating previous experiments in other languages such as English as in the works of (LEE et al., 2019), (ALSENTZER et al., 2019), and (LI et al., 2019)).

Also, by providing a contextualized word representation and using the Transformer architecture, BERT-based language models had a positive impact on the results when compared to traditional machine learning algorithms and word embeddings, used in the work of (SOUZA et al., 2019) and (LOPES; TEIXEIRA; OLIVEIRA, 2019). For example, in (LOPES; TEIXEIRA; OLIVEIRA, 2019) the authors used BiLSTM-CRF and fastText on the PortugueseClinicalNER corpus, achieving an F1 of 0.759 with an in-domain model, while BioBERTpt(clin) achieved 0.926. In general, all BERT-based models performed better compared to the previous baselines, for both corpora, even the generic BERT models (out-of-domain).



https://github.com/HAILab-PUCPR/BioBERTpt

README.md

** new: how to use the model for NER task, see below **

BioBERTpt

BioBERTpt - Portuguese Clinical and Biomedical BERT

This repository contains fine-tuned BERT models trained on the clinical domain for Portuguese language. Pre-trained BERT-multilingual-cased were fine-tuned with clinical narratives from Brazilian hospitals and abstracts of scientific papers from Pubmed and Scielo.

Table of contents

- Table of contents
  - Contributors
  - Attention-head view
  - BioBERTpt on Hugging Faces repo
  - NER Experiment in SemClinBr Corpora
  - How to use the model
  - How to use the model for the NER task
  - Prerequisite
  - Reproduce BioBERTpt

Releases

No releases published

Packages

No packages published

Contributors 3

lisaterumi Elisa Terumi Rubel Schneider

JoaoAndrioli João Vitor Andrioli de Sou...

hailabpucpr

Languages

● Jupyter Notebook 100.0%

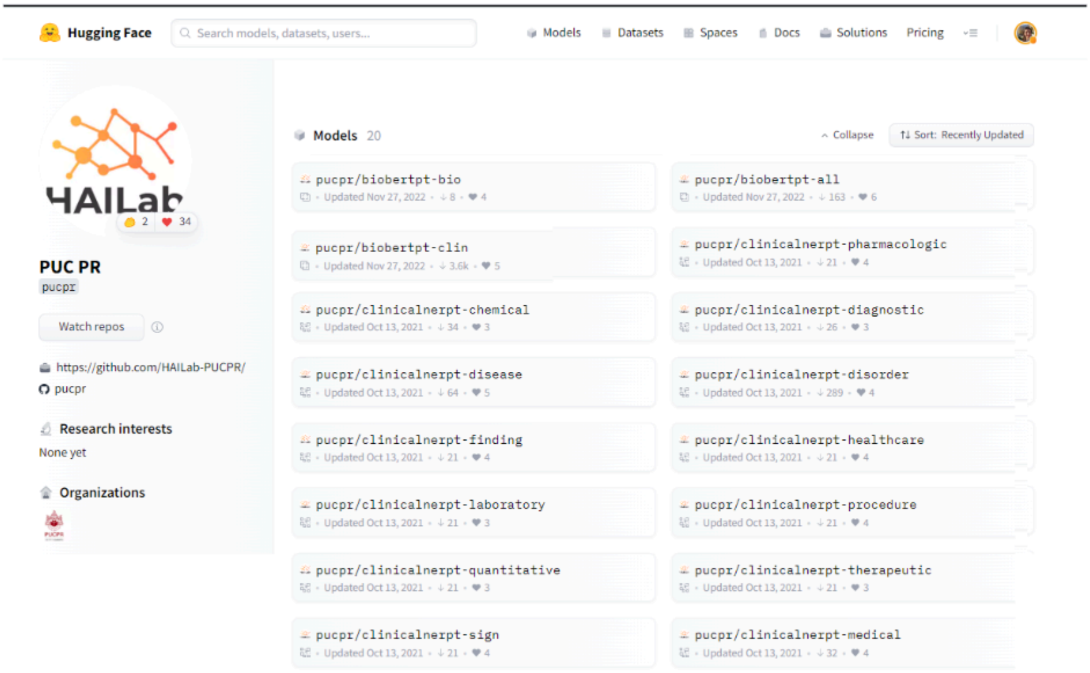
Figure 7.2: Repository of generated resources, on GitHub.

Our clinical and biomedical BERT-based models have the potential to support clinical NLP tasks for Portuguese, a language with relatively lower resources, especially in the health domain. The World Health Organization [4] (WHO) has released a list of 13 urgent health challenges the world will face over the next decade, and to face these challenges, access to quality health information is essential. Since extracting structured information from clinical documents can provide health care assistance, support other biomedical tasks, and contribute to urgent health challenges, we release publicly BioBERTpt and all NER models developed in our research (13 in total), for researchers and for the Portuguese-speaking community. All models, source code to replicate the work, usage instructions, and the complete results of the experiments performed are in our research groups GitHub repository [5] , as shown in Figure 7.2. Also, we have made the language models available in the Hugging Faces repository, a community-based repository for open-source machine learning technology, un-

[4] https://www.who.int/

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

der the username of our university [6] (PUCPR). Figure 7.3 shows the language models generated in this work and available on the HuggingFaces platform, also showing the total accesses (downloads) in the last month. As we can see, the models are being useful to the community, with the BioBERTpt(clin) model having approximately 3.6 thousand accesses in the last month (March 2023 information).



PUC PR
pucpr

https://github.com/HAILab-PUCPR/
pucpr

Research interests
None yet

Organizations

Models 20

pucpr/biobertpt-bio
Updated Nov 27, 2022 • ↓8 • ♥ 4

pucpr/biobertpt-all
Updated Nov 27, 2022 • ↓163 • ♥ 6

pucpr/biobertpt-clin
Updated Nov 27, 2022 • ↓3.6k • ♥ 5

pucpr/clinicalnerpt-pharmacologic
Updated Oct 13, 2021 • ↓21 • ♥ 4

pucpr/clinicalnerpt-chemical
Updated Oct 13, 2021 • ↓34 • ♥ 3

pucpr/clinicalnerpt-diagnostic
Updated Oct 13, 2021 • ↓26 • ♥ 3

pucpr/clinicalnerpt-disease
Updated Oct 13, 2021 • ↓64 • ♥ 5

pucpr/clinicalnerpt-disorder
Updated Oct 13, 2021 • ↓289 • ♥ 4

pucpr/clinicalnerpt-finding
Updated Oct 13, 2021 • ↓21 • ♥ 4

pucpr/clinicalnerpt-healthcare
Updated Oct 13, 2021 • ↓21 • ♥ 4

pucpr/clinicalnerpt-laboratory
Updated Oct 13, 2021 • ↓21 • ♥ 3

pucpr/clinicalnerpt-procedure
Updated Oct 13, 2021 • ↓21 • ♥ 4

pucpr/clinicalnerpt-quantitative
Updated Oct 13, 2021 • ↓21 • ♥ 3

pucpr/clinicalnerpt-therapeutic
Updated Oct 13, 2021 • ↓21 • ♥ 3

pucpr/clinicalnerpt-sign
Updated Oct 13, 2021 • ↓21 • ♥ 4

pucpr/clinicalnerpt-medical
Updated Oct 13, 2021 • ↓32 • ♥ 4

Figure 7.3: Repository of generated resources, on HuggingFaces.

[6] https://huggingface.co/pucpr

8

Experiments, Results, and Discussion

In this chapter, we detail the experiments performed with BioNestedNER, in the selected corpora. First, we present the architecture, tools, and parameters used, some evaluation details and comparative models, and then we show the results for each corpus and a discussion about the obtained results. In the end, we revisit the research objectives and hypothesis of this thesis.

8.1 Architecture Definition, Tools, and Parameters

By studying and experimenting, we defined the architecture used in our experiments with BioNestedNER as models based on BERT (Phase 1 of the method), for generating the representation of words and for task fine-tuning. BERT uses the Transformer architecture, where its self-attention mechanism allows the model to capture dependencies between distant words in a sentence, improving long-range dependencies. Also, pre-trained BERT models on a large amount of data allow the learning of general features of language that can be applied to a wide variety of NLP tasks, making it versatile and effective for language processing. Furthermore, its bidirectional approach to language modeling captures a more nuanced understanding of language. It performs highly on tasks that require a deeper understanding of context.

To make a fair comparison between our method and other methods in the literature, we used only BERT-based models. We prioritized the use of the same pre-trained weights (checkpoint) in all methods, BioBERT (LEE et al.,

Table 8.1: Main hyper-parameter settings of the experiments.

Hyper-parameter	Value
Training batch size	32, 16, and 8
Evaluation batch size	32, 16, and 8
Learning rate	3e-5 and 2e-05
Embedding size	768
Dropout probability	0.1
Transformer blocks	12
Attention heads (for each attention layer)	12
Optimizer	AdamW
Activation function	GELU
Sentence maximum length	180
Maximum number of epochs	30 and 10

2019) in the experiments in English and BioBERTpt (SCHNEIDER et al., 2020) in the experiments in the Portuguese language. Therefore, we can compare the methods without taking into account the differences between the language models. In addition, we also performed a small few-shot experiment with GPT-3, which is explained later.

In the implementation, we used the Python programming language, version 3, the PyTorch version of the "Transformers" library provided by the Hugging Face API (WOLF et al., 2020), and the "Crfsuite" package from "Sklearn" (PEDREGOSA et al., 2011).

In terms of hardware, we have used: a) NVIDIA T4 Tensor Core GPU with CUDA version 11.2, 15 GB of GPU memory, and up to 32 GB of RAM, service accessed in the cloud and provided by Google Colab Pro [1], b) NVIDIA Geforce RTX 2060 SUPER, with CUDA version 12.0, 8 GB of GPU memory, and an Intel i7 with 16 GB of RAM, and c) NVIDIA Geforce GTX TITAN X, with CUDA version 11.6, 12 GB of GPU memory, and an Intel Xeon E5-1620 v4 with 16 GB of RAM.

For the experiments, we used the configuration shown in Table 8.1.

[1] https://colab.research.google.com/

8.2 Evaluation Details and Comparative Models

Following other works in the literature, we consider micro metrics and report only the exact matches, i.e., both entity type and boundaries must be correct. In experiments run locally, we also show the accuracy (ACC) of nested entities (NE), discontinuous entities (DE), and multi-type entities (ME), representing the proportion of correctly classified instances in relation to the total number of instances.

In some cases, we have trained baselines using BERT models to compare with our method. In cases of the corpus with nested and multi-type entities, we have performed a binary relevance (BR), i.e., we have trained a specific model for each type of entity and then joined the results.

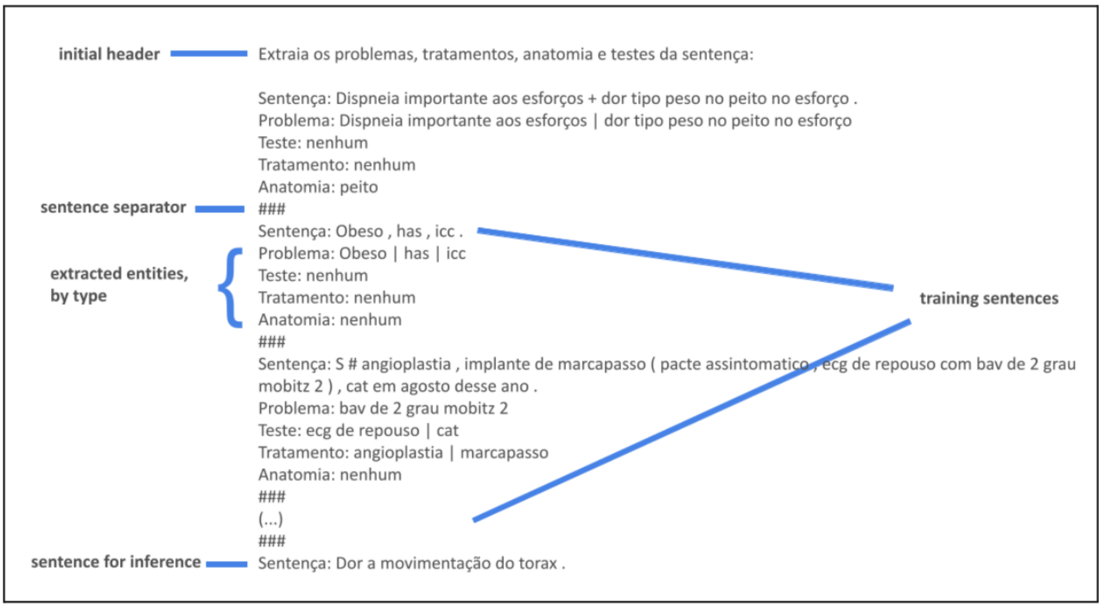
Besides BR, we also trained local models with the MRC (LI et al., 2020), PIQN (SHEN et al., 2022), and QA-NER (BANERJEE et al., 2021) (the original and with an adapted version without the CNN) methods, which are similar to ours, in order to be able to compare the results. We used the same settings presented above but with the following changes (proposed by the authors in their repository): for MRC and PIQN, the maximum number of epochs was *30*, for MRC we used dropout as *0.2* and weight decay as *0.002*, and for PIQN we used learning rate as *2e-05*.

In view of the success of ChatGPT (OPENAI, 2023), in experiments E3.1 (NestedClinBr) and E5.3 (GENIA few shot) we also did a small few-shot experiment with GPT-3 model, Davinci (BROWN et al., 2020), via a prompt [2]. GPT-3 has been trained on an enormous amount of data and has a high capacity to generalize to new tasks. Its pre-training includes exposure to a broad range of natural language processing tasks, which makes it able to perform well on tasks it has never seen before. As demonstrated by (BROWN et al., 2020), GPT-3 has the ability to perform well on few-shot learning tasks, having a high capacity to generalize to new tasks as its pre-training includes exposure to a broad range of natural language processing tasks. Also, GPT-3 includes a large number of parameters (175 billion), which allows it to capture complex patterns in data and learn from a few examples. We have experimented with a few-shot training with GPT-3 with just 15 and 20 examples of input, following (BROWN et al.,

[2] Few-shot prompting is a technique where the model is given a small number of samples, in order to quickly adapt to new examples.

2020) where they suggest a value between 10 to 100.

In Figure 8.1 we show an example of the prompt used in the few-shot training.



```
Extraia os problemas, tratamentos, anatomia e testes da sentença:

Sentença: Dispneia importante aos esforços + dor tipo peso no peito no esforço .
Problema: Dispneia importante aos esforços | dor tipo peso no peito no esforço
Teste: nenhum
Tratamento: nenhum
Anatomia: peito
###
Sentença: Obeso , has , icc .
Problema: Obeso | has | icc
Teste: nenhum
Tratamento: nenhum
Anatomia: nenhum
###
Sentença: S # angioplastia , implante de marcapasso ( pacte assintomatico , ecg de repouso com bav de 2 grau
mobitz 2 ) , cat em agosto desse ano .
Problema: bav de 2 grau mobitz 2
Teste: ecg de repouso | cat
Tratamento: angioplastia | marcapasso
Anatomia: nenhum
###
(...)
###
Sentença: Dor a movimentação do torax .
```

Figure 8.1: An example of the prompt for few-shot ChatGPT training.

We trained several variants of our main model to identify the best settings for balancing and loss reduction during training, and in the case of the CRF, to find the best number of clusters and thresholds. An explanation of each variant follows:

- BioNestedNER(classWeight+sum): means that we trained the model using the class balancing with sum reduction in loss function;
- BioNestedNER(classWeight+sum+CRF): means that we trained the model using the class balancing with sum reduction in loss function and we combined it with the results of the best trained CRF model;
- BioNested-NER(classWeight+sum+CRF ensemble): means that we trained the model using the class balancing with sum reduction in loss function and we combined it with the results of an ensemble formed by all trained CRF models;
- BioNestedNER(binaryWeight+sum): means that we trained the model using the binary balancing with sum reduction in loss function;
- BioNestedNER(binaryWeight+sum+CRF): means that we trained the model using the binary balancing with sum reduction in loss function and we combined it with the results of the best trained CRF model;

- BioNestedNER(binaryWeight+sum+CRF ensemble): means that we trained the model using the binary balancing with sum reduction in the loss function, and we combined it with the results of an ensemble formed by all trained CRF models;
- BioNestedNER(classWeight+mean): means that we trained the model using the class balancing with a mean reduction in loss function;
- BioNested-NER(classWeight+mean+CRF): means that we trained the model using the class balancing with a mean reduction in loss function and we combined it with the results of the best trained CRF model;
- BioNestedNER(classWeight+mean+CRF ensemble): means that we trained the model using the class balancing with a mean reduction in the loss function, and we combined it with the results of an ensemble formed by all trained CRF models;
- BioNestedNER(binaryWeight+mean): means that we trained the model using the binary balancing with a mean reduction in loss function;
- BioNestedNER(binaryWeight+mean+CRF): means that we trained the model using the binary balancing with a mean reduction in the loss function, and we combined it with the results of the best trained CRF model;
- BioNestedNER(binaryWeight+mean+CRF ensemble): means that we trained the model using the binary balancing with a mean reduction in loss function and combined it with the results of an ensemble formed by all trained CRF models.

In the comparison tables with other models, we follow this pattern:

- BioNestedNER (QA-NER): the model generated in the first phase of the method, that is, the model based on QA-NER with the best balancing configuration and loss reduction;
- BioNestedNER (QA-NER + CRF): the same as above combined with the best trained CRF model results;
- BioNestedNER (QA-NER end-to-end): the model generated in the first phase of the method, using end-to-end modification, that is, the model based on QA-NER with the best balancing configuration and loss reduction;
- BioNestedNER (QA-NER end-to-end + CRF): the same as above combined with the best trained CRF model results;
- BioNestedNER (CRF multi-label): the model generated in the second phase of the method, that is, the CRF multi-label with the best-selected features (e.g., number of clusters) and threshold value;
- BioNestedNER (CRF ensemble): an ensemble formed by all trained CRF models.

Following the same guidelines as the methods in the literature, all reported results are using micro metrics, calculating the aggregated performance of the model across all classes, appropriate when the class distribution is imbalanced and the main interest is to evaluate the model's overall performance across all classes. We apply the non-parametric Friedman test (an alternative to the repeated measures ANOVA) to verify if there are statistical differences between results and processing time, followed by a Nemenyi post-hoc test to see which groups are different. The Friedman test was selected as it is commonly used in comparing multiple models (such as neural networks) in experiments using the same dataset, followed by post-hoc Nemenyi when the null hypothesis is rejected (DEMŠAR, 2006). In our experiments, we consider a statistically significant result when *p-value* < *0.05*.

8.3 Results and Discussion

This section provides a comprehensive presentation of the evaluation results for all experimented corpora, accompanied by a discussion and analysis of the outcomes.

8.3.1 NestedClinBr Corpus (E3)

Tables 8.2 and 8.3 show the results in NestedClinBr, a new proposed corpus with Brazilian-Portuguese clinical notes. As it is a new corpus, proposed in this work, it is not yet possible to compare it with results from the literature, so we trained binary relevance models with BioBERTpt, BERTimbau, and mBERT. We have trained models with MRC, PIQN, and QA-NER, with CNN and without it, using BioBERTpt as a base model.

Since the methods being compared with BioNestedNER were not trained to recognize discontinuous entities, in order to make a fair comparison, we performed an experiment containing only the corpus with flat and nested entities, removing the discontinuous entities, and another experiment that also considers the discontinuous ones.

In both experiments with NestedClinBr corpus, our method achieved the SOTA in F1-score, although there was no statistical difference. As the method based on QA-NER is very similar, it obtained very similar results, with our method only 0.19 points ahead in the first experiment and 1.44 in the second

Table 8.2: Results in the NestedClinBr corpus, without discontinuous entities (experiment E3.1).

1.NestedClinBr - only flat and nested entities				
Model	Recall	Precision	F1-score	Acc NE
Baseline (binary relevance)				
BioBERTpt	0.7309	0.8333	0.7787	0.4907
BERTimbau	0.6253	0.7709	0.6905	0.2870
mBERT	0.5455	0.6959	0.6116	0.1296
Literature				
MRC	0.8134	0.8093	0.8113	-
PIQN	0.7661	0.8766	0.8176	-
QA-NER (CNN)	0.8599	0.8895	0.8744	0.6481
QA-NER (w/o CNN)	0.8589	**0.8931**	0.8756	0.6667
GPT-3 (few shot)	0.5396	0.6598	0.5936	-
Ours				
BioNestedNER (CRF multi-label)	0.7033	0.8357	0.7638	0.2870
BioNestedNER (QA-NER)	0.8629	0.8926	**0.8775**	0.6389
BioNestedNER (QA-NER + CRF)	**0.9043**	0.8346	0.8681	**0.7037**

Table 8.3: Results in the NestedClinBr corpus, with discontinuous entities (experiment E3.2).

2.NestedClinBr - flat, nested, and discontinuous entities					
Model	Recall	Precision	F1-score	Acc NE	Acc DE
Baseline (binary relevance)					
BioBERTpt	0.7085	0.8219	0.7571	0.3443	0
BERTimbau	0.6057	0.7535	0.6716	0.2203	0
mBERT	0.5301	0.6802	0.5959	0.1041	0
Literature					
MRC	0.7929	0.7728	0.7827	-	-
PIQN	0.7456	0.8351	0.7878	-	-
QA-NER (CNN)	0.8287	0.8575	0.8420	0.5056	0
QA-NER (w/o CNN)	0.8288	0.8686	0.8482	0.5279	0.0227
Ours					
BioNestedNER (QA-NER)	0.8542	**0.8711**	**0.8626**	0.6394	**0.2727**
BioNestedNER (QA-NER + CRF)	**0.8873**	0.8176	0.8510	**0.6617**	**0.2727**

one. However, when analyzing the accuracy for complex entities, our accuracy of nested entities is 0.7037, versus 0.6667 of QA-NER, in experiment 1, and 0.6617 versus 0.5279 in experiment 2. BioNestedNER was able to find 27.27% of discontinuous entities, and although this percentage is still conservative, it demonstrates that the method has the capacity to find this kind of entities, a plus compared to most other complex entity recognition methods that are not able to recognize this type of entity.

Compared to baselines with binary relevance, the method was 9.88 points ahead of the best model, which uses the same checkpoint model (BioBERTpt), indicating that the method was more effective than the traditional NER algorithm. Further, with BioNestedNER only one Transformer-based model needed to be trained, unlike the baseline where one model had to be trained for each entity type. Our method also performed better than MRC and PIQN, similar to our method as they also adapted the QA task to find entities. The MRC and PIQN methods were also trained using BioBERTpt as a checkpoint, but they were not as effective in F1-score. This may be due to the size of the corpus, which, being small, may have affected the performance of these methods. Also, in this corpus, we performed a few-shot experiment with GPT-3, in which it presented impressive results in view of the small number of input samples. Although it presented inferior results, it demonstrates the potential of this model in NLP tasks, requiring further tests and experiments with GPT-3.

In Appendix 10.3.1 we can see BioNestedNER results by entity type, where we noticed that in general, the model performed reasonably in all.

As expected, BioNestedNER model formed only by the multi-label CRF did not perform well alone, but when added to the BERT model, it improved recall and the number of nested entities found.

Analyzing the results in the recognition of discontinuous entities, using the necessary adaptation to recognize this type of entity, out of 44 discontinuous entities in the test set, our method found 24, and of these, only 12 are strictly correct. In Table 8.4, we can check some of the errors when identifying this type of entity. A large part of the errors concerns the lack of marking one or more words, or the inclusion of a word, causing all tags to be considered an error, as we use the restricted metric.

For an ablation study, we presented in Table 8.5 all experiments performed with the method, with the configuration variants, in the nested corpus. As we can see, in this corpus the best result of F1 was with class balancing using the

Table 8.4: Examples of matches and errors of discontinuous entities with the BioNestedNER method.

Gold	Predict	Expression	Free translation
Matches			
MMII edema	*MMII edema*	*MMII sem edema*	LL without edema
AD aumentado	*AD aumentado*	*AD = aumentado*	AD = increased
abdome massas	*abdome massas*	*abdome globoso normotenso indolor ausência de massas*	abdomen globus normotensive painless absence of masses
VTri refluxo discreto	*VTri refluxo discreto*	*VTri = refluxo discreto com PSAP 56mmHg.*	VTri = discrete reflux with PSAP 56mmHg.
Errors			
síncope mais com o calor	*síncope com o calor*	*síncope inicio em janeiro mais com o calor*	syncope beginning in January more with the heat
MMII EMPASTAMENTO	*MMII EDEMA EMPASTAMENTO*	*MMII : SEM EDEMA OU EMPASTAMENTO*	*LL: WITHOUT EDEMA OR CAMPING*
Microalbuminúria	*Microalbuminúria creatinina*	*Microalbuminúria 12 / 04: 10.64 / G de creatinina*	Microalbuminuria 12/04: 10.64/G creatinine
dor de cabeça melhora com analgésico comum	*dor de cabeça melhora*	*dor de cabeça melhora com analgésico comum*	headache improves with common pain reliever
AO CUSPIDES CALCIFICADAS	*CUSPIDES LESAO*	*AO : CUSPIDES CALCIFICADAS , COM DUPLA LESAO .*	AO: CALCIFICATED SCUPS, WITH DOUBLE LESION.
Decréscimo FC	*Decréscimo normal da FC*	*Decréscimo normal da FC no 1° minuto de recuperação*	Normal decrease in HR in the 1st minute of recovery
VMi folhetos espessados	*VMi*	*VMi = folhetos espessados , abertura preservada , refluxo discreto*	VMi = thickened leaflets, preserved opening, mild reflux

mean of losses in Cross-Entropy. The best precision value was obtained using the binary weights and the average of the CE losses. When adding the CRF results, we obtain the highest value of matches for nested entities (78.70%), with an increase in recall but at the expense of accuracy. This must be taken into account in each task, depending on the task and application and the consequences of false positives or false negatives.

Analyzing only the results of the CRF models, we found that the best configuration was using the division into 100 clusters, using our trained clinical Word2vec model (compared to the generic Portuguese model trained by (NILC, 2023)), with a threshold of 0.35, achieving 0.7033 in recall, 0.8357 in precision and 0.7638 in F1, and 0.2870 of nested accuracy. In Appendix 10.3.2 we have the complete results of the tests performed with CRF, where we can verify that with

Table 8.5: Experiments performed with all method variants

Experiments performed with all BioNestedNER variants				
Metrics	Recall	Precision	F1	Acc NE
BioNestedNER(classWeight+sum)	0.8589	0.8829	0.8707	0.6204
BioNestedNER(classWeight+sum+CRF)	0.8957	0.8299	0.8616	0.6852
BioNestedNER(classWeight+sum+CRF ensemble)	0.9354	0.6749	0.7841	0.7500
BioNestedNER(binaryWeight+sum)	0.8619	0.8787	0.8702	0.6759
BioNestedNER(binaryWeight+sum+CRF)	0.9021	0.8247	0.8616	0.7407
BioNestedNER(binaryWeight+sum+CRF ensemble)	0.9387	0.6775	0.7870	**0.7870**
BioNestedNER(classWeight+mean)	0.8629	0.8926	**0.8775**	0.6389
BioNestedNER(classWeight+mean+CRF)	0.9043	0.8346	0.8681	0.7037
BioNestedNER(classWeight+mean+CRF ensemble)	**0.9393**	0.6807	0.7892	0.7407
BioNestedNER(binaryWeight+mean)	0.8377	**0.9023**	0.8688	0.6389
BioNestedNER(binaryWeight+mean+CRF)	0.8800	0.8349	0.8569	0.7130
BioNestedNER(binaryWeight+mean+CRF ensemble)	0.9318	0.6757	0.7833	**0.7870**
BioNestedNER(CRF multi-label)	0.7033	0.8357	0.7638	0.2870

our clinical Word2Vec model, we obtained the best results.

8.3.2 SemClinBr Corpus (E4)

SemClinBr (OLIVEIRA et al., 2022) is a semantically annotated corpus for Brazilian-Portuguese clinical NER, containing 1,000 labeled clinical notes and 43,659 entities compatible with UMLS standard. SemClinBr has an interesting characteristic, the fact that each mention can have more than one label associated (which occurs in 14% of entities), making it ideal for our work of identifying multi-type entities. We perform experiments using the same hold-out split (60%-20%-20%) of previous works such as (SCHNEIDER et al., 2020), (SOUZA et al., 2019), and (SOUZA et al., 2021), as well as grouping entities in categories ("Disorder", "ChemicalDrugs", "Procedures" and "Abbreviation"), making possible comparisons with previous work. As a baseline, we have trained models with binary relevance using BioBERTpt, BERTimbau, and mBERT, as well as the MRC, PIQN, and QA-NER literature methods, all of which were initialized with BioBERTpt weights.

Our BioNestedNER method obtained superior results in F1 and recall met-

Table 8.6: Results in the SemClinBr corpus.

Model	Recall	Precision	F1-score	Acc ME
Baseline (binary relevance)				
BioBERTpt	0.5951	**0.7588**	0.6671	0.4969
BERTimbau	0.5568	0.7426	0.6365	0.4177
mBERT	0.5329	0.7259	0.6146	0.4015
Literature				
MRC	0.6352	0.5983	0.6162	-
PIQN	0.4738	0.6720	0.5558	-
QA-NER (CNN)	0.5975	0.7469	0.6639	0.5071
QA-NER (w/o CNN)	0.6003	0.7410	0.6633	0.5071
Ours				
BioNestedNER (CRF multi-label)	0.7110	0.6347	0.6707	0.5062
BioNestedNER (QA-NER)	0.6689	0.7312	0.6987	0.5471
BioNestedNER (QA-NER + CRF)	**0.8798**	0.7042	**0.7822**	**0.6989**

rics (but not statistically significant), although in precision BioBERTpt with BR obtained better results, as can be seen in Table 8.6. Our model reached the state-of-the-art in the F1 metric, with 0.7822 (11.51 above BioBERTpt, the second best placed) and 0.8798 in recall (24.46 above MRC, the second best placed). In terms of precision, BioBERTpt with binary relevance had better results with 0.7588. In this corpus, since it is multi-label and CRF has been trained in a multi-label way, it exhibited competitive results alone, ahead of the BR methods and the literature in terms of F1. We found that the ensemble with all CRF models was superior to a single isolated CRF model. When added with our QA-NER method, it increased by 8.35 F1 points and 15.18 in multi-type matches.

In Appendix 10.4.1 we can see BioNestedNER results by entity type, where we noticed that in general, the model performed well at finding entities like "ChemicalDrugs", with 0.9149 of F1, but not so well at recognizing entities like "Procedures". This may be due to these classes having different linguistic characteristics, for example, "ChemicalDrugs" entities may have more distinct textual patterns or specific keywords that facilitate their detection.

Analyzing multi-type entities, the model correctly found 69,89% of occurrences, versus 50,71% of the original QA-NER. The test dataset has a total of 2,252 entities with more than one type, most formed by a medical concept + "Abbreviation" (examples: "*HAS*", an "Abbreviation" + "Disorder", "*AAS*", an "Abbreviation" + "Chemical&Drug", and "*CAT*", an "Abbreviation" + "Procedure"). In a few cases, mention has two medical concepts like "*Valvoplastia Aortica*" (Aortic Valvoplasty) which is both a "Disorder" and a "Procedure".

Of the 2,252 multi-type entities, the model correctly found 1,574, which corresponds to approximately 69,89% of matches. Most of the errors are related to the model having recognized only one type of entity, as in some examples in Table 8.7.

Table 8.7: Examples of error in recognizing multi-type entities with the BioNestedNER method.

Expression	Free translation	Gold	Predict
PURAN	PURAN	Abbreviation, ChemicalDrug	ChemicalDrug
LAB	LAB	Abbreviation, Procedure	Abbreviation
DCA CHAGASICA	CHAGASICAL DCA	Abbreviation, Disorder	Disorder
TAC	TAC	Abbreviation, ChemicalDrug	ChemicalDrug
TX RENAL	RENAL TX	Abbreviation, Disorder, Procedure	Abbreviation
vacina de bcg	bcg vaccine	Abbreviation, Procedure, ChemicalDrug	(none)

All the experiments performed in this corpus are shown in Table 8.8.

Table 8.8: Experiments performed with all method variants.

Experiments performed with all BioNestedNER variants				
Metrics	Recall	Precision	F1	Acc ME
BioNestedNER(classWeight+sum)	0.6483	0.7303	0.6869	0.5355
BioNestedNER(classWeight+sum+CRF ensemble)	0.8730	0.7027	0.7787	0.6847
BioNestedNER(binaryWeight+sum)	0.6462	**0.7361**	0.6882	0.5231
BioNestedNER(binaryWeight+sum+CRF ensemble)	0.8697	0.7038	0.7780	0.6794
BioNestedNER(classWeight+mean)	0.6689	0.7312	0.6987	0.5471
BioNestedNER(classWeight+mean+CRF ensemble)	**0.8798**	0.7042	**0.7822**	**0.6989**
BioNestedNER(binaryWeight+mean)	0.6421	0.7307	0.6835	0.5275
BioNestedNER(binaryWeight+mean+CRF ensemble)	0.8705	0.7020	0.7772	0.6785
BioNestedNER(CRF ensemble)	0.7110	0.6347	0.6707	0.5062

We can see that the best configuration for F1, recall, and multi-type matches is the class-balanced method using the average in CE, plus the CRF ensemble results.

Analyzing only the results of the CRF models, we found that the best configuration was using the division into 50 clusters, using our trained clinical Word2vec model (compared to the generic Portuguese model trained by Nilc (NILC, 2023)), with a threshold of 0.15, achieving 0.5141 in recall, 0.6545 in precision and 0.5759 in F1, and 31.26% of multi-type hits. In Appendix 10.4.2 we have the complete results of the tests performed with CRF. We report the results of the individual CRFs models, although in this corpus, the ensemble formed by all CRFs together obtained the best results.

8.3.3 GENIA Corpus (E5)

The GENIA corpus (KIM et al., 2003) was created to develop and evaluate molecular biology information retrieval systems, with 2,000 PubMed abstracts based on the three medical subject heading terms: human, blood cells, and transcription factors. GENIA is one of the most used corpora to evaluate biomedical nested recognition models, as it contains 18,546 sentences with 56,870 entities, in which 31.64% approximately are nested and 3.65% are discontinuous entities (CHEN et al., 2020).

Although this dataset has been annotated with 36 types of entities, researchers usually use only "DNA", "RNA", "Protein", "Cell line" and "Cell type" entities, grouping the granular classes and ignoring all others. Also, in both training and evaluation, usually, the discontinuous entities are discarded since the research's focus commonly is on identifying nested entities. The authors preferentially follow the same division of data (the first 90% of sentences used in training and validation, and the rest for testing). We follow the same dataset split used in several works such as (FINKEL; MANNING, 2009), (LU; ROTH, 2015), and (MUIS; LU, 2017). In order to maintain compatibility with works in the literature and allow a fair comparison, we also adopted these configurations, although we also report results with our method in the original corpus, containing the discontinuous entities.

To facilitate, we divided this section into three parts: initially, we report our results in the corpus with nested entities (experiment E5.1), comparing the method with others in the literature. Next, we report the results in the complete corpus, containing the discontinuous entities (experiment E5.2). And at the end, we show a “few-shot” experiment performed with this corpus (experiment E5.3), executing several experiments with smaller parts of the corpus, in order to verify if the method can quickly adapt and learn more abstract and transferable representations of the data, in addition to measure and compare the processing time.

Genia with nested entities (experiment E5.1)

Table 8.9 shows the results of our method, in comparison with recent selected methods from the literature.

Although BioNestedNER did not reach the highest values of F1, it proved competitive with recent methods for nested entities. To find whether the differences are statistically relevant, we used the non-parametric statistical Friedman test with these results, followed by the Nemenmy post-hoc test to find the groups of data that differ from the rest. For this set of evaluated models, no statistical differences were found for the F1-score (*p-value = 0.456836*), indicating that our method presented competitive results with those of the literature.

Analyzing our model's performance in recognizing nested entities, we noticed that of the 1,111 nested entities that the test dataset had, our model was able to find 378, which represents 34.02

All the experiments performed in this corpus are shown in Table 8.10.

The corpus does not have discontinuous entities, however, it has nested entities of the same type, so the version with the end-to-end model presented the best results. We can see that the best configuration for F1 was with the binary balanced method using the mean in CE. Although the multi-label CRF alone did not present satisfactory results, as expected (with 0.6648 of F1), when combined with the Transformer-based model, it helped to increase the accuracy of nested entities found from 0.2592 to 0.6067.

In Appendix 10.5.1 we can see BioNestedNER results by entity type, where we noticed that the model performed well at finding all entities in general, being better on entities "Protein" and "RNA". "Cell line" and "Cell type", appearing in similar contexts, may be more difficult to distinguish from one another.

Analyzing only the results of the CRF models, we found that the best configuration was using the division into 10 clusters, with a threshold of 0.2, achieving 0.6148 in recall, 0.7385 in precision, and 0.6148 in F1, and 16.02% of nested hits.

Table 8.9: Results in the GENIA corpus, with flat and nested entities.

Model	F1	Precision	Recall
Locate and Label (SHEN et al., 2021)	0.8054	0.8019	**0.8089**
Multi-agent Comm (LI et al., 2021)	0.7650	0.7820	0.7480
Pyramid Layered (WANG et al., 2020)	0.7931	0.8031	0.7833
Multi-head Pyramid Layered (CUI; JOE, 2022)	0.8010	0.7947	0.7979
MRC (LI et al., 2020) *	0.7612	0.7751	0.7475
Excluding Best Path (WANG et al., 2021)	0.7858	0.7621	0.7737
BioBERT+TreeCRFs (FU1 CHUANQI TAN, 2020)	0.7820	0.7820	0.7820
Dependency Parsing (YU; BOHNET; POESIO, 2020)	0.8050	0.8180	0.7930
TCSF (SUN et al., 2020)	0.7730	0.7820	0.7650
PIQN (SHEN et al., 2022) *	0.8013	0.8110	0.7919
QA-NER (BANERJEE et al., 2021) *	0.7591	0.8118	0.7128
MTL-BAM (W LI Y, 2022)	**0.8065**	0.8062	0.8068
Labeling Gaps (MUIS; LU, 2017)	0.7080	0.7540	0.6680
Hypergraph RNN (KATIYAR; CARDIE, 2018)	0.7380	0.7670	0.7670
NER layers LSTM+ CRF (JU; MIWA; ANANIADOU, 2018)	0.7470	0.7850	0.7130
Deep exhaustive (SOHRAB; MIWA, 2018)	0.7710	**0.9320**	0.6400
Boundary-aware model (ZHENG et al., 2019a)	0.7470	0.7590	0.7360
Segmental hypergraphs (WANG; LU, 2018)	0.7700	0.7330	0.7510
Linearization (LUAN et al., 2019)	0.7830	-	-
MLC+Flair (ROJAS; BRAVO-MARQUEZ; DUNSTAN, 2022)	0.7760	0.8010	0.7520
BioNestedNER (OURS)	0.7913	0.8154	0.7686

* Results obtained by the author of this thesis, which can differ from those reported by the original authors, in view of the difference between the parameters used.

Table 8.10: Experiments performed with all method variants.

Experiments performed with all BioNestedNER variants				
Metrics	Recall	Precision	F1	Acc NE
BioNestedNER(classWeight+sum)	0.7316	0.8095	0.7686	0.2592
BioNestedNER(classWeight+sum)+CRF (best)	0.7815	0.7788	0.7801	0.3582
BioNestedNER(classWeight+sum)+CRF (ensemble)	0.8904	0.6648	0.7613	**0.6067**
BioNestedNER(binaryWeight+sum)	0.7208	0.8039	0.7601	0.2376
BioNestedNER(binaryWeight+sum)+CRF (best)	0.7865	0.7648	0.7755	0.3564
BioNestedNER(binaryWeight+sum)+CRF (ensemble)	0.8869	0.6643	0.7596	0.5851
BioNestedNER(classWeight+mean)	0.7424	0.8057	0.7728	0.2664
BioNestedNER(classWeight+mean)+CRF (best)	0.7894	0.7728	0.7810	0.3510
BioNestedNER(classWeight+mean)+CRF (ensemble)	0.8900	0.6650	0.7613	0.5923
BioNestedNER(binaryWeight+mean)	0.7386	0.8116	0.7734	0.2502
BioNestedNER(binaryWeight+mean)+CRF (best)	0.7967	0.7723	0.7843	0.3726
BioNestedNER(binaryWeight+mean)+CRF (ensemble)	**0.8925**	0.6672	0.7636	0.5995
BioNestedNER-end-to-end (binaryWeight+mean)	0.7686	**0.8154**	**0.7913**	0.3402
BioNestedNER-end-to-end (binaryWeight+mean)+CRF (best)	0.8014	0.7803	0.7907	0.4068
BioNestedNER-end-to-end (binaryWeight+mean)+CRF (ensemble)	0.8507	0.6935	0.7641	0.5149
CRF multi-label (best)	0.6683	0.6613	0.6648	0.2196

Genia with discontinuous entities (experiment E5.2)

Next, we report on Table 8.11 the results with the complete GENIA corpus, containing the discontinuous entities as well. In this case, we could only compare with the MRC, PIQN, and QA-NER methods, which we executed locally. As these literature methods are not prepared to deal with discontinuous entities, we do not train new models but used previously trained models with nested entities. We also didn't train specific CRF models, for the same reason, as our method using CRF doesn't allow us to find discontinuous entities. In this way, we used the same trained CRF models previously (to find nested entities). Our method, trained to recognize discontinuous entities, presented a higher F1 value, with 0.7800, followed by QA-NER (w/o CNN) with 0.7400, and a higher recall with 0.7832 followed by 0.7094 by MRC. Our method also managed to find more nested entities than the original QA-NER method (34.81% versus 16.02%), in addition to finding 31.30% of discontinuous entities.

Table 8.11: Results in the complete GENIA corpus, with flat, nested, and discontinuous entities.

Model	Recall	Precision	F1-score	Acc NE	Acc DE
Literature					
MRC	0.7094	0.7571	0.7325	-	-
PIQN	0.7009	0.6639	0.6829	-	-
QA-NER (CNN)	0.6739	0.8041	0.7333	0.1529	0
QA-NER (w/o CNN)	0.6809	0.8104	0.7400	0.1602	0
Ours					
BioNestedNER (QA-NER)	0.7484	**0.8093**	0.7777	0.2957	**0.3130**
BioNestedNER (QA-NER + CRF)	**0.7832**	0.7769	**0.7800**	**0.3481**	**0.3130**

We only consider discontinuous entities, mentions with one level of discontinuity (spacing) and one level of overlap between tokens, obtaining 115 discontinuous entities in our test dataset. Of this total, our method correctly found 36 entities, with strict mach, equivalent to 31.30%. In Table 8.12, we show some examples of mistakes and successes.

Our model predicted 84 discontinuous entities in total, but only 36 had the mention's boundaries + type correct (strict match), i.e. 31.30% of the predicted entities. Although the method had 36 matches for discontinuous entities, it was penalized in the overall performance by false positives and missing other discontinuous entities (false negatives).

Table 8.12: Examples of matches and errors of discontinuous entities with the BioNestedNER method.

Gold	Predict	Expression
Matches		
B cells	B cells	B and T cells
Human B lymphocytes	Human B lymphocytes	Human T and B lymphocytes
Human T lymphocytes	Human T lymphocytes	Human T and B lymphocytes
X boxes	X boxes	X and Y boxes
interleukin - 1 genes	interleukin - 1 genes	interleukin - 1 and MHC class II genes
megakaryocytic cell lines	megakaryocytic cell lines	megakaryocytic and erythroid cell lines
quiescent cells	quiescent cells	quiescent and stimulated cells
actin mRNA	actin mRNA	actin and fibronectin receptor mRNA
immunoglobulin heavy - chain genes	immunoglobulin heavy - chain genes	immunoglobulin heavy - and light - chain genes
immunoglobulin heavy chain gene enhancers	immunoglobulin heavy chain gene enhancers	immunoglobulin heavy and kappa light chain gene enhancers
Errors		
Human Rhom - 2	Human	Human and mouse Rhom - 2
human immunodeficiency virus promoters	human immunodeficiency virus interferon promoters	human immunodeficiency virus and beta interferon promoters
cytoskeletal genes	cytoskeletal	cytoskeletal, and extracellular matrix genes
heavy chain enhancer	heavy light chain enhancer	heavy and kappa light chain enhancers
human chromosomes 11p13	human chromosomes	human chromosomes 11p15 and 11p13
purified human macrophages	macrophages	purified human monocytes and macrophages
mitogen - treated peripheral blood lymphocytes	mitogen -	mitogen - and anti - CD3 - treated peripheral blood lymphocytes
human fibroblastic cells	*(none)*	human fibroblastic or keratinocyte - derived human cells
human papillomavirus transformed cells	human papillomavirus cells	human papillomavirus - or adenovirus - transformed cells
positive cis - acting DNA elements	cis - acting DNA elements	positive and negative cis - acting DNA elements

GENIA FEW SHOT (EXPERIMENT E5.3)

We performed five experiments with only 15% of the GENIA corpus, to compare the results and also the execution speed with other methods in the literature. In this cut of the database, we worked with 2,500 examples in training and 350 for validation. We created five smaller datasets, with this amount of examples, and performed five experiments with BioBestedNER, PIQN, MRC, QA-NER (CNN), and QA-NER (w/o CNN), as they are methods that use a similar approach to ours, all using BioBERT as a checkpoint. It is worth mentioning that the five small datasets are balanced according to the original dataset. We measured the micro F1 metric in the same test set for all experiments, with 1,854 samples, and computed the runtime as well. We only considered the execution times per epoch, and not the total time, since in MRC and PIQN the maximum number of epochs (with early stop) was 20 and 30, respectively (according to default values). In the case of the CRF execution time, we consider the entire training time (with 100 interactions), since there is no epoch count.

Table 8.13 shows the results of our few-shots experiments, in terms of the micro F1-score metric.

Table 8.13: Micro F1 results in the GENIA few-shot corpus for the five experiments.

Method	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)	Avg	ACC NE
Literature							
MRC	0.7316	0.7345	0.7364	0.7294	0.7309	0.7325	-
PIQN	0.7162	0.7184	0.7132	0.7279	0.7148	0.7181	-
QA-NER (CNN)	0.7396	0.7434	0.7490	0.7355	0.7319	0.7399	0.1919
QA-NER (w/o CNN)	0.7260	0.7393	0.7409	0.7305	0.7388	0.7351	0.1843
Ours							
BioNestedNER (CRF multi-label)	0.6647	0.6718	0.6660	0.6680	0.6648	0.6671	0.2272
BioNestedNER (QA-NER)	0.7650	0.7600	0.7642	0.7632	0.7630	0.7631	0.2707
BioNestedNER (QA-NER + CRF)	**0.7663**	0.7623	0.7623	0.7623	**0.7629**	0.7632	**0.3550**
BioNestedNER (QA-NER end-to-end)	0.7597	**0.7628**	**0.7722**	**0.7679**	0.7584	**0.7642**	0.2880
BioNestedNER (QA-NER end-to-end + CRF)	0.7612	0.7618	0.7693	0.7658	0.7588	0.7634	0.3510

It is remarkable that our models trained with only 15% of the training dataset achieved similar results to the models trained with 100% of the data, with only 2.71 points difference in F1-score on average. This may be explained by the fact that a small portion of the GENIA data may be representative enough to capture the main patterns and features of the task. Our model trained with the end-to-end format obtained the best result on average, with 0.7642 of F1, being 2.43 points ahead of the original model QA-NER, with 0.7399. Furthermore, our models were also able to find more nested entities (35.50% versus 19.19% of the original QA-NER).

We performed the Friedman statistical test with the F1 values of the five experiments, to see if there is a difference between the results. We found that our models (BioNestedNER (QA-NER)+CRF, BioNestedNER (QA-NER end-to-end), and BioNestedNER (QA-NER)) had statistically better results than the PIQN and BioNestedNER (multi-label CRF) models (Figure 8.2). Our CRF model alone was already expected to present inferior results since it is a complement to the method. PIQN is a model that presented better results using the complete GENIA corpus, but its performance may suffer when using a small dataset for training due to overfitting or poor generalization to new data.

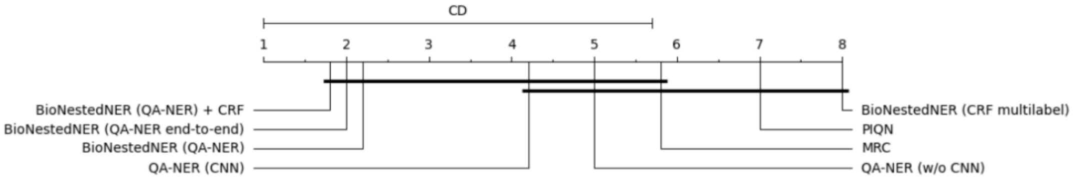


Figure 8.2: Statistical result of GENIA few-shot experiments using F1-scores.

We also did an experiment with GPT-3 (Davinci) (BROWN et al., 2020), which achieved 0.4126 on F1-score. We cannot compare the results with others because we trained with ChatGPT using only 20 example sentences and evaluating on only 719 instances (limit reached). Still, it was interesting to test how the model reacts in a biomedical experiment being trained with very few input examples, exploring the potential of this tool in NLP tasks. Table 8.14 shows the results of our few-shots experiments, in terms of training time for epoch (TPE).

Table 8.14: Training time in the GENIA few-shot corpus for the five experiments.

Method	TPE(1)	TPE(2)	TPE(3)	TPE(4)	TPE(5)	Avg
Literature						
MRC	00:04:40	00:04:51	00:04:47	00:04:50	00:04:23	00:04:42
PIQN	00:07:53	00:08:17	00:08:20	00:07:59	00:08:00	00:08:06
QA-NER (CNN)	00:04:43	00:04:45	00:04:51	00:04:47	00:04:43	00:04:46
QA-NER (w/o CNN)	00:04:24	00:04:28	00:04:32	00:04:29	00:04:32	00:04:29
Ours						
BioNestedNER (CRF multi-label)	**00:00:20**	**00:00:16**	**00:00:20**	**00:00:19**	**00:00:19**	**00:00:19**
BioNestedNER (QA-NER)	00:04:21	00:04:23	00:04:25	00:04:15	00:04:16	00:04:20
BioNestedNER (QA-NER + CRF)	00:04:41	00:04:39	00:04:55	00:04:34	00:04:35	00:04:39
BioNestedNER (QA-NER end-to-end)	00:04:56	00:04:56	00:05:02	00:05:02	00:05:01	00:05:00
BioNestedNER (QA-NER end-to-end + CRF)	00:05:16	00:05:12	00:05:22	00:05:21	00:05:20	00:05:19

The CRF training was faster, with 19 seconds in general, as is a simpler model with fewer parameters to train compared to deep learning models. We calculated the statistical differences, transforming the times per epoch into seconds (Figure 8.3), and found that BioNestedNER (multi-label CRF), BioNestedNER (QA-NER) and QA-NER (w/o CNN) were statistically faster than PIQN. We noticed that the end-to-end models, although they improve in F1, tend to make the processing slower, with BioNestedNER (QA-NER) being faster than BioNestedNER (QA-NER end-to-end) + CRF, in addition to PIQN.

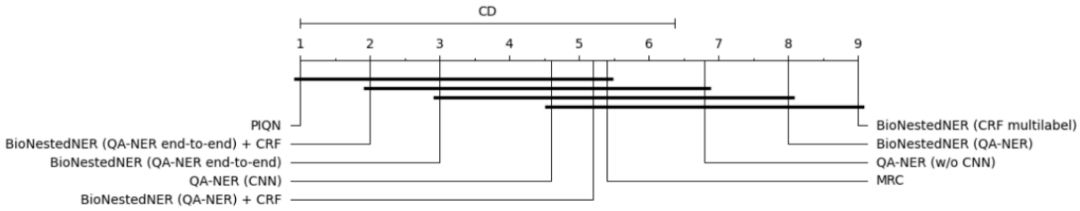


Figure 8.3: Statistical result of GENIA few-shot experiments using time values per epochs.

Our method presented competitive training time to other similar methods (MRC and QA-NER) and was even faster than PIQN, which may indicate that the method requires similar or low computational complexity.

8.3.4 Rare Disease Corpus (E6)

Next, we report our results in Rare Disease, an annotated corpus in English of rare diseases and their clinical manifestations, containing 9,141 sentences and approximately 9,300 annotated entities, including nested and discontinuous entities (MARTíNEZ-DEMIGUEL et al., 2022). Rare Disease also includes the annotation of relations between entities, outside of this research scope. Unfortunately, at the time of running the experiments, we didn't have access to the test dataset, so we had to move about 25% of the training dataset to use as validation. For this reason, it will not be possible to perform a comparison with the baseline of (SEGURA-BEDMAR; CAMINO-PERDONAS; GUERRERO-ASPIZUA, 2021) since our results are unofficial. In addition, the authors of the baseline did not deal with nested and discontinuous entities, appearing as future works in their paper. Nevertheless, we report our results along with some literature methods, locally trained, in Table 8.15.

Table 8.15: Results in the Rare Disease corpus.

Model	Recall	Precision	F1-score	Acc NE
Literature				
MRC	0.7169	0.7051	0.7196	-
PIQN	0.7216	0.7430	**0.7322**	-
QA-NER (CNN)	0.6699	0.7643	0.7140	0.0180
QA-NER (w/o CNN)	0.6697	**0.7694**	0.7161	0.0180
Ours				
BioNestedNER (CRF multi-label)	0.6196	0.6968	0.6560	0
BioNestedNER (QA-NER)	0.6932	0.7587	0.7245	**0.0360**
BioNestedNER (QA-NER + CRF)	**0.7309**	0.7271	0.7290	0.0180

In this corpus, the PIQN method had better results in micro F1, with 0.7322, 0.32 points ahead of BioNestedNER, which, in compensation, obtained a higher recall value of 0.7309. Our method did not have superior results, but it was above the original QA-NER method, demonstrating that the modifications in the approach had an effect here. Furthermore, in the baseline of (SEGURA-BEDMAR; CAMINO-PERDONAS; GUERRERO-ASPIZUA, 2021), the authors report 0.7181 of F1, 1.09 points behind BioNestedNER. Even though the results cannot be compared for the reasons explained above, we can have an idea of the competitiveness of our method.

In Appendix 10.6.1 we can see BioNestedNER results by entity type, with our results being similar to (SEGURA-BEDMAR; CAMINO-PERDONAS; GUERRERO-ASPIZUA, 2021), where the "Rare Disease" entity, with one of the largest support values, had the highest values of precision, recall, and F1. On the other hand, "Sign" and "Symptom" entities showed the lowest F1-score, which may be because these mentions are usually nominal phrases, unlike a few technical terms. Moreover, the support value of the "Symptom" entity is small, making learning more difficult.

Analyzing the performance in recognizing nested entities, we noticed that of the 111 nested entities that the test dataset had, our model was able to find just 2, which represents 1.8%. We noticed that the model had difficulties in distinguishing "Rare disease" from only diseases, and also "Disease" from "Sign". For example, "progressive arthritis of the spine" is annotated with a "Sign", and "progressive arthritis", a nested entity being a "Disease". The model only predicted "progressive arthritis of the spine" as a "Sign", dismissing the "Disease" inside it. The same occurred in "hereditary ataxia", which is a "Rare disease", and "ataxia", a "Sign". The model found only "hereditary ataxia", however as a "Disease" (instead as a "Rare disease"). In another example, "Infectious arthritis" is a "Rare disease", while "arthritis" is a "Sign". The model did not find either of the two entities. As these clinical entities have very similar meanings, they generally are more complex and harder to detect. Additionally, as exhibited in Figure 8.4, the imbalanced class distribution may affect the performance of the model, both in general and in the search for nested entities.

All the experiments performed in this corpus are shown in Table 8.16.

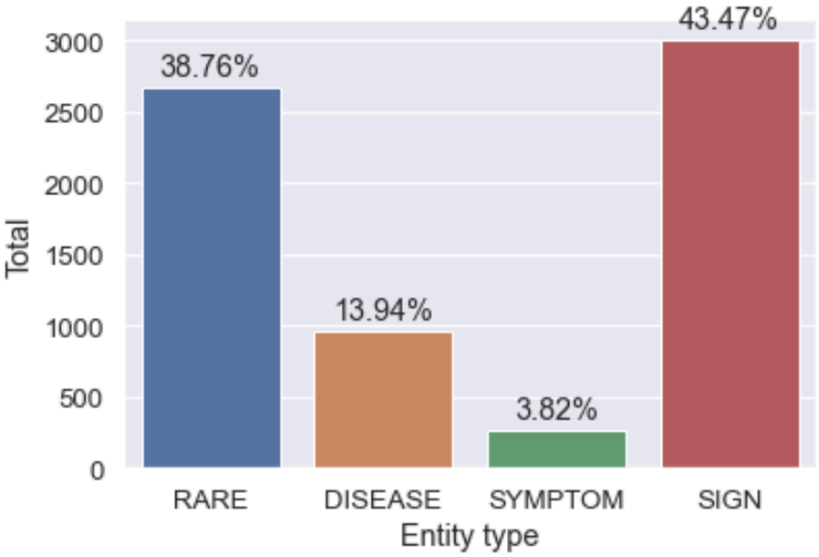


Figure 8.4: Balancing classes in the Rare Disease corpus used in our training.

Table 8.16: Experiments performed with all method variants.

Experiments performed with all BioNestedNER variants				
Metrics	Recall	Precision	F1	Acc NE
BioNestedNER(classWeight+sum)	0.6710	**0.7713**	0.7176	0.0180
BioNestedNER(classWeight+sum+CRF)	0.7258	0.7291	0.7274	0.0180
BioNestedNER(classWeight+sum+CRF ensemble)	0.8743	0.5298	0.6598	0.3063
BioNestedNER(binaryWeight+sum)	0.6846	0.7670	0.7234	0
BioNestedNER(binaryWeight+sum+CRF)	0.7309	0.7271	**0.7290**	0.0180
BioNestedNER(binaryWeight+sum+CRF ensemble)	0.8765	0.5323	0.6624	0.3063
BioNestedNER(classWeight+mean)	0.6897	0.7463	0.7169	0.0180
BioNestedNER(classWeight+mean+CRF)	0.7364	0.7061	0.7210	0.0180
BioNestedNER(classWeight+mean+CRF ensemble)	**0.8826**	0.5270	0.6600	**0.3243**
BioNestedNER(binaryWeight+mean)	0.6932	0.7587	0.7245	0.0360
BioNestedNER(binaryWeight+mean+CRF)	0.7397	0.7175	0.7284	0.0541
BioNestedNER(binaryWeight+mean+CRF ensemble)	0.8806	0.5279	0.6600	**0.3243**
BioNestedNER(CRF multi-label)	0.6196	0.6968	0.6560	0
BioNestedNER(CRF ensemble)	0.8078	0.5270	0.6379	0.2883

We can see that the best configuration for F1 is the binary class-balanced method using the sum in CE, plus the best CRF result. However, if we consider the number of matches in nested entities, the best configurations are our QA-NER model with class balancing and binary class balancing using the average of losses added with the CRF ensemble, with 32.43% of nested accuracy. Although there is a loss in precision, this result can be interesting when the objective of the application is to minimize false negatives, i.e. when it is more important to identify all positive instances, even at the cost of some false positives. This is often the case in applications such as medical diagnosis, where missing a positive instance can have serious consequences. We can see that although the CRF alone does not perform well, when performing an ensemble with all the trained CRFs, we reached 28.83% of the nested entities in the corpus. Analyzing the results of each CRF model, we found that the best configuration was using the division into 50 clusters, with a threshold of 0.3, achieving 0.6196 in recall, 0.6968 in precision, and 0.6560 in F1.

8.3.5 PortugueseClinicalNER Corpus (E7)

PortugueseClinicalNER (LOPES; TEIXEIRA; OLIVEIRA, 2019) is a collection of 281 clinical texts in Portuguese manually annotated for named entities in the biomedical domain. The corpus has only flat entities, important for our research since we also want to evaluate the method on a corpus containing only flat entities. We trained models to be baselines, a BioBERTpt traditional NER model, MRC, PIQN, and QA-NER methods, all using BioBERTpt as a checkpoint. Table 8.17 exhibits the results, where our method has reached the state-of-the-art for this corpus in micro F1 with 0.9482, being even better than the simple NER with BioBERTpt with 0.9332.

This result is relevant, since this corpus has 13 types of entities, demonstrating that our method is also effective when there is a relatively large number of entities. Unexpectedly, our CRF model alone achieved better results than the baseline using BiLSTM+CRF (LOPES; TEIXEIRA; OLIVEIRA, 2019). This may have occurred due to the size of the dataset, which has only 281 clinical texts which may not be sufficient to train a complex model like BiLSTM. The feature engineering and the model label dependencies of CRF models may have affected the results as well.

In Appendix 10.8.1 we can see BioNestedNER results by entity type, where

Table 8.17: Results in the PortugueseClinicalNER corpus.

Model	Recall	Precision	F1
Baseline (traditional NER)			
BiLSTM+CRF (out-of-domain) (LOPES; TEIXEIRA; OLIVEIRA, 2019)	0.7335	0.7506	0.7419
BiLSTM+CRF (in-domain) (LOPES; TEIXEIRA; OLIVEIRA, 2019)	0.7448	0.7525	0.7486
BioBERTpt	0.9139	0.9533	0.9332
Literature			
MRC	0.9184	0.8778	0.8976
PIQN	0.7749	0.9356	0.8477
QA-NER (CNN)	0.9280	0.9604	0.9439
QA-NER (w/o CNN)	0.9295	0.9625	0.9457
Ours			
BioNestedNER(CRF multi-label)	0.8438	0.7841	0.8438
BioNestedNER(QA-NER)	0.9290	**0.9680**	0.9481
BioNestedNER(QA-NER + CRF)	**0.9437**	0.9528	**0.9482**

we noticed that in general, the model performed well at finding all entities, reaching value 1 in F1 for "Genetic" and "Via" entities.

Analyzing only the results of the CRF models, we found that the best configuration was using the division into 100 clusters, using our trained clinical Word2vec model (compared to the generic Portuguese model trained by Nilc (NILC, 2023), with a threshold of 0.35, achieving 0.8438 in recall, 0.7841 in precision and 0.8438 in F1. In Appendix 10.8.2 we have the complete results of the tests performed with CRF, where we noticed that in general, the model performed reasonably in all.

All the experiments performed in this corpus are shown in Table 8.18.

Table 8.18: Experiments performed with all method variants

Experiments performed with all BioNestedNER variants			
Metrics	Recall	Precision	F1
BioNestedNER(classWeight+sum)	0.9189	0.9631	0.9405
BioNestedNER(classWeight+sum+CRF)	0.9351	0.9485	0.9417
BioNestedNER(binaryWeight+sum)	0.9290	**0.9680**	0.9481
BioNestedNER(binaryWeight+sum+CRF)	0.9437	0.9528	**0.9482**
BioNestedNER(classWeight+mean)	0.9285	0.9584	0.9432
BioNestedNER(classWeight+mean+CRF)	0.9391	0.9463	0.9427
BioNestedNER(binaryWeight+mean)	0.9290	0.9550	0.9418
BioNestedNER(binaryWeight+mean+CRF)	**0.9408**	0.9446	0.9427
BioNestedNER(CRF multi-label)	0.8438	0.7841	0.8438

We can see that the best configuration for F1 is the binary class-balanced method using the sum in CE, plus the best CRF result.

8.3.6 JNLPBA Corpus (E8)

JNLPBA is a biomedical dataset originated from GENIA corpus (version 3.02) (COLLIER; KIM, 2004), where nested and discontinuous entities were removed. Like the experiment with PortugueseClinicalNER, we use this corpus to validate the method in a corpus containing only flat entities, but this time in English.

In our research, we used the original corpus, containing the tags in BIO format plus the type of entity. However, there is a corpus variation where only the BIO tags are present, without the entity type. For this reason, in our comparative table, we will not include all results reported in JNLPBA, but only those that were performed with the same original corpus. Our method did not reach the state of the art in the JNLPBA, but presented competitive results, being behind the first place by only 1.61 in F1 (Table 8.19).

Table 8.19: Results in the JNLPBA corpus.

Model	Recall	Precision	F1
Literature			
Deep exhaustive (SOHRAB; MIWA, 2018)	0.6680	**0.9640**	**0.7840**
MRC (LI et al., 2020)	0.7789	0.7009	0.7378
PIQN (SHEN et al., 2022)	0.7572	0.7399	0.7484
QA-NER (CNN) (BANERJEE et al., 2021)	0.7547	0.7480	0.7513
QA-NER (w/o CNN) (BANERJEE et al., 2021)	0.7486	0.7514	0.7500
LSTM+CNN (WEI et al., 2019)	0.7622	0.7137	0.7372
NERBio (TSAI et al., 2006)	0.7398	0.7201	0.7298
Gimli (CAMPOS; MATOS; OLIVEIRA, 2013)	0.7162	0.7285	0.7223
Multi-task LSTM (WANG et al., 2019)	0.7634	0.7091	0.7352
Gram-CNN (ZHU et al., 2017)	-	-	0.7257
Ours			
BioNestedNER (CRF multi-label)	0.6738	0.7190	0.6956
BioNestedNER (QA-NER)	0.7744	0.7482	0.7611
BioNestedNER (QA-NER + CRF)	**0.8157**	0.7254	0.7679

Regarding the compared methods, our method reached the best recall value, with 0.8157, 3.68 points ahead of second place, MRC. Although the method focuses on nested, discontinuous, and multi-type entities, these results demonstrate that BioNestedNER can also be applied to corpora containing flat NER, with results similar to other methods in the literature.

In Appendix 10.7.1 we can see BioNestedNER results by entity type, where we noticed that in general, the model performed well at finding all entities, with the exception of "Cell line". As in GENIA, the model may have difficulty distinguishing "Cell line" from Cell type", prioritizing "Cell type" that has greater support.

All the experiments performed in this corpus are shown in Table 8.20.

Table 8.20: Experiments performed with all method variants

Experiments performed with all BioNestedNER variants			
Metrics	Recall	Precision	F1
BioNestedNER(classWeight+sum)	0.7659	**0.7549**	0.7603
BioNestedNER(classWeight+sum+CRF)	0.8094	0.7302	0.7678
BioNestedNER(binaryWeight+sum)	0.7488	0.7543	0.7516
BioNestedNER(binaryWeight+sum+CRF)	0.8020	0.7302	0.7644
BioNestedNER(classWeight+mean)	0.7744	0.7482	0.7611
BioNestedNER(classWeight+mean+CRF)	**0.8157**	0.7254	**0.7679**
BioNestedNER(binaryWeight+mean)	0.7687	0.7471	0.7578
BioNestedNER(binaryWeight+mean+CRF)	0.8135	0.7252	0.7668
BioNestedNER(CRF multi-label)	0.6738	0.7190	0.6956

We can see that the best configuration for F1 is the class-balanced method using the mean in CE, plus the CRF results.

Analyzing only the results of the CRF models, we found that the best configuration was using the division into 300 clusters, with a threshold of 0.35, achieving 0.6738 in recall, 0.7190 in precision, and 0.6956 in F1.

8.4 Revisiting Research Objectives and Hypothesis

In this section, we correlate the research goals with the developed methods and results, describing how this work responds to the research questions. This section can be used as a summary of all the achievements and findings of this thesis.

8.4.1 Research Objectives Achievement

First, we bring up the research goals and summarize how they were accomplished.

Main objective

The development of a named entity recognition method that also considers nested, discontinuous and multi-type entities, using state-of-the-art architecture for NLP, such as Transformer architecture and deep learning.

The main objective of this work was to develop a NER method that can recognize nested, discontinuous, and multi-type entities (in addition to the flat ones), since traditional NER methods cannot identify these entities, and ignoring them can cause relevant information to be lost.

In this study, we explored several approaches to recognize these complex entities, developing a flexible and efficient method that we call BioNestedNER. The evaluation protocol included comparing the results against other literature methods through experiments with six NER corpora resulting in nine experiments. This objective was reached with our method using deep learning and Transformer architecture and presenting state-of-the-art results in six experiments and competitive results in the others.

Specific objective 1

To study existing approaches that address the recognizing of complex named entities to compare with the proposed method.

We performed a study in the literature, according to the Related Work section, to understand the problem and recognize the gaps and limitations of existing works. We found similar works, based on the same strategy and compared them with our method to find the benefits and limitations of each one.

Specific objective 2

To search available NER corpora containing complex entities in English and Portuguese languages to perform experiments.

In the exploratory phase of the research, in addition to searching for methods adapted for complex entities, we also searched for available corpora containing this type of entity. We found 12 corpora used by the researchers, of which three were selected for the research because they are in English or Portuguese languages, in the clinical or biomedical domains, and have public access available. We also identified a gap during the research: no clinical corpus in Portuguese containing nested and discontinuous entities was found.

Specific objective 3

To develop a two-phase method that combines the QA-NER approach with CRF to recognize nested, discontinuous, and multi-type entities, while achieving competitive results without the high computational demands associated with exhaustive methods.

Our method uses the QA-based approach, being simple and not demanding high computational resources as in exhaustive methods, which first needs to list all candidate entities through a combination of all tokens, and then classify each one. It also requires training a unique Transformer-based model, which returns all the entities in a single passage, as opposed to methods that first find the entity boundaries and then classify them. Although the method is composed of two phases, as the CRF has a relatively simple model structure with fewer parameters, it does not impact the complexity of the method.

Specific objective 4

To develop a guideline for human annotations of nested and discontinuous entities in clinical texts in Portuguese.

We developed a guideline for annotating clinical entities, using the guideline by (GUMIEL et al., 2023) as a basis. Our guidelines will be available to the community, which may be helpful for building a larger corpus, containing nested and discontinuous entities. Although built for the Portuguese language, it can be applied in other languages in the medical domain, generating corpus for training and evaluating machine learning models and thereby helping the extraction of clinical information.

Specific objective 5

To build a corpus with clinical texts in Brazilian Portuguese containing nested and discontinuous entities.

We built and will make available NestedClinBr, a Brazilian-Portuguese clinical corpus, annotated twice by different annotators. We measured the IAA based on F1 and reached 94.08% in general, meeting the expected results.

Specific objective 6

To train clinical and biomedical Transformer-based models for the Portuguese language.

We have trained clinical and biomedical language models using BERT architecture, which reached state-of-the-art in many NLP tasks. In total, we trained three generic (no-task) models for the medical domain and 13 specific models for entity extraction. As Portuguese is a low-resource language, we made all our models freely available to the community.

Specific objective 7

To evaluate the proposed method in both English and Portuguese in clinical and biomedical domains.

We evaluated our method in six corpora, in Portuguese (NestedClinBr, SemClinBr, and PortugueseClinicalNER) and English languages (GENIA, JNLPBA, and Rare Disease). Our method reached SOTA in all corpora in Portuguese in micro F1-score, and competitive results in English corpora.

8.4.2 Research Hypotheses Response

This section revisits the presented research hypotheses and, based on the research and results of this study, aims to answer them.

Research hypothesis H1

A new NLP task, which combines aspects of both NER and QA, allows the successful recognition of nested, multi-type, and discontinuous entities, yielding competitive results with literature methods.

We verify that this hypothesis is true since BioNestedNER was able to find the complex entities simply and efficiently. This has been proven through experiments: E3.1 and E3.2 (NER experiment in the NestedClinBr corpus), E4 (NER experiment in the SemClinBr corpus), E5.1, E5.2, and E5.3 (NER experiment in the GENIA corpus), and E6 (NER experiment in the Rare Disease corpus).

In addition to the previous contributions by (BANERJEE et al., 2021) in demonstrating the feasibility of recognizing nested and multi-type entities, our research extends these findings by showcasing the recognition of discontinuous entities, in experiments E3.2 and E5.2.

Research hypothesis H2

By incorporating a multi-label CRF model into the Transformer-based model, the method improves the coverage of nested and multi-type entities.

We verified that by combining the results of the Transformer-based model with those of the CRF, there is an increase in the accuracy of the nested and multi-type entities found, in addition to an increase in recall in general.

In the corpora SemClinBr (E4), GENIA with discontinuous (E5.2), Rare Disease (E6), PortugueseClinicalNER (E7), and JNLPBA (E8), we noticed an increase in the value of F1-score with the addition of the CRF. Moreover, the accuracy value of nested and multi-type entities increased with the addition of CRF in all evaluated scenarios.

Research hypothesis H3

H3: We hypothesize that our method achieves state-of-the-art performance in NER task, when performed in corpora containing complex entities.

The hypothesis was partially confirmed. In experiments E3.1 and E3.2 (NER experiment in the NestedClinBr corpus), E4 (NER experiment in the SemClinBr corpus), and E5.2 (NER experiment in the GENIA corpus) our method presented the best F1-score results.

In E5.3, we performed a few-shot experiment with 15% of the training corpus. Our method has reached state-of-the-art F1-score results compared to similar methods (same approach), statistically superior to the PIQN method

In other experiments performed with the GENIA corpus (E5.1) and Rare Disease corpus (E6), our method did not reach the state-of-the-art, however, obtained competitive results (statistically similar) to the literature methods.

9 Conclusion

In several situations, named entities in a text can be formed by nested, overlapping, multi-type, or discontinuous mentions. However, traditional NER methods are not able to capture these complex entities, which can lead to the loss of relevant information, especially in the clinical and biomedical domains. Although most recent works in NLP present the use of Transformer architecture, deep learning, and contextualized models such as BERT, few methods focus on the recognition of complex entities. Moreover, less attention has been given to lower-resource languages, such as Portuguese.

This thesis has explored the challenges and opportunities associated with nested NER, including the development of a new method to improve its efficiency, addressing studies in Portuguese and English. We have identified several key factors that can impact the performance of complex NER systems, including the use of contextualized language models, Transformer architecture, and the use of in-domain models as checkpoints. Additionally, we proposed a new method formed by two phases, called BioNestedNER. The method is based on a QA approach which proved to be efficient in recognizing nested, multi-type, and discontinuous entities, combined with a multi-label CRF model. Combining different methods can improve the coverage, accuracy, and robustness of NLP tasks as NER since it is possible to obtain a broader and more complete coverage of the text.

To support experiments in the Portuguese language, we have generated several models for the clinical and biomedical domains, the BioBERTpt models, which are publicly available to the research community. We have also constructed a clinical corpus and plan to release it openly. As far as our knowledge extends, this will be the first clinical corpus in Brazilian Portuguese to include nested and discontinuous entities.

For the feature extraction, to train the CRF models, we also have trained a POS-tagger model, based on BioBERTpt weights, and a Word2Vec model with clinical data, used to generate word clusters based on their similarity. Compared to the generic model (trained by (NILC, 2023)), our clinical Word2vec model showed improvements in the results of CRF models, as can be seen in Appendices 10.3.2, 10.4.2, and 10.8.2. We also intend to make all these resources public.

Overall, our results suggest that the proposed method is a promising approach for complex NER and related NLP tasks, such as information extraction, question-answering, and multi-label tasks. Our method reached SOTA in terms of micro F1 in six experiments out of nine, involving the search for flat, nested, discontinuous, multi-type entities and a few-shot scenario. In addition, it can be applied in several domains and languages, with the potential for further improvements. The findings of this thesis are likely to be of interest to researchers and practitioners working in the field of NLP and related areas.

9.1 Contributions

This thesis contributes to clinical NLP research by exploring a critical task, the recognition of complex entities, very common in clinical and biomedical texts. Accurate information retrieval could increase, directly or indirectly, the quality of patient care. Furthermore, the production and availability of resources and models of this study should contribute to the development and evaluation of several new methods and models. As seen in the “Introduction” section, our main contributions are a new method for complex NER, resources for Portuguese clinical and biomedical domains, and a new Brazilian Portuguese clinical corpus containing nested and discontinuous entities.

9.2 Future Work

In this section, we recommend several research directions that could be explored in further studies:

- **Improvements on NestedClinBr corpus.** It would be convenient to increase the size of the corpus with more labeled clinical notes, using the defined guideline, allowing to train and evaluate more robust models. Also, like SemClinBr and TempClinBr, it would be interesting to label the negations, to extract the most accurate information possible.

- **Conducting experiments in other domains and languages.** The developed method has been evaluated in the clinical and biomedical domains, in two different languages, but has the potential to be applied to other domains and languages. Since Spanish is a language similar to Portuguese and also belongs to the Romance language group, which originates from the Latin languages, it will be interesting to evaluate the model on the NLPMedTerm and CWLCE corpora. This will allow us to assess our method in a language that shares a common origin. It will also be interesting to test with other language models such as RoBERTa.

- **Improve recognition of discontinuous entities.** Identifying discontinuous entities is itself a considerable challenge. Our method can be improved in the future, targeting this type of entity.

- **Explore more techniques for data imbalance.** We also would like to explore more techniques for dealing with data imbalance, in order to avoid class bias, improve generalization and enhance the machine learning models' accuracy.

- **Develop and make available more resources for the Portuguese language, in the health area.** As we still have few resources in the clinical domain for the Portuguese language, new language models will be very useful. We intend to train a new version of the BioBERTpt models, using the BERTimbau model as a checkpoint and more clinical and biomedical data. In the course of this work, we also trained a biomedical GPT-2 model for Portuguese and new BERT models with cardiology clinical data. Also, it would be interesting to generate clinical models based on other architectures less explored in Portuguese, such as RoBERTa, DistilBERT, etc.

- **More experiments with generative algorithms** We plan to conduct further experiments involving generative algorithms such as GPT-3, GPT-4 [1], and BARD [2], enabling new comparisons, since these models were not available at the time of writing this document.

[1] https://openai.com/research/gpt-4

[2] https://bard.google.com/

10 Appendix

10.1 Search Queries

The full search queries used in the scientific databases are presented below.

10.1.1 PubMed

English

(nested [Title/ Abstract] OR "complex entity" [Title/ Abstract] OR "complex named entity" [Title/ Abstract] OR "complex named entities" [Title/ Abstract] OR "complex entities" [Title/ Abstract] OR "multi type entity" [Title/ Abstract] OR "multi type entities" [Title/ Abstract] OR "overlapping entity" [Title/ Abstract] OR "overlap entity" [Title/ Abstract] OR "overlapped entity" [Title/ Abstract] OR "discontinuous entity" [Title/ Abstract] OR "discontiguous entity" [Title/ Abstract] OR "multilabel entity" [Title/ Abstract] OR "multi label entity" [Title/ Abstract] OR "structured entity" [Title/ Abstract] OR "structured entities" [Title/ Abstract] OR "structured named entity" [Title/ Abstract] OR "structured named entities" [Title/ Abstract] OR "irregular entity" [Title/ Abstract] OR "irregular entities" [Title/ Abstract] OR "cascaded entity" [Title/ Abstract] OR "cascaded entities" [Title/ Abstract]) AND (ner [Title/ Abstract] OR "named entities" [Title/ Abstract] OR "named entity" [Title/ Abstract] OR "entity recognition" [Title/ Abstract]) NOT (nucleotide excision repair [Title/ Abstract])

Portuguese

(aninhada [Title/ Abstract] OR aninhadas [Title/ Abstract] OR "entidade complexa" [Title/ Abstract] OR "entidade nomeada complexa" [Title/ Abstract] OR "entidades nomeadas complexas" [Title/ Abstract] OR "entidades complexas" [Title/ Abstract] OR "entidade multitipo" [Title/ Abstract] OR "entidades multitipos" [Title/ Abstract] OR sobreposta [Title/ Abstract] OR sobreposta [Title/ Abstract] OR sobrepor [Title/ Abstract] OR descontinuas [Title/ Abstract] OR descontinuadas [Title/ Abstract] OR "multirrótulo " [Title/ Abstract] OR "multi rótulo" [Title/ Abstract] OR "entidade estruturada" [Title/ Abstract] OR "entidades estruturadas" [Title/ Abstract] OR "entidade nomeada estruturada" [Title/ Abstract] OR "entidaded nomeadas estruturadas" [Title/ Abstract] OR "entidade irregular" [Title/ Abstract] OR "entidades irregulares" [Title/ Abstract] OR "entidade encadeada" [Title/ Abstract] OR "entidades encadeadas" [Title/ Abstract]) AND ("entidade nomeada" [Title/ Abstract] OR "entidades nomeadas" [Title/ Abstract] OR "reconhecimento de entidade" [Title/ Abstract])

10.1.2 ACM Digital Library

English

(Abstract: (ner) OR Abstract: ("named entity") OR Abstract: ("named entities") OR Abstract: ("entity recognition")) AND (Abstract: (nested) OR Abstract: ("complex entity") OR Abstract: ("complex entities") OR Abstract: ("complex named entity") OR Abstract: ("complex named entities") OR Abstract: ("multi type entity") OR Abstract: ("multi type entities") OR Abstract: ("overlapping") OR Abstract: ("overlap") OR Abstract: ("overlapped") OR Abstract: ("discontinuous") OR Abstract: ("discontiguous") OR Abstract: ("multilabel entity") OR Abstract: ("multi-label entity") OR Abstract: ("structured entity") OR Abstract: ("structured entities") OR Abstract: ("structured named entity") OR Abstract: ("structured named entities") OR Abstract: ("irregular entities") OR Abstract: ("irregular entity") OR Abstract: ("cascaded entities") OR Abstract: ("cascaded entity"))

Portuguese

(Abstract: (ren) OR Abstract: ("entidade nomeada") OR Abstract: ("entidades nomeadas") OR Abstract: ("reconhecimento de entidade")) AND (Abstract: (aninhada) OR Abstract: ("entidade complexa") OR Abstract: ("entidades complexas") OR

Abstract: ("entidade nomeada complexa") OR Abstract: ("entidades nomeadas complexas") OR Abstract: ("entidade multitipo") OR Abstract: ("entidades multitipo") OR Abstract: ("sobreposta") OR Abstract: ("sobrepostas") OR Abstract: ("sobrepor") OR Abstract: ("descontinuas") OR Abstract: ("descontinuadas") OR Abstract: ("multirrótulo") OR Abstract: ("multi rótulo") OR Abstract: ("entidade estruturada") OR Abstract: ("entidades estruturadas") OR Abstract: ("entidade nomeada estruturada") OR Abstract: ("entidades nomeadas estruturadas") OR Abstract: ("entidade irregular") OR Abstract: ("entidades irregulares") OR Abstract: ("entidade encadeada") OR Abstract: ("entidades encadeadas"))

10.1.3 Science Direct

English

(ner OR "named entity" OR "named entities" OR "entity recognition") AND (nested OR "complex entity" OR "complex entities" OR "complex named entity" OR "complex named entities") + (ner OR "named entity" OR "named entities" OR "entity recognition") AND ("multi type entity" OR "multi type entities" OR "overlapping" OR "overlap" OR "overlapped") + (ner OR "named entity" OR "named entities" OR "entity recognition") AND ("discontinuous" OR "discontiguous" OR "multilabel" OR "multi-label" OR "structured entity") + (ner OR "named entity" OR "named entities" OR "entity recognition") AND ("structured entities" OR "structured named entity" OR "structured named entities" OR "irregular entities" OR "irregular entity") + (ner OR "named entity" OR "named entities" OR "entity recognition") AND ("cascaded entities" OR "cascaded entity")

Portuguese

(ren AND "entidade nomeada" OR "entidades nomeadas" OR "reconhecimento de entidade") AND ("aninhadas" OR "entidade complexa" OR "entidades complexas" OR "entidade nomeada complexa") + (ren AND "entidade nomeada" OR "entidades nomeadas" OR "reconhecimento de entidade") AND ("entidades nomeadas complexas" OR "entidade multitipo" OR "entidades multitipo") + (ren AND "entidade nomeada" OR "entidades nomeadas" OR "reconhecimento de entidade") AND ("sobrepostas" OR "descontinuas" OR "estruturada" OR "entidade irregular") + (ren AND "entidade nomeada" OR "entidades nomeadas" OR "reconhecimento de entidade") AND ("entidades irregulares" OR "entidade encadeada" OR "entidades encadeadas")

10.1.4 Springer Link

English

("named entity" OR "named entities" OR "entity recognition") AND ("nested entity" OR "nested entities" OR "complex entity" OR "complex entities" OR "complex named entity" OR "complex named entities" OR "multi type entity" OR "multi type entities" OR "entity overlapping" OR "entity overlap" OR "entity overlapped" OR "discontinuous entity" OR "discontiguous entity" OR "structured entity" OR "structured entities" OR "structured named entity" OR "structured named entities" OR "irregular entities" OR "irregular entity" OR "cascaded entities" OR "cascaded entity")

Portuguese

("entidade nomeada" OR "entidades nomeadas" OR "reconhecimento de entidade") AND ("aninhada" OR "aninhadas" OR "entidade complexa" OR "entidades complexas" OR "entidade nomeada complexa" OR "entidades nomeadas complexas" OR "entidade multitipo" OR "entidades multitipo" OR "sobreposta" OR "sobrepostas" OR "sobrepor" OR "descontinuas" OR "descontinuadas" OR "entidade estruturada" OR "entidades estruturadas" OR "entidade nomeada estruturada" OR "entidades nomeadas estruturadas" OR "entidade irregular" OR "entidades irregulares" OR "entidade encadeada" OR "entidades encadeadas")

10.1.5 IEEE Xplore

English

("Abstract":ner OR "Abstract": "entities recognition" OR "Abstract":"entity recognition" OR "Abstract":"named entity" OR "Abstract":"named entities") AND (("Abstract":"nested entity") OR ("Abstract":"complex entity") OR ("Abstract":"multi type entity") OR ("Abstract":"overlap entity") OR ("Abstract":"discontinuous entity") OR ("Abstract":"discontiguous entity") OR ("Abstract":"structured entity") OR ("Abstract":"irregular entity") OR ("Abstract":"cascaded entity") OR ("Abstract":"nested entities") OR ("Abstract":"complex entities") OR ("Abstract":"multi type entities") OR ("Abstract":"overlap entities") OR ("Abstract":"discontinuous entities") OR ("Abstract":"discontiguous entities") OR ("Abstract":"structured entities") OR ("Abstract":"irregular entities") OR ("Abstract":"cascaded entities"))

Portuguese

("Abstract":ren OR "Abstract": "reconhecimento de entidade" OR "Abstract": "reconhecimento de entidades" OR "Abstract": "entidade nomeada" OR "Abstract": "entidades nomeadas") AND (("Abstract":aninhada) OR ("Abstract":complexa) OR ("Abstract": "multitipo") OR ("Abstract":sobreposta) OR ("Abstract":descontinua) OR ("Abstract":descontinuada) OR ("Abstract":estruturada) OR ("Abstract":irregular) OR ("Abstract":encadeada))

10.1.6 ACL Anthology

English

("nested entity") OR ("complex entity") OR ("overlap entity") OR ("discontinuous entity") OR ("discontiguous entity") or ("structured entity") OR ("irregular entity") OR ("cascaded entity") OR ("multitype entity") OR ("nested entities") OR ("complex entities") OR ("overlap entities") OR ("discontinuous entities") OR ("discontiguous entities") OR ("structured entities") OR ("irregular entities") OR ("cascaded entities") OR ("multitype entities") AND (ner OR "entity recognition")

Portuguese

entidade AND (aninhada OR "entidade complexa" OR multitipo OR sobreposta OR descontinua OR estruturada OR irregular OR encadeada)

10.2 Example of Samples in the QA-based Format

We provide examples of sentences in the QA-based input format (in JSON), which will be used by the model.

Nested Entities

We present a sentence with nested entities, from the NestedClinBr corpus. The same sentence is sent and processed four times, once for each entity type. In this example, we have four entity types: Problem, Test, Treatment, and Anatomy.

```
{"qid":"1001",
"text":"Dispneia importante aos esforcos + dor tipo peso no peito no
   esforco .",
"question":"Problema",
"answer":[ "Dispneia importante aos esforcos","dor tipo peso no peito
    no esforco"]}
{"qid":"1002",
"text":"Dispneia importante aos esforcos + dor tipo peso no peito no
   esforco .",
"question":"Teste",
"answer":[]}
{"qid":"1003",
"text":"Dispneia importante aos esforcos + dor tipo peso no peito no
   esforco .",
"question":"Tratamento",
"answer":[]}
{"qid":"1004",
"text":"Dispneia importante aos esforcos + dor tipo peso no peito no
   esforco .",
"question":"Anatomia",
"answer":[ "peito"]}
```

Discontinuous Entities

We present a sentence with nested and discontinuous entities, from the NestedClinBr corpus, to be processed by the model adapted for the recognition of discontinuous entities. The same sentence is sent and processed four times, once for each entity type. In this example, we have four entity types: Problem, Test, Treatment, and Anatomy.

```
{"qid":"182",
"text":"O # BEG , LOTE , CORADO , HIDRATADO , EUPNEICO , AFEBRIL , PA
    150 / 80 , FC 74 , CP : SP , ACV : BCRNF 2T SS , AP : MV +
   REDUZIDO DIFUSAMENTE , SEM RA , ABD : SP , MMII : PULSOS REDUZIDOS
    BILAT , SEM EDEMA OU EMPASTAMENTO , LAB 13 / 01 / 14 : GLICOSE
   304 ; GLICOSE P S - PRANDIAL 309 ; CT 119 ; HDL 21 ; TG 214 ; TGO
    19 ; HBA1C 6 , 70 ; CPK 72 ; CR 1 , 00 ; K 4 , 7 ; UR 30 ;
   MICROALBUMINURIA 24 HS ( 2114 MG ) ; PU ( GLICOSE + + ; LEUC 2 )
   .",
"question":"Problema",
"answer":[ "AFEBRIL","SS","MV + REDUZIDO DIFUSAMENTE","RA","MMII
   PULSOS REDUZIDOS BILAT","MMII EDEMA","MMII EMPASTAMENTO"],
"answer_indices":[ [12],[30],[34, 35, 36, 37],[40],[46, 48, 49, 50],[
   46, 53],[46, 55]]}
{"qid":"183",
"text":"O # BEG , LOTE , CORADO , HIDRATADO , EUPNEICO , AFEBRIL , PA
    150 / 80 , FC 74 , CP : SP , ACV : BCRNF 2T SS , AP : MV +
   REDUZIDO DIFUSAMENTE , SEM RA , ABD : SP , MMII : PULSOS REDUZIDOS
    BILAT , SEM EDEMA OU EMPASTAMENTO , LAB 13 / 01 / 14 : GLICOSE
   304 ; GLICOSE P S - PRANDIAL 309 ; CT 119 ; HDL 21 ; TG 214 ; TGO
    19 ; HBA1C 6 , 70 ; CPK 72 ; CR 1 , 00 ; K 4 , 7 ; UR 30 ;
   MICROALBUMINURIA 24 HS ( 2114 MG ) ; PU ( GLICOSE + + ; LEUC 2 )
   .",
"question":"Teste",
"answer":[ "PA","FC","LAB","GLICOSE","GLICOSE P S - PRANDIAL","CT","
   HDL","TG","TGO","HBA1C","CPK","CR","K","UR","MICROALBUMINURIA 24
   HS","PU","GLICOSE","LEUC"],
"answer_indices":[ [14],[19],[57],[64],[67, 68, 69, 70],[73],[76],[79
   ],[82],[85],[90],[93],[98],[103],[106, 107, 108],[114],[116],[120
   ]]}
{"qid":"184",
"text":"O # BEG , LOTE , CORADO , HIDRATADO , EUPNEICO , AFEBRIL , PA
    150 / 80 , FC 74 , CP : SP , ACV : BCRNF 2T SS , AP : MV +
   REDUZIDO DIFUSAMENTE , SEM RA , ABD : SP , MMII : PULSOS REDUZIDOS
    BILAT , SEM EDEMA OU EMPASTAMENTO , LAB 13 / 01 / 14 : GLICOSE
   304 ; GLICOSE P S - PRANDIAL 309 ; CT 119 ; HDL 21 ; TG 214 ; TGO
    19 ; HBA1C 6 , 70 ; CPK 72 ; CR 1 , 00 ; K 4 , 7 ; UR 30 ;
   MICROALBUMINURIA 24 HS ( 2114 MG ) ; PU ( GLICOSE + + ; LEUC 2 )
   .",
"question":"Tratamento",
"answer":[],
"answer_indices":[]}
{"qid":"185",
"text":"O # BEG , LOTE , CORADO , HIDRATADO , EUPNEICO , AFEBRIL , PA
    150 / 80 , FC 74 , CP : SP , ACV : BCRNF 2T SS , AP : MV +
   REDUZIDO DIFUSAMENTE , SEM RA , ABD : SP , MMII : PULSOS REDUZIDOS
    BILAT , SEM EDEMA OU EMPASTAMENTO , LAB 13 / 01 / 14 : GLICOSE
   304 ; GLICOSE P S - PRANDIAL 309 ; CT 119 ; HDL 21 ; TG 214 ; TGO
    19 ; HBA1C 6 , 70 ; CPK 72 ; CR 1 , 00 ; K 4 , 7 ; UR 30 ;
   MICROALBUMINURIA 24 HS ( 2114 MG ) ; PU ( GLICOSE + + ; LEUC 2 )
   .",
"question":"Anatomia",
"answer":[ "CP","ACV","ABD","MMII"],
"answer_indices":[ [22],[26],[42],[46]]}
```

10.3 Complete Results in NestedClinBr Corpus

10.3.1 Metrics per Entity

In table 10.1, we report the results of the BioNestedNER (with the best configuration) method in the NestedClinBr corpus by entity type, in addition to the micro, macro and weighted metrics.

Table 10.1: Results in the NestedClinBr corpus.

	Precision	Recall	F1	Support
Metrics per entity				
Problem	0.8549	0.8195	0.8369	338
Test	0.9087	0.8975	0.9031	244
Treatment	0.8976	0.8598	0.8783	214
Anatomy	**0.9312**	**0.8980**	**0.9143**	196
Avg metrics				
Micro avg	0.8926	0.8629	0.8775	992
Macro avg	0.8981	0.8687	0.8831	992
Weighted avg	0.8924	0.8629	0.8774	992

10.3.2 CRF results

In table 10.2, we report the results of all CRF models trained for NestedClinBr corpus, with different configurations: number of generated clusters, Word2vec model used to generate the clusters, generic Portuguese (NILC, 2023) vs clinical (ours), with UMLS concept, and better threshold setting.

Table 10.2: Results of CRF models in NestedClinBr corpus

Model	Recall	Precision	F1-score	Acc NE
With generic Word2vec				
cluster-5 (threshold-0.3)	0.6552	0.7738	0.7096	0.2963
cluster-10 (threshold-0.35)	0.6492	0.7970	0.7156	0.2593
cluster-50 (threshold-0.35)	0.6603	0.8017	0.7242	0.2037
cluster-100 (threshold-0.35)	0.6623	0.8002	0.7248	0.2130
cluster-300 (threshold-0.35)	0.6704	0.8022	0.7304	0.2315
With clinical Word2vec				
cluster-5 (threshold-0.3)	0.6936	0.8019	0.7438	0.2870
cluster-10 (threshold-0.30)	0.6895	0.8172	0.7480	0.2685
cluster-50 (threshold-0.25)	0.6976	0.8103	0.7497	0.2963
cluster-100 (threshold-0.35)	0.6835	0.8188	0.7451	0.2778
cluster-300 (threshold-0.35)	0.6875	0.8287	0.7515	0.2593
With clinical Word2vec and UMLS concepts				
cluster-5 (threshold-0.3)	0.6936	0.8019	0.7438	0.2870
cluster-10 (threshold-0.4)	0.6966	0.8350	0.7577	0.2593
cluster-50 (threshold-0.3)	0.7026	0.8200	0.7568	**0.3056**
cluster-100 (threshold-0.35)	**0.7033**	**0.8357**	**0.7638**	0.2870
cluster-300 (threshold-0.4)	0.6855	0.8354	0.7530	0.2037

10.4 Complete Results in SemClinBr Corpus

10.4.1 Metrics per Entity

In table 10.3, we report the results of the BioNestedNER (with the best configuration) method in the SemClinBr corpus by entity type, in addition to the micro, macro, and weighted metrics.

Table 10.3: Results in the SemClinBr corpus.

	Precision	Recall	F1	Support
Metrics per entity				
Procedures	0.5803	0.790188	0.6692	1916
Disorders	0.6821	0.8843	0.7701	3173
ChemicalsDrugs	**0.8919**	**0.9390**	**0.9149**	984
Abbreviation	0.7553	0.9074	0.8244	3575
Avg metrics				
Micro avg	0.7042	0.8798	0.7822	9648
Macro avg	0.7274	0.8802	0.7946	9648
Weighted avg	0.7104	0.8798	0.7850	9648

10.4.2 CRF results

In table 10.4, we report the results of all CRF models trained for SemClinBr corpus, with different configurations: number of generated clusters, Word2vec model used to generate the clusters, generic Portuguese (NILC, 2023) vs clinical (ours), with UMLS concept, and better threshold setting.

Table 10.4: Results of CRF models in SemClinBr corpus

Model	Recall	Precision	F1-score	Acc ME
With generic Word2vec				
cluster-5 (threshold-0.15)	0.4894	0.6421	0.5554	0.2895
cluster-10 (threshold-0.2)	0.4742	0.6582	0.5513	0.2371
cluster-50 (threshold-0.15)	0.4944	0.6328	0.5551	0.3011
cluster-100 (threshold-0.2)	0.4782	0.6590	0.5542	0.2567
cluster-300 (threshold-0.15)	0.4896	0.6385	0.5502	0.2815
With clinical Word2vec				
cluster-5 (threshold-0.2)	0.4877	0.6697	0.5644	0.2513
cluster-10 (threshold-0.15)	0.5019	0.6477	0.5655	0.3108
cluster-50 (threshold-0.15)	**0.5141**	0.6545	**0.5759**	0.3126
cluster-100 (threshold-0.15)	0.5084	0.6555	0.5727	**0.3171**
cluster-300 (threshold-0.15)	0.5058	0.6523	0.5698	0.3064
With clinical Word2vec and UMLS concepts				
cluster-5 (threshold-0.2)	0.4632	**0.6806**	0.5644	0.2513
cluster-10 (threshold-0.15)	0.5019	0.6477	0.5655	0.3108
cluster-50 (threshold-0.15)	**0.5141**	0.6545	**0.5759**	0.3126
cluster-100 (threshold-0.15)	0.5084	0.6555	0.5727	**0.3171**
cluster-300 (threshold-0.15)	0.5058	0.6524	0.5698	0.3064

10.5 Complete Results in GENIA Corpus

10.5.1 Metrics per Entity

In table 10.5, we report the results of the BioNestedNER (with the best configuration) method in the GENIA corpus by entity type, in addition to the micro, macro, and weighted metrics.

Table 10.5: Results in the GENIA corpus.

	Precision	Recall	F1	Support
Metrics per entity				
DNA	0.7680	0.7371	0.7523	1244
RNA	**0.8942**	**0.8532**	**0.8732**	109
Cell line	0.8353	0.6384	0.7237	437
Cell type	0.7594	0.7368	0.7479	604
Protein	0.8409	0.8033	0.8216	3065
Avg metrics				
Micro avg	0.8154	0.7686	0.7913	5459
Macro avg	0.8196	0.7538	0.7838	5459
Weighted avg	0.8159	0.7686	0.7909	5459

10.6 Complete Results in Rare Disease Corpus

10.6.1 Metrics per Entity

In table 10.6, we report the results of the BioNestedNER (with the best configuration) method in the Rare Disease corpus by entity type, in addition to the micro, macro and weighted metrics.

Table 10.6: Results in the Rare Disease corpus.

	Precision	Recall	F1	Support
Metrics per entity				
Rare disease	**0.8292**	**0.9023**	**0.8642**	522
Disease	0.6627	0.6111	0.6358	180
Symptom	0.4865	0.6207	0.5455	29
Sign	0.6583	0.6216	0.6394	592
Avg metrics				
Micro avg	0.7271	0.7309	0.7290	1323
Macro avg	0.6592	0.6889	0.6712	1323
Weighted avg	0.7226	0.7309	0.7256	1323

10.7 Complete Results in JNLPBA Corpus

10.7.1 Metrics per Entity

In table 10.7, we report the results of the BioNestedNER (with the best configuration) method in the JNLPBA corpus by entity type, in addition to the micro, macro, and weighted metrics.

Table 10.7: Results in the JNLPBA corpus.

	Precision	Recall	F1	Support
Metrics per entity				
DNA	0.7133	0.7871	0.7484	958
RNA	0.6528	0.8174	0.7259	115
Cell type	0.7454	0.7230	0.7340	1668
Cell line	0.5527	0.7302	0.6291	467
Protein	**0.7440**	**0.8650**	**0.8000**	4489
Avg metrics				
Micro avg	0.7254	0.8157	0.7679	7697
Macro avg	0.6816	0.7845	0.7275	7697
Weighted avg	0.7275	0.8156	0.7678	7697

10.8 Complete Results in PortugueseClinicalNER Corpus

10.8.1 Metrics per Entity

In table 10.8, we report the results of the BioNestedNER (with the best configuration) method in the PortugueseClinicalNER corpus by entity type, in addition to the micro, macro, and weighted metrics.

Table 10.8: Results in the PortugueseClinicalNER corpus.

	Precision	Recall	F1	Support
Metrics per entity				
Caracterização	0.8623	0.8095	0.8351	147
Teste	0.9482	0.9794	0.9636	243
Evolução	0.9770	0.9341	0.9551	91
Genética	**1.0000**	**1.0000**	**1.0000**	8
Anatomia	0.9679	0.9679	0.9679	343
Negação	0.9790	1.0000	0.9894	93
Observações	0.9667	0.7838	0.8657	37
Condição	0.9390	0.9550	0.9469	467
Resultados	0.9582	0.9347	0.9463	245
Data	0.9817	0.9699	0.9758	166
Terapêutica	0.9873	0.8864	0.9341	88
Valor	0.9630	0.9630	0.9630	54
Via	**1.0000**	**1.0000**	**1.0000**	6
Avg metrics				
Micro avg	0.9528	0.9437	0.9482	1988
Macro avg	0.9639	0.9372	0.9494	1988
Weighted avg	0.9527	0.9437	0.9477	1988

10.8.2 CRF RESULTS

In table 10.9, we report the results of all CRF models trained for PortugueseClinicalNER corpus, with different configurations: number of generated clusters, Word2vec model used to generate the clusters, generic Portuguese (NILC, 2023) vs clinical (ours), with UMLS concept, and better threshold setting.

Table 10.9: Results of CRF models in PortugueseClinicalNER corpus

Model	Recall	Precision	F1-score
With generic Word2vec			
cluster-5 (threshold-0.25)	0.8644	0.7764	0.8180
cluster-10 (threshold-0.3)	0.8559	0.7839	0.8183
cluster-50 (threshold-0.35)	0.8545	0.7835	0.8174
cluster-100 (threshold-0.35)	0.8550	0.7825	0.8171
cluster-300 (threshold-0.35)	0.8539	0.7865	0.8188
With clinical Word2vec			
cluster-5 (threshold-0.3)	0.8540	0.7855	0.8183
cluster-10 (threshold-0.3)	0.8458	0.7853	0.8145
cluster-50 (threshold-0.25)	**0.8609**	0.7849	0.8212
cluster-100 (threshold-0.2)	0.8438	0.7841	0.8128
cluster-300 (threshold-0.3)	0.8565	0.7857	0.8196
With clinical Word2vec and UMLS concepts			
cluster-5 (threshold-0.4)	0.8515	**0.7909**	0.8200
cluster-10 (threshold-0.35)	0.8458	0.7853	0.8145
cluster-50 (threshold-0.35)	**0.8609**	0.7849	0.8212
cluster-100 (threshold-0.35)	0.8438	0.7841	**0.8438**
cluster-300 (threshold-0.35)	0.8565	0.7857	0.8196

References

ALEX, B.; HADDOW, B.; GROVER, C. Recognising nested named entities in biomedical text. In: *Biological, translational, and clinical language processing*. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 65–72. Disponível em: <https://aclanthology.org/W07-1009>.

ALSENTZER, E.; MURPHY, J.; BOAG, W.; WENG, W.-H.; JINDI, D.; NAUMANN, T.; MCDERMOTT, M. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 72–78. Disponível em: <https://aclanthology.org/W19-1909>.

AMOR, M. B.; GRANITZER, M.; MITROVIć, J. *Impact of Position Bias on Language Models in Token Classification*. 2023.

ARTEMOVA, E.; ZMEEV, M.; LOUKACHEVITCH, N. V.; ROZHKOV, I.; BATURA, T.; IVANOV, V.; TUTUBALINA, E. Runne-2022 shared task: Recognizing nested named entities. *ArXiv*, abs/2205.11159, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:248987620>.

BÁEZ, P.; VILLENA, F.; ROJAS, M.; DURÁN, M.; DUNSTAN, J. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, 2020. p. 291–300. Disponível em: <https://aclanthology.org/2020.clinicalnlp-1.32>.

BANERJEE, P.; PAL, K. K.; DEVARAKONDA, M.; BARAL, C. Biomedical named entity recognition via knowledge guidance and question answering. *ACM Trans. Comput. Healthcare*, Association for Computing Machinery, New York, NY, USA, v. 2, n. 4, jul 2021. ISSN 2691-1957. Disponível em: <https://doi.org/10.1145/3465221>.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JANVIN, C. A neural probabilistic language model. *J. Mach. Learn. Res.*, JMLR.org, v. 3, n. null, p. 1137–1155, mar 2003. ISSN 1532-4435.

BENIKOVA CHRIS BIEMANN, M. K. D.; PADO., S. Germeval 2014 named entity recognition shared task: Companion paper. In: *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*. [S.l.: s.n.], 2014.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, 07 2016.

BOJAR, O.; CHATTERJEE, R.; FEDERMANN, C.; GRAHAM, Y.; HADDOW, B.; HUCK, M.; YEPES, A. J.; KOEHN, P.; LOGACHEVA, V.; MONZ, C.; NEGRI, M.; NÉVÉOL, A.; NEVES, M.; POPEL, M.; POST, M.; RUBINO, R.; SCARTON, C.; SPECIA, L.; TURCHI, M.; VERSPOOR, K.; ZAMPIERI, M. Findings of the 2016 conference on machine translation. In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 131–198. Disponível em: <https://aclanthology.org/W16-2301>.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>.

BYRNE, K. Nested named entity recognition in historical archive text. In: *International Conference on Semantic Computing (ICSC 2007)*. [S.l.: s.n.], 2007. p. 589–596.

CAMPILLOS-LLANOS L., V.-M. A. C.-C. A. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine.s. *BMC Med Inform Decis Mak 21, 69*, 2021.

CAMPOS, D.; MATOS, S.; OLIVEIRA, J. Gimli: Open source and high-performance biomedical name recognition. *BMC bioinformatics*, v. 14, p. 54, 02 2013.

CAMPOS, D.; MATOS, S.; OLIVEIRA, J. L. Biomedical named entity recognition: A survey of machine-learning tools. In: SAKURAI, S. (Ed.). *Theory and Applications for Advanced Text Mining*. Rijeka: IntechOpen, 2012. cap. 8. Disponível em: <https://doi.org/10.5772/51066>.

CHAN, S.-K.; LAM, W.; YU, X. An online cascaded approach to biomedical named entity recognition. 01 2008.

CHEN, Y.; HU, Y.; LI, Y.; HUANG, R.; QIN, Y.; WU, Y.; ZHENG, Q.; CHEN, P. A boundary assembling method for nested biomedical named entity recognition. *IEEE Access*, v. 8, p. 214141–214152, 2020.

CHINCHOR, N.; SUNDHEIM, B. MUC-5 evaluation metrics. In: *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*. [s.n.], 1993. Disponível em: <https://aclanthology.org/M93-1007>.

CHIU, B.; CRICHTON, G.; KORHONEN, A.; PYYSALO, S. How to train good word embeddings for biomedical NLP. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 166–174. Disponível em: <https://aclanthology.org/W16-2922>.

COLLIER, N.; KIM, J.-D. Introduction to the bio-entity recognition task at JNLPBA. In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*. Geneva, Switzerland: COLING, 2004. p. 73–78. Disponível em: <https://aclanthology.org/W04-1213>.

CUI, S.; JOE, I. A multi-head adjacent attention-based pyramid layered model for nested named entity recognition. *Neural Comput. Appl.*, Springer-Verlag, Berlin, Heidelberg, v. 35, n. 3, p. 2561–2574, sep 2022. ISSN 0941-0643. Disponível em: <https://doi.org/10.1007/s00521-022-07747-8>.

DAI, H.; LAI, P.-T.; CHANG, Y.-C.; TSAI, R. T.-H. Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *Journal of cheminformatics*, v. 7, p. S14, 01 2015.

DAI, X. Recognizing complex entity mentions: A review and future directions. In: *Proceedings of ACL 2018, Student Research Workshop*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 37–44. Disponível em: <https://aclanthology.org/P18-3006>.

DALIANIS, H. Clinical text mining: Secondary use of electronic patient records. *Springer, Cham*, 2018.

DELEGER, L.; LI, Q.; LINGREN, T.; KAISER, M.; MOLNAR, K.; STOUTENBOROUGH, L.; KOURIL, M.; MARSOLO, K.; SOLTI, I. et al. Building gold standard corpora for medical natural language processing tasks. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2012. v. 2012, p. 144.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, n. Jan, p. 1–30, 2006.

DEVLIN, J.; CHANG, M.-W.; 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. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>.

DOGAN, R.; LEAMAN, R.; LU, Z. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, v. 47, 01 2014.

DU, X.; YUXIANG, J.; HONGYING, Z. MRC-based medical NER with multi-task learning and multi-strategies. In: *Proceedings of the 21st Chinese National Conference on Computational Linguistics*. Nanchang, China: Chinese Information Processing Society of China, 2022. p. 836–847. Disponível em: <https://aclanthology.org/2022.ccl-1.74>.

DUDCHENKO, A.; GANZINGER, M.; KOPANITSA, G. Machine learning algorithms in cardiology domain: A systematic review. *The Open Bioinformatics Journal*, v. 13, p. 25–40, 04 2020.

FEI, H.; REN, Y.; JI, D. Dispatched attention with multi-task learning for nested mention recognition. *Information Sciences*, v. 513, p. 241–251, 2020. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025519310333>.

FINKEL, J. R.; MANNING, C. D. Nested named entity recognition. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2009. p. 141–150. Disponível em: <https://aclanthology.org/D09-1015>.

FISHER, J.; VLACHOS, A. Merge and label: A novel neural network architecture for nested NER. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 5840–5850. Disponível em: <https://aclanthology.org/P19-1585>.

FREITAS, C.; MOTA, C.; SANTOS, D.; OLIVEIRA, H. G.; CARVALHO, P. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf>.

FU1 CHUANQI TAN, M. C. S. H. F. H. Y. Nested named entity recognition with partially-observed treecrfs. In: . Online: [s.n.], 2020.

GRANCHAROVA, M.; BERG, H.; DALIANIS, H. Improving named entity recognition and classification in class imbalanced swedish electronic patient records through resampling. In: . [S.l.]: Eighth Swedish Language Technology Conference (SLTC 2020), 2020.

GU, B. Recognizing nested named entities in GENIA corpus. In: *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*. New York, New York: Association for Computational Linguistics, 2006. p. 112–113. Disponível em: <https://aclanthology.org/W06-3318>.

GU, Y.; TINN, R.; CHENG, H.; LUCAS, M.; USUYAMA, N.; LIU, X.; NAUMANN, T.; GAO, J.; POON, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, Association for Computing Machinery, New York, NY, USA, v. 3, n. 1, oct 2021. ISSN 2691-1957. Disponível em: <https://doi.org/10.1145/3458754>.

GUMIEL, Y. B.; OLIVEIRA, L. E.; SOUZA, J. V. de; SCHNEIDER, E. T.; FURLAN, L. H.; PARAISO, E. C.; MORO, C.; CARVALHO, D. R. Novel annotation schema for improved temporal reasoning over cardiology notes. 2023. Disponível em: <https://github.com/HAILab-PUCPR/TempClinBr>.

GURULINGAPPA, H.; MATEEN, A.; ROBERTS, A.; FLUCK, J.; HOFMANN-APITIUS, M.; TOLDO, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, http://dx.doi.org/10.1016/j.jbi.2012.04.008, 04 2012.

HERRERO-ZAZO, M.; SEGURA-BEDMAR, I.; MARTíNEZ, P.; DECLERCK, T. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, v. 46, n. 5, p. 914–920, 2013. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046413001123>.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 328–339. Disponível em: <https://aclanthology.org/P18-1031>.

JU, M.; MIWA, M.; ANANIADOU, S. A neural layered model for nested named entity recognition. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 1446–1459. Disponível em: <https://aclanthology.org/N18-1131>.

KATIYAR, A.; CARDIE, C. Nested named entity recognition revisited. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 861–871. Disponível em: <https://aclanthology.org/N18-1079>.

KIM, J.-D.; OHTA, T.; TATEISI, Y.; TSUJII, J. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, v. 19 Suppl 1, p. i180–2, 02 2003.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 282–289. ISBN 1558607781.

LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; SO, C. H.; KANG, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, v. 36, n. 4, p. 1234–1240, 09 2019. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btz682>.

LI, C.; WANG, G.; CAO, J.; CAI, Y. A multi-agent communication based model for nested named entity recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 29, p. 2123–2136, 2021.

LI, D.; SAVOVA, G.; KIPPER-SCHULER, K. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Columbus, Ohio: Association for Computational Linguistics, 2008. p. 94–95. Disponível em: <https://aclanthology.org/W08-0615>.

LI, F.; JIN, Y.; LIU, W.; RAWAT, B. P. S.; 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*, v. 7, n. 3, p. e14830, Sep 2019. ISSN 2291-9694. Disponível em: <http://medinform.jmir.org/2019/3/e14830/>.

LI, J.; SUN, A.; HAN, J.; LI, C. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, v. 34, n. 1, p. 50–70, 2022.

LI, R.; MO, T.; YANG, J.; LI, D.; JIANG, S.; WANG, D. Bridge inspection named entity recognition via bert and lexicon augmented machine reading comprehension neural model. *Adv. Eng. Inform.*, Elsevier Science Publishers B. V., NLD, v. 50, n. C, oct 2021. ISSN 1474-0346. Disponível em: <https://doi.org/10.1016/j.aei.2021.101416>.

LI, X.; FENG, J.; MENG, Y.; HAN, Q.; WU, F.; LI, J. A unified MRC framework for named entity recognition. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 5849–5859. Disponível em: <https://aclanthology.org/2020.acl-main.519>.

LIN, H.; LU, Y.; HAN, X.; SUN, L.; DONG, B.; JIANG, S. Gazetteer-enhanced attentive neural networks for named entity recognition. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 6232–6237. Disponível em: <https://aclanthology.org/D19-1646>.

LIN, H.; LU, Y.; HAN, X.; SUN, L. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 5182–5192. Disponível em: <https://aclanthology.org/P19-1511>.

LONG, X.; NIU, S.; LI, Y. Hierarchical region learning for nested named entity recognition. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 4788–4793. Disponível em: <https://aclanthology.org/2020.findings-emnlp.430>.

LOPES, F.; TEIXEIRA, C.; OLIVEIRA, H. G. Contributions to clinical named entity recognition in Portuguese. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, 2019. p. 223–233. Disponível em: <https://aclanthology.org/W19-5024>.

LOUKACHEVITCH, N.; MANANDHAR, S.; BARAL, E.; ROZHKOV, I.; BRASLAVSKI, P.; IVANOV, V.; BATURA, T.; TUTUBALINA, E. NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities. *Bioinformatics*, v. 39, n. 4, 04 2023. ISSN 1367-4811. Btad161. Disponível em: <https://doi.org/10.1093/bioinformatics/btad161>.

LU, W.; ROTH, D. Joint mention extraction and classification with mention hypergraphs. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 857–867. Disponível em: <https://aclanthology.org/D15-1102>.

LUAN, Y.; WADDEN, D.; HE, L.; SHAH, A.; OSTENDORF, M.; HAJISHIRZI, H. A general framework for information extraction using dynamic span graphs. 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. 3036–3046. Disponível em: <https://aclanthology.org/N19-1308>.

LUO, Y.; ZHAO, H. Bipartite flat-graph network for nested named entity recognition. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 6408–6418. Disponível em: <https://aclanthology.org/2020.acl-main.571>.

MA, X.; HOVY, E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1064–1074. Disponível em: <https://aclanthology.org/P16-1101>.

MADY, L.; AFIFY, y.; BADR, N. Nested biomedical named entity recognition. *International Journal of Intelligent Computing and Information Sciences*, Ain Shams University, Faculty of Computer and Information Science, v. 22, n. 1, p. 98–107, 2022. ISSN 1687-109X. Disponível em: <https://ijicis.journals.ekb.eg/article_219235.html>.

MALMASI, S.; FANG, A.; FETAHU, B.; KAR, S.; ROKHLENKO, O. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, 2022. p. 1412–1437. Disponível em: <https://aclanthology.org/2022.semeval-1.196>.

MANSOURI, A.; AFFENDEY, L.; MAMAT, A. Named entity recognition approaches. *Int J Comp Sci Netw Sec*, v. 8, 01 2008.

MARINHO, Z.; MENDES, A.; MIRANDA, S.; NOGUEIRA, D. Hierarchical nested named entity recognition. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 28–34. Disponível em: <https://aclanthology.org/W19-1904>.

MARTíNEZ-DEMIGUEL, C.; SEGURA-BEDMAR, I.; CHACóN-SOLANO, E.; GUERRERO-ASPIZUA, S. The raredis corpus: A corpus annotated with rare diseases, their signs and symptoms. *Journal of Biomedical Informatics*, v. 125, p. 103961, 2022. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046421002902>.

MCDONALD, R.; CRAMMER, K.; PEREIRA, F. Flexible text segmentation with structured multilabel classification. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005. p. 987–994. Disponível em: <https://aclanthology.org/H05-1124>.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, v. 2013, 01 2013.

MITCHELL ALEXIS, e. a. Ace 2004 multilingual training corpus. In: . [S.l.: s.n.], 2005.

MUIS, A. O.; LU, W. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In: *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 2608–2618. Disponível em: <https://aclanthology.org/D17-1276>.

NEVES, M.; ŠEVA, J. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, v. 22, n. 1, p. 146–163, 12 2019. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bbz130>.

NILC. Interinstitutional center for computational linguistics. 2023. Disponível em: <https://sites.google.com/view/nilc-usp/>.

OLIVEIRA, L. E. S. e; PETERS, A. C.; SILVA, A. M. P. da; GEBELUCA, C. P.; GUMIEL, Y. B.; CINTHO, L. M. M.; CARVALHO, D. R.; HASAN, S. A.; MORO, C. M. C. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, Springer Science and Business Media LLC, v. 13, n. 1, maio 2022. Disponível em: <https://doi.org/10.1186/s13326-022-00269-1>.

OPENAI. Introducing chatgpt - openai. 2023. Disponível em: <https://chat.openai.com/chat>.

OUCHI, H.; SUZUKI, J.; KOBAYASHI, S.; YOKOI, S.; KURIBAYASHI, T.; KONNO, R.; INUI, K. Instance-based learning of span representations: A case study through named entity recognition. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 6452–6459. Disponível em: <https://aclanthology.org/2020.acl-main.575>.

PANCHAPAGESAN, S.; SUN, M.; KHARE, A.; MATSOUKAS, S.; MANDAL, A.; HOFFMEISTER, B.; VITALADEVUNI, S. Multi-task learning and weighted cross-entropy for dnn-based keyword spotting. In: . [S.l.: s.n.], 2016. p. 760–764.

PATIL, N.; PATIL, A.; PAWAR, B. Named entity recognition using conditional random fields. *Procedia Computer Science*, v. 167, p. 1181–1188, 2020. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050920308978>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods* *in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162>.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 2227–2237. Disponível em: <https://aclanthology.org/N18-1202>.

PIROVANI, J.; OLIVEIRA, E. Portuguese named entity recognition using conditional random fields and local grammars. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1705>.

PONOMAREVA, N.; ROSSO, P.; PLA, F.; MARCO, A. M.; CAMINO; VALENCIA, I. F.; SPAIN. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. 09 2007.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019.

RAMSHAW, L.; MARCUS, M. Text chunking using transformation-based learning. In: *Third Workshop on Very Large Corpora*. [s.n.], 1995. Disponível em: <https://aclanthology.org/W95-0107>.

RINGLAND, N.; DAI, X.; HACHEY, B.; KARIMI, S.; PARIS, C.; CURRAN, J. Nne: A dataset for nested named entity recognition in english newswire. In: . [S.l.: s.n.], 2019.

ROJAS, M.; BRAVO-MARQUEZ, F.; DUNSTAN, J. Simple yet powerful: An overlooked architecture for nested named entity recognition. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. p. 2108–2117. Disponível em: <https://aclanthology.org/2022.coling-1.184>.

SANG, E. F. T. K. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. [s.n.], 2002. Disponível em: <https://www.aclweb.org/anthology/W02-2024>.

SANG, E. F. T. K.; MEULDER, F. D. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. [s.n.], 2003. p. 142–147. Disponível em: <https://aclanthology.org/W03-0419>.

SANG, E. F. T. K.; VEENSTRA, J. Representing text chunks. In: *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*. USA: Association for Computational Linguistics, 1999. (EACL '99), p. 173–179. Disponível em: <https://doi.org/10.3115/977035.977059>.

SCHNEIDER, E. T. R.; GUMIEL, Y. B.; OLIVEIRA, L. F. A. de; MONTENEGRO, C. de O.; BARZOTTO, L. R.; MORO, C.; PAGANO, A.; PARAISO, E. C. Developing a transformer-based clinical part-of-speech tagger for brazilian portuguese. In: *XIX Congresso Brasileiro de Informática em Saúde*. Online: [s.n.], 2022.

SCHNEIDER, E. T. R.; SOUZA, J. V. A. de; KNAFOU, J.; OLIVEIRA, L. E. S. e.; COPARA, J.; GUMIEL, Y. B.; OLIVEIRA, L. F. A. d.; PARAISO, E. C.; TEODORO, D.; BARRA, C. M. C. M. 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. Disponível em: <https://www.aclweb.org/anthology/2020.clinicalnlp-1.7>.

SCHULZ, S.; ŠEVA, J.; RODRIGUEZ, S.; OSTENDORFF, M.; REHM, G. Named entities in medical case reports: Corpus and experiments. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 4495–4500. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.553>.

SEGURA-BEDMAR, I.; CAMINO-PERDONAS, D.; GUERRERO-ASPIZUA, S. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. *BMC Bioinformatics*, v. 23, 2021.

SEGURA-BEDMAR, I.; MARTÍNEZ, P.; HERRERO-ZAZO, M. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, 2013. p. 341–350. Disponível em: <https://aclanthology.org/S13-2056>.

SHEN, D.; ZHANG, J.; ZHOU, G.; SU, J.; TAN, C.-L. Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain. In: *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. Sapporo, Japan: Association for Computational Linguistics, 2003. p. 49–56. Disponível em: <https://aclanthology.org/W03-1307>.

SHEN, Y.; MA, X.; TAN, Z.; ZHANG, S.; WANG, W.; LU, W. Locate and label: A two-stage identifier for nested named entity recognition. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

Papers). Online: Association for Computational Linguistics, 2021. p. 2782–2794. Disponível em: <https://aclanthology.org/2021.acl-long.216>.

SHEN, Y.; WANG, X.; TAN, Z.; XU, G.; XIE, P.; HUANG, F.; LU, W.; ZHUANG, Y. Parallel instance query network for named entity recognition. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 947–961. Disponível em: <https://aclanthology.org/2022.acl-long.67>.

SHESKIN, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*. 4. ed. [S.l.]: Chapman amp; Hall/CRC, 2007. ISBN 1584888148.

SHIBUYA, T.; HOVY, E. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, MIT Press, Cambridge, MA, v. 8, p. 605–620, 2020. Disponível em: <https://aclanthology.org/2020.tacl-1.39>.

SOHRAB, M. G.; MIWA, M. Deep exhaustive model for nested named entity recognition. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 2843–2849. Disponível em: <https://aclanthology.org/D18-1309>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019. Disponível em: <http://arxiv.org/abs/1909.10649>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8.

SOUZA, J. V. A. d.; SCHNEIDER, E. T. R.; CEZAR, J. O.; OLIVEIRA, L. E. S. e.; GUMIEL, Y. B.; PARAISO, E. C.; TEODORO, D.; BARRA, C. M. C. M. A multilabel approach to portuguese clinical named entity recognition. *Journal of Health Informatics*, v. 12, mar. 2021. Disponível em: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/840>.

SOUZA, J. V. de; GUMIEL, Y.; OLIVEIRA, L. E.; MORO, C. M. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In: *Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2019. p. 318–323. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/6269>.

STENETORP, P.; PYYSALO, S.; TOPIĆ, G.; OHTA, T.; ANANIADOU, S.; TSUJII, J. Brat: a web-based tool for nlp-assisted text annotation. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2012. p. 102–107.

STRAKOVÁ, J.; STRAKA, M.; HAJIC, J. Neural architectures for nested NER through linearization. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 5326–5331. Disponível em: <https://aclanthology.org/P19-1527>.

SUN, L.; SUN, Y.; JI, F.; WANG, C. Joint learning of token context and span feature for span-based nested ner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 28, p. 2720–2730, 2020.

TALAFHA, B. Ner shared task 2023 - subtask 2: Nested ner. In: . [s.n.], 2023. Disponível em: <https://dlnlp.ai/st/wojood/>.

TAN, C.; QIU, W.; CHEN, M.; WANG, R.; HUANG, F. Boundary enhanced neural span classification for nested named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 05, p. 9016–9023, Apr. 2020. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/6434>.

TSAI, R. T.-H.; SUNG, C.-L.; DAI, H.-J.; HUNG, H.-C.; SUNG, T.-Y.; HSU, W.-L. Nerbio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, v. 7, p. S11 – S11, 2006.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, v. 3, p. 1–13, 09 2009.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

W LI Y, G. X. C. S. Z. S. G. Research on named entity recognition based on multi-task learning and biaffine mechanism. *Comput Intell Neurosci.*, 2022.

WALKER CHRISTOPHER, e. a. Ace 2005 multilingual training corpus. In: *Philadelphia: Linguistic Data Consortium, 2006*. [S.l.: s.n.], 2006.

WANG, B.; LU, W. Neural segmental hypergraphs for overlapping mention recognition. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 204–214. Disponível em: <https://aclanthology.org/D18-1019>.

WANG, B.; LU, W.; WANG, Y.; JIN, H. A neural transition-based model for nested mention recognition. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 1011–1017. Disponível em: <https://aclanthology.org/D18-1124>.

WANG, J.; SHOU, L.; CHEN, K.; CHEN, G. Pyramid: A layered model for nested named entity recognition. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 5918–5928. Disponível em: <https://aclanthology.org/2020.acl-main.525>.

WANG, X.; ZHANG, Y.; REN, X.; ZHANG, Y.; ZITNIK, M.; SHANG, J.; LANGLOTZ, C.; HAN, J. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, Oxford University Press, v. 35, n. 10, p. 1745–1752, 2019.

WANG, Y.; HOU, Y.; CHE, W.; LIU, T. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, v. 11, p. 1611–1630, 2020.

WANG, Y.; LI, Y.; TONG, H.; ZHU, Z. HIT: Nested named entity recognition via head-tail pair and token interaction. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. p. 6027–6036. Disponível em: <https://aclanthology.org/2020.emnlp-main.486>.

WANG, Y.; SHINDO, H.; MATSUMOTO, Y.; WATANABE, T. Nested named entity recognition via explicitly excluding the influence of the best path. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021. p. 3547–3557. Disponível em: <https://aclanthology.org/2021.acl-long.275>.

WANG, Y.; TONG, H.; ZHU, Z.; LI, Y. Nested named entity recognition: A survey. *ACM Trans. Knowl. Discov. Data*, Association for Computing Machinery, New York, NY, USA, v. 16, n. 6, jul 2022. ISSN 1556-4681. Disponível em: <https://doi.org/10.1145/3522593>.

WEI, H.; GAO, M.; ZHOU, A.; CHEN, F.; QU, W.; WANG, C.; LU, M. Biomedical named entity recognition via a hybrid neural network model. In: *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. [S.l.: s.n.], 2019. p. 455–462.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. M. 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. Disponível em: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

XIA, C.; ZHANG, C.; YANG, T.; LI, Y.; DU, N.; WU, X.; FAN, W.; MA, F.; YU, P. Multi-grained named entity recognition. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 1430–1440. Disponível em: <https://aclanthology.org/P19-1138>.

XU, M.; JIANG, H.; WATCHARAWITTAYAKUL, S. A local detection approach for named entity recognition and mention detection. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 1237–1247. Disponível em: <https://aclanthology.org/P17-1114>.

YU, J.; BOHNET, B.; POESIO, M. Named entity recognition as dependency parsing. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 6470–6476. Disponível em: <https://aclanthology.org/2020.acl-main.577>.

ZHANG, H.; GUO, J.; WANG, Y.; ZHANG, Z.; ZHAO, H. Judicial nested named entity recognition method with mrc framework. *International Journal of Cognitive Computing in Engineering*, v. 4, p. 118–126, 2023. ISSN 2666-3074. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666307423000128>.

ZHANG, J.; SHEN, D.; ZHOU, G.; SU, J.; TAN, C.-L. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, v. 37, n. 6, p. 411–422, 2004. ISSN 1532-0464. Named Entity Recognition in Biomedicine. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046404000838>.

ZHANG, S.; ELHADAD, N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, v. 46, n. 6, p. 1088–1098, 2013. ISSN 1532-0464. Special Section: Social Media Environments. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046413001196>.

ZHANG, Y.; XU, G.; WANG, Y.; LIN, D.; LI, F.; WU, C.; ZHANG, J.; HUANG, T. A question answering-based framework for one-step event argument extraction. *IEEE Access*, v. 8, p. 65420–65431, 2020.

ZHENG, C.; CAI, Y.; XU, J.; LEUNG, H.-f.; XU, G. A boundary-aware neural model for nested named entity recognition. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 357–366. Disponível em: <https://aclanthology.org/D19-1034>.

ZHENG, C.; CAI, Y.; XU, J.; LEUNG, H.-f.; XU, G. A boundary-aware neural model for nested named entity recognition. In: *Proceedings of the 2019 Conference* *on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 357–366. Disponível em: <https://aclanthology.org/D19-1034>.

ZHOU, G. Recognizing names in biomedical texts using mutual information independence model and svm plus sigmoid. *International Journal of Medical Informatics*, v. 75, n. 6, p. 456–467, 2006. ISSN 1386-5056. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1386505605001218>.

ZHOU, G.; ZHANG, J.; JIAN, S.; SHEN, D.; TAN, C. L. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics (Oxford, England)*, v. 20, p. 1178–90, 06 2004.

ZHOU, Y.; LIU, L.; CHEN, Y.; HUANG, R.; QIN, Y.; LIN, C. A novel mrc framework for evidence extracts in judgment documents. *Artificial Intelligence and Law*, 01 2023.

ZHU, Q.; LI, X.; CONESA, A.; PEREIRA, C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, v. 34, n. 9, p. 1547–1554, 12 2017. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btx815>.

ZHUANG, L.; WAYNE, L.; YA, S.; JUN, Z. A robustly optimized BERT pre-training approach with post-training. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Huhhot, China: Chinese Information Processing Society of China, 2021. p. 1218–1227. Disponível em: <https://aclanthology.org/2021.ccl-1.108>.