Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5790

BioNestedNER: A Hybrid Language Model Approach for Recognizing Nested, Discontinuous, and Multi-Type Named Entities

Elisa Terumi Rubel Schneider [**Pontifícia Universidade Católica do Paraná**| ***lisa.terumi@gmail.com***]
Yohan Bonescki Gumiel [**Instituto do Coração - InCor/HC FMUSP**| ***yohan.gumiel@gmail.com***]
Paloma Martínez [**Computer Science and Engineering Department, Universidad Carlos III de Madrid**| ***pmf@inf.uc3m.es***]
Claudia Moro [**Pontifícia Universidade Católica do Paraná** | ***claudia.moro@gmail.com***]
Emerson Cabrera Paraiso [**Pontifícia Universidade Católica do Paraná** | ***paraiso@ppgia.pucpr.br***]

Graduate Program in Informatics (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR, Brazil.

Received: 30 March 2025 • **Accepted:** 29 September 2025 • **Published:** 05 April 2026

Abstract Named Entity Recognition (NER) is essential in Natural Language Processing (NLP) for extracting pertinent information from unstructured data. Traditional NER approaches assume continuous and non-overlapping entities, which can be limiting in real-world scenarios. This research introduces **BioNestedNER**, a hybrid method for nested, discontinuous, and multi-type entity recognition, with a focus on clinical and biomedical domains. Our approach employs a language model (encoder-only Transformer-based model) using a machine reading comprehension strategy, treating NER as a question-answering-like task. A Conditional Random Field also addresses multi-label sequence labeling for handling nested entities as multi-type entities. Evaluation in Portuguese demonstrated state-of-the-art performance in micro F1-Scores across two clinical corpora. In *NestedClinBr*, featuring nested and discontinuous entities, our method achieved an F1-Score of 0.863, surpassing the second-place result by 2.1%. In *SemClinBr*, with multi-type entities, an F1-Score of 0.782 was achieved, surpassing the second-place result by 11.5%. This paper also presents a new clinical corpus in Brazilian Portuguese annotated with nested and discontinuous entities, offering a valuable resource for developing and evaluating models handling these complex entities. In conclusion, BioNestedNER presents an adaptable and effective NER solution for nested, discontinuous, and multi-type entities, with the potential to benefit various clinical applications.

Keywords: Natural language processing, Named entity recognition, Language models, Clinical corpus, Transformer architecture, Machine learning, Question Answering

1 Introduction

Natural Language Processing (NLP) is a field that enables interactions between computers and humans using natural language. With the widespread adoption of Electronic Health Record (EHR) systems, NLP techniques have become essential in the medical domain, facilitating the extraction of valuable patient information from unstructured clinical narratives and enhancing decision-making in healthcare.

Named Entity Recognition (NER) is one of the most widely used NLP tasks, allowing machines to extract knowledge from unstructured text by identifying and classifying meaningful entities. In the clinical domain, NER plays a fundamental role in detecting medical concepts such as diseases, symptoms, and medications, serving as a foundation for downstream tasks such as clinical event prediction and entity relationship extraction. Traditionally, NER is formulated as follows: given a sequence of tokens (i.e., contiguous character sequences representing semantic units in text), the model returns a list of tuples $<Is, Ie, t>$, where *Is* and *Ie* denote the start and end indices of the entity mention, respectively, and *t* represents the entity type from a predefined category. This formulation is based on two key assumptions: (1) an entity mention consists of a continuous sequence of words within the interval *[Is, Ie]*, and (2) entity mentions do not overlap [Dai, 2018]. This conventional entity representation is known as a "flat entity".

Beyond flat entities, Dai [2018] defines "complex entities" as those that exhibit nested, incorporated, overlapped, discontinuous, and/or multi-type structures, deviating from the traditional NER assumptions. These entities are particularly prevalent in clinical and biomedical texts [Finkel and Manning, 2009] and hold valuable information for downstream tasks. For instance, biological entities often encompass one another, as seen in proteins, genes, and chemical substances [Wang and Lu, 2018; Alex *et al*., 2007; Lu and Roth, 2015]). This is illustrated in the *GENIA* corpus, as shown in **Figure 1** (a), where the *Protein* mention "OTF-2" is embedded within the broader *DNA* entity "OTF-2 expression vector". Discontinuous entities emerge when entity mentions are composed of non-sequential words in a text. **Figure 1** (b) shows a discontinuous case, where the entity *Cell line* is split across the tokens "HeLa" and "K562 cells". Another complex scenario involves multi-type entities, where a single mention is associated with multiple entity categories. **Figure 1** (c) provides a

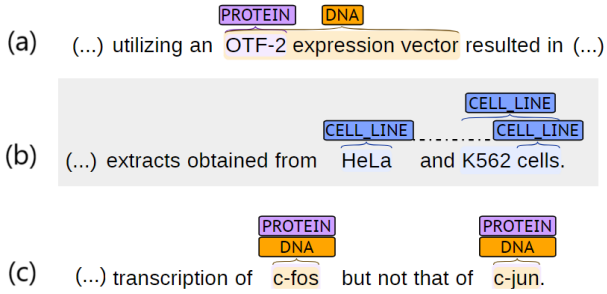

Figure 1. Examples of nested (a), discontinuous (b), and multi-type (c) entities from *GENIA* corpus.

example from *GENIA*, with the mentions "c-fos" and "c-jun" annotated as both *DNA* and *Protein*.

Despite their frequent occurrence in real-world clinical and biomedical texts, traditional NER models are not inherently designed to handle these complexities. Several approaches have been adapted to recognize complex entities [Alex *et al*., 2007; Finkel and Manning, 2009; Katiyar and Cardie, 2018; Lu and Roth, 2015; Rivera-Zavala and Martínez, 2021]. However, as highlighted by Dai [2018], Ji *et al*. [2025], and Alhassan *et al*. [2025], these methods face some challenges:

- **Limited expressivity**: Token-level approaches often rely on tagging schemes with inherent restrictions, limiting their ability to capture complex entity structures.
- **Computational complexity and scalability issues**: Span-based, graph-based, and generative methods often require costly enumeration, decoding, or autoregressive generation steps.
- **Exposure bias and error propagation**: Sequential and generative models are affected by reliance on gold spans during training and may suffer from cascading errors during inference, reducing generalization.
- **Structural robustness challenges**: Hypergraph and grid tagging approaches may produce ambiguous decodings or spurious relations, especially in the presence of overlapping and nested entities.

Although recent LLM approaches have started addressing complex entities, their adoption remains limited. While generative models are widely used for NLP tasks, they can present limitations in clinical NER, particularly when dealing with complex entity structures such as nested and discontinuous entities, as demonstrated by Naguib *et al*. [2024]. Their findings indicate that masked language models, when fine-tuned with limited data, tend to outperform generative models, offering superior precision and recall. Additionally, while generative models often incur higher computational costs, BERT-based models are more efficient, requiring less computational power, making them a more suitable choice for our approach.

Therefore, this paper introduces **BioNestedNER**, a hybrid method featuring two modules designed for the recognition of nested, discontinuous, and multi-type entities, with a focus on the clinical and biomedical domains. The first module leverages a Transformer-based [Vaswani *et al*., 2017] language model, drawing inspiration from Machine Reading Comprehension, and frames NER as a Question-Answering (QA) task, as seen in Zhang *et al*. [2020] and Banerjee *et al*. [2021]. The second module integrates a Conditional Random Field (CRF) [Lafferty *et al*., 2001] approach, adapted for multi-label sequence labeling. The final output is obtained by combining the predictions from both models. While CRF alone may not outperform Transformer architectures, its integration with other models proves advantageous in enhancing overall performance. This approach has been successfully utilized in previous studies, such as Lopes *et al*. [2020], where Bi-LSTM was combined with CRF. In our method, the multi-label CRF addresses the challenge of identifying entities with multiple types or nested structures. Evaluation on Portuguese clinical corpora has shown that our method achieves state-of-the-art results, outperforming existing similar approaches.

Given the prevalence of complex entities in health science texts, as demonstrated by Finkel and Manning [2009] and Wang and Lu [2018], we conducted experiments on *SemClinBr* [e Oliveira *et al*., 2022], a Portuguese-language clinical corpus, where it achieved state-of-the-art performance in terms of F1-Score. Additionally, we tested our method on *GENIA* [Kim *et al*., 2003], an English biomedical corpus, and obtained competitive results.

Another gap identified is the scarcity of annotated corpora featuring such complex entities in the health domain. To the best of our knowledge, no corpus in the Portuguese language contains nested and discontinuous entities within the clinical domain. To further validate the effectiveness of our method and contribute to the advancement of research in this domain, we have annotated a corpus named **NestedClinBr**, which contains clinical notes labeled with complex entities in Brazilian Portuguese. This corpus aims to foster the development of corpora and models for low-resource languages. Our method also achieved the best performance on this newly proposed dataset.

2 Related Work

Several methods have emerged to address nested NER, generally belonging into one of the following categories based on their model architectures: early rule-based, layered, region-based, hypergraph, transition-based, QA-based (MRC), grid tagging, generation-based (seq2seq), graph-based, and LLM-based.

Early rule-based methods tackle nested NER by relying on handcrafted rules and post-processing. For instance, Shen *et al*. [2003] presented a rule-based solution to extract biomedical nested entities, utilizing deterministic, morphological, part-of-speech (POS), and semantic features. Approaches by Zhou *et al*. [2004] and Zhou [2006] also employed rule-based techniques, enhancing post-processing by automatically generating rules from training data.

Layered-based approaches treat nested NER as a sequence of flat NER tasks arranged in a cascading structure with interconnected layers. Each layer typically identifies a specific group of entities, often based on hierarchical levels. For example, Zhang *et al*. [2004] proposed a two-Hidden Markov Model (HMM) approach, where the first HMM extracts short embedded entities and the second HMM extends them. Alex *et al*. [2007] introduced a cascaded method utilizing three

CRF models, incorporating inside-out and outside-in CRFs. Ju *et al*. [2018] proposed a Bidirectional Long Short-Term Memory (BiLSTM) encoder with a CRF decoder to identify nested entities, stacking flat NER layers in an inside-out configuration. Wang *et al*. [2020] introduced Pyramid, a neural layered model with interconnected layers designed to predict whether a text region constitutes an entity.

Region-based nested NER methods formalize the nested NER task as a multiclass classification problem, where each potential region is classified into one of the predefined classes. Byrne [2007] introduced a method that transforms the input sentence into potential entities by concatenating adjacent tokens, up to a length of six. In the work of Sohrab and Miwa [2018], a neural exhaustive model was proposed, where BiLSTM was first used to encode the input sentence, and then each region representation was classified as either a specific class or non-entity. Lin *et al*. [2019] developed GEANN, a Gazetteer-Enhanced Attentive Neural Network, which models both the candidate region and its contextual information using a Multi-Layer Perceptron (MLP) classifier. Additionally, Tan *et al*. [2020] introduced a nested NER approach based on boundary detection, where two token-wise classifiers determine whether a token is at the beginning or end of an entity, predicting its entity category.

Hypergraph-based approaches utilize the hypergraph structure to represent the nested nature of entities, with hyperarcs used to indicate that tokens belong to different entities. Lu and Roth [2015] proposed a hypergraph for both boundary detection and category prediction, with five types of nodes representing entities from different semantic categories and boundaries. Katiyar and Cardie [2018] also employed a hypergraph structure, using the BILOU tag scheme[1], where an LSTM-based sequence tagging model learns the hypergraph representation.

Transition-based nested NER approaches, inspired by transition-based dependency parsers, build a tree structure through greedy decoding, one action at a time. Finkel and Manning [2009] introduced a discriminative constituent parser for recognizing nested entities, extracting a constituency-based parsing tree from a sentence that captures its nested structure. Furthermore, Marinho *et al*. [2019] proposed HNNER, the Hierarchical and Nested Named Entity Recognition model, which was designed to handle entities with varying levels of nesting.

QA-based, or **Machine Reading Comprehension (MRC)** approaches, reframe the NER task as a question answering problem, where entities are extracted as responses to queries about the text. In Li *et al*. [2020], the NER dataset was transformed into *QUESTION, ANSWER, CONTEXT* tuples, allowing entity extraction in the same fashion as standard QA tasks. Shen *et al*. [2022] introduced PIQN (Parallel Instance Query Network), which employs global and learnable instance queries to extract entities in parallel from a sentence. Zhang *et al*. [2020] and Banerjee *et al*. [2021] proposed hybrid approaches that combine NER and QA: the model receives an entity type (as a question) and a sentence (as context), but instead of predicting answer spans, it performs token-level classification in the traditional NER format.

Grid tagging is a NER approach that models sentences as a 2D grid, where each cell represents a relation between word pairs. Neural networks assign tags to these cells to capture entity structures, including nested and discontinuous entities. Li *et al*. [2022] introduced two relation tags for labelling the matrix, later extended by Liu *et al*. [2023] to improve syntactic diversity coverage. Though still emerging, grid tagging has shown potential in complex NER tasks.

Generation-based approaches have reframed NER as a sequence generation task. The work ok Yan *et al*. [2021] utilized a BART-based Seq2Seq model to generate entity spans, demonstrating its effectiveness in handling flat, nested, and discontinuous entities. Other works also incorporated enhancements such as pointer networks or calibration techniques to guide the model towards more accurate identification of complex entity structures, as Fei *et al*. [2021].

Graph-based methods, particularly Graph Neural Networks (GNNs), capture relational structure in text. BiFlaG, proposed by Luo and Zhao [2020], modeled entities and their dependencies via heterogeneous graphs, enhancing performance in nested NER. Yuan *et al*. [2020] applied a "parallelism principle" to guide edge updates, capturing syntactic patterns in elliptical expressions. Although GNNs can model complex entity relationships, this approach tends to involve costly preprocessing and high computational demands, and may underperform on sparse data [Ji *et al*., 2025].

LLM-based approaches leverage large pre-trained language models for NER, commonly using prompting or fine-tuning techniques. For instance, Wang *et al*. [2025] proposed GPT-NER, which employes GPT-3 with prompt-guided entity extraction, showing robustness especially in low-resource scenarios. Similarly, prompt-based learning methods, such as InstructionNER [Wang *et al*., 2022], and 2INER [Zhang *et al*., 2023], fine-tune models with minimal data using instructional prompts. These techniques represent a paradigm shift by unifying recognition and generation while effectively capturing complex entity structures. However, they also tend to require higher computational resources and face challenges related to domain adaptation and interpretability [Alhassan *et al*., 2025; Ji *et al*., 2025].

In our work, we propose a unified approach for recognizing nested, discontinuous, and multi-type entities, building upon a **QA-based** framework known for its simplicity and efficiency. To further enhance the model's capability, we incorporated a CRF layer to improve the coverage and precision of nested entity recognition.

3 A new Portuguese clinical corpus

In this section, we present the development of **NestedClinBr**, a new corpus annotated with nested and discontinuous entities within Brazilian Portuguese clinical narratives. The primary goal of *NestedClinBr* is to provide a human-annotated resource that can be leveraged for training and evaluating machine learning models focused on extracting relevant medical information from Portuguese-language texts, with an emphasis on handling nested and discontinuous entities.

[1] Refers to a labeling convention distinguishing between the Beginning (B), Inside (I), Last (L), Outside (O), and Unit (U) parts of entities in text data.

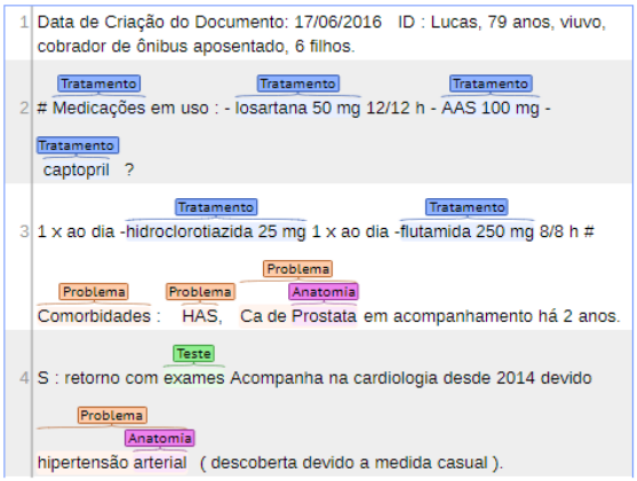

Figure 2. Example of entity annotations in the *NestedClinBr* corpus, in Brazilian Portuguese.

3.1 Data acquisition

We used data from *TempClinBr* [Gumiel *et al*., 2023], a corpus containing clinical notes in the cardiology domain in Portuguese, annotated for entity recognition and temporal relations. Out of 2,094,929 clinical notes from hospitals in Brazil, 126 notes were selected, representing 2,347 sentences and 20,907 tokens. All clinical narratives were properly de-identified to ensure compliance with patient privacy regulations, as outlined by the Brazilian General Data Protection Law (LGPD[2]). The research received approval from the Ethical Committee, with Certificate of Presentation for Ethical Appreciation number 51376015.4.0000.0020.

TempClinBr is originally annotated with entities of the following types: *Problem*, *Treatment*, *Test*, *Evidence*, *Occurrence*, and *Clinical Department*, in addition to annotations for polarity and temporal relations. For the creation of *NestedClinBr*, we retained only the *Problem*, *Treatment*, and *Test* entities, which were reviewed and annotated for nested and discontinuous mentions. Additionally, following the examples of Campillos-Llanos L. [2021] and Báez *et al*. [2020], we introduced the *Anatomy* entity, which represents the location of the human body.

3.2 Annotation Guidelines

To perform the manual annotations, we used the BRAT rapid annotation tool [Stenetorp *et al*., 2012], a web-based platform for text annotation. An example of entity annotation in the *NestedClinBr* corpus using BRAT is shown in **Figure 2**. Each text was annotated by two independent annotators, with any discrepancies resolved by a third annotator (the adjudicator), resulting in a gold-standard corpus. This double annotation process helps prevent bias and allows for the evaluation of annotation quality by measuring the agreement between annotators. We calculated the Inter-Annotator Agreement (IAA) for all annotated data, using the F1-Score at the token level, as done in Deleger *et al*. [2012] and Martínez-deMiguel *et al*. [2022]. The overall F1-Score achieved was 94.8%, indicating substantial agreement between annotators.

The annotation guidelines provide detailed instructions on how to annotate each concept, with a set of useful examples to guide annotators. An entity was annotated as nested when one or more entities were semantically and syntactically embedded within another (e.g., "replacement of the mitral valve" where "mitral valve" is an *Anatomy* entity nested within a *Treatment* mention). Discontinuous annotations were used when a single concept was split by intervening words or punctuation in the sentence, but still referred to the same clinical concept (e.g., "pain in the chest and back", where "pain in the back" is a *Problem* entity). These annotation decisions were driven by the goal of creating a corpus capable of representing complex clinical mentions, as observed in actual narrative texts, rather than strictly matching the exact span to a clinical term. Moreover, all entity types were mapped to semantic categories in the Unified Medical Language System (UMLS)[3], focusing on high-level semantic groups rather than specific Concept Unique Identifiers (CUIs). Instead of full concept linking, we aligned each entity type (e.g., *Problem*, *Treatment*, etc.) with its corresponding semantic group in the UMLS (e.g., *Disorders*, *Procedures*), as shown in **Table 1**. This mapping aimed to ensure conceptual consistency across entities without requiring exact CUI matches. Following the approaches in Martínez-deMiguel *et al*. [2022]; e Oliveira *et al*. [2022]; Gumiel *et al*. [2023]; Dogan *et al*. [2014], our guidelines offer precise definitions of entities along with illustrative examples. **Table 1** summarizes the definitions and examples for each entity tag, with translations provided in English (the original content is in Portuguese).

3.3 Final Considerations

We propose *NestedClinBr*, a Brazilian-Portuguese corpus annotated with clinical concepts in flat, nested, and discontinuous formats. **Table 2** presents corpus statistics, divided into training and testing sets. *NestedClinBr* can be considered a gold-standard resource, as it was manually annotated and its quality verified through IAA scores.

4 Proposed Method

We propose **BioNestedNER**, a hybrid method for recognizing nested, discontinuous, and multi-type named entities, which combines a language model based on BERT architecture with CRF models. The first module employs the QA-based NER approach, handling nested entities by using queries to extract entities. The second module concerns training CRF models adapted to handle multi-label outputs. Finally, the results of the two modules are combined, generating the final result.

4.1 Module 1: QA-based approach

The first module involves recognizing named entities through the QA-based approach, similar to the works of Zhang *et al*. [2020] and Banerjee *et al*. [2021], where a BERT model is trained to return QA-results in NER format. This format includes tagging schemes such as IOB (*Inside-Outside-Begin*)

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

[3] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

Table 1. Entity types from *NestedClinBr* corpus with their respective description and examples.

Entity type	Definition	UMLS group	Translated Examples
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.

Table 2. Statistics of the *NestedClinBr* corpus.

Item	Training	Test	Total
Nested	778	267	1,045
% of total entities	24.4%	27.2%	25.1%
Discontinuous	89	44	133
% of total entities	2.8%	4.5%	3.2%
Total entities	3,186	981	4,167
Balancing (*entity* vs *O*)	30.8%	30.1%	30.45%
Entity avg. length	1.9	1.9	–
Max. tokens per sentence	192	146	169.5

or IOBES (*Inside-Outside-Begin-End-Single*), which offer a structured representation of named entities in text, in contrast to traditional QA methods that return only the start and end indices of entities. Given an input sequence $X = x_1, x_2, ..., x_n$, where n represents the sequence length in tokens, the model is trained to assign a label $y \in Y$ to each token x, where Y corresponds to the entity boundary tagging label in the sentence. In the IOBES scheme, Y could be defined as Y = *B-ENT, I-ENT, E-ENT, S-ENT, O*, while for IOB2, Y = *B-ENT, I-ENT, O*. The model's output (y_n) provides only the start, continuation, and end of the entity, omitting the entity type, as this is addressed by the query referring to the class.

The model training consists of three steps, outlined in **Figure 3**. The first step involves a pre-processing phase, where the corpus is transformed into a QA-based format. To illustrate, consider the sentence "Hemoglobin and insulin are vital proteins for life.", in which "Hemoglobin" and "insulin" are entities of type *Protein*. Suppose we are working with the entity types *Protein*, *DNA*, and *RNA*. This sentence would be converted into the following QA-based instances:

> {"text":"Hemoglobin and insulin are vital proteins for life.", "question":"Protein", "answer":["Hemoglobin", "insulin"]}
>
> {"text":"Hemoglobin and insulin are vital proteins for life.", "question":"DNA", "answer":[]}
>
> {"text":"Hemoglobin and insulin are vital proteins for life.", "question":"RNA", "answer":[]}

For each QA pair, the input is formatted by prepending a special token ([CLS]) to denote the beginning, followed by the query (i.e., a word or a brief set of words describing the entity type), a separator token ([SEP]), and then the full sentence. During training, token-level labels are assigned similarly to standard NER approaches, but only for the second part of the input, i.e., the sentence tokens. This process is repeated for each entity type, generating multiple instances from a single sentence. The resulting training instances for the example above would be:

- Input: [CLS] Protein [SEP] Hemoglobin and insulin are vital proteins for life . Labels: (ignored), (ignored), (ignored), 1, 0, 1, 0, 0, 0, 0 [4]
- Input: [CLS] DNA [SEP] Hemoglobin and insulin are vital proteins for life . Labels: (ignored), (ignored), (ignored), 0, 0, 0, 0, 0, 0, 0
- Input: [CLS] RNA [SEP] Hemoglobin and insulin are vital proteins for life . Labels: (ignored), (ignored), (ignored), 0, 0, 0, 0, 0, 0, 0

Moving to the second step, contextual representations for each word in the sentence are generated using pre-trained language models. Adjustments are made to the segment embeddings to signal sentence breaks in the input text (E_A corresponds to tokens of the query, and E_B to tokens of the second sentence).

The final step consists of fine-tuning the model for token-level classification. A linear layer is added on top of the encoder, and only the outputs corresponding to the sentence tokens are used, excluding the [CLS] token and the query, similar to the QA paradigm. The resulting predictions follow the NER format, identifying the spans in the text that correspond to the queried entity type. The architecture of this step is illustrated in **Figure 4**. Although our implementation relies on the BERT architecture [Devlin *et al*., 2019], the method is model-agnostic and can be adapted to other Transformer-based architectures.

4.1.1 Adaptation to Find Discontinuous Entities

We also adapted the QA-based method to recognize discontinuous entities, common in clinical and biomedical texts. We proposed an end-to-end model, with two classifiers trained to identify both regular entities and discontinuous ones simultaneously, sharing the same embeddings and the loss during training, as shown in **Figure 5** (a). In the pre-processing step, we send two labels for each token, one for each classifier,

[4] Where "1" indicates a token belonging to an entity of the current query type; "0" otherwise.

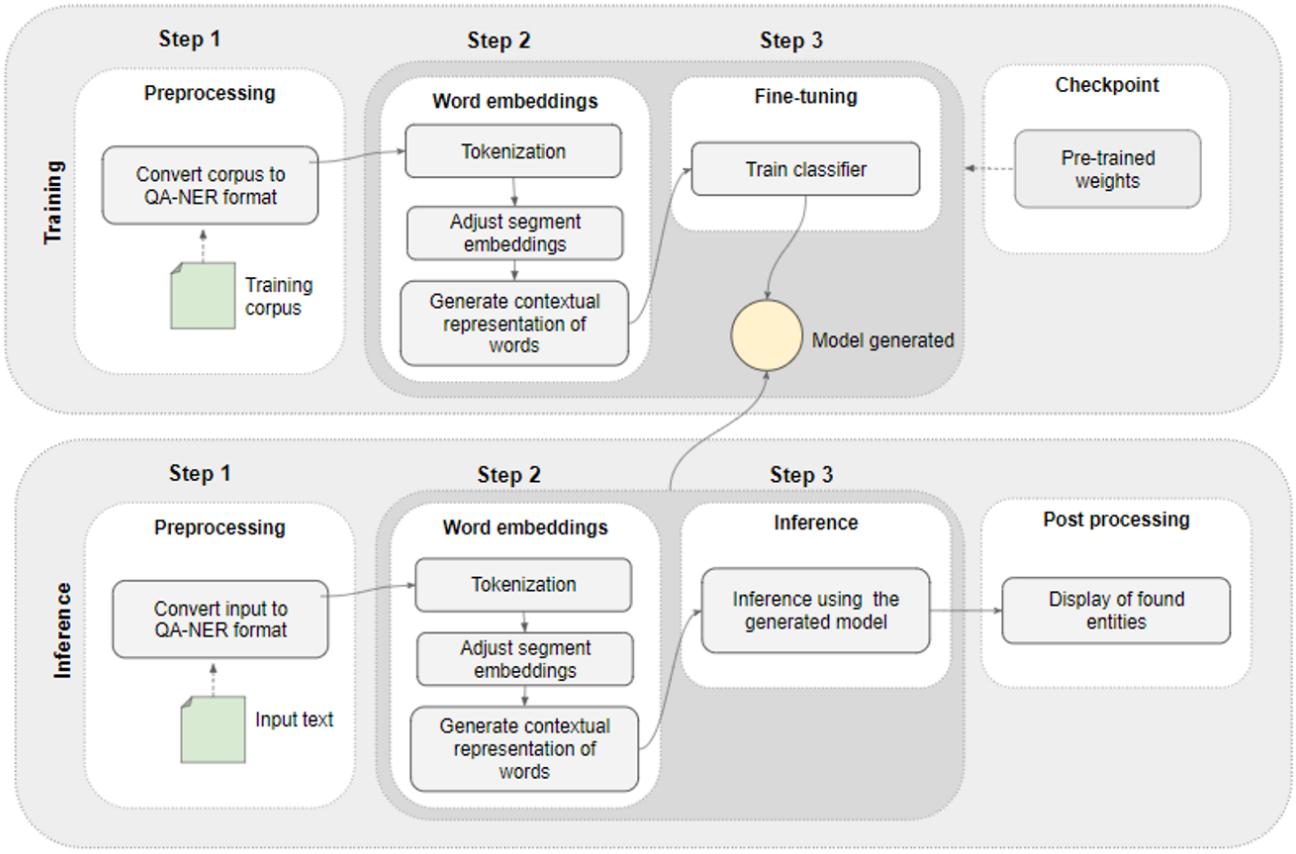

Figure 3. Overview of the first module of BioNestedNER, a QA-based approach.

in NER format (e.g. IOBES). In the training step, each classifier will learn to classify its specific sub-task, and during inference, both types of entities are identified. We also made an adaptation to extract 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** (b).

4.1.2 Handling Class Imbalance

In this setting, the number of non-entity ("O") tokens is higher than that of entity tokens, which amplifies the class imbalance compared to traditional NER tasks. To address this, we adapted the standard cross-entropy loss function by incorporating class weights, thus increasing the importance of underrepresented classes during training. Specifically, the class weights were computed using the `compute_class_weight` utility from the `Scikit-learn` library[5] (with the `balanced` mode), and applied directly to the weighted cross-entropy loss. This adaptation ensures that less frequent classes contribute more to the overall loss, improving their representation in the model's learning process. This approach assigns to each class a weight inversely proportional to its frequency in the training set. The formula is shown in Equation (1), where $n_{samples}$ is the total number of samples, $n_{classes}$ is the number of unique classes, and `np.bincount(y)` returns the frequency of each label (y) in the training labels y:

$$class_{\text{weight}} = \frac{n_{\text{samples}}}{(n_{\text{classes}} * np.bincount(y))} \quad (1)$$

These weights were then used in the weighted cross-entropy loss function, as illustrated in Equation (2):

$$l(x, y) = -\sum_{c=1}^{C} w_{\text{c}} log \frac{exp(x_{\text{n,c}})}{\sum_{i=1}^{C} exp(x_{\text{n,i}})} y_{\text{n,c}} \quad (2)$$

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 as well as $d1, ..., dk$ for the K-dimensional case.

4.2 Module 2: Multi-label CRF

In the second module, we train CRF models utilizing a range of features commonly employed in biomedical NER, drawing inspiration from Mady *et al*. [2022]; Zhang *et al*. [2004]; Campos *et al*. [2012]. These features include morphological, orthographic, contextual, part-of-speech (POS), and semantic information. The morphological features analyze the internal structure of words, capturing their constituent parts and their interactions to identify structural similarities. Orthographic features group words with similar forms, providing insight into word formation patterns. Contextual features consider the surrounding words - both to the left and right of a token - within a defined window of four tokens, to help determine the

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

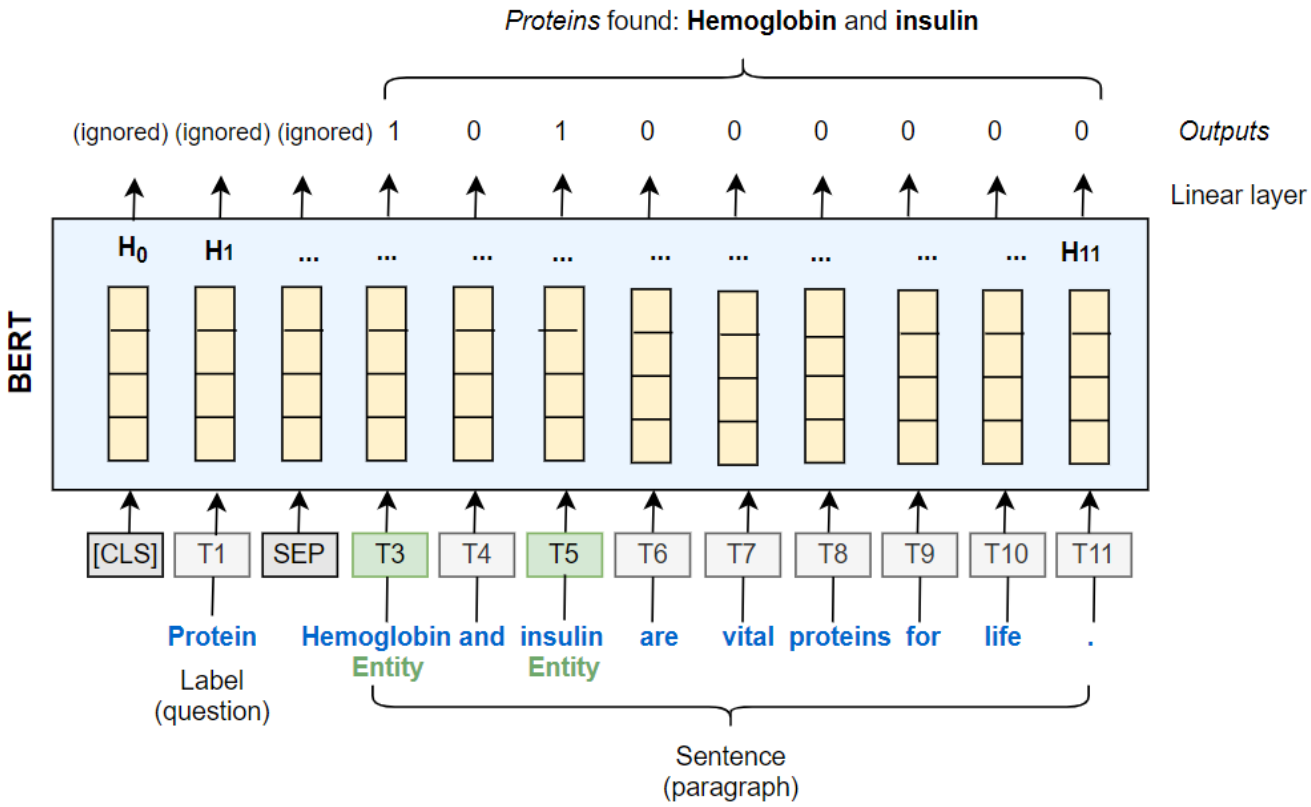

Figure 4. The architecture of the fine-tuning step in BioNestedNER, a hybrid approach that combines elements of NER and QA.

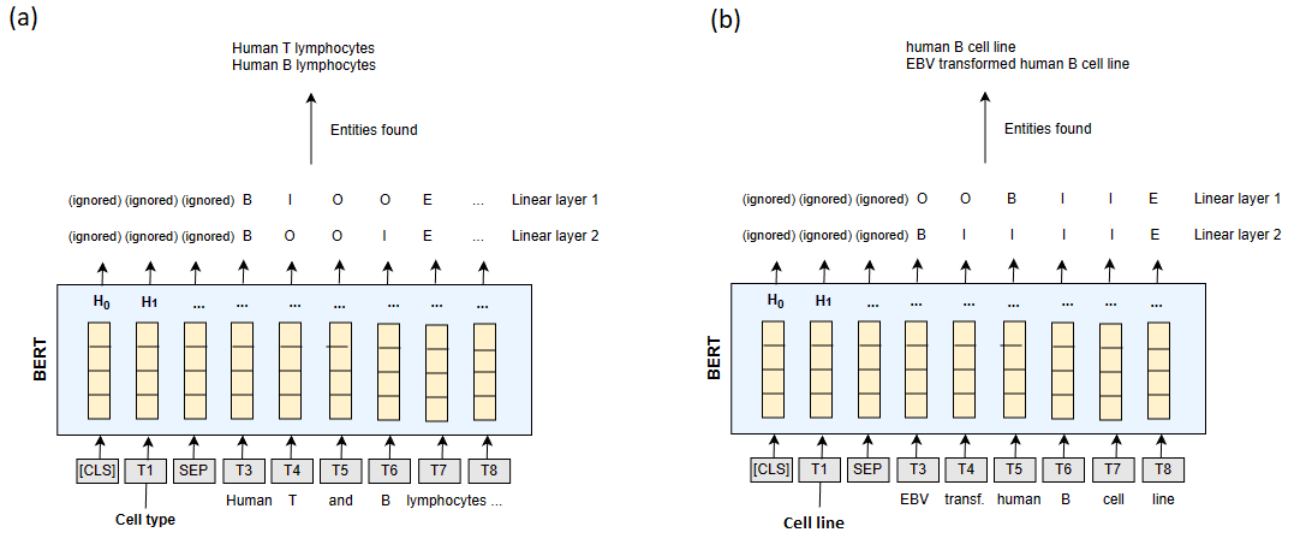

Figure 5. Adaptation to find discontinuous entities (a) and nested entities of the same type (b).

token's class label. POS features leverage morpho-syntactic information, as nouns are often strong candidates for named entities, while verbs and prepositions typically signal entity boundaries. Semantic features focus on the meaning of words and their relationships within a sentence or passage. Additionally, we incorporate a clustering-based feature, following the approach in Mady *et al*. [2022], where clustering algorithms are applied to the text to extract additional features.

Nested entities present a challenge for traditional NER models, as a single token may simultaneously belong to multiple entity types. To address this complexity, we modified the CRF training and inference procedures to handle multi-label scenarios. We employed the Problem Transformation Method 5 (PT5) strategy, as proposed by Tsoumakas and Katakis [2009], which decomposes each example (x, Y) into $|Y|$ separate examples (x, l) for each label $l \in Y$, allowing the model to learn a single-label classifier for each transformed instance.

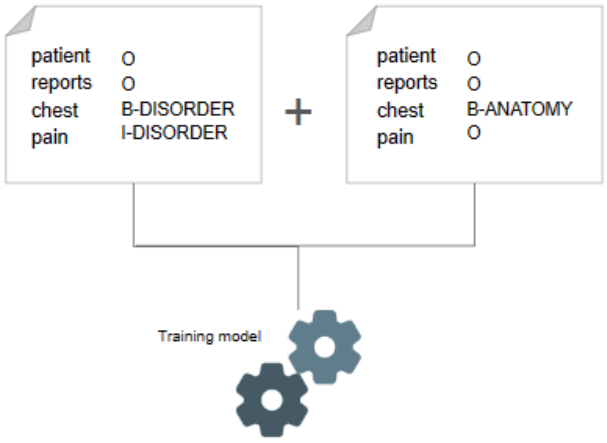

Figure 6. Illustration of example decomposition for multi-label CRF training using the PT5 strategy. An original multi-labeled phrase is transformed into separate training instances, each featuring the full phrase but annotated for a single specific label type at a time.

In practice, for every multi-label token within an input sentence, the entire sentence is replicated for each of its associated labels. For instance, for the expression "patient reports chest pain", as "chest" has labels *B-ANATOMY* and *B-DISORDER* (in IOB format), two distinct training examples will be created: one where "chest" is specifically annotated as *B-ANATOMY* (with other tokens as "O" or their respective single labels) and another where "chest" is specifically annotated as *B-DISORDER*. These distinct examples, while sharing the same underlying text, allow the model to learn contextually for each specific label type (as illustrated in **Figure 6**). The CRF model then returns a probability distribution over the possible labels, indicating the likelihood that each token belongs to each entity class. To adapt CRF for multi-label NER, we define a threshold value to classify the model's output into binary categories (positive or negative) for each label. Threshold values ranging from 0.15 to 0.8 were empirically tested across various corpora to determine the optimal threshold for each dataset.

Finally, we combine the outputs of the QA-based model with the CRF model, resulting in a more comprehensive and accurate coverage of the text, enhancing our entity recognition capabilities.

5 Experiments

This section presents some details about the experiments conducted with BioNestedNER.

5.1 Datasets

We used three datasets to evaluate our method: *NestedClinBr* (our proposed corpus), *SemClinBr* [e Oliveira *et al*., 2022], and *GENIA* [Kim *et al*., 2003].

SemClinBr is a semantically annotated corpus for the Portuguese language, containing 1,000 clinical notes humanly labeled, with 43,659 entities compatible with UMLS standard. In *SemClinBr*, each mention can have more than one label associated (which occurs in 14% of entities), making it ideal for identifying multi-type entities. We performed experiments using a hold-out split (60%-20%-20%) and grouped entities in categories (*Disorder*, *ChemicalDrugs*, *Procedures* and *Abbreviation*).

GENIA was created to develop and evaluate molecular biology information retrieval systems, with 2,000 PubMed abstracts based on human, blood cells, and transcription factors subjects. *GENIA* is one of the most used corpora to evaluate biomedical nested recognition models, as it contains 18,546 sentences with 56,870 entities, of which approximately 31.64% are nested and 3.65%, discontinuous entities according to Chen *et al*. [2020]. We follow the same dataset split used in several works such as Finkel and Manning [2009]; Lu and Roth [2015], where the first 90% of sentences are used in training and validation, and the rest for testing. As in other research, we only used *DNA*, *RNA*, *Protein*, *Cell line* and *Cell type*, ignoring all other entities, in order to maintain compatibility with related works. Unlike most other studies that typically discard discontinuous entities, we report results using the original corpus, which includes both nested and discontinuous entities.

5.2 Baselines

We have trained baselines using BERT models to compare with our method, using binary relevance (BR) technique, i.e., we have trained a specific model for each type of entity and then joined the results. For the experiments in Portuguese, we use the checkpoints of BioBERTpt [Schneider *et al*., 2020], BERTimbau [Souza *et al*., 2020], and BERT-multilingual [Devlin *et al*., 2019], which we refer to as mBERT. For the English experiment, we use BioBERT [Lee *et al*., 2019]. We also trained local models with similar methods: Li *et al*. [2020], referred to here as MRC; Shen *et al*. [2022] referred to here as PIQN; and Banerjee *et al*. [2021], referred to here as QA-NER, to compare the results.

5.3 Evaluation Metrics

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, without considering partial matches. Precision (P), Recall (R), and F1-Score were calculated as normal classification tasks in an entity-based way (not token-based). 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.

5.4 Implementation details

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]). We employed consistent hyper-parameter settings across our experiments and baseline models, including batch sizes ranging from `8` to `32`, a learning rate of`3e-5`, a dropout rate of

`0.1`, a maximum of `10` training epochs, and the `AdamW` optimizer. For the MRC and PIQN baselines, adjustments were made according to their original implementations: both were trained for up to `30` epochs; the MRC model used an increased dropout rate of `0.2` and a weight decay of `0.002`; the PIQN model employed a learning rate of `2e-5`. 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[6], 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.

6 Results and Discussions

In *NestedClinBr* corpus, our proposed method achieved an F1-Score score of 0.851, surpassing other models, with an accuracy for nested entities of 0.662. Concerning the discontinuous entities, a challenging entity type, our model correctly identified 27.3%. Compared to binary relevance trained models, BioNestedNER, which also uses BioBERTpt as a checkpoint, outperformed BioBERTpt by 10.55 points, indicating that the method performed better than the traditional NER algorithm. **Table 3** shows the results on the *NestedClinBr* corpus.

Table 3. Results in the *NestedClinBr* corpus, with flat, nested, and discontinuous entities. (NE: Nested Entities; DE: Discontinuous Entities).

Model	R	P	F1	Acc NE	Acc DE
BioBERTpt (BR)	0.709	0.822	0.757	0.344	0.000
BERTimbau (BR)	0.606	0.754	0.672	0.220	0.000
mBERT (BR)	0.530	0.680	0.596	0.104	0.000
MRC	0.793	0.773	0.783	-	-
PIQN	0.746	0.835	0.788	-	-
QA-NER	0.829	0.858	0.842	0.506	0.000
BioNestedNER (QA)*	0.854	**0.871**	**0.863**	0.639	**0.273**
BioNestedNER (full)*	**0.887**	0.818	0.851	**0.662**	**0.273**

***** *BioNestedNER (QA)* refers to the QA-only module, and *BioNestedNER (full)* includes the CRF layer.

In the *SemClinBr* corpus, BioNestedNER achieved state-of-the-art results with an F1 score of 0.782, outperforming BioBERTpt (second place) by 11.5 points. The model had 0.880 in recall and could identify 69.9% of multi-type entities, as shown in **Table 4**. The improvement in recall underscores the model's proficiency in capturing a broader range of relevant information over the text. This suggests that the model's architecture is well-suited to capture a more extensive set of relevant concepts in biomedical and clinical text analysis, contributing to real-world scenarios.

In the *GENIA* corpus, BioNestedNER outperformed other similar methods, achieving an F1-Score of 0.780. The model correctly recognized 34.8% of nested entities and 31.3% of discontinuous entities, as can be seen in **Table 5**.

[6] `https://colab.research.google.com/`

Table 4. Results in the *SemClinBr* corpus, with flat and multi-type entities. (ME: Multi-Type Entities).

Model	R	P	F1	Acc ME
BioBERTpt (BR)	0.595	**0.759**	0.667	0.497
BERTimbau (BR)	0.557	0.743	0.637	0.418
mBERT (BR)	0.533	0.726	0.615	0.402
MRC	0.635	0.598	0.616	-
PIQN	0.474	0.672	0.556	-
QA-NER	0.598	0.747	0.664	0.507
BioNestedNER (QA)*	0.669	0.731	0.699	0.547
BioNestedNER (full)*	**0.880**	0.704	**0.782**	**0.699**

***** *BioNestedNER (QA)* refers to the QA-only module, and *BioNestedNER (full)* includes the CRF layer.

Table 5. Results in the *GENIA* corpus, with flat, nested, and discontinuous entities. (NE: Nested Entities; DE: Discontinuous Entities).

Model	R	P	F1	Acc NE	Acc DE
MRC	0.709	0.757	0.733	-	-
PIQN	0.701	0.664	0.683	-	-
QA-NER	0.674	0.804	0.733	0.153	0
BioNestedNER (QA)*	0.748	**0.809**	0.778	0.296	**0.313**
BioNestedNER (full)*	**0.783**	0.777	**0.780**	**0.348**	**0.313**

***** *BioNestedNER (QA)* refers to the QA-only module, and *BioNestedNER (full)* includes the CRF layer.

Comparing the performance of our method with the QA-NER [Banerjee *et al*., 2021], which also employs a similar strategy but without improvements and the CRF module, BioNestedNER achieved superior results in the three experiments. This demonstrates the effectiveness of the specific adaptations we introduced, with a positive impact on the overall performance.

The results shown from BioNestedNER are promising, as comprehensive entity recognition is critical for powering clinical decision-making processes. The ability to extract not only flat entities but also nested, multi-type, and discontinuous entities, which are frequent in clinical and biomedical texts, is a relevant feature. Moreover, the model's strong performance in handling clinical texts, as evidenced by its performance on the *SemClinBR* and *NestedClinBR* corpora — which are based on Brazilian Portuguese clinical texts — highlights its efficiency in diverse linguistic environments. This adaptability suggests that BioNestedNER could be effectively applied across various languages and specialized domains, opening up possibilities for its use in a wide array of research settings.

7 Error Analysis

To gain a deeper understanding of the model's limitations and inform future improvements, we performed a detailed error analysis of the BioNestedNER method across all evaluated corpora: *NestedClinBr*, *SemClinBr*, and *GENIA*. The analysis focuses on the three categories of complex entities tackled by our model - nested, discontinuous, and multi-type entities.

7.1 Recognizing Nested Entities

Nested entity recognition was evaluated in the *NestedClinBr* and *GENIA* corpora. While BioNestedNER achieved com-

petitive results and demonstrated improvements over several baseline and literature methods in both corpora, a detailed inspection reveals important challenges that persist:

- Nesting of entities with the same type (e.g., *Protein* within *Protein*) in *GENIA* was particularly challenging.
- Overlapping type confusion, especially for biologically related categories such as *Cell type* and *Cell line*.
- Incorrect nesting depth or flat predictions - cases where the model predicted only the outer or inner entity, failing to capture the full hierarchical structure.
- Boundary mismatches, particularly in nested multi-word expressions with modifiers or ambiguous phrasing, leading to partial or shifted spans.

Such limitations were observed even in instances where flat entity spans were correctly identified, highlighting the need for more advanced structural modeling and improved boundary resolution strategies.

7.2 Recognizing Multi-Type Entities

In the *SemClinBr* corpus, the model successfully identified 69.9%, substantially improving over QA-NER (50.7%). However, the remaining 30% of errors were primarily caused by:

- Only one label being predicted, missing additional types (e.g., predicting only *Disorder* for an entity labeled as *Disorder* and *Abbreviation*).
- Incorrect label combinations, especially in borderline cases like *Procedure* vs. *ChemicalDrug*.

For instance, the term “PURAN”, annotated as both *Abbreviation* and *ChemicalDrug*, was predicted solely as *ChemicalDrug*. Similarly, “TAC”, which should be labeled as *Abbreviation* and *ChemicalDrug*, was assigned only the latter. Another example is “TX RENAL”, a complex entity annotated with three types —*Abbreviation*, *Disorder*, and *Procedure* — but the model recognized only the *Abbreviation* type. All these examples are clinical terms in Brazilian Portuguese, extracted from the *SemClinBr* corpus.

This suggests that the model may tend to default to the most salient or frequent label, potentially overlooking secondary types — a limitation for tasks that require complete and accurate semantic labeling.

7.3 Recognizing Discontinuous Entities

Discontinuous entities represent challenging structures in clinical and biomedical text. In the *NestedClinBr* and *GENIA* corpora, the BioNestedNER method showed a moderate ability to recognize such entities, correctly identifying 27.27% of discontinuous entities in *NestedClinBr* and 31.30% in *GENIA*, using a strict match criterion.

Errors were typically due to:

- Partial span detection, where one or more subcomponents of the entity were missed.
- Over-inclusion, where extra words were mistakenly considered part of the entity.
- Boundary mismatches, caused by linguistic ambiguity or overlapping token structures.

Table 6 illustrates representative examples of correct predictions and common error types encountered by the model in *GENIA* corpus.

These errors reveal that while the model can capture certain patterns of discontinuity, it still faces challenges with complex sentence constructions, particularly those involving coordinated structures or implicit references.

8 Ablation Studies - Effect of CRF

In our method, we incorporated a multi-label CRF model to improve the coverage of nested and multi-type entities. Despite the recent advances in contextual models, CRFs remain an important tool in NLP, providing a way to model sequential data taking into account the dependencies between labels of adjacent tokens. Also, CRFs can be trained efficiently with small amounts of labeled data, making them useful in low-resource settings.

To evaluate the impact of the multi-label CRF component in our pipeline, we conducted ablation studies across the three corpora: *NestedClinBr*, *SemClinBr*, and *GENIA*. In the experiment in the *NestedClinBr* corpus, combining the CRF results did not improve performance in terms of F1-Score, decreasing by 1.16 points. However, it improved recall (from 0.854 to 0.887) and the number of nested entities found, from 63.9% to 66.2% (**Table 3** - BioNestedNER (full)). In the *SemClinBr* corpus, adding the CRF model improved the F1-Score result by 8.3 points and recall by 21.1. Moreover, the percentage of multi-type entities found increased from 54.7% to 69.9% with CRF (**Table 4** - BioNestedNER (full)). Finally, in the *GENIA* corpus, by incorporating the CRF, the F1-Score increased by 0.2, and the percentage of nested entities increased from 29.6% to 34.8% (**Table 5** - BioNestedNER (full)).

Overall, the results indicate that CRFs can be helpful in improving recall and capturing complex entities, although their effectiveness may differ across datasets.

9 Limitations

Although our approach does not require computational power as the exhaustive or LLM-based 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. Our 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. However, this limitation may have limited impact in practice, since such cases are relatively rare in common biomedical corpora. For instance, only 3.51% of all annotated entities in the *GENIA* corpus fall into this category, and in the *NestedClinBr* corpus the percentage is even lower, at 1.45%. Finally, the CRF models alone show inferior results compared to deep learning models trained with Transformer architecture, serving only as a complement to the method.

Table 6. Examples of matches and errors of discontinuous entities with the BioNestedNER method.

Gold Annotation	Predicted	Context	Error Type
Matches			
Human B lymphocytes	Human B lymphocytes	Human T and B lymphocytes	-
interleukin - 1 genes	interleukin - 1 genes	interleukin - 1 and MHC class II genes	-
immunoglobulin heavy - chain genes	immunoglobulin heavy - chain genes	immunoglobulin heavy - and light - chain genes	-
Errors			
cytoskeletal genes	cytoskeletal	cytoskeletal, and extracellular matrix genes	Truncated entity span
heavy chain enhancer	heavy light chain enhancer	heavy and kappa light chain enhancers	Over-inclusion of unrelated tokens
human chromosomes 11p13	human chromosomes	human chromosomes 11p15 and 11p13	Missed specific locus detail

One of the limitations of the new corpus proposed 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.

10 Conclusion

In several situations, entities 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 medical domains. Moreover, less attention has been given to lower-resource languages, such as Portuguese. This work has explored the challenges and opportunities associated with these complex entities, including the development of a new method for nested, discontinuous, and multi-type entity recognition, evaluated in Portuguese and English. We proposed **BioNestedNER**, a new method formed by two modules, inspired on a QA-based approach which proved to be efficient in recognizing these complex entities, combined with a multi-label CRF model. We have propose **NestedClinBr** corpus, to the best of our knowledge, the first clinical corpus in Brazilian Portuguese containing nested and discontinuous entities. In future work, we would like to increase the size of the *NestedClinBr* corpus with more labeled clinical notes and conduct more experiments in other domains, languages, and architectures. We also want to improve the recognition of discontinuous entities, targeting this type of entity, and explore more techniques for data imbalance.

Declarations

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Fundação Araucária - Brazil and Grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN AI) by MICIU/AEI/ 10.13039/501100011033 and FEDER/UE.

Authors' Contributions

ETRS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Yohan Bonescki Gumiel**: Data curation, Writing – original draft. **Paloma Martínez**: Validation, Investigation, Methodology, Writing – review & editing. **Claudia Moro**: Formal Analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Emerson Cabrera Paraiso**: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The *NestedClinBr* corpus is publicly available and can be accessed at `https://github.com/HAILab-PUCPR/NestedClinBr` and `https://huggingface.co/datasets/pucpr-br/nestedclinbr`. Additionally, the source code for replicating the experiments and the BioNestedNER method is open source and available at `https://github.com/HAILab-PUCPR/BioNestedNER`.

References

Alex, B., Haddow, B., and Grover, C. (2007). Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics. DOI: 10.3115/1572392.1572404.

Alhassan, A., Schlegel, V., Aloud, M., Batista-Navarro, R., and Nenadic, G. (2025). Discontinuous named entities in clinical text: A systematic literature review. *Journal of Biomedical Informatics*, 162. DOI: 10.1016/j.jbi.2025.104783.

Báez, P., Villena, F., Rojas, M., Durán, M., and Dunstan, J. (2020). 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*, pages 291–300, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.clinicalnlp-1.32.

Banerjee, P., Pal, K. K., Devarakonda, M., and Baral, C. (2021). Biomedical named entity recognition via knowledge guidance and question answering. *ACM Trans. Comput. Healthcare*, 2(4). DOI: 10.1145/3465221.

Byrne, K. (2007). Nested named entity recognition in historical archive text. In *International Conference on Semantic Computing (ICSC 2007)*, pages 589–596. DOI: 10.1109/ICSC.2007.107.

Campillos-Llanos L., Valverde-Mateos A., C.-C. A. (2021). A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine.s. *BMC Med Inform Decis Mak 21, 69*. DOI: 10.1186/s12911-021-01395-z.

Campos, D., Matos, S., and Oliveira, J. L. (2012). Biomedical named entity recognition: A survey of machine-learning tools. In Sakurai, S., editor, *Theory and Applications for Advanced Text Mining*, chapter 8. IntechOpen, Rijeka. DOI: 10.5772/51066.

Chen, Y., Hu, Y., Li, Y., Huang, R., Qin, Y., Wu, Y., Zheng, Q., and Chen, P. (2020). A boundary assembling method for nested biomedical named entity recognition. *IEEE Access*, 8:214141–214152. DOI: 10.1109/ACCESS.2020.3040182.

Dai, X. (2018). Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics. DOI: 10.18653/v1/P18-3006.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., Solti, I., *et al*. (2012). Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association. Available at:`https://pubmed.ncbi.nlm.nih.gov/23304283/`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Dogan, R., Leaman, R., and lu, Z. (2014). Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47. DOI: 10.1016/j.jbi.2013.12.006.

e Oliveira, L. E. S., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Hasan, S. A., and Moro, C. M. C. (2022). SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1). DOI: 10.1186/s13326-022-00269-1.

Fei, H., Ji, D., Li, B., Liu, Y., Ren, Y., and Li, F. (2021). Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12785–12793. DOI: 10.1609/aaai.v35i14.17513.

Finkel, J. R. and Manning, C. D. (2009). Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics. DOI: 10.3115/1699510.1699529.

Gumiel, Y. B., Oliveira, L. E., de Souza, J. V., Schneider, E. T., Furlan, L. H., Paraiso, E. C., Moro, C., and Carvalho, D. R. (2023). Novel annotation schema for improved temporal reasoning over cardiology notes. Available at:`https://github.com/HAILab-PUCPR/TempClinBr`.

Ji, L., Dang, Y., Du, Y., Gao, W., and Zhang, H. (2025). Nested named entity recognition: A survey of latest research. *Expert Systems*, 42(7):e70052. e70052 EXSY-Jan-25-037.R1. DOI: 10.1111/exsy.70052.

Ju, M., Miwa, M., and Ananiadou, S. (2018). 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)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics. DOI: 10.18653/v1/N18-1131.

Katiyar, A. and Cardie, C. (2018). 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)*, pages 861–871, New Orleans, Louisiana. Association for Computational Linguistics. DOI: 10.18653/v1/N18-1079.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, 19 Suppl 1:i180–2. DOI: 10.1093/bioinformatics/btg1023.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Available at:`https://dl.acm.org/doi/10.5555/645530.655813`.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. DOI: 10.1093/bioinformatics/btz682.

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D., and Li, F. (2022). Unified named entity recognition as word-word relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10965–10973. DOI: 10.1609/aaai.v36i10.21344.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., and Li, J. (2020). A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics. DOI:

10.18653/v1/2020.acl-main.519.

Lin, H., Lu, Y., Han, X., Sun, L., Dong, B., and Jiang, S. (2019). 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)*, pages 6232–6237, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-1646.

Liu, J., Ji, D., Li, J., Xie, D., Teng, C., Zhao, L., and Li, F. (2023). Toe: A grid-tagging discontinuous ner model enhanced by embedding tag/word relations and more fine-grained tags. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:177–187. DOI: 10.1109/TASLP.2022.3221009.

Lopes, F., Gonçalo Oliveira, H., and Teixeira, C. (2020). Comparing different methods for named entity recognition in portuguese neurology text. *Journal of Medical Systems*. DOI: 10.1007/s10916-020-1542-8.

Lu, W. and Roth, D. (2015). Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Lisbon, Portugal. Association for Computational Linguistics. DOI: 10.18653/v1/D15-1102.

Luo, Y. and Zhao, H. (2020). Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. DOI: 10.18653/v1/2020.acl-main.571.

Mady, L., afify, y., and Badr, N. (2022). Nested biomedical named entity recognition. *International Journal of Intelligent Computing and Information Sciences*, 22(1):98–107. DOI: 10.21608/ijicis.2022.104170.1134.

Marinho, Z., Mendes, A., Miranda, S., and Nogueira, D. (2019). Hierarchical nested named entity recognition. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 28–34, Minneapolis, Minnesota, USA. Association for Computational Linguistics. DOI: 10.18653/v1/W19-1904.

Martínez-deMiguel, C., Segura-Bedmar, I., Chacón-Solano, E., and Guerrero-Aspizua, S. (2022). The raredis corpus: A corpus annotated with rare diseases, their signs and symptoms. *Journal of Biomedical Informatics*, 125:103961. DOI: 10.1016/j.jbi.2021.103961.

Naguib, M., Tannier, X., and Névéol, A. (2024). Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-emnlp.400.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. Available at:`https://jmlr.org/papers/v12/pedregosa11a.html`.

Rivera-Zavala, R. M. and Martínez, P. (2021). Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. *BMC bioinformatics*, 22(1):1–23. DOI: 10.1186/s12859-021-04247-9.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics. Available at:`https://www.aclweb.org/anthology/2020.clinicalnlp-1.7`.

Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C.-L. (2003). 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*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics. DOI: 10.3115/1118958.1118965.

Shen, Y., Wang, X., Tan, Z., Xu, G., Xie, P., Huang, F., Lu, W., and Zhuang, Y. (2022). 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)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.67.

Sohrab, M. G. and Miwa, M. (2018). Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1309.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-$8_2$8.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). 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*, pages 102–107. DOI: 10.1016/0020-7519(95)00001-i.

Tan, C., Qiu, W., Chen, M., Wang, R., and Huang, F. (2020). Boundary enhanced neural span classification for nested named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9016–9023. DOI: 10.1609/aaai.v34i05.6434.

Tsoumakas, G. and Katakis, I. (2009). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13. DOI: 10.4018/jdwm.2007070101.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.65215/nxvz2v36.

Wang, B. and Lu, W. (2018). Neural segmental hypergraphs for overlapping mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 204–214, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1019.

Wang, J., Shou, L., Chen, K., and Chen, G. (2020). Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.525.

Wang, L., Li, R., Yan, Y., Yan, Y., Wang, S., Wu, W., and Xu, W. (2022). Instructionner: A multi-task instruction-based generative framework for few-shot ner. DOI: 10.48550/arxiv.2203.03903.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., and Guo, C. (2025). GPT-NER: Named entity recognition via large language models. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2025.findings-naacl.239.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-demos.6.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., and Qiu, X. (2021). A unified generative framework for various ner subtasks. pages 5808–5822. DOI: 10.18653/v1/2021.acl-long.451.

Yuan, C., Wang, Y., Shang, N., Li, Z., Zhao, R., and Weng, C. (2020). A graph-based method for reconstructing entities from coordination ellipsis in medical text. *Journal of the American Medical Informatics Association*, 27(9):1364–1373. DOI: 10.1093/jamia/ocaa109.

Zhang, J., Liu, X., Lai, X., Gao, Y., Wang, S., Hu, Y., and Lin, Y. (2023). 2INER: Instructive and in-context learning on few-shot named entity recognition. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3940–3951, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.259.

Zhang, J., Shen, D., Zhou, G., Su, J., and Tan, C.-L. (2004). Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6):411–422. Named Entity Recognition in Biomedicine. DOI: 10.1016/j.jbi.2004.08.005.

Zhang, Y., Xu, G., Wang, Y., Lin, D., Li, F., Wu, C., Zhang, J., and Huang, T. (2020). A question answering-based framework for one-step event argument extraction. *IEEE Access*, 8:65420–65431. DOI: 10.1109/ACCESS.2020.2985126.

Zhou, G. (2006). Recognizing names in biomedical texts using mutual information independence model and svm plus sigmoid. *International Journal of Medical Informatics*, 75(6):456–467. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue. DOI: 10.1016/j.ijmedinf.2005.06.012.

Zhou, G., Zhang, J., Jian, S., Shen, D., and Tan, C. L. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics (Oxford, England)*, 20:1178–90. DOI: 10.1093/bioinformatics/bth060.