

# Assessing the Limits of Straightforward Models for Nested Named Entity Recognition in Spanish Clinical Narratives

Matías Rojas<sup>1</sup>, Casimiro Pio Carrino<sup>2</sup>, Aitor Gonzalez-Agirre<sup>2</sup>  
Jocelyn Dunstan<sup>1</sup>, and Marta Villegas<sup>2</sup>

<sup>1</sup>Center for Mathematical Modeling, University of Chile

<sup>2</sup>Text Mining Unit, Barcelona Supercomputing Center

## Abstract

Nested Named Entity Recognition (NER) is an information extraction task that aims to identify entities that may be nested within other entity mentions. Despite the availability of several corpora with nested entities in the Spanish clinical domain, most previous work has overlooked them due to the lack of models and a clear annotation scheme for dealing with the task. To fill this gap, this paper provides an empirical study of straightforward methods for tackling the nested NER task on two Spanish clinical datasets, Clinical Trials, and the Chilean Waiting List. We assess the advantages and limitations of two sequence labeling approaches; one based on Multiple LSTM-CRF architectures and another on Joint labeling models. To better understand the differences between these models, we compute task-specific metrics that adequately measure the ability of models to detect nested entities and perform a fine-grained comparison across models. Our experimental results show that employing domain-specific language models trained from scratch significantly improves the performance obtained with strong domain-specific and general-domain baselines, achieving state-of-the-art results in both datasets. Specifically, we obtained  $F_1$  scores of 89.21 and 83.16 in Clinical Trials and the Chilean Waiting List, respectively. Interestingly enough, we observe that the task-specific metrics and analysis properly reflect the limitations of the models when recognizing nested entities. Finally, we perform a case study on an aggregated NER dataset created from several clinical corpora in Spanish. We highlight how entity length and the simultaneous recognition of inner and outer entities are the most critical variables for the nested NER task.

## 1 Introduction

Named Entity Recognition (NER) is a widely studied task that seeks to identify text spans associated with predefined categories. Nested Named Entity

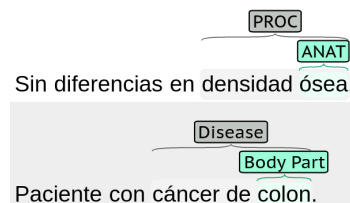


Figure 1: Example of nested entities in the Clinical Trials and Chilean Waiting List datasets.

Recognition is a particular case of NER, where entities can be nested within each other (Finkel and Manning, 2009), such as the example in Figure 1. Traditional NER models simplify nested entities through predetermined rules, such as keeping the most external entity and ignoring inner ones. This simplified problem is better known as flat NER and allows solving the task using traditional sequence labeling architectures such as the BiLSTM-CRF (Lample et al., 2016) approach or fine-tuning transformer-based models (Vaswani et al., 2017).

Regarding the Spanish language, there are several biomedical and clinical datasets containing nested entities, such as the Spanish radiology corpus (Cotik et al., 2017), NUBes (Lima Lopez et al., 2020), the Chilean Waiting List (Báez et al., 2020), Clinical Trials (Campillos-Llanos et al., 2021). However, most previous works transformed the task into a flat NER. As mentioned in Wang et al. (2022), this simplification is due to technological rather than ideological reasons, mainly explained by the difficulty of representing nested entities with the traditional annotation scheme, for example, with the IOB2 sequence labeling format. We argue that treating the nested NER task as flat NER is not optimal since removing part of the entities could result in a loss of information previously annotated by humans, wasting time and resources, and harming the model’s performance.

This paper explores simple neural network-based models as a proxy to address the challenging nested

NER task. Specifically, we revisited the Multiple LSTM-CRF (MLC) and the Joint Labeling architectures and performed experiments on two Spanish clinical corpora. The former consists of training a flat NER model for each entity type following the IOB2 format, while the latter transforms the nested NER task into a flat NER using an annotation scheme that allows preserving the nested entities. We analyze the impact of using pre-trained language models trained on specific domains compared to general-domain ones.

To evaluate the performance of our models, we provide a detailed analysis of task-specific evaluation metrics that adequately measure the effectiveness of the models in recognizing nested entities, considering variables such as entity length, the nesting depth level, and the different types of nested entities. In addition, to better understand the limitations of these models, we created an aggregated corpus formed from several Spanish clinical NER corpora.

In summary, the main contributions of our work are the following:

- We show that straightforward architectures leveraging domain-specific models can tackle the nested NER task, achieving state-of-the-art performances on two clinical datasets in Spanish.
- We conduct an empirical study that compares the impact of using domain-specific language models against general-domain ones, either by using contextualized embeddings or fine-tuning the model in the task.
- We performed an in-depth analysis of the advantages and limitations of the previous approaches by testing our models on an aggregated clinical corpus in Spanish exhibiting complex annotations.

## 2 Related Work

The nested nature of named entities has recently gained special attention from the NLP research community. Several models have been proposed to handle the nesting problem, which can be mainly divided into three categories: region-based, hypergraph-based, and sequence labeling-based models.

Region-based models list potential span candidates and then classify them into predefined categories. In [Sohrab and Miwa \(2018\)](#), they used an

exhaustive neural model enumerating all possible spans within a limited length and then predicted the entity types of those regions using boundary and average internal token representation. [Zheng et al. \(2019\)](#) used a sequence labeling layer to identify candidate spans and then classified the selected regions into their entity category labels. Another region-based model was proposed by [Yu et al. \(2020\)](#), who used contextual representations models to encode sentences and two separate MLPs to create start and end token representations. They then ranked all possible start-end regions in the sentence using nested constraints to predict the labels. Recently, [Shen et al. \(2021\)](#) used a two-stage identifier, using a filter and a regressor to identify high-quality candidate spans and then classifying them into their entity types.

Hypergraph-based models learn the nested structure of entities in the sentence through hypergraphs. The aim is to capture the relations between inner and outer entities to leverage the extraction of nested entities. In [Lu and Roth \(2015\)](#), they proposed a mention hypergraph representation for both extracting entity boundaries and predicting entity labels. Similarly, [Katiyar and Cardie \(2018\)](#) designed a directed hypergraph using LSTM features to learn the nesting structure. In [Luo and Zhao \(2020\)](#), they used a flat NER module for recognizing the most external entities and a graph module for inner entities.

Sequence labeling-based models formulate the nested NER task as several flat NER models. Early work from [Alex et al. \(2007\)](#) introduced three CRF-based methods to reduce the nested NER as several IOB2 tagging problems. [Ju et al. \(2018\)](#) took advantage of inner entity information to improve outer entity recognition. They dynamically stacked LSTM-CRF layers predicting entities in an inside-to-outside manner. In contrast, [Shibuya and Hovy \(2020\)](#) recognized entities from outermost to inner ones using a recursive method based on separate CRFs. This method was improved in [Wang et al. \(2021\)](#), demonstrating that inner to outermost recognition is best for modeling this task. Finally, [Wang et al. \(2020\)](#) recursively introduced the embedding of tokens and regions into flat NER layers simulating the shape of a pyramid and extracting nested entities from the innermost to the outermost entities. The models used in our experiments fall into this category.

### 3 Nested NER Models

In recent years, contextual representational models have improved the performance of many neural network-based models, making it possible to achieve state-of-the-art in several NLP tasks. Unlike traditional word embeddings, language models can represent words according to the sentence-level context. Regarding the NER task, using contextual word embeddings or fine-tuning a pre-trained language model to a specific domain has boosted the performance of models in datasets from several domains.

Previous work in clinical NER showed that using domain-specific language models improves results considerably compared to general-domain language models. However, no studies show this behavior occurs when there is a nested structure in the entities, especially in low-resource languages such as Spanish. In this work, we study whether this trend is confirmed in nested NER datasets using two sequence labeling-based architectures, the Joint Labeling, and the Multiple LSTM-CRF models.

#### 3.1 Joint Labeling Model

The Joint Labeling architecture (Agrawal et al., 2022) consists of formulating nested NER as a flat NER task using an appropriate annotation scheme. Since nested entities allow a token to have more than one entity type, all the token labels are merged into a single token label using a delimiter. This scheme allows solving the problem using traditional sequence labeling architectures that treat the problem as a token-level classification.

We decided to use this architecture due to its high performance on the nested NER task in other languages, such as English and German. Therefore, it is interesting to study the performance of this approach on Spanish datasets, which have been less explored. To solve the token-level classification, we followed the classic approach of fine-tuning transformer-based language models on the NER task. In other words, we fine-tuned language models trained on giant text corpora and added a linear layer to perform the token-level classification.

#### 3.2 Multiple LSTM-CRF

The second approach uses the Multiple LSTM-CRF (MLC) architecture (Rojas et al., 2022a), which trains separate flat NER models for each entity type. The predicted labels of the input sentences corre-

spond to the union of the outputs of each model, thus retrieving both nested entities and text spans tagged with multiple labels.

Each flat NER module consists of three main layers: the embedding layer, the encoding layer with a BiLSTM, and the classification layer, where the most likely sequence of labels is obtained using the CRF algorithm. Regarding the embedding layer, we incorporated contextualized word representations retrieved from a language model, replacing traditional representations such as word and character-level embeddings.

As for the previous model, we tested several domain-specific and general-domain transformer-based language models. The vector representation of words was computed by averaging the representations retrieved from all hidden states. Since BERT-based language models use WordPiece tokenization, we calculated word embeddings using the embedding of the first subtoken. In addition, we tested Clinical Flair (Rojas et al., 2022b), a character-level language model trained on Spanish clinical narratives. Being a character-level model, it is particularly effective for handling out-of-vocabulary and misspelled words, which are very common in clinical texts.

## 4 Experiments

In this section, we present the datasets, settings, and evaluation metrics used in our experiments.

### 4.1 Datasets

We conducted our experiments with two corpora containing nested entities.

- **Chilean Waiting List**<sup>1</sup> (Báez et al., 2020): clinical corpus annotated from real diagnoses of the Chilean healthcare system. It is composed of 87,024 entity mentions and seven entity types. From a nested NER point of view, it is a good resource since 48.23% of the entities are involved in nesting.
- **Clinical Trials**<sup>2</sup> (Campillos-Llanos et al., 2021): clinical corpus created from 500 abstracts of journal articles about clinical trials and 700 announcements of trial protocols. It consists of 46,518 entity mentions and four

<sup>1</sup><https://zenodo.org/record/3926705>

<sup>2</sup>[http://www.111f.uam.es/ESP/nlpmmedterm\\_en](http://www.111f.uam.es/ESP/nlpmmedterm_en)

	Chilean Waiting List			Clinical Trials		
	Train	Test	Dev	Train	Test	Dev
tokens	291,561	36,963	34,987	202,541	67,281	67,661
sentences	15,290	1,912	1,911	7,604	2,522	2,550
avg sent len	19.07	19.33	18.31	26.64	26.68	26.53
entities	69,847	8,837	8,340	27,967	8,940	9,611
avg entity len	2.73	2.71	2.74	1.89	1.86	1.88
nested entities	33,667	4,182	4,126	7,373	2,333	2,580
nested entities (%)	48.20	47.32	49.47	26.36	26.10	26.84

Table 1: Statistics of the datasets used in our experiments.

entity types, which belong to a subset of semantic groups from the Unified Medical Language System (UMLS).

Table 1 shows the overall statistics for each corpus. Compared to other well-known nested NER datasets such as GENIA (Kim et al., 2003) and GermEval (Benikova et al., 2014), where the nesting percentage is less than 20%, these two datasets are a valuable resource for the nested NER task. Especially the Chilean Waiting List corpus, which contains more than twice as much nesting compared to the datasets mentioned above.

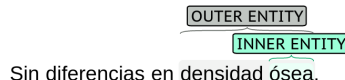
## 4.2 Settings

To analyze the impact of domain-specific language models in Spanish, we used the biomedical version of RoBERTa (*bsc-bio-es*<sup>3</sup>) and the clinical version of RoBERTa (*bsc-bio-ehr-es*<sup>4</sup>) (Carrino et al., 2022). We compared these models with a general-domain Spanish model (*BETO*) (Cañete et al., 2020), a multilingual model (*mBERT*) (Devlin et al., 2019), and two domain-specific models based on continuous pre-training: *mBERT-Galén* (based on mBERT) and *BETO-Galén* (based on BETO) (López-García et al., 2021). As previously mentioned, the MLC model uses these models as contextualized embeddings, while for Joint Labeling, we used them to perform a fine-tuning and solve the token-level classification task.

To train the Joint Labeling model, we used the Adam optimizer and searched for an optimal learning rate out of 1e-5, 5e-5, 5e-6, and 1e-6, with linear decay and no warm-up steps. We trained the model up to a maximum of 20 epochs using a batch size of 8 sequences with a maximum length of 512 tokens and a gradient accumulation of 2 steps, resulting in a total batch size of 16. The training took

<sup>3</sup><https://huggingface.co/PlanTL-GOB-ES/bsc-bio-es>

<sup>4</sup><https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es>



Sin diferencias en densidad ósea.

Figure 2: Example of different types of entities.

approximately 45 minutes for each dataset, using 2 AMD MI50 GPUs with 32 GB of VRAM each.

Regarding the MLC architecture, to train the model of each entity type, we used the SGD optimizer to a maximum of 100 epochs, with mini-batches of size 16 and a learning rate of 0.1. We set the number of RNN layers to 1 and the hidden size to 256. To control overfitting, we employed a learning rate scheduler and an early stopping strategy based on the performance of the validation partition. We also applied dropout regularization after the embedding layer and BiLSTM. The training for each entity type took at most 7 hours under the same hardware settings as Joint Labeling. Since the model of each entity type is independent of the others, this allows us to perform parallel training, reducing the computational cost of this approach.

## 4.3 Metrics

To evaluate the performance of our models, we computed the micro-average precision, recall, and  $F_1$  score over all entities, which is the standard metric used by the research community for evaluating NER systems. In this context, precision is the percentage of entities found by our system that belonged to the test set, while recall is the percentage of entities from the test set found by our system. This metric follows a strict evaluation approach since an entity is considered correct when both entity types and boundaries are predicted correctly. However, one of the main drawbacks of the above metrics is that they do not differentiate nested entities from flat entities. Since flat entities are the most frequent in nested NER datasets, this could overestimate the model’s performance on the task.

	Chilean Waiting List			Clinical Trials		
	$P$	$R$	$F_1$	$P$	$R$	$F_1$
Joint Labeling w/ mBERT cased	74.33 <sub>0.84</sub>	78.08 <sub>1.02</sub>	76.16 <sub>0.93</sub>	83.34 <sub>0.64</sub>	85.97 <sub>0.55</sub>	84.64 <sub>0.55</sub>
Joint Labeling w/ mBERT-Galén	75.16 <sub>0.56</sub>	79.24 <sub>0.33</sub>	77.15 <sub>0.43</sub>	82.53 <sub>0.48</sub>	84.83 <sub>0.37</sub>	83.67 <sub>0.41</sub>
Joint Labeling w/ BETO	75.93 <sub>0.89</sub>	79.10 <sub>0.52</sub>	77.48 <sub>0.67</sub>	84.96 <sub>0.43</sub>	87.19 <sub>0.17</sub>	86.06 <sub>0.21</sub>
Joint Labeling w/ BETO-Galén	74.52 <sub>0.46</sub>	78.79 <sub>0.39</sub>	76.59 <sub>0.10</sub>	82.47 <sub>0.38</sub>	85.49 <sub>0.13</sub>	83.95 <sub>0.15</sub>
Joint Labeling w/ Biomedical RoBERTa	76.55 <sub>0.23</sub>	80.32 <sub>0.33</sub>	78.39 <sub>0.24</sub>	87.92 <sub>0.14</sub>	90.20 <sub>0.24</sub>	89.04 <sub>0.10</sub>
Joint Labeling w/ Clinical RoBERTa	77.31 <sub>0.40</sub>	81.27 <sub>0.46</sub>	79.24 <sub>0.39</sub>	88.03 <sub>0.34</sub>	<b>90.43</b> <sub>0.12</sub>	<b>89.21</b> <sub>0.14</sub>
MLC w/ mBERT cased	79.41 <sub>0.16</sub>	71.31 <sub>0.34</sub>	75.14 <sub>0.15</sub>	84.67 <sub>0.11</sub>	83.90 <sub>0.17</sub>	84.28 <sub>0.05</sub>
MLC w/ mBERT-Galén	78.94 <sub>0.18</sub>	75.56 <sub>0.09</sub>	77.21 <sub>0.13</sub>	84.99 <sub>0.26</sub>	81.67 <sub>0.30</sub>	83.29 <sub>0.24</sub>
MLC w/ BETO	79.33 <sub>0.58</sub>	72.26 <sub>0.25</sub>	75.63 <sub>0.40</sub>	86.04 <sub>0.81</sub>	81.02 <sub>0.73</sub>	83.46 <sub>0.76</sub>
MLC w/ BETO-Galén	79.14 <sub>0.30</sub>	74.67 <sub>0.17</sub>	76.84 <sub>0.23</sub>	85.91 <sub>0.18</sub>	82.21 <sub>0.26</sub>	84.02 <sub>0.22</sub>
MLC w/ Bio RoBERTa	80.30 <sub>0.19</sub>	75.40 <sub>0.35</sub>	77.77 <sub>0.27</sub>	87.97 <sub>0.06</sub>	84.84 <sub>0.43</sub>	86.37 <sub>0.20</sub>
MLC w/ Clinical RoBERTa	80.71 <sub>0.51</sub>	76.13 <sub>1.09</sub>	78.35 <sub>0.82</sub>	<b>88.80</b> <sub>0.23</sub>	85.90 <sub>0.07</sub>	87.32 <sub>0.13</sub>
MLC w/ Clinical Flair	<b>84.31</b> <sub>0.37</sub>	<b>82.04</b> <sub>0.68</sub>	<b>83.16</b> <sub>0.28</sub>	88.38 <sub>0.13</sub>	85.21 <sub>0.13</sub>	86.76 <sub>0.06</sub>

Table 2: Overall results on two nested NER datasets. The reported results correspond to the average of three evaluation rounds using different seeds. Subscript numbers indicate the standard deviations.

To address the above issue, we compute task-specific metrics proposed in Rojas et al. (2022a) that allow analyzing the predictions in detail according to the nested NER task. Specifically, we compute a score for entities not involved in nestings ( $m_{flat}$ ), entities involved in nestings ( $m_{nested}$ ), inner entities in nestings ( $m_{inner}$ ), outer entities in nestings ( $m_{outer}$ ), and complete nestings ( $m_{nesting}$ ). In this context, a nesting is composed of inner and outer entities, and  $m_{nested}$  encompasses the  $m_{inner}$  and  $m_{outer}$  metrics.

These task-specific metrics were calculated using micro-average precision, recall, and  $F_1$  score. Using Figure 2 as an example to better understand the different types of entities, the inner entity is *ósea*, while the outer entity is *densidad ósea*. Both inner and outer entities compose a nesting of depth 2, and there are no flat entities to measure. All experiments and models are freely available to ensure reproducibility<sup>5</sup>.

## 5 Overall Results

Table 2 shows the overall results of our experiments. We observe that across all the experiments, the Joint Labeling model obtains a lower precision than the recall, while in the case of the MLC model, the opposite occurs. As expected, in both models and datasets, the incorporation of domain-specific contextual representation models contributes to significant improvements in the performance compared to general-domain models. However, in some cases, it occurred that the BETO-Galén and mBERT-Galén models did not provide improvements over the general-domain base models. One plausible reason

may be found in the domain-specific vocabulary since the Galén model was trained with the continuous training technique, unlike the RoBERTa-based models, which were trained from scratch.

Although the MLC and Joint Labeling architectures appear to be simple approaches for solving nested NER, we observe that their results are pretty high. Specifically, the best setting for the Chilean Waiting List corpus is the MLC model with embeddings retrieved from the Clinical Flair model. Using the same data splits, we obtained state-of-the-art results with an improvement of almost three micro  $F_1$  points over the best system to date, as reported in Báez et al. (2022), where they achieved a micro  $F_1$  score of 80.27. This excellent performance could be explained since Clinical Flair is a character-level language model, particularly beneficial in datasets with many misspelled and out-of-vocabulary words, such as diagnoses from public hospitals.

On the other hand, the best setting in Clinical Trials is the Joint Labeling approach with the clinical version of RoBERTa. To date, the only result reported on Campillos-Llanos et al. (2021) achieved a micro  $F_1$  score of 86.74 without considering the nested entities. In contrast, we obtained a micro  $F_1$  score of 89.21, achieving state-of-the-art in the corpus and demonstrating the importance of considering nested entities.

## 6 Discussion and Analysis

### 6.1 Nested NER Performance

For a more detailed analysis of the above results, we employ the metrics introduced in Section 4.3 that decompose the model’s performances for different types of nested entities. Table 3 shows the

<sup>5</sup><https://github.com/TeMU-BSC/clinical-nested-ner>

	Chilean Waiting List					Clinical Trials				
	$m_{flat}$	$m_{inner}$	$m_{outer}$	$m_{nested}$	$m_{nesting}$	$m_{flat}$	$m_{inner}$	$m_{outer}$	$m_{nested}$	$m_{nesting}$
Joint Labeling w/ mBERT cased	76.64 <sub>0.89</sub>	82.90 <sub>0.40</sub>	65.95 <sub>1.89</sub>	75.62 <sub>0.97</sub>	54.81 <sub>1.60</sub>	84.59 <sub>0.38</sub>	87.84 <sub>0.89</sub>	81.32 <sub>1.72</sub>	84.79 <sub>1.24</sub>	72.17 <sub>1.70</sub>
Joint Labeling w/ mBERT-Galén	77.06 <sub>1.00</sub>	83.44 <sub>0.56</sub>	69.11 <sub>0.31</sub>	77.27 <sub>0.23</sub>	57.25 <sub>0.17</sub>	83.67 <sub>0.43</sub>	87.13 <sub>0.51</sub>	79.73 <sub>0.57</sub>	83.64 <sub>0.47</sub>	70.55 <sub>1.06</sub>
Joint Labeling w/ BETO	78.22 <sub>1.15</sub>	83.05 <sub>0.29</sub>	68.23 <sub>0.10</sub>	76.64 <sub>0.13</sub>	56.35 <sub>0.21</sub>	86.11 <sub>0.13</sub>	89.09 <sub>0.61</sub>	82.32 <sub>0.36</sub>	85.93 <sub>0.48</sub>	74.07 <sub>0.43</sub>
Joint Labeling w/ BETO-Galén	76.18 <sub>0.41</sub>	83.22 <sub>0.74</sub>	68.81 <sub>0.32</sub>	77.07 <sub>0.42</sub>	57.13 <sub>0.71</sub>	83.85 <sub>0.12</sub>	88.06 <sub>0.29</sub>	79.95 <sub>0.32</sub>	84.25 <sub>0.24</sub>	71.57 <sub>0.18</sub>
Joint Labeling w/ Bio RoBERTa	78.40 <sub>0.19</sub>	85.05 <sub>0.12</sub>	69.42 <sub>0.74</sub>	78.37 <sub>0.36</sub>	58.80 <sub>0.34</sub>	89.09 <sub>0.23</sub>	91.54 <sub>0.40</sub>	85.95 <sub>0.21</sub>	88.91 <sub>0.29</sub>	78.62 <sub>0.62</sub>
Joint Labeling w/ Clinical RoBERTa	79.50 <sub>0.56</sub>	84.76 <sub>0.33</sub>	71.22 <sub>0.73</sub>	78.94 <sub>0.27</sub>	59.89 <sub>0.40</sub>	<b>89.16</b> <sub>0.15</sub>	<b>91.85</b> <sub>0.16</sub>	<b>86.58</b> <sub>0.28</sub>	<b>89.36</b> <sub>0.19</sub>	<b>78.94</b> <sub>0.68</sub>
MLC w/ mBERT cased	75.57 <sub>0.30</sub>	82.45 <sub>0.17</sub>	64.32 <sub>0.28</sub>	74.66 <sub>0.03</sub>	52.30 <sub>0.14</sub>	84.63 <sub>0.10</sub>	85.89 <sub>0.17</sub>	80.41 <sub>0.28</sub>	83.30 <sub>0.10</sub>	67.93 <sub>0.18</sub>
MLC w/ mBERT-Galén	78.13 <sub>0.41</sub>	82.63 <sub>0.58</sub>	67.75 <sub>0.18</sub>	76.20 <sub>0.41</sub>	53.82 <sub>0.58</sub>	83.60 <sub>0.16</sub>	84.49 <sub>0.12</sub>	80.15 <sub>0.88</sub>	82.43 <sub>0.47</sub>	67.71 <sub>0.70</sub>
MLC w/ BETO	76.43 <sub>0.29</sub>	81.12 <sub>0.52</sub>	66.32 <sub>0.65</sub>	74.73 <sub>0.58</sub>	52.05 <sub>0.59</sub>	83.90 <sub>0.82</sub>	84.38 <sub>0.92</sub>	79.61 <sub>0.96</sub>	82.15 <sub>0.72</sub>	66.88 <sub>2.36</sub>
MLC w/ BETO-Galén	77.46 <sub>0.35</sub>	82.47 <sub>0.58</sub>	67.81 <sub>0.29</sub>	76.15 <sub>0.21</sub>	53.80 <sub>0.50</sub>	84.51 <sub>0.22</sub>	84.19 <sub>0.25</sub>	80.89 <sub>0.25</sub>	82.62 <sub>0.25</sub>	68.42 <sub>0.39</sub>
MLC w/ Bio RoBERTa	78.47 <sub>0.45</sub>	83.70 <sub>0.28</sub>	68.15 <sub>0.05</sub>	77.00 <sub>0.15</sub>	55.52 <sub>0.28</sub>	86.59 <sub>0.20</sub>	87.68 <sub>0.16</sub>	83.60 <sub>0.33</sub>	85.76 <sub>0.24</sub>	73.06 <sub>0.61</sub>
MLC w/ Clinical RoBERTa	79.34 <sub>0.73</sub>	83.73 <sub>0.70</sub>	68.71 <sub>1.37</sub>	77.26 <sub>0.94</sub>	55.72 <sub>1.67</sub>	87.47 <sub>0.12</sub>	89.46 <sub>0.20</sub>	84.04 <sub>0.18</sub>	86.90 <sub>0.18</sub>	74.72 <sub>0.19</sub>
MLC w/ Clinical Flair	<b>84.11</b> <sub>0.27</sub>	<b>88.62</b> <sub>0.19</sub>	<b>73.41</b> <sub>0.85</sub>	<b>82.09</b> <sub>0.34</sub>	<b>62.82</b> <sub>0.86</sub>	86.69 <sub>0.03</sub>	90.69 <sub>0.29</sub>	82.76 <sub>0.16</sub>	86.98 <sub>0.23</sub>	74.77 <sub>0.48</sub>

Table 3: Task-specific metrics for nested NER.

results according to task-specific metrics. Interestingly, we note that the nesting metric score, which consists of simultaneously recognizing inner and outer entities, is between 10 and 20  $F_1$  points lower than the standard  $F_1$  metric across models and datasets. In fact, in all cases, the models fail more in recognizing outermost entities than inner ones, suggesting that straightforward methods for nested NER cannot correctly model existing relations between the components of a nested entity. Presumably, since outermost entities are longer in the number of tokens, it is easier for the model to make mistakes when using a strict evaluation metric. Therefore, despite the high score obtained with the standard  $F_1$  metric (see Table 2), this finding points out the importance of using suitable metrics to test the limitations of nested NER approaches. Finally, we can notice that the best models, according to the standard metric, also get the best results according to the nested metrics, proving that the standard metric is consistent but insufficient according to the above findings.

Another point to analyze is the multilabel entities. These entities correspond to text spans associated with more than one entity type, as in the case of the medical term *HTN*, which is both an Abbreviation and Disease. In the Chilean Waiting List corpus, 1,030 entities participate in this type of nesting. Considering only the  $F_1$  score of both models on these types of entities, the MLC approach with Clinical Flair obtained 85.1, while Joint Labeling with Clinical RoBERTa obtained 84.21. Therefore, the difference in the standard metric cannot be explained by the performance of these types of nestings. In the following sections, we perform a detailed analysis of the model predictions, looking for information that explains the difference in performance between the Joint Labeling and MLC approaches beyond the domain in

Level removed	MLC [CF]	Joint Labeling [CR]	$\Delta F_1$
None	<b>83.16</b> <sub>0.28</sub>	79.24 <sub>0.39</sub>	3.92
$\geq 3$ (88)	<b>82.86</b> <sub>0.29</sub>	78.96 <sub>0.35</sub>	3.90
$\geq 2$ (1,875)	<b>79.08</b> <sub>0.55</sub>	75.91 <sub>0.49</sub>	3.17

Table 4: Overall results of our two best models in the Chilean Waiting List when removing deeper entities.  $\Delta F_1$  corresponds to the subtraction in the performance between two models. Here, CF stands for Clinical Flair, while CR is Clinical RoBERTa. The values in parentheses correspond to the support.

which they were trained.

## 6.2 Nesting Depth

An interesting point to analyze between both approaches is the variation in the standard metric when deeper nesting level entities are removed. In Table 4, we show the results in the Chilean Waiting List when entities of depths 2, 3, and 4 entities are removed. Here, depth 1 are the outermost entities, while entities in level 4 are the innermost. First, we notice that by removing nested entities of depths 3 and 4, the  $\Delta F_1$  score between both models remains similar. However, when we removed entities of depth 2, the difference was reduced by 1  $F_1$  point. This might suggest that removing inner entities within a nesting implies a higher decay in MLC performance compared to the Joint Labeling approach. To support this hypothesis, we will analyze the performance of both architectures according to entity length.

## 6.3 Entities of Different Length

In Figure 3, we separate the results obtained in Table 2 depending on the entity’s length. The left side of the figure shows that when the entity length increases, the MLC curve gets closer to the Joint Labeling curve, suggesting that the performance on shorter entities is better for MLC. This finding is confirmed when observing the Clinical Trials

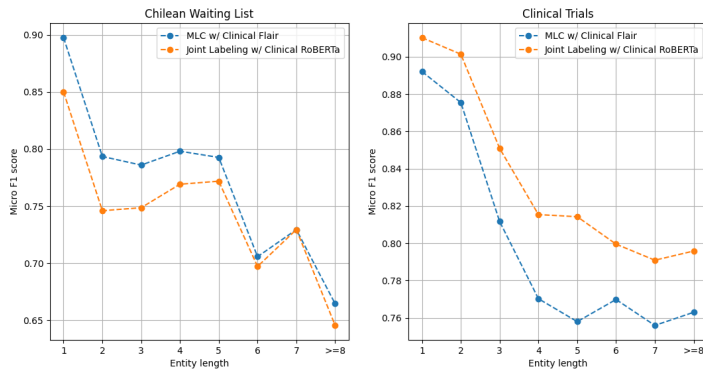


Figure 3: Results of our models according to the entity length.

Length	$\Delta F_1$ Chilean Waiting List	$\Delta F_1$ Clinical Trials
1	4.76 (4, 198)	-1.82 (5, 312)
2	4.75 (1, 522)	-2.59 (1, 780)
3	3.74 (976)	-3.93 (917)
4	2.90 (667)	-4.51 (442)
5	2.08 (470)	-5.62 (207)
6	0.86 (289)	-2.97 (116)
7	-0.01 (223)	-3.49 (61)
$\geq 8$	1.8 (492)	-3.28 (105)

Table 5:  $\Delta F_1$  Score between MLC with Clinical Flair and Joint Labeling with Clinical RoBERTa depending on the length of entities. The values in parentheses correspond to the support.

figure, where the curves move further apart as the length increases.

In Table 5, we see this behavior more explicitly using the  $\Delta F_1$ , which corresponds to the subtraction of the  $F_1$  scores of both models. In the case of MLC, we can see that the most significant difference in the Chilean Waiting List occurs in shorter entities, which could be influencing the standard NER metric. In contrast, in Clinical Trials, although the MLC approach does not outperform Joint Labeling according to the standard metric, the  $\Delta F_1$  score decreases as the entities become smaller.

In the following section, we perform a case study on a synthetic dataset created from several clinical corpora in Spanish. The aim is to study if this behavior is repeated in a dataset containing a similar percentage of nested entities compared to the Chilean Waiting List. Note that this dataset was not used in the experiments section since it is not publicly available for privacy reasons; thus, future works could not reproduce the experiments.

## 7 Case Study

In order corroborate the conclusions presented above, we have created a synthetic nested NER cor-

	Train	Test	Dev
tokens	240,381	29,600	31,364
sentences	9,482	1,120	1,230
avg sent len	25.35	26.43	25.50
entities	18,912	2,283	2,597
avg entity len	2.15	2.21	2.14
nested entities	8,167	1,019	1,147
- entities at level 1	6,577	827	938
- entities at level 2	1,572	191	209
- entities at level 3	18	1	0

Table 6: Statistics of the SPACCC Aggregated dataset.

pus by aggregating the datasets from the PharmaCoNER (Gonzalez-Agirre et al., 2019), CODIESP (Miranda-Escalada et al., 2020), and the recent DisTEMIST (Miranda-Escalada et al., 2022) shared tasks. These datasets are based on the SPACCC corpus<sup>6</sup>, a collection of 1,000 clinical cases from SciELO. Since all the datasets are annotated on the same plain text, merging the annotation of the different tasks is possible. The aggregated dataset is composed of seven entity types, where three are from the PharmaCoNER corpus, two from CODIESP, and one from DisTEMIST.

To generate the aggregated dataset, some important factors have been considered. First, CODIESP is not a NER task but a clinical coding task. However, the authors annotated not only the ICD-10 codes but also the textual evidence that supports the assigned codes. For this experiment, we used the textual evidence from CODIESP as if they were named entities. Secondly, we have found that some textual evidences are either discontinuous or partially contained within other evidences, better known as crossing entities. Both cases are beyond the scope of this research, so we decided to discard them. Thirdly, DisTEMIST is an ongoing task, and we do not have access to the test set

<sup>6</sup><https://zenodo.org/record/2560316>

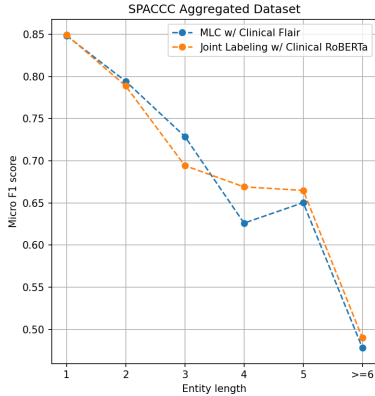


Figure 4: Results of both models on the SPACCC aggregated dataset depending on the entity length.

Metric	MLC	Joint Labeling	Support
<i>standard F<sub>1</sub></i>	<b>78.53</b> <sub>0.16</sub>	78.25 <sub>0.09</sub>	2,283
<i>m<sub>flat</sub></i>	<b>77.99</b> <sub>0.48</sub>	77.48 <sub>0.12</sub>	1,264
<i>m<sub>inner</sub></i>	<b>79.27</b> <sub>0.10</sub>	76.44 <sub>0.41</sub>	520
<i>m<sub>outer</sub></i>	79.07 <sub>0.60</sub>	<b>82.21</b> <sub>0.31</sub>	499
<i>m<sub>nested</sub></i>	79.18 <sub>0.24</sub>	<b>79.23</b> <sub>0.36</sub>	1,019
<i>m<sub>nesting</sub></i>	63.76 <sub>0.60</sub>	<b>64.32</b> <sub>0.08</sub>	499
<i>m<sub>level<sub>1</sub></sub></i>	72.68 <sub>0.03</sub>	<b>77.08</b> <sub>0.32</sub>	827
<i>m<sub>level<sub>2</sub></sub></i>	<b>51.57</b> <sub>2.83</sub>	48.71 <sub>0.67</sub>	191

Table 7: Standard and nested metrics on the SPACCC aggregated dataset.

annotations. For this reason, we only have annotations for 750 clinical cases of the SPACCC corpus. Finally, once the three datasets were aggregated, we found that there were discontinuous annotations between CODIESP and DisTEMIST in 17 of the documents. We removed these documents from the corpus, leaving us with 733 documents. The dataset was divided into 80% for training, 10% for validation, and 10% for testing. The statistics of the corpus are shown in Table 6, and in Appendix A, we show examples of nested entities in this corpus.

According to the standard NER metric, the results for the MLC and Joint Labeling approaches are 78.53 and 78.25, respectively. Although the performance was comparable between both models, analyzing Figure 4, we note that the behavior in the two previous datasets is repeated. The MLC curve is higher than the Joint Labeling curve for the smallest entities, but as the number of tokens increases, the Joint Labeling model obtains slightly better results than MLC.

Considering the  $m_{nested}$  and  $m_{nesting}$  metrics shown in Table 7, we see that Joint Labeling achieves 79.23 and 64.32  $F_1$  scores, while MLC obtains 79.18 and 63.76. Therefore, the former architecture handles better the nested entities in this

corpus. One possible reason why MLC performs better on the standard evaluation metric is that this model achieves the best results according to the  $m_{inner}$  metric by a wide margin, obtaining 79.27 versus 76.44. In contrast, using the  $m_{outer}$  metric, MLC achieves 3.14 points less than Joint Labeling. These findings reaffirm our hypothesis that MLC is better at recognizing smaller entities. For example, if we analyze the metrics in each nesting depth level ( $m_{level_1}$  and  $m_{level_2}$ ), we can see how the MLC model obtains better results in recognizing entities in level 2, which are the innermost entities within a nesting. Finally, and as seen in the other corpora, the results according to the  $m_{nesting}$  metric are low, and the standard metric cannot reflect this limitation.

## 8 Conclusions and Future Work

Since most previous works on nested NER have focused on solving the task in English, this paper contributes to the exploration of diverse models for solving the task on two Spanish clinical datasets, resulting in the state-of-the-art in both corpora. Specifically, we explore the advantages and limitations of the Multiple LSTM-CRF approach, which consists of training one model for each entity type, and the Joint Labeling approach, which through an appropriate annotation scheme, allows solving the task by fine-tuning transformer-based models.

To assess the limitations, we studied task-specific metrics for the nested NER task, which consider variables such as the entity position in the nesting, the impact of nesting depth, and entity length. Although our approaches achieve high results according to the standard metric, we found limitations concerning the recognition of nested entities. The main drawbacks of these architectures are the low performance when recognizing complete nestings and the outermost entities of a nesting. In addition, the MLC approach combined with a character-level language model performs less when recognizing entities with many tokens.

We believe this work can contribute to the NLP community to re-think how the nested NER task is being evaluated, considering task-specific metrics beyond the traditional micro  $F_1$  score. Furthermore, our case study on the SPACCC aggregated dataset points out many of the challenges of the nested NER task, especially when complex annotations are allowed due to the aggregation pro-



cess. Therefore, future work will analyze the performance of other existing architectures beyond the sequence labeling-based approach and compare their performance against our models. We also plan to propose new methods to treat the cases of discontinuous entities and crossing entities, which are entities that overlap others but are not fully contained, to address the nested NER task fully.

## Limitations

Although both approaches achieved excellent results across all the datasets in this research, they have clear limitations. The main drawback is that both models cannot handle the case of nested entities of the same type. This is explained since the file format used for training these architectures cannot incorporate this type of nesting. The second major limitation of both models is that they cannot capture the existing relations between inner and outer entities, leading to poor performance in recognizing complete nestings. These limitations could be addressed by using architectures that separate the problem of detecting entity boundaries from classifying the entity type or hypergraph-based models.

Another significant limitation of the MLC architecture is the high computational cost. Although the models of each entity type can be trained in parallel, when scaling to a dataset with many entity types, the training and inference time could increase considerably compared with other models. On the other hand, we have shown that using character-level language models in this architecture obtains low performance when recognizing longer entities.

Finally, despite the Joint labeling approach employing one model for all the entities, its label space increase exponentially with the number of entities involved, resulting in a bigger classification layer and thus requiring more computational resources than standard NER classification layers.

## Ethics Statement

We defend that this work meets EMNLP Code of Ethics requirements. Our findings have been corroborated by creating an aggregated dataset and replicating the experiments. Furthermore, the aggregated corpus has been generated using other corpora with a free license (Creative Commons Attribution 4.0 International),<sup>7</sup> and the documents

<sup>7</sup><https://creativecommons.org/licenses/by/4.0/legalcode>

where the annotation quality might have been compromised have been removed from the final corpus.

## Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM); Millennium Science Initiative Program ICN17\_002 (IMFD) and ICN2021\_004 (iHealth), and Fondecyt grant 11201250. In addition, it was funded by the Spanish State Secretariat for Digitalization and Artificial Intelligence (SEDIA) within the framework of the Plan-TL<sup>8</sup>. Regarding hardware, the research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02) and the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042).

## References

- Ankit Agrawal, Sarsij Tripathi, Manu Vardhan, Vikas Sihag, Gaurav Choudhary, and Nicola Dragoni. 2022. [Bert-based transfer-learning approach for nested named-entity recognition using joint labeling](#). *Applied Sciences*, 12:976.
- Beatrice Alex, Barry Haddow, and Claire Grover. 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.
- Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. [Automatic extraction of nested entities in clinical referrals in spanish](#). *ACM Trans. Comput. Healthcare*, 3(3).
- Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 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.
- Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. [NoSta-D named entity annotation for German: Guidelines and dataset](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. [A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine](#). *BMC Medical Informatics and Decision Making*, 21.

<sup>8</sup><https://plantl.mineco.gob.es/Paginas/index.aspx>

- Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. [Pretrained biomedical language models for clinical NLP in Spanish](#). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PMLADC at ICLR 2020*.
- Viviana Cotik, Darío Filippo, Roland Roller, Hans Uszkoreit, and Feiyu Xu. 2017. [Annotation of entities and relations in Spanish radiology reports](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 177–184, Varna, Bulgaria. INCOMA Ltd.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.
- Jenny Rose Finkel and Christopher D. Manning. 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.
- Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Itxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. [PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track](#). In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.
- Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 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.
- Arzoo Katiyar and Claire Cardie. 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.
- Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. [Genia corpus—a semantically annotated corpus for bio-textmining](#). *Bioinformatics (Oxford, England)*, 19 Suppl 1:i180–2.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. [NUBes: A corpus of negation and uncertainty in Spanish clinical texts](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France. European Language Resources Association.
- Wei Lu and Dan Roth. 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.
- Ying Luo and Hai Zhao. 2020. [Bipartite flat-graph network for nested named entity recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418, Online. Association for Computational Linguistics.
- Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. [Transformers for clinical coding in spanish](#). *IEEE Access*, 9:72387–72397.
- Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, , and Martin Krallinger. 2022. Overview of DISTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multi-lingual resources.
- Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. [Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020](#). In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022a. [Simple yet powerful: An overlooked architecture for nested named entity recognition](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

- Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022b. [Clinical flair: A pre-trained language model for Spanish clinical natural language processing](#). In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.
- Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. [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)*, pages 2782–2794, Online. Association for Computational Linguistics.
- Takashi Shibuya and Eduard H. Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.
- Mohammad Golam Sohrab and Makoto Miwa. 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 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.
- Yiran Wang, Hiroyuki Shindo, Yuji Matsumoto, and Taro Watanabe. 2021. [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)*, pages 3547–3557, Online. Association for Computational Linguistics.
- Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. [Nested named entity recognition: A survey](#). *ACM Trans. Knowl. Discov. Data*. Just Accepted.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. [Named entity recognition as dependency parsing](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.
- Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. [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)*, pages 357–366, Hong Kong, China. Association for Computational Linguistics.

## A Examples from the SPACCC Aggregated Dataset

As discussed in Section 7, the SPACCC aggregated dataset represents a challenging case study since it may pose severe limitations to straightforward approaches addressing nested NER tasks, mainly due to entity annotations such as discontinuous entities, nested entities, and different entity types. To better visualize such complex entity annotations, we selected some sentences from the SPACCC aggregated dataset before we removed them to perform our experiments. Specifically, Figure 5 shows three different entities, namely, disease entity (DIS\_ENFERMEDAD), ICD diagnosis (CIE\_DIAGNOSTICO), and protein names (PHA\_PROTEINAS), from the PharmaCoNER, CODIESP, and DisTEMIST datasets are presented in different colors to highlight the amount of overlap and crossing between them.

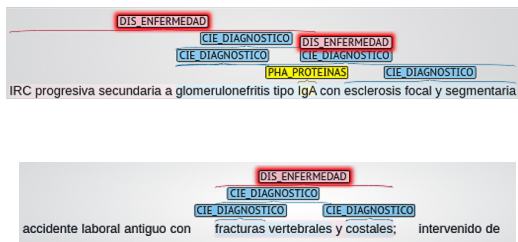


Figure 5: Example of annotations from the SPACCC aggregated dataset with different types of entities belonging to the PharmaCoNER, CODIESP, and DisTEMIST datasets.