SINAI@MEDDOPLACE: Detecting, Normalizing, and Classifying Places and Related Information in Spanish Medical Texts

Mariia Chizhikova[1], Jaime Collado Montañez[1], Manuel Carlos Díaz-Galiano[1], Luis Alfonso Ureña-López[1] and María Teresa Martín Valdivia[1]

[1]*Department of Computer Science, University of Jaén, Campus Las Lagunillas, s/n, Jaén, 23071, Spain*

Abstract

This paper presents the participation of the SINAI team in the MEDDOPLACE shared task, held on IberLEF2023, focusing on detecting, normalizing, and classifying various types of places and related information in Spanish clinical case reports. Our team tackled all four sub-tasks. For named entity recognition, recurrent classifiers trained on contextual embeddings from a pre-trained RoBERTa language model outperformed the baseline, achieving F1-score of 0.8512. In the second sub-task, a combination of string matching, transformer-based embeddings, TF-IDF character-based n-gram, and Levenshtein distances yielded promising results, albeit with room for further analysis and improvement. On the entity classification sub-task we achieved a micro-average accuracy of 0.7694 by fine-tuning the RoBERTa model. Our team also presented three pipeline versions for end-to-end classification, incorporating variations in the NER component. Notably, the best-performing pipeline achieved F1-score of 0.6272.

Keywords

Clnical Natural Language Processing, Section Identification, RoBERTa language model,

1. Introduction

Electronic health records (EHRs) constitute a valuable source of clinical information. The endorsement of electronic health record (EHR) adoption has been put forth as a pivotal solution for addressing issues pertaining to the quality of care, clinical decision support, and the seamless exchange of reliable information among individuals and departments involved in patient care [1]. The fact that clinical narrative, which is considered to be the most important part of an EHR, is presented in free-text format challenges information extraction - the key step towards leveraging this data.

One of the ways of structuring free-text clinical narratives if through recognizing relevant entities and mapping them to codes from a controlled vocabulary. Such curation requires significant time and monetary investment, as the meticulous annotation process needs to be performed by professional clinical coders. For these reasons, automation of clinical Named

IberLEF 2023, September 2023, Jaén, Spain

mchizhik@ujaen.es (M. Chizhikova); jcollado@ujaen.es (J. C. Montañez); mcdiaz@ujaen.es (M. C. Díaz-Galiano); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. M. Valdivia)

0000-0002-0302-912X (M. Chizhikova); 0000-0002-9672-6740 (J. C. Montañez); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. M. Valdivia)

CEUR Workshop Proceedings (CEUR-WS.org)

Entity Recognition (NER) has been attracting the attention of the research community for years since one of the first clinical coding systems that employed rule-based techniques was announced in 1968 [2].

In recent years many advances were made in the field of clinical Natural Language Processing (NLP) as well as in the specific task of NER. Systems were proposed for detection, normalization and, in some cases relation extraction of such entities as drug [3], protein [4], diseases [5], tumor morphology [6], etc. Initially focused mainly on the English language, nowadays the field has opened notably to working on other languages thus opening opportunities for accessing information regarding many diverse cohorts of patients from non English speaking countries [7]. In the case of Spanish language clinical NLP, this impulse of development derives, for the most part, from the organization of shared tasks that bring the community effort to develop systems able to accurately recognize relevant terms and map them to codes from a controlled vocabulary like International Classification of Diseases (ICD-10) or Systematized Nomenclature of Medicine (SNOMED).

Nevertheless, there are many open challenges and tasks to be resolved. While there is a general consensus that the named entities detected by automatic systems can be either general (person, organization, location) or domain-specific (disease or drug mentions) [8], the definition of the set of entities to detect depends on each independent use case. Thus, the entities considered general, such as allusions to locations, can become clinically relevant in the context of infectious disease, for example. Detection of location mentions can assess not only clinical practice but also health management and retrospective studies within medical research. Moreover, place-related mentions in clinical narratives can be both general (geographical and geopolitical entities) and domain-specific (facilities within a hospital), which presents an additional difficulty.

The MEDical DOcument PLAce-related Content Extraction (MEDDOPLACE) shared task organized on the Iberian Language Evaluation Forum 2023 held on SEPLN 2023 aims at fostering the development of systems capable of accurately detecting, classifying and normalizing 10 different types of place-related entities by providing an extensively annotated corpus of clinical case reports written in Spanish. This challenge consisted in four sub-tasks each one with its own particular objective.

The location entity recognition sub-task required the development of a system that would retrieve the start and the end positions of the entities from free-text clinical reports. The geographic normalization sub-task raised the challenge of mapping the location mentions to their corresponding code from GeoNames, PlusCodes or SNOMED-CT in terms of the type of each entity. Within the third subtask, participants were faced with the problem of categorizing location entities (namely geopolitical (GPE), geographical (GEO) entities and facility mentions (FAC)) into four distinct classes depending on whether they were related to the patient's origin place; the patient's residence's location; a place where the patient has traveled to or from; or a place where the patient has received medical attention. The last sub-task consisted of an end-to-end evaluation of recognition, normalization and classification systems.

The main objective of this paper is to present the system developed by the SINAI team for the MEDDOPLACE shared task. Our team took part in all four sub-tasks. For the first one, we present an approach to multi-class token classification that leverages the benefits of the widely used transformer encoder language model pre-trained on a combination of biomedical and clinical corpora [9] in combination with recurrent neural network classifiers. For the second sub-task, we developed a combination of literal, fuzzy and transformer-based matching techniques, depending on the coding source. The entity classification sub-task was approached as a sequence classification problem for which we opted for a fine-tuning of the same pre-trained model used in sub-task 1. Finally, the last sub-task gave us the opportunity of a joint evaluation of all three systems.

The remainder of the paper is organized as follows: Section 2 offers a brief description of the dataset provided by the organizers of the shared task; Section 3 contains the description of all systems implemented for the four MEDDOPLACE sub-tasks; Section 4 discloses the results obtained during the official evaluation. Finally, Section 5 summarizes the conclusions drawn from the work carried out and suggests possible lines of future research.

2. Data

The MEDDOPLACE corpus is a collection of 1,000 Spanish clinical case reports manually selected in order to ensure their relevance for the task of place-related information extraction. These reports were annotated by clinical experts following the MEDDOPLACE guidelines [10, 11].

In order to be able to perform an objective evaluation, the corpus was divided by the organizers into training and test sub-sets, and only the former, comprised of 743 clinical cases, was available throughout the development phase of the competition.

In summary, the corpus encompasses close to 10,000 annotations, categorized with 10 distinct labels, namely GPE_NOM (geopolitical named entity), GPE_GEN (general geopolitical entity), GEO_NOM (geographical named entity), GEO_GEN (general geographical entity), FAC_NOM (named human-made constructions and buildings mentions), FAC_GEN (general facilities mentions), DEPARTAMENTO (hospital department mentions), TRANSPORTE (transport), COMUNIDAD (sociodemographic information), IDIOMA (language(s) spoken by the patient). Figure 1 illustrates the label distribution in the training dataset which appears to be quite unbalanced with a major presence of general facilities mentions (FAC_GEN), named geopolitical entities (GPE_NOM) and department mentions, while the other labels have a considerably lower representation, most notably the named geographical entities (GEO_NOM) and language mentions that sum up only to 37 occurrences for each one.

The entries of the training sub-set contained from 1 to 79 annotations, being the average number of entities per report 9.41 with a standard deviation of 8.28. Figure 2 shows the 10 most frequently encountered entities in the training subset. As for the length of the reports measured in tokens extracted by the tokenizer of the transformer model that we based our systems on, the shortest report was tokens 92 long, while the longest summed up to 5201 tokens, being the mean length equal to 814.41 with a standard deviation of 609.82. There was noted a positive correlation between the length of a report and the number of annotations it contained.

As we mentioned in the introduction, the second sub-task consisted in normalizing the entities according to one of the following coding standards depending on their type: GeoNames (https://www.geonames.org/) for GPE_NOM and GEO_NOM; PlusCodes (https://maps.google.com/pluscodes/) (also known as Open Location Codes or OLCs) for FAC_NOM and SNOMED-CT (https://www.snomed.org/get-snomed) for the remaining entity types. Figure 3 displays the most frequently observed codes from the GeoNames and SNOMED-CT normalization of the

Figure 1: NER label distribution across the MEDDOPLACE training subset

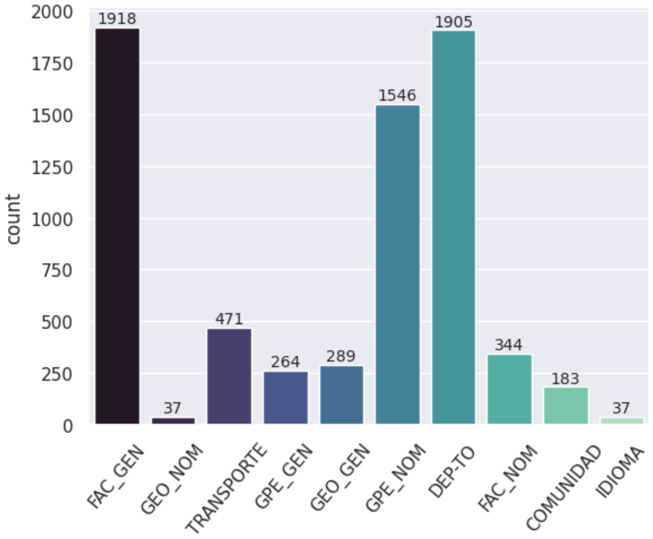
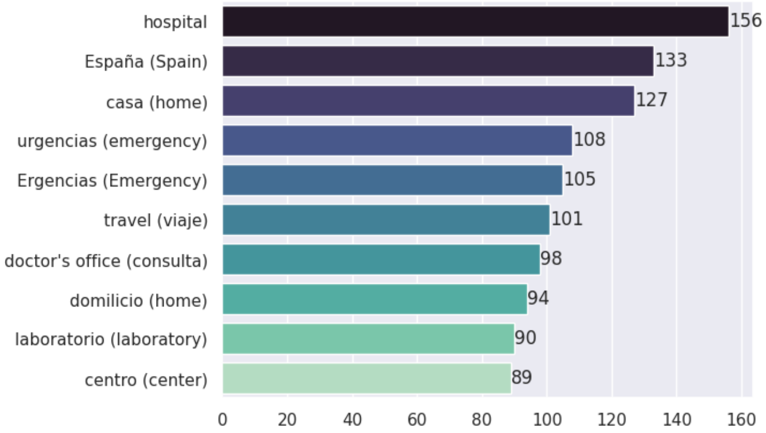


Figure 2: 10 most present entity spans in the MEDDOPLACE training subset. *Translations made only to ease the reading*

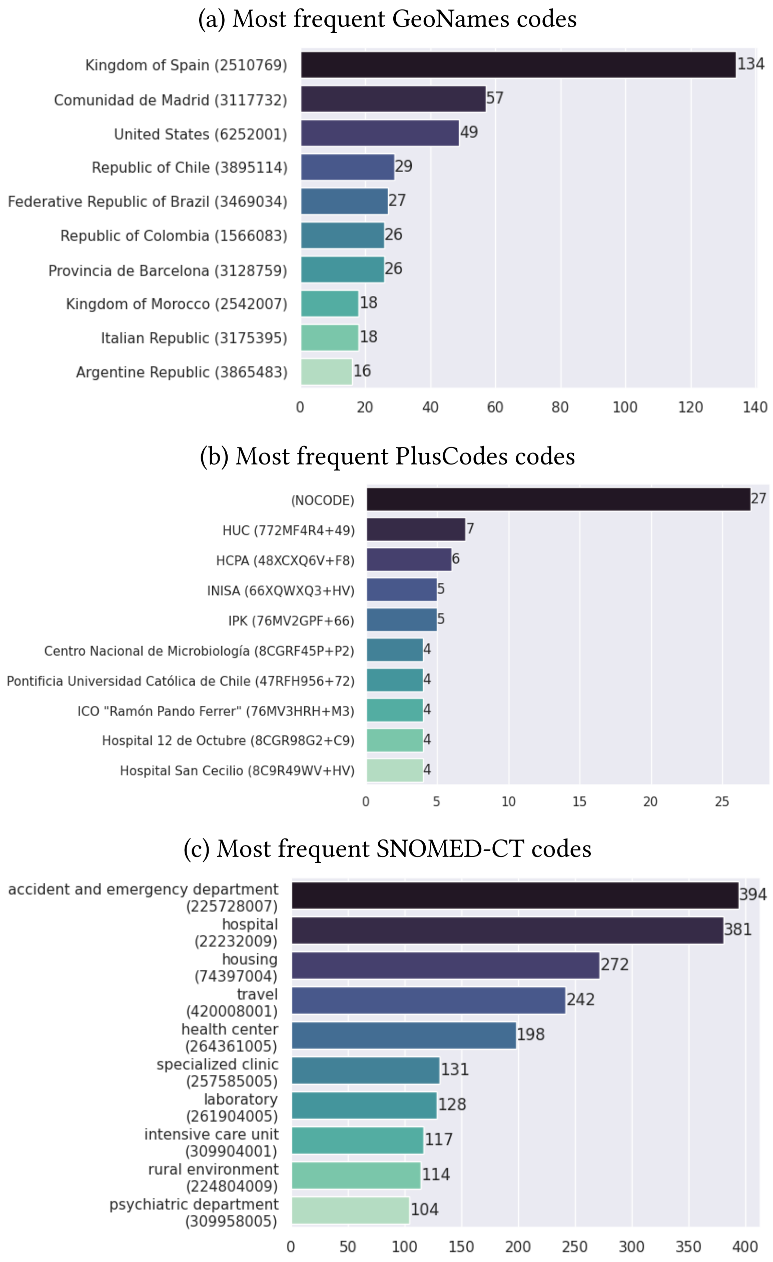


training set.

The third sub-task posed the challenge of classifying location entities, specifically geopolitical (GPE), geographical (GEO) entities, and facility mentions (FAC), into four categories based on their association with the patient's origin, residence, travel destinations, or medical care received. Figure 4 illustrates the frequency distribution of the labels in the training data for the third sub-task.

In order to be able to perform an in-house evaluation of the model, we shuffled and then randomly split the training set so that 24% of the reports formed the validation set.

Figure 3: 10 most frequently observed codes for sub-tasks 2.1 and 2.2. *Translations made only to ease the reading*


(a) Most frequent GeoNames codes

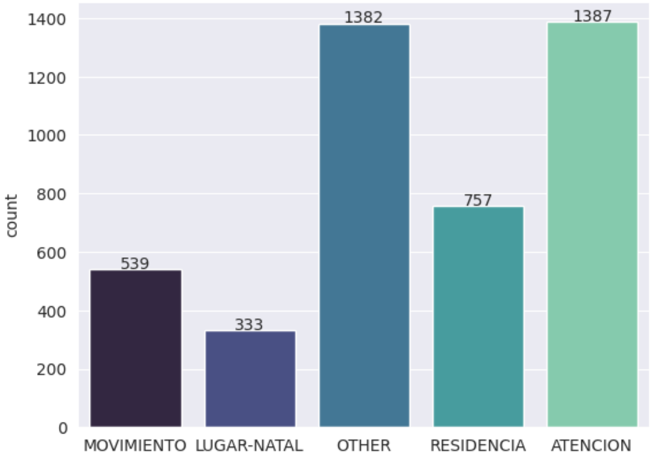
(b) Most frequent PlusCodes codes

(c) Most frequent SNOMED-CT codes

3. System description

This section provides a description of systems presented by our team for each of the four sub-tasks. We also delve into the experimentation carried out during the development process

Figure 4: Label distribution in the training data for sub-task 3



and justify the decisions we made during the whole implementation process.

3.1. Sub-task 1: Location Entity Recognition

As we have mentioned in the introduction to the paper, we formulated the problem of Location Entity Recognition as a multiclass token classification task. In the following paragraphs we describe the methods that underlie each of the presented NER systems.

3.1.1. Baseline

Considering our previous experience in clinical NER shared tasks, our baseline followed the approach that was originally presented for the Living NER shared task at IberLEF 2022 [12]. A model of RoBERTa architecture pre-trained on biomedical and clinical corpora [9] was fine-tuned by adding a linear layer on top of a dropout layer to perform token classification at sentence level. The need of performing sentence-level NER arises from the fact that 455 texts from the training subset (63% of the total) and 163 from the evaluation dataset (65,2% of the total) exceed the maximum input length for the selected transformer model. This approach used the BIO scheme for entity labeling that distinguishes between tokens that begin a named entity from tokens that lay inside and outside. We will refer to this system as ‘run1’ in the remainder of the paper.

In order to maximize the resulting performance we performed a hyperparameter optimization based on the Optuna framework [13]. We optimized the parameters of learning rate, batch size, weight decay, epsilon of the AdamW optimizer and the number of warm up steps. As for the number of training epochs, it was selected through an early stopping strategy that interrupted the fine-tuning and restored the best model after 5 epochs without improvements of the reference metric (macro F1-score) during the evaluation on the validation set that was performed after each epoch. Table 1 summarizes the hyperparameter search space used in the optimization and the values selected during the best trial. The optimization was performed on a single Tesla V100-PCIE-32GB GPU and took approximately 14 hours and 10 minutes.

Table 1
Hyperparameter search space and the selected values for the optimization of the RoBERTa model

Parameter	Search space	Selected value
Learning rate	Float value between $3e-5$ and $5e-5$	$3,8e-5$
Training batch size	8 or 16	16
Weight decay	Float value between $1e-12$ and $1e-1$	$7.7e-5$
AdamW ϵ	Float value between $1e-10$ and $1e-6$	$1.1e-9$
Warmup steps	Integer value between 0 and 1000	948
Number of training epochs	–	12

3.1.2. Recurrent classifiers

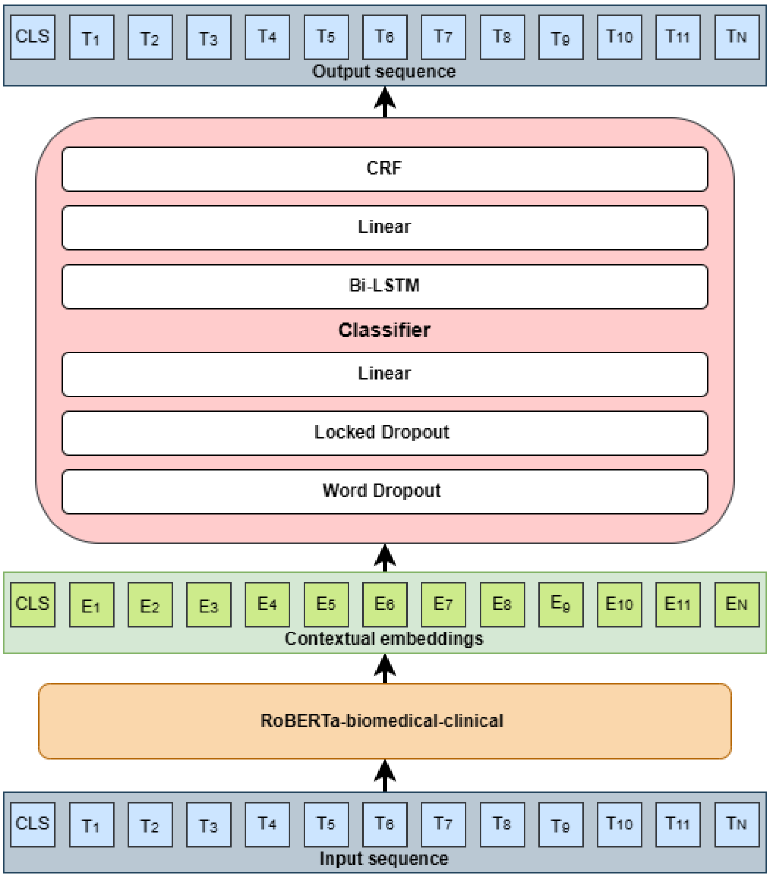
In order to outperform the baseline NER system, we considered implementing a more sophisticated classifier than the RoBERTa token classification head that consisted of a 0.1 dropout layer and a linear layer. Recently, novel approaches have emerged that leverage the advantages of combining transformers and recurrent models, such as Long Short-Term Memory Networks, to enhance the quality of NER [14]. However, the inherent recurrence in these classifiers makes them susceptible to the vanishing gradient problem, particularly when confronted with exceedingly lengthy sequences [15].

Given that a combination of Bidirectional LSTM layers and Conditional Random Field (CRF) classifiers was proven to give promising results when implemented for the task of clinical NER [16], we experimented with training a classifier that incorporates these layers as core elements. In contrast to the state-of-the-art (SOTA) approach of fine-tuning transformer models by incorporating a classification head, often comprising a linear layer along with regularization techniques like dropout [17], employing recurrent classifiers such as LSTM layers or Gated Recurrent Unit (GRU) layers offers the advantage of not necessitating neither inputs with identical dimensions nor restrictions on the maximum input length a priori [15, 18].

In this approximation we used the BIOES labeling scheme, which is slightly more sophisticated than the BIO one, as it distinguishes between single-token entities and the end of an entity. In this approach text representation relied on the same pre-trained model, RoBERTa-biomedical-clinical, used for the baseline system, but without any kind of fine-tuning. Text representations were acquired by averaging the outputs of the model's last 4 layers. These embeddings were extracted in a contextual manner, using a sliding window technique: each token embedding corresponds to its representation in a context of 64 preceding and 64 following tokens. The generated embeddings are then passed to the classifier that was comprised of the following parts:

- Word dropout layer that masked with a probability of 0.05 embeddings of the input tokens with vector corresponding to <UNK>
- Locked dropout layer that sets to zero 10% of the output values from the previous layer. Locked dropout applies the same dropout mask at every step of the training in order to avoid compromising the recurrent classifier performance [19].

Figure 5: Overview of the recurrent classifier architecture


CLS T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 TN
Output sequence
CRF
Linear
Bi-LSTM
Classifier
Linear
Locked Dropout
Word Dropout
CLS E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 EN
Contextual embeddings
RoBERTa-biomedical-clinical
CLS T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 TN
Input sequence

- Linear layer that re-projects the embeddings
- Bi-LSTM layer
- Linear layer that returns a vector of dimensions equal to the number of classes
- CRF layer

Figure 5 provides a graphical overview of the classifier. The implementation and training of all models of recurrent architecture relied on FLAIR framework [20].

In order to evaluate the impact of the length of the input sequences, we performed two experiments, one that performed token classification at sentence level (run2) and one that took the full text of clinical reports as input (run3). The training was run on the same Tesla V100-PCIE-32GB GPU and took approximately 3 hours and 11 minutes for run2 and 5 hours and 15 minutes for run3.

3.1.3. Classifier ensemble for nested entity recognition

Exploratory descriptive analysis revealed that the corpus contained nested entities. Nested entities are multi-word expressions that constitute one large named entity that contain another one, e.g. ‘hospital Clinic de Barcelona’ is a named entity of the FAC_NOM type that contains ‘Barcelona’ which is also a GPE_NOM entity. Both recurrent classifiers and the baseline systems were not trained to recognize nested entities, as the tokens were labeled in a multiclass manner. This led to a decrease in performance in classes that tend to overlap. After a thorough dataset analysis, we considered ensembling two classifiers of the same recurrent architecture described in the previous Section: one trained to recognize more general labels (GPE_NOM, GPE_GEN, COMUNIDAD, GEO_NOM and GEO_GEN) and the other for more domain-specific labels (FAC_NOM, FAC_GEN, DEPARTAMENTO, TRANSPORTE).

As in the case of not ensembled recurrent classifiers, two experiments were carried out: sentence-level (run4) and text-level (run5). All models were trained on hardware identical to the other experiments, and took 2 hours 50 minutes and 2 hours 19 minutes for each model ensembled for run4, and 3 hours and 37 minutes and 4 hours 2 minutes for both models of run5.

3.2. Sub-task 2: Geographic Normalization

In this sub-task, we had to assign either GeoNames, PlusCodes or SNOMED-CT codes to the entities provided in the testing set, depending on their entity type. Thus, this sub-task is further split into three different tracks where we always perform the same first step, which consists of looking for perfect matches against the training set and assigning the codes to each found entity. After that, the following steps slightly differ depending on the nature of the coding sources, which we describe below.

3.2.1. Geocoding to GeoNames

GeoNames is a gazetteer where over twelve million named places are identified with a unique code. In addition, each place is provided with alternate names and different features indicating which kind of place the term is referring to.

After the previously mentioned string matching against the training set, we processed the provided gazetteer and then repeated the process of string matching, now against said processed ontology. Thus, we first filtered out all the rows where the columns *asciiname*, *feature code*, *country code*, *cc2* and *admin1 code* were all duplicated while keeping the ones with the most frequent *feature class* value in the training set, which is given in the following order: A, P, L, T, H, S, R, U, V. Then, we expanded the gazetteer leaving one row per name variant (i.e. *name*, *ascii name* or alternate name inside *alternatenames*), which resulted in 21777082 possible terms to be matched with.

Finally, for those mentions that could not be found in the training set or the processed gazetteer, we performed a two-step fuzzy matching against the ontology. In the first place, the codes assigned to each mention were the ones from the closest terms computed as the cosine similarity distance between their character-level TF-IDF vectors. As a second and final step, entities with no close terms in the gazetteer (i.e. cosine distance < 0.75) were assigned the code of the closest term as computed by the Levenshtein or edit distance, which is the minimum number of single-character edits required to transform one string into the other.

3.2.2. Geocoding to PlusCodes

In this track we were required to encode coordinates from the entities in the testing set into PlusCode format, which is a technology designed to replace street addresses by coding space areas as a sequence of digits where similar codes are located closer together than codes that are different.

Our approach to this problem relied on the use of the search tool Nominatim (https://nominatim.org/), which we used to extract coordinates based on the entities provided in the testing set that could not be found in the training set. Lots of these entities included the place's abbreviation inside parentheses (e.g. *Hospital Central de la Defensa (HCD)*), which could make the search harder for the tool. For this reason, we split these entities and made the search with both strings separately (*Hospital Central de la Defensa* / *HCD*). Then, all the entities successfully found in Nominatim were encoded with the OpenLocationCode library (https://github.com/google/open-location-code) in order to obtain its PlusCode.

For those entities that could not be found, we applied a Spacy (https://spacy.io/) pipeline to get sub-named entities that could be useful to extract some geopolitical information (e.g. *Hospital General Universitario de Valencia* should extract *Valencia*, which is a city in Spain) and then assigned the codes of such entities as an approximation to the original ones.

Any entity that remained unlabelled was assigned the default code of Madrid's Kilometre Zero (8CGRC78W+MF), which marks the geographical center of Spain, as it is the country where most of the missing entities were from.

3.2.3. Normalization to SNOMED-CT

In order to provide SNOMED-CT codes to the test entities that could not be extracted from the training set, we developed a system similar to the one described in subsection 3.2.1. First, we looked for string matches against the gazetteer provided by the organizers, which consisted of about 27000 different terms. Then, for those entities that did not appear exactly in the ontology, we extracted their embeddings with the cross-lingual language model SapBERT-XLMR [21], trained with UMLS [22] 2020AB using XLM-RoBERTa [23] as the base model. We also calculated the embeddings for all the terms in the gazetteer and then looked for the closest one for each entity by calculating the Euclidean distance between them.

3.3. Sub-task 3: Entity Classification

We approached named Entity classification sub-task as a multiclass sentence classification problem. This approach was possible due to the fact that in the whole training dataset none of the GPE, GEO and FAC entities fell within the same sentence. We performed a fine-tuning of the same pre-trained model of RoBERTa architecture for the sequence classification task by adding a classification implemented in the Huggigface's transformers Python library [24]. Following the same logic as in the NER case, we performed 5 trials of hyperparameter optimization within an identical search space with only one modification: batch size could be chosen between 8, 16 and 32. We also used the early stopping strategy to prevent overfitting. The selected parameters can be seen in the table 2

Table 2
Hyperparameter search results for the optimization of the RoBERTa sequence classification model

Parameter	Selected value
Learning rate	$4, 2e - 5$
Training batch size	32
Weight decay	$3.e - 4$
AdamW ϵ	$5.85e - 7$
Warmup steps	90

The optimization was performed on a single NVIDIA A100-SXM4-40GB and took 1 hour 27 minutes to complete.

3.4. Sub-task 4: End-to-end Evaluation

For the end-to-end evaluation we presented all the sentence-level NER systems described in Section 3.1. Regarding the second sub-task, the systems developed were the same as the ones described in Section 3.2 with a slight difference in the SNOMED-CT track given that we had a few days more to develop our systems for the previous sub-task. This difference relied on the last step of our matching process where, instead of finding the closest embedding in the gazetteer provided by SapBERT, we implemented the same fuzzy matching technique we used in the GeoNames track. Finally, named entity classification was performed as described in Section 3.3.

A total of 3 runs were submitted for this sub-task, for each one of the sentence-level NER systems: fine-tuned RoBERTa (run4-1), Bi-LSTM classifier (run4-2) and the ensemble of Bi-LSTM classifiers (run4-3).

4. Results

4.1. Sub-task 1: Location Entity Recognition

The metric selected for official evaluation of NER systems are the ones that are commonly used to report the performance of token classification systems: precision, recall and F1-score. Table 3 summarizes the score obtained by all 5 presented systems during the official evaluation on the held-out test data subset.

Table 3
Official evaluation results for the NER sub-task

System	Precision	Recall	F1-score
run1	0.8162	**0.8458**	0.8307
run2	0.8421	0.8342	0.8381
run3	0.8373	0.8111	0.8240
run4	**0.8639**	0.8391	**0.8512**
run5	0.8503	0.8208	0.8353

As it can be inferred from the results, sentence-level classifiers (run1, run2 and run4) are the ones that achieve better performance than the text-level (run4 and run5), which suggest that despite of the absence of a-priory input length restrictions the implemented recurrent classifiers suffer from vanishing gradient problem when dealing with longer input sequences. The best performance in terms of the reference metric of this competition, F1-score, was achieved by the ensemble of recurrent classifiers that operated on the sentence level. This fact highlights the benefits of our approach to nested entity recognition. Notably, in terms of recall, the best performing system is the RoBERTa based sentence-level token classifier.

4.2. Sub-task 2: Geographic Normalization

The normalization subtask required to develop three different systems, one per each coding source as described in section 2: GeoNames, PlusCodes and SNOMED-CT.

In the first one, 1583 manually annotated entities were provided by the organizers in the training set while the testing set contained 510 mentions to be annotated by our system. 302 of them were exactly found on the training set, 156 out of the remaining ones were matched literally with some gazetteer term and 43 codes were assigned with the TF-IDF fuzzy matching technique. The last 9 entities were coded using the Levenshtein distance against the ontology. This system achieved an accuracy of 0.7308 as reported by the organizers.

The PlusCodes track was composed of 344 entities for the training set and 94 for the testing one. Only seven out of the 94 were literally detected in the training set while 61 were found in Nominatim. From the remaining entities, 8 were found with the Spacy's approach and the rest were labelled with the default code. An accuracy of 0.3261 was obtained for this approach.

Regarding the last dataset, which was released with 5066 and 2074 entities in train and test sets respectively, 1582 mentions were matched in the first step. Only 71 were literally found in the SNOMED-CT gazetteer and the remaining 421 entities were matched to the closest embedding in the ontology. This system reported an accuracy of 0.7819 according to the results provided by the organizers.

4.3. Sub-task 3: Entity Classification

The metric used to evaluate the performance of the entity classification system in this competition is micro-averaged accuracy. The presented system achieved a promising result of 0.7694. Table 4 summarizes the performance of the presented system per each of the classes. The lower performance on detecting the MOVIMIENTO and RESIDENCIA labels can be related to the scarcity of training examples, which, however, is not the case for the LUGAR_NATAL label that presents the summed up the smallest amount of examples amount the training data without being this an impediment for relatively good model performance.

4.4. Sub-task 4: End-to-end Evaluation

As can be seen in the Table 5, the best performance on the end-to-end evaluation was achieved by the third model that gives the highest score on the NER sub-task - the ensemble of two recurrent classifiers that take as input contextual embeddings from the RoBERTa model.

Table 4
Metrics scored by the presented system per each of the classes in the sub-task 3

Label	Micro-avg accuracy
OTHER	0.7796
MOVIMIENTO	0.64
ATENCION	0.8426
RESIDENCIA	0.6931
LUGAR-NATAL	0.8053

Table 5
Official evaluation results for the sub-task 4

System	Task-1	Task 2.1	Task 2.2	Task 2.3	Task 3	Avg
run4-1	0.8307	0.6549	0.1915	0.6820	0.6290	0.5976
run4-2	0.8381	0.6601	0.2041	0.6889	0.64475	0.6072
run4-3	**0.8512**	0.6660	0.2660	0.6910	0.6619	**0.6272**

5. Conclusions and future work

This paper covers the participation of the SINAI team in the MEDDOPLACE shared task at the IberLEF2023 that focuses on the detection, normalization and classification of different kinds of places and related types of information such as nationalities or patient movements in medical documents in Spanish.

Our team took part in all four subtasks. For the named entity recognition challenge, we experimented with training recurrent classifiers that took as input contextual embeddings extracted from the last 4 layers of a RoBERTa architecture language model pre-trained on a combination of biomedical and clinical corpora. In order to make the resulting system capable of recognizing nested sentences we trained two separate recurrent classifiers: one that would recognize more general named entities such as geographical and geopolitical mentions as well as sociodemographic information; and the other one that would recognize more domain-specific mentions such as facilities, departments within hospitals and means of transport. When trained to perform sentence-level token classification, this system outperforms the baseline set by fine-tuning the same RoBERTa language model for the same task and overall shows the best result (0.8512 of F1-score) out of all presented systems.

Regarding the second sub-task, we developed a combination of exact string matching with transformer-based embeddings, TF-IDF character-based n-gram and Levenshtein distances that reported good results in terms of accuracy for the first and third tracks (0.7308 and 0.7819 accuracy respectively). However, the metrics reported in the second one (0.3261) suggest the need for an in-depth analysis of the results predicted for this track that might be caused by the scarce matching against the training set.

For the entity classification sub-task we developed a system based on fine-tuning the same RoBERTa language model employed in the first subtask, that assigns labels to the entities after processing the sentence in which the entity is detected. This approach achieved 0.7694 micro-average accuracy.

For the end-to-end classification we presented three versions of pipelines that differed from each other in the NER component: each pipeline corresponding to each sentence-level NER system presented for the first sub-task. For the normalization part of this sub-task, we developed a very similar system to the one presented in the second sub-task, where the only difference relied on the replacement of the fuzzy TF-IDF matching with a transformer approach to find the remaining unmatched entities for the SNOMED-CT part. The best performing pipeline reached 0.6272 of micro-averaged F1-score.

Regarding future work, a more thorough performance and error analysis is required to identify the limitations of the implemented systems.

Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

[1] S. R. Kundeti, J. Vijayananda, S. Mujjiga, M. Kalyan, Clinical named entity recognition: Challenges and opportunities, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 1937–1945.

[2] A. M. Scott, Automatic coding of a diagnosis, in: G. McLachlan, R. Shegog (Eds.), Computers in the Service of Medicine, volume 2, Oxford University Press London, 1968, p. 89.

[3] R. M. Murphy, J. E. Klopotowska, N. F. de Keizer, K. J. Jager, J. H. Leopold, D. A. Dongelmans, A. Abu-Hanna, M. C. Schut, Adverse drug event detection using natural language processing: A scoping review of supervised learning methods, Plos one 18 (2023) e0279842.

[4] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, A. Valencia, Overview of the protein-protein interaction annotation extraction task of biocreative ii, Genome biology 9 (2008) 1–19.

[5] S. Kaewphan, K. Hakala, F. Ginter, Utu: Disease mention recognition and normalization with crfs and vector space representations, in: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), 2014, pp. 807–811.

[6] G. K. Savova, E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris, H. Hochheiser, C. Lin, G. Chavan, et al., Deepphe: a natural language processing system for extracting cancer phenotypes from clinical records, Cancer research 77 (2017) e115–e118.

[7] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[8] 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 (2020) 50–70.

[9] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, arXiv preprint arXiv:2109.03570 (2021).

[10] S. Lima-López, E. Farré-Maduell, V. Brivá-Escalada, L. Gascó, M. Krallinger, MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts, Procesamiento del Lenguaje Natural 71 (2023).

[11] S. L. López, E. F. Maduell, V. B. Iglesias, L. G. Sánchez, M. Krallinger, MEDDOPLACE Guidelines (Spanish): Annotation, Normalization and Classification of Locations and Related-Information in Medical Documents, 2023. URL: https://doi.org/10.5281/zenodo.7775235. doi:10.5281/zenodo.7775235.

[12] M. Chizhikova, J. Collado-Montañez, P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Sinai at livingner shared task 2022: Species mention recognition and normalization using transfer learning and string matching techniques (2022).

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[14] J. Li, T. Wang, W. Zhang, An improved chinese named entity recognition method with tb-lstm-crf, in: 2020 2nd Symposium on Signal Processing Systems, 2020, pp. 96–100.

[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[16] P. López-Úbeda, M. C. Díaz-Galiano, M. T. Martín-Valdivia, L. A. U. López, Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings., in: IberLEF@ SEPLN, 2020, pp. 324–334.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[19] S. Merity, N. S. Keskar, R. Socher, Regularizing and optimizing lstm language models, arXiv preprint arXiv:1708.02182 (2017).

[20] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.

[21] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: Proceedings of ACL-IJCNLP 2021, 2021, pp. 565–574.

[22] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology., Nucleic Acids Res. 32 (2004) 267–270. URL: http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04.

[23] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116.

arXiv:1911.02116.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.