

# FLERT-Matcher: A Two-Step Approach for Clinical Named Entity Recognition and Normalization

Matías Rojas<sup>1,2</sup>, Jose Barros<sup>1,2</sup>, Mauricio Araneda<sup>2,3,4</sup> and Jocelyn Dunstan<sup>1,5,6</sup>

<sup>1</sup>*Center for Mathematical Modeling (CMM) - CNRS IRL 2807, University of Chile, Chile.*

<sup>2</sup>*Department of Computer Sciences, University of Chile, Chile.*

<sup>3</sup>*Millennium Institute for Foundational Research on Data (IMFD), Chile.*

<sup>4</sup>*National Center for Artificial Intelligence (CENIA), Chile.*

<sup>5</sup>*Initiative for Data & Artificial Intelligence, University of Chile, Chile.*

<sup>6</sup>*Millenium Institute for Intelligent Healthcare Engineering (iHealth), ANID, Chile.*

## Abstract

In recent years, the appearance of pre-trained language models has boosted the performance of several Natural Language Processing (NLP) models, achieving state of the art in many NLP tasks. Previous work in Named Entity Recognition (NER) has shown that using sentence-level context is not always enough to obtain high-quality contextualized representations while using document-level context contributes to significant improvements in the task. In this paper, we compared the performance of several domain-specific and general-domain language models to identify species mentions on the LivingNER shared task. Specifically, we fine-tuned these models using document-level context with the FLERT approach, which consists of creating the representation based on the context of the actual sentence and its neighboring sentences. Then, to obtain the codes of each entity mention, we used the output of the FLERT model and a Levenshtein distance-based approach. Finally, we trained NER models for real clinical use cases using a similar two-step system and combined these results to perform document-level classification and coding. Our submission results show that our models' performance is far superior to the average of other systems proposed, thus being an important contribution to species recognition and normalization. To reproduce our experiments, the source code of the system is freely available at <https://github.com/plncmm/flert-matcher>.

## Keywords

Named Entity Recognition, Entity Linking, Language Models

## 1. Introduction

Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP) that seeks to identify sequences of words (entities) expressing references to predefined categories such as person names, locations, and organizations [1]. NER, or in general the task of recognizing entity mentions [2], has drawn the attention of the research community due to its relevance in several NLP applications such as relation extraction [3], entity linking [4] and co-reference resolution [5].

---

*IberLEF 2022, September 2022, A Coruña, Spain.*

✉ matias.rojas.g@ug.uchile.cl (M. Rojas); jose.barros.s@ug.uchile.cl (J. Barros); mauricio.araneda@ug.uchile.cl (M. Araneda); jdunstan@uchile.cl (J. Dunstan)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

The NER task has recently been extended to several domains and applications, such as clinical texts. In this context, the automatic recognition of species is critical for scientific disciplines, such as medicine, biology, nutrition, and agriculture. Due to the lack of annotated corpora, most previous work has focused on English datasets. However, since these models are commonly based on rule-based systems, their adaptation to other languages is not trivial as the grammatical and semantic rules change between languages.

LivingNER [6] is the first track aiming to recognize living species and normalization in Spanish clinical case reports. Since this corpus was released in different languages, it is possible to explore the usage of transfer learning with state-of-the-art models for English and then adapt them for languages such as Spanish. Specifically, the LivingNER task is divided into three independent subtasks:

- Subtask 1 - LivingNER-Species NER track: Given a collection of clinical case report documents, this task consists of extracting mentions of human and non-human species.
- Subtask 2 - LivingNER-Species Norm track: Given a collection of clinical case report documents, this task aims to identify species mentions and their corresponding NCBI taxonomy concept identifiers.
- Subtask 3 - LivingNER-Clinical IMPACT track: Given a collection of plain text documents, this task consists of performing a binary classification according to information relevant to real-world clinical use cases of high impact. Then, it seeks to identify the list of NCBI Taxonomy identifies that support the binary classification.

This paper describes our system proposal for the LivingNER shared task, the FLERT-Matcher model. We approach the first subtask using FLERT document-level features, which consist of fine-tuning a clinical version of RoBERTa in Spanish, considering the context of the actual sentence and a window of tokens from the neighbors' sentences. Then, for the second subtask, using the predictions of the previous model and the Levenshtein distance, we normalize the entity mentions into their corresponding concept identifiers. Finally, we use this two-step system for the third subtask to extract relevant information in a real-world clinical use case of high impact.

## 2. Related Work

NER systems have been widely used to identify entities in clinical reports. Previous work can be divided into three main approaches: rule-based methods, traditional machine learning models, and deep learning. However, deep learning techniques applied to NER have shown a substantial performance improvement compared to rule-based approaches. Current Deep Learning approaches can be analyzed as a three-part pipeline [7]: Input Representation, Context Encoder, and Tag Decoder.

The Input Representation module represents words in a document as dense real vectors. Common input representations are based on character-level, word-level, and sentence-level embedding techniques. However, recent work from Schweter and Akbik [8] has demonstrated that using a sentence-level contextualized representation is not enough to obtain the entire

**Table 1**

Statistics of training and validation partitions of the LivingNER documents and annotations used for Subtask 1.

|                                  | Train  | Val    |
|----------------------------------|--------|--------|
| Documents                        | 1000   | 500    |
| Sentences                        | 27264  | 12597  |
| Sentence avg token len           | 23.27  | 23.54  |
| Tokens                           | 642813 | 296161 |
| Entities                         | 16097  | 7106   |
| - Human mentions                 | 7007   | 3289   |
| - Human mentions avg token len   | 1.09   | 1.56   |
| - Species mentions               | 9090   | 3817   |
| - Species mentions avg token len | 1.09   | 1.52   |

context of the word. They proposed a method to obtain the input representations using information from the actual sentence, the last sentence, and the following sentence, achieving state-of-the-art in several NLP tasks.

Context encoders mostly use some Recurrent Neural Network (RNN) variants to identify relevant parts of the document, although Convolutional Neural Networks (CNN) and Multi-Layer Perceptron (MLP) have been used as well.

Finally, Tag Decoders transform the representation of these vectors into a classification among a set of classes. They often use Conditional Random Fields (CRF) to predict the tag for each span. However, Softmax and RNN-based strategies have been used too.

Entity linking, which aims to link entity mentions detected in a document to their related concepts in a given knowledge base or an ontology, is one of the fundamental tasks in information extraction [9]. Entity linking is highly relevant because it can facilitate many tasks such as knowledge base population, question answering, and information integration. In the biomedical domain, entity linking is known as entity normalization or encoding. This task has been solved using different approaches, including rule-based systems [10], machine learning [11, 12] and deep learning [13].

Based on the latest corpora created for NER and normalization task in the clinical and biomedical domain (PharmaCoNER [14], eHealth-KD [15], eHealth CLEF [16], Chilean Waiting List [17], Cantemist [18]), LivingNER is the first shared task explicitly focusing on the extraction of animal species. Precisely, it consists of identifying and normalizing clinical species concepts.

Inspired by previous work on NER, we define our main solution to subtask 1 as a combination of FLERT, a clinical version of RoBERTa, and linear classifiers. Then, to address the Normalization task, we used the predictions of the NER model combined with Levenshtein distance to assign the codes to entity mentions.

### 3. Dataset

The LivingNER Gold Standard consists of a collection of 2000 clinical case reports distributed in plain text, where each clinical case is stored as a single file. The annotation files comprise the character offsets of the entity mentions in TSV (tab-separated values) files and their corre-

**Table 2**

Overall codes distribution for subtask 2.

|                     | Train | Val  |
|---------------------|-------|------|
| Total               | 16097 | 7106 |
| unique codes        | 887   | 552  |
| top-1 most frequent | 7007  | 3289 |
| top-5 most frequent | 8435  | 3929 |
| isComplex           | 630   | 229  |
| isH                 | 730   | 262  |
| isN                 | 125   | 52   |

sponding NCBI Taxonomy code annotations. The corpus was manually annotated by clinical experts following annotation guidelines created specifically for the task and openly distributed at <https://zenodo.org/record/6424678>. It was originally annotated in Spanish and subsequently expanded as a multilingual corpus with neural machine translation strategies.

From Table 1 we can see that the average length of both human and species mentions is similar. Since the NER metrics depend on entity length, the performance for both types of mentions should not vary much. In Table 2 we can see that the distribution of codes is highly concentrated in the five most repeated codes, accumulating more than 50% of the dataset, both for training and validation partitions. Furthermore, we can observe that class representation for isComplex, isH, and isN do not exceed 5% of the data. This is challenging due to the underrepresentation of both codes and positive document classes.

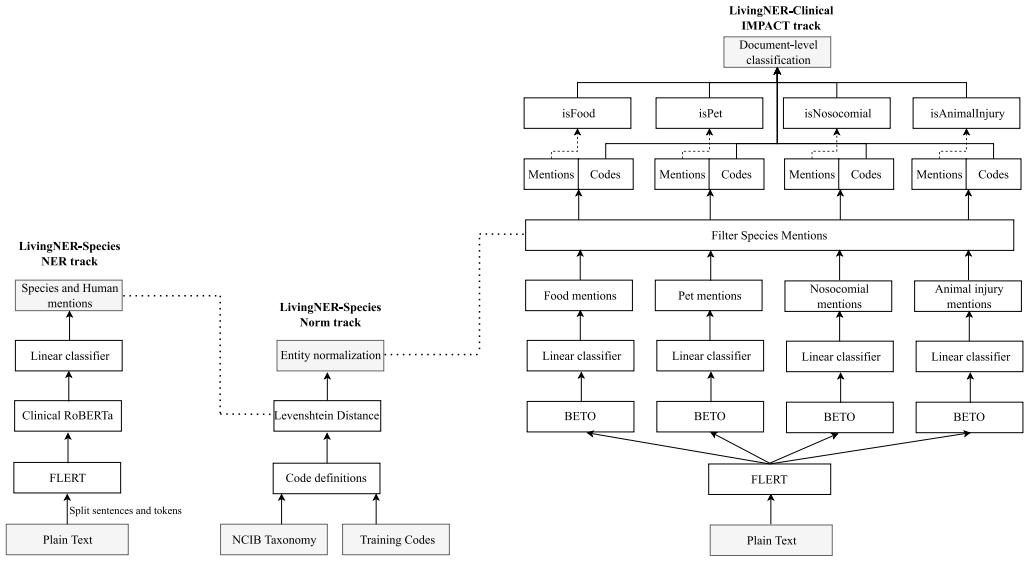
## 4. Methodology

This section provides an overview of the proposed models for each LivingNER subtask mentioned in the introduction. In Figure 1, we show an overview of our architecture for the complete sequence of tasks.

### 4.1. Subtask 1: LivingNER-Species NER track

As shown in the left side of Figure 1, we created a NER system based on the FLERT approach to address the first task. Specifically, this model fine-tunes a transformer-based model but considers the document-level context instead of the sentence-level context. For this purpose, we added a window of 64 tokens from the previous sentence and 64 tokens from the following sentence. We refer to this approach as the FLERT module. We consider it relevant to implement this methodology since these clinical notes are extensive and contain many sentences. This implies that two contiguous sentences most likely describe the same clinical finding, and it is important to consider the word’s context in neighbor sentences.

To analyze the impact of domain-specific language models in Spanish, we measured the performance of the FLERT module using the biomedical version of RoBERTa (*bsc-bio-es*) and the clinical version of RoBERTa (*bsc-bio-ehr-es*) [19]. To compare these models with general-domain ones, we used Spanish BERT (*BETO*) [20] and Spanish RoBERTa (*roberta-base-bne*) [21].



**Figure 1:** Overview of the FLERT-Matcher system architecture.

#### 4.2. Subtask 2: LivingNER-Species Norm track

Figure 1 shows that after identifying species mentions with the FLERT module, we designed a matching system to assign their codes. We refer to this algorithm as the Matcher module. We built a reference dictionary by adding all  $(code, span)$  pairs present in the NCBI Taxonomy dictionary and the training data. Then, we group them by code, resulting in a list of spans for each code. The matching is performed by computing the minimum Levenshtein Distance between each span in the testing partition and our reference dictionary, selecting the code with the lowest value. Note that this operation applies Levenshtein Distance  $nm$  times, where  $n$  is the span size in the reference dictionary, and  $m$  is the span size of the validation data.

To reduce the computational cost of the algorithm, we omit the distance computation for two pairs if the current computed value is larger than the global minimum previously obtained. In addition, to speed up the process, we created a hash map where the already assigned mentions were mapped to the code, and we consulted this map before calculating the Levenshtein distance.

#### 4.3. Subtask 3: LivingNER-Clinical IMPACT track

As shown on the right side of Figure 1, we combined the FLERT-Matcher approach used in the previous tasks to detect species, and each one of the categories (Animal Injury, Food, Pet, and Nosocomial) mentioned with their respective codes. Unlike the traditional systems that treat this problem as a text classification task, we formulated the problem as a NER-normalization task and then used the output of this system to perform the document-level classification.

Since each binary classification problem must be supported with the codes of each entity mention, we cannot separate the problem into classification and coding. If we use one model to classify the document into the respective categories and another to identify the codes of that document, we cannot identify to which category each code belongs. Therefore, it is necessary to set up a two-step system, where the NER model gives us the document category, and the normalization module gives us the codes associated with those mentions.

To combine the results of each FLERT-Matcher model, we decided to adopt a conservative, hierarchical approach; we set the root of the hierarchy to be the species found in subtask 2 and then performed an inner join with each of the other models (Animal Injury, Food, Pet, and Nosocomial), keeping only the mentions of entities that had previously been recognized as a species. Then, if the document contained an entity of the specific *Category*, we manually classified the document with a True value in the *isCategory* column. Note that a document may belong to more than one category since we have independent classifications that depend only on the NER model of the entity type, and the species mentions found in subtask 2. Finally, the codes that support this binary classification are the ones obtained with the Matcher system of subtask 2.

#### 4.4. Experiments

Since the 1000 documents in the training folder did not have a large number of annotations to train the NER models of subtask 3, we decided to merge the texts from the training and validation folder, resulting in a total of 39861 sentences to train our NER models of subtasks 1 and 3. In addition, this allows us to have a more significant amount of data for the hyperparameter search. We split the sentences into 60% for training, 20% for validation, and 20% for testing. The validation partition was used for hyperparameter search, early stopping techniques, and learning rate schedules. The reported test set results were obtained after training the model with the training and validation data on the best hyperparameter configuration. We searched for an optimal learning rate out of 1e-5, 5e-5, 5e-6, and 1e-6. For brevity, we report only the best model, trained with a learning rate of 5e-6. To train each NER model, we used the Adam optimizer with linear decay and no warm-up steps. The models were trained for 20 epochs using a batch size of 16 sequences with a maximum length of 512 tokens. The training of each model took approximately 3 hours using a Tesla V100 GPU.

To evaluate the performance of our models, we computed the micro-average precision, recall, and  $F_1$  score over all entities, which are the standard metric used by the research community for evaluating NER systems. In this context, precision is the percentage of correct entities found by our system, while recall is the percentage of entities in the corpus found by our system. An entity is considered correct when both entity types and boundaries are predicted correctly.

Table 3 shows the results of the four language models used to train our FLERT-based NER models. We observe that the best results are obtained with Clinical RoBERTa achieving a micro  $F_1$  score of 0.965. This is expected since this model was trained on clinical texts, the same domain as the LivingNER texts. On the other hand, comparing the results for each entity type, we notice that it is more challenging to recognize species than humans and that the domain of the language model does not influence the recognition performance of the human entity type.

To evaluate the performance of the Matcher module, we performed an experiment in which

**Table 3**

Overall results in Subtask 1 using domain-specific and general-domain language models.

| Model              | Overall Results |              |                | Species      |              |                | Human        |              |                |
|--------------------|-----------------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
|                    | P               | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> |
| RoBERTa            | 0.956           | 0.953        | 0.954          | 0.937        | 0.932        | 0.934          | 0.978        | 0.977        | 0.977          |
| BETO               | 0.961           | 0.958        | 0.959          | 0.943        | 0.936        | 0.940          | 0.980        | 0.983        | 0.982          |
| Biomedical RoBERTa | 0.963           | 0.967        | 0.965          | <b>0.952</b> | 0.952        | 0.952          | 0.976        | <b>0.985</b> | 0.981          |
| Clinical RoBERTa   | <b>0.964</b>    | <b>0.968</b> | <b>0.966</b>   | 0.950        | <b>0.954</b> | <b>0.952</b>   | <b>0.980</b> | 0.984        | <b>0.982</b>   |

**Table 4**

Overall results in Subtask 2 using FLERT-Matcher approach.

| Model            | P            | R            | F <sub>1</sub> |
|------------------|--------------|--------------|----------------|
| Clinical RoBERTa | <b>0.934</b> | <b>0.916</b> | <b>0.925</b>   |

**Table 5**

Overall results of NER modules on Subtask 3.

| Model              | Food         |              |                | Pet          |              |                | Nosocomial   |              |                | Animal Injury |              |                |
|--------------------|--------------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|---------------|--------------|----------------|
|                    | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P             | R            | F <sub>1</sub> |
| RoBERTa            | 0.824        | <b>0.807</b> | 0.815          | 0.722        | <b>0.813</b> | 0.765          | 0.571        | 0.588        | 0.580          | 0.425         | 0.500        | 0.460          |
| BETO               | <b>0.880</b> | 0.785        | <b>0.830</b>   | <b>1.000</b> | 0.688        | <b>0.815</b>   | <b>0.588</b> | <b>0.588</b> | <b>0.588</b>   | <b>0.611</b>  | <b>0.647</b> | <b>0.629</b>   |
| Clinical RoBERTa   | 0.802        | 0.785        | 0.794          | 0.743        | 0.748        | 0.746          | 0.444        | 0.471        | 0.457          | 0.295         | 0.529        | 0.379          |
| Biomedical RoBERTa | 0.773        | 0.807        | 0.790          | 0.724        | 0.741        | 0.733          | 0.241        | 0.382        | 0.296          | 0.261         | 0.529        | 0.350          |

we fit the Matcher to the definitions of the NCBI taxonomy codes and the mentions in the training subset. Then we predicted the validation subset and calculated precision, recall, and  $F_1$ . We only performed this experiment on the best model from subtask 1, taking into account that the objective of this experiment was to weigh how much error propagation would occur. As shown in Table 4, the error propagation is very low, passing from a  $F_1$  score of 0.96 to 0.92. This shows that if the NER step is successful enough, a simple method such as Levenshtein distance for NCBI taxonomy codes performs well on the LivingNER normalization track.

Table 5 shows the results of the four language models used to train our FLERT-based approach for each one of the categories in subtask 3. Although the domain-specific language models outperform the domain-general ones on subtask 1 by a small margin, BETO and Roberta significantly outperform the biomedical models on this subtask. This can be explained for two main reasons; first, the general-domain language models were trained with three times the data than the domain-specific ones, and second, different from subtask 1, the entity types of this subtask were nevertheless clinical, as in the case of food and pets. Therefore, we can conclude that the best option for these NER models is using a general domain language model, such as BETO. Finally, it is also relevant to note that the performance varies significantly for each category. For example, Food has a  $F_1$  score of 0.830 while Nosocomial has a  $F_1$  score of 0.588.

#### 4.5. Submission

Our experimental results allow us to conclude two important things. First, using domain-specific models is important in species recognition, whereas using a domain-general model is essential

**Table 6**

Submission results in Subtask 1.

| Model               | Overall Results |              |                | Species      |              |                | Human        |              |                |
|---------------------|-----------------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
|                     | P               | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> |
| Ours                | <b>0.946</b>    | <b>0.937</b> | <b>0.941</b>   | <b>0.923</b> | <b>0.904</b> | <b>0.913</b>   | <b>0.976</b> | <b>0.983</b> | <b>0.980</b>   |
| Other systems (avg) | 0.876           | 0.808        | 0.824          | 0.812        | 0.758        | 0.778          | 0.931        | 0.875        | 0.885          |

for recognizing entities in a real clinical use-case context. Second, the Levenshtein distance algorithm is sufficient to generate a high-quality normalization of the entity mentions. Due to this, the system submitted in the shared task is as follows:

- For subtask 1, we used the FLERT approach with Clinical RoBERTa (*bsc-bio-ehr-es*) and a linear classifier to recognize species and human mentions.
- For subtask 2, we compared each mention from subtask 1 with NCBI dictionary definitions and mentions in the corpus. The code assigned to the entity mentions was the minimum Levenshtein distance.
- For each NER model in subtask 3, we used the FLERT approach with *BETO* and a linear classifier to recognize each entity type. The outputs of these models were combined as discussed in 4.3 to get the final document-level predictions and codes.

## 5. Results

Tables 6, 7, and 8 show the results of our models for Subtasks 1, 2, and 3, respectively. In addition, in order to measure the quality of our models, the average results obtained by the other competitors are added.

### 5.1. Subtask 1

Table 6 shows the results obtained by our system compared to the average results of the other systems submitted. We can see that for each metric, our results are above average by a large margin. Regarding the entity types, it was easier to recognize humans than species. This can be explained by the fact that this entity type is of a more general domain and that there are not many different spans of text associated with this category.

Another important point to consider is the length of the entities. In the case of the human entity type, the average length in the data provided is 1.09, while the average length for the species type is 1.59. This may explain the model’s low performance in species since using a strict evaluation metric makes it easier for the model to make errors in terms of entity boundaries. Finally, another observation is that our recall and precision metrics are much more balanced than the average of the other systems, where precision is, in all cases, higher.

### 5.2. Subtask 2

Table 7 shows the results of our FLERT-Matcher model for the normalization task. We can see that we achieve an  $F_1$  score of 0.910, outperforming by 0.083 points the average score of

**Table 7**

Submission results in Subtask 2.

| Model               | Overall Results |              |                | Species      |              |                | Human        |              |                |
|---------------------|-----------------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
|                     | P               | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> |
| Ours                | <b>0.914</b>    | <b>0.906</b> | <b>0.910</b>   | <b>0.867</b> | <b>0.849</b> | <b>0.858</b>   | <b>0.976</b> | <b>0.983</b> | <b>0.980</b>   |
| Other systems (avg) | 0.849           | 0.807        | 0.827          | 0.760        | 0.692        | 0.723          | 0.959        | 0.962        | 0.960          |

**Table 8**

Submission results in Subtask 3 (Classification and Coding).

| Model               | Food         |              |                | Pet          |              |                | Nosocomial |   |                | AnimalInjury  |              |                |
|---------------------|--------------|--------------|----------------|--------------|--------------|----------------|------------|---|----------------|---------------|--------------|----------------|
|                     | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P          | R | F <sub>1</sub> | P             | R            | F <sub>1</sub> |
| Ours                | <b>0.020</b> | <b>0.385</b> | <b>0.038</b>   | <b>0.032</b> | <b>0.364</b> | <b>0.058</b>   | 0          | 0 | 0              | 0             | 0            | 0              |
| Other systems (avg) | 0.009        | 0.154        | 0.016          | 0.009        | 0.102        | 0.017          | 0          | 0 | 0              | <b>0.0001</b> | <b>0.021</b> | <b>0.0002</b>  |

the systems. We believe that the high performance of our model is due to two main reasons: the high performance of our NER model in the previous subtask and the consideration of the training set codes for our matching algorithm.

In the case of the human entity type, we can see that the performance is far superior to that of the species type. This is because we considered that humans had a single code, so not obtaining 100% is due to the error propagated from the previous subtask, where there were incorrectly recognized mentions of humans. Regarding the species entity type, the performance is lower since it is a more challenging problem to recognize the codes of this entity, and the error propagated from subtask 1 is much higher. One of the greatest challenges is the recognition of composite and nosocomial codes. Anyway, our performance results obtained are high and demonstrate the quality of our FLERT-Matcher model for entity extraction and subsequent normalization.

### 5.3. Subtask 3

Table 8 shows the results of our submission for the binary classification and coding, while Table 9 only shows the results for classification. We notice that the performance of our model and the other systems is very low for both the classification and coding tasks. In fact, despite obtaining high recall scores for some of the classifications, the precision scores are deficient, generating a low performance according to the F<sub>1</sub> score.

One possible explanation for this behavior is error propagation. This is because the result of document classification and code identification in this subtask depends on the following factors: the performance of the NER models in subtask 3, the performance of the Matcher module in subtask 2, which provides the codes of species mentions, and in turn, the performance of the FLERT module in subtask 1, which in turn influences the results of the Matcher module.

## 6. Limitations

Our system has two main limitations. Firstly, we may lose important information because we cannot recognize nested entities. This can be addressed using simple sequence labeling-based

**Table 9**

Submission results in Subtask 3 (Only Classification).

| Model               | Food         |              |                | Pet          |              |                | Nosocomial   |            |                | AnimalInjury |            |                |
|---------------------|--------------|--------------|----------------|--------------|--------------|----------------|--------------|------------|----------------|--------------|------------|----------------|
|                     | P            | R            | F <sub>1</sub> | P            | R            | F <sub>1</sub> | P            | R          | F <sub>1</sub> | P            | R          | F <sub>1</sub> |
| Ours                | <b>0.048</b> | <b>0.923</b> | <b>0.091</b>   | <b>0.040</b> | <b>0.417</b> | <b>0.073</b>   | <b>0.006</b> | <b>0.5</b> | <b>0.012</b>   | <b>0.028</b> | <b>0.5</b> | <b>0.053</b>   |
| Other systems (avg) | 0.018        | 0.391        | 0.034          | 0.020        | 0.292        | 0.037          | 0.001        | 0.208      | 0.003          | 0.009        | 0.333      | 0.018          |

architecture, such as the work presented in Báez et al. [22], where they recognized nested entities by training one NER model for each entity type. In this work, we decided not to use this approach since there were few cases of nested entities. Second, we may design a mechanism to mitigate the problem of error propagation between different subtasks. As mentioned above, the outcome of our models in each subtask is conditional on the quality of the predictions obtained in the previous tasks, with subtask 3 being a clear example of the high degree of influence of having a cascaded system.

## 7. Conclusion and Future Work

In this paper, we described our FLERT-Matcher system for the LivingNER shared task, which extracts and normalizes entity mentions from Spanish clinical documents. Our system succeeded in two of the three tasks at hand, obtaining high  $F_1$  scores for subtask 1 and subtask 2. Regarding subtask 3, our system had a high recall but low precision, which explains the drop in performance in subtask 3. Nevertheless, we obtained above-average results in the three subtasks at hand, proving that our choice of architecture was a correct approach.

Regarding the NER systems, we have proven that document-level embedding following the FLERT approach obtains an excellent performance, especially when using domain-specific language models. On the other hand, the Levenshtein distance as a matcher of previously identified mentions proved successful, avoiding high error propagation and obtaining results very similar to those obtained in the NER task.

For future work, we would like to take advantage of the multilingual characteristic of this corpus. Specifically, we want to use the FLERT-Matcher approach but use English texts and language models since it is a language more explored by the NLP community.

## Acknowledgments

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM) and FB210017 (CENIA), Millennium Science Initiative Program Code ICN17\_002 (IMFD) and ICN2021\_004 (iHealth), and Fondecyt 11201250. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02) and the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042). We also thank Mircea Petrache and Claudio Aracena for proofreading this article.

## References

- [1] N. Chinchor, P. Robinson, MUC-7 named entity task definition, in: Proceedings of the 7th Conference on Message Understanding, volume 29, 1997, pp. 1–21.
- [2] R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, S. Roukos, A statistical model for multilingual entity detection and tracking, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 1–8. URL: <https://www.aclweb.org/anthology/N04-1001>.
- [3] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 1003–1011. URL: <https://www.aclweb.org/anthology/P09-1113>.
- [4] S. Guo, M.-W. Chang, E. Kiciman, To link or not to link? a study on end-to-end tweet entity linking, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1020–1030. URL: <https://www.aclweb.org/anthology/N13-1122>.
- [5] K.-W. Chang, R. Samdani, D. Roth, A constrained latent variable model for coreference resolution, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 601–612. URL: <https://www.aclweb.org/anthology/D13-1057>.
- [6] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources, Procesamiento del Lenguaje Natural (2022).
- [7] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 50–70. doi:[10.1109/TKDE.2020.2981314](https://doi.org/10.1109/TKDE.2020.2981314).
- [8] S. Schweter, A. Akbik, FLERT: Document-level features for named entity recognition, 2020. [arXiv:2011.06993](https://arxiv.org/abs/2011.06993).
- [9] Z. Ji, Q. Wei, H. Xu, Bert-based ranking for biomedical entity normalization, AMIA Summits on Translational Science Proceedings 2020 (2020) 269.
- [10] J. D’Souza, V. Ng, Sieve-based entity linking for the biomedical domain, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 297–302.
- [11] J. Xu, H.-J. Lee, Z. Ji, J. Wang, Q. Wei, H. Xu, Uth\_ccb system for adverse drug reaction extraction from drug labels at tac-adr 2017., in: TAC, 2017.
- [12] F. Villena, P. Báez, S. Peñafiel, M. Rojas, I. Paredes, J. Dunstan, Automatic support system for tumor coding in pathology reports in spanish, SSRN Electronic Journal (2021). URL: <https://doi.org/10.2139/ssrn.3982259>. doi:[10.2139/ssrn.3982259](https://doi.org/10.2139/ssrn.3982259).

- [13] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang, D. Huang, Cnn-based ranking for biomedical entity normalization, *BMC bioinformatics* 18 (2017) 79–86.
- [14] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: <https://aclanthology.org/D19-5701>. doi:10.18653/v1/D19-5701.
- [15] A. P.-M. y Suilan Estevez-Velarde y Yoan Gutierrez y Yudivian Almeida-Cruz y Andrés Montoyo y Rafael Muñoz, Overview of the ehealth knowledge discovery challenge at iberlef 2021, *Procesamiento del Lenguaje Natural* 67 (2021) 233–242. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392>.
- [16] L. Goeuriot, H. Suominen, L. Kelly, L. Alemany, N. Brew-Sam, V. Cotik, D. Filippo, G. Gonzalez-Sáez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, J. Vivaldi, M. Viviani, C. Xu, CLEF eHealth Evaluation Lab 2021, 2021, pp. 593–600. doi:10.1007/978-3-030-72240-1\_69.
- [17] P. Báez, F. Villena, M. Rojas, M. Durán, J. Dunstan, 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, 2020, pp. 291–300.
- [18] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [19] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estabé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: <https://aclanthology.org/2022.bionlp-1.19>.
- [20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
- [21] A. Gutiérrez-Fandiño, J. Armengol-Estabé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
- [22] P. Báez, F. Bravo-Marquez, J. Dunstan, M. Rojas, F. Villena, Automatic extraction of nested entities in clinical referrals in spanish, *ACM Transactions on Computing for Healthcare (HEALTH)* 3 (2022) 1–22.