Fusion @ BioASQ MedProcNER: Transformer-based Approach for Procedure Recognition and Linking in Spanish Clinical Text

Notebook for the BioASQ Lab at CLEF 2023

Sylvia Vassileva[1,*], Georgi Grazhdanski[1], Svetla Boytcheva[1,2] and Ivan Koychev[1]

[1] *Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria*
[2] *Ontotext, Sofia, Bulgaria*

Abstract

The paper presents an approach for solving subtasks 1 and 2 from the MedProcNER shared task, part of the BioASQ challenge - detecting and linking procedures in Spanish medical documents. The system consists of separate named entity recognition and entity linking modules and uses different BERT-based models for each module. For the NER subtask, we pre-train a Spanish RoBERTa model with additional Spanish clinical data and fine-tune the model for token classification. After the entities are detected, they are passed to the entity linker which uses the cross-lingual SapBERT XLMR-large to generate entity and mention representations and generates the candidate entity using cosine similarity. When evaluated with the test set, our system shows 0.71 F1 score on the NER subtask and 0.53 F1 score on the linking subtask.

Keywords

Named entity recognition (NER), Entity linking, Biomedical NLP, Clinical terms extraction, Clinical terms linking, Spanish clinical NER

1. Introduction

This paper presents our approach to BioASQ Task 3 - MedProcNER: Spanish Medical Procedures Named Entity Recognition, Linking, and Indexing Shared Task ([1], [2]). The challenge focuses on the concept of a clinical procedure - *'a set of actions, interventions, or treatments that are carried out by healthcare professionals to diagnose, treat, or manage a patient's medical condition'*[1]. Automatic identification and normalization of clinical procedure mentions in unstructured text is crucial for knowledge discovery, facilitating clinical research, building, and integrating systems that ultimately improve the quality of provided healthcare. There are three subtasks[2]:

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.
svasileva@fmi.uni-sofia.bg (S. Vassileva); ggrazhdans@uni-sofia.bg (G. Grazhdanski); svetla@uni-sofia.bg (S. Boytcheva); koychev@fmi.uni-sofia.bg (I. Koychev)
0000-0002-2257-0659 (S. Vassileva); 0000-0002-5542-9168 (S. Boytcheva); 0000-0003-3919-030X (I. Koychev)

[1] Clinical procedure definition by the MedProcNER task.
[2] MedProcNER subtask descriptions.

- **Clinical Procedure Recognition** - a named entity recognition task where mentions of clinical procedures must be identified in unstructured Spanish clinical case texts.
- **Clinical Procedure Normalization** - assigning SNOMED CT codes to the clinical procedure mentions identified in the NER subtask.
- **Clinical Procedure-based Document Indexing** - assigning SNOMED CT codes to the full clinical report texts, so that they could be indexed.

We take part in the first two subtasks - Clinical Procedure Recognition, and Procedure Normalization. The system performs named entity recognition for procedures using a pre-trained Spanish RoBERTa model [3], and entity linking using cross-lingual SapBERT XLMR-large [4]. We perform several different experiments to compare the performance of different clinical BERT-based language models, including models trained in English, Spanish, or cross-lingual. We investigate the effect of additional pre-training of the Spanish RoBERTa model using task-specific Spanish data from the procedures gazetteer, as well as the effect of input preprocessing, and we find performance improvements when using both. We also attempt to train an Adapter over the BioM-ALBERT-Large model, which will adapt the English model to Spanish, however, the results shown are lower than the fully trained Spanish RoBERTa model.

The code related to this task is available on GitHub[3].

2. Related Work

2.1. Subtask 1 - Named Entity Recognition

Named entity recognition (NER) is a fundamental subtask in multiple NLP challenges for Spanish biomedical and clinical texts. Deep neural models are predominantly used to tackle the NER task. In the DisTEMIST [5] task, the best-performing NER model, developed by Moscato et al. [6], is based on PlanTL-GOB-ES/roberta-base-biomedical-clinical-es [3], with a classification head on top. Xiong et al. [7] view the NER task as a machine reading comprehension (question answering) problem by using the definition of the entity as the question, and part of the clinical text as the segment, and training a joint model for the NER and entity linking subtasks. Xiong et al. [8] achieve the best NER result in the PharmaCoNER [9] task using a BERT-based [10] model and conditional random fields.

2.2. Subtask 2 - Entity Linking

Entity linking (EL) is a crucial task for extracting structured data from clinical texts. It usually relies on the result from NER and consists of two steps - generating candidates from a knowledge base of entities, and ranking and selecting the best candidate. The EL approach can take advantage of the global context information to inform the linking decisions based on neighboring linked entities.

A common approach used for biomedical entity linking is using deep neural networks to generate entity representations in an embedding space and searching for the closest entities based on a distance metric like cosine similarity. SapBERT and cross-lingual SapBERT were

[3] https://github.com/svassileva/clef2023-medprocner

pre-trained using UMLS and have shown very good results for linking for both English and other languages ([11], [4]). In the case of Spanish biomedical entity linking, the best model on the DisTEMIST dataset used an ensemble of cross-lingual SapBERT and TF-IDF vectorizer based on character n-gram features for candidate generation which scored F1 0.5657 [12]. Other participants in the competition used FastText embeddings and approximate nearest neighbor similarity and scored F1 0.4987 [13]. Due to the lack of large corpora of biomedical training data in many languages, approaches using exact match and surface form similarity are still used, for example, several papers in the CanTEMIST competition used the Levenshtein distance metric ([14], [15]).

In some cases, EL and NER can be combined and solved as sequence labeling task, as long as the entity identifier is composed of different hierarchical components, like the tumor codes (ICD-O-3) in CanTEMIST. The highest scoring systems in CanTEMIST use sequence labeling and predict the different tumor code components for each word ([16], [17]).

3. Data

3.1. MedProcNER Dataset

The MedProcNER corpus [18] contains 1,000 clinical case reports in Spanish annotated with procedure mentions and normalized to SNOMED CT codes. The corpus consists of a fully annotated train set, a smaller test set, as well as a gazetteer of SNOMED-CT codes and different aliases. Statistics about the train and test sets are shown in Table 1.

The train set contains 4,857 annotated entities, 1,630 unique entity codes, 3,086 unique entity mentions, 106 nested mentions, and 510 abbreviations. The majority of entities have only one SNOMED-CT code, but there are 51 mentions which have no code selected, and 125 which have more than one code. There are very few ambiguous mentions which are labeled with different codes based on the context - a total of 12.

The Spanish SNOMED-CT gazetteer contains a total of 234,675 aliases for terms in multiple categories, including procedures, substance, clinical drug, and others. The most important category for this task is procedures. The gazetteer contains 61,714 unique SNOMED-CT codes of procedures and 94,133 total different aliases for procedure terms. The average number of procedure aliases per code is 1.52.

The corpus has a test set of 250 documents for testing purposes and the annotated version is not public as of the moment of writing this paper.

Table 1
MedProcNER dataset metrics.

Metric	Train	Test
Documents	750	250
Sentences	11,884	3,986
Tokens	323,192	107,551

3.2. Language Pre-training Dataset

We experimented with two different pre-training datasets:

3.2.1. MedProcNER Gazetteer Dataset

The dataset consists of all of the terms in the initial version of the MedProcNER gazetteer, each as a separate example. This was the dataset of choice for our pre-trained PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model submission.

3.2.2. Custom Spanish Medical Procedure Dataset

This dataset was compiled to include clinical procedure term descriptions from CIE-10-ES 2022 Procedimientos Tabla de Referencia[4], and all term descriptions from the Spanish release of SNOMED CT [19] which contain ’procedimiento’ and do not contain ’RETIRADO’ (deprecated). It also includes all sentences from the MedProcNER train and test files, and all unique values in the *COMPONENT*, *SYSTEM*, and *METHOD_TYP*[5] columns of the Spanish (Spain) linguistic variant of LOINC v2.74. Since this set is larger than the gazetteer one, it is only used for training a language adapter for one of the models - a less computationally expensive alternative to full pre-training.

3.3. Unified Medical Language System Procedures (UMLS) Procedures

For the purposes of entity linking, we have extracted all procedures from UMLS (2023AA edition) [20] which are in Spanish, have a Spanish SNOMED CT code, and that code is part of the gazetteer procedure codes. We extracted 80,681 procedure aliases from UMLS, however, about half of them were already present in the MedProcNER gazetteer. The number of new aliases extracted from UMLS was 45,010.

We construct the entity-linking knowledge base by combining the UMLS and the gazetteer procedures. The total number of procedure aliases in the knowledge base is 125,691.

4. Methods

The system performs named-entity recognition and entity linking of procedures in Spanish clinical texts. The overall system architecture is shown on Figure 1 and consists of separate steps for named-entity recognition and entity linking of the procedures.

After preprocessing, the NER module identifies the spans in the text which contain procedures and they are passed to the Entity Linking module to predict their SNOMED-CT identifiers. As a result, a list of procedure spans and their corresponding identifiers is returned.

[4] CIE-10-ES 2022 Procedimientos Tabla de Referencia

[5] LOINC Table Structure

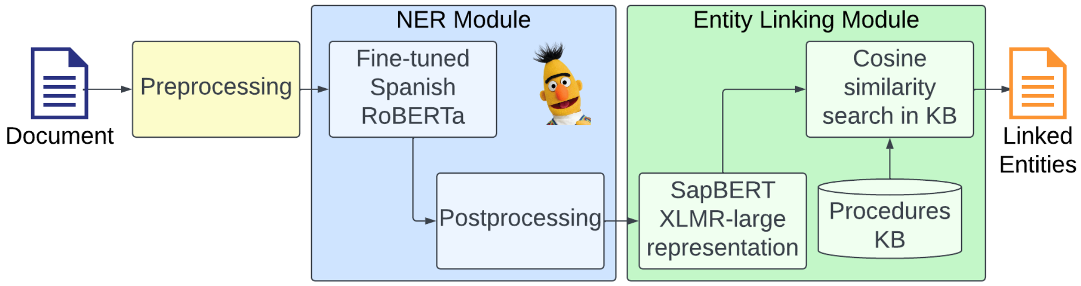

Figure 1: End-to-end architecture of the system - the preprocessed clinical case text is passed to a token classifier. Classified tokens are merged into the clinical procedure mentions in the postprocessing step. Finally, an entity linker uses the identified mentions as input and assigns the corresponding SNOMED CT codes.

BERT image: https://www.iconspng.com/image/146543/sesame-street-bert-standing

4.1. Preprocessing

Due to the fact that more than 33% of the clinical case texts exceed the 512 token limit of our models, we split the texts into sentences, and use each sentence as a separate example for fine-tuning. We use the SPACCC Sentence Splitter [6]. We also replace any number with the ’NUMBER’ literal, and remove any punctuation, special characters, and symbols matching the following regex:

```
[~:\’+\[\\@^{%(\-"*|,&<‘}._=\]!>;?#$)/®©℠]
```

4.2. Subtask 1 - Named Entity Recognition

For the NER subtask, we fine-tune a large transformer model with a classification head on the preprocessed MedProcNER train dataset, postprocessing the results to reconstruct the final procedure from tokens. The classification head labels correspond to the 3 classes in the IOB2 annotation scheme [21].

The process of pre-training the model using the MedProcNER gazetteer and then fine-tuning it on the MedProcNER train set is shown on figure 2.

4.2.1. Language Model Pre-training

We pre-train the model using the masked language model objective and the standard BERT architecture [10] for two epochs using the HuggingFace Transformers library [7]. We use the default value for the probability of masking a token - 15%. The MedProcNER gazetteer dataset from Section 3.2 is used to pre-train the model on Spanish medical procedure vocabulary. Table 2 shows the hyperparameter values used for pre-training. They are the same as the hyperparameters used for pre-training RoBERTa (base) in the original paper [22].

The resulting pre-trained language model is used for the downstream NER task.

[6] SPACCC Sentence Splitter

[7] HuggingFace Transformers Library

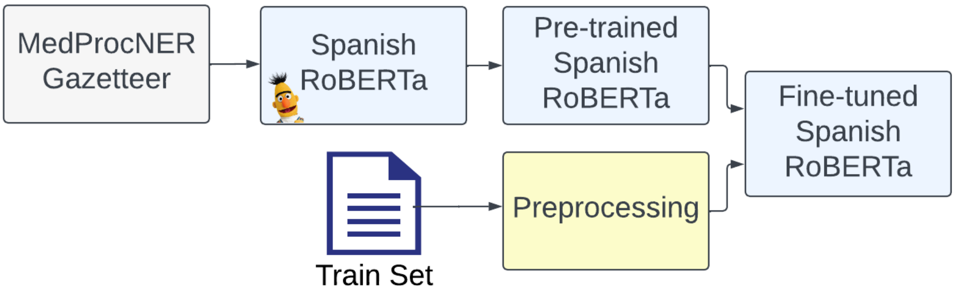

Figure 2: Language model pre-training using the Gazetteer terms pre-training dataset from Section 3.2 and subsequent fine-tuning.

BERT image: https://www.iconspng.com/image/146543/sesame-street-bert-standing

Table 2
Hyperparameter values used during pre-training of Spanish RoBERTa-Biomedical-Clinical.

Learning Rate	Weight Decay	Adam ϵ	Adam β_1	Adam β_2	Warmup Ratio	Batch Size	Gradient Accumulation Steps	Epochs
6e-4	0.01	1e-6	0.9	0.98	0.048	80	100	2

4.3. Classification Model Selection

We experiment with the following models for the token classifier:

- **PlanTL-GOB-ES/roberta-base-biomedical-clinical-es** [3] - a RoBERTa-based language model, trained on a large Spanish biomedical-clinical corpus of more than 1B tokens. Systems based on this model have achieved competitive results on previous Spanish biomedical-clinical tasks. We further pre-train the model on the **Gazetteer** dataset.
- **CLIN-X-ES** [23] - a cross-lingual language model, based on XLM-RoBERTa (large), pre-trained on the MeSpEN [24] dataset, and Spanish clinical documents from the Scielo archive[8]. We use CLIN-X-ES to explore the cross-lingual knowledge transfer capabilities of a *multilingual* model in the clinical domain. No additional pre-training is done.
- **BioM-BERT-Large** [25] - a BERT-based model (with ELECTRA architecture) for the biomedical domain, pre-trained exclusively on an English corpus - PubMed Abstracts + PMC + general domain vocab (EN Wiki + Books). It achieves state-of-the-art results on certain bio text classification tasks such as ChemProt. With no further pre-training, BioM BERT serves as a good baseline, proving that cross-lingual knowledge transfer in the biomedical-clinical context is possible to a certain extent even with a monolingual model, likely due to the Latin and Greek etymology of clinical terms.
- **BioM-ALBERT-Large** [25] - an ALBERT-based model, pre-trained on English PubMed abstracts only. We train a language adapter with the Multiple ADapters for Cross-lingual transfer (MAD-X) architecture which has shown very good results for cross-lingual

[8] Scielo archive

adaptation [26]. We train the adapter layers on the MLM objective using the Custom Spanish Medical Procedure dataset from Section 3.2 for 1 epoch. The core idea is to augment the original model with a new set of parameters, by inserting bottleneck feed-forward layers in each layer of the transformer model, and only train these new parameters on the masked language modeling objective for Spanish. This not only speeds up the training process, but also improves modularity and reusability, as multiple adapters can be composed for different tasks, and the adapter parameters are saved and shared separately. Table 3 shows the hyperparameter values used when training the adapter. We use the default value for masking probability - 15%.

Table 3
Hyperparameter values used during training of a language adapter for BioM-ALBERT-Large.

Learning Rate	**Weight Decay**	**Adam** ϵ	**Adam** β_1	**Adam** β_2	**Warmup Ratio**	**Batch Size**	**Gradient Accumulation Steps**	**Epochs**
1.76e-3	0	1e-8	0.9	0.999	0	32	256	1

4.3.1. Postprocessing

Since our models are token classifiers, to get the final clinical procedure mentions in the required format for the NER task, we must merge the token class predictions into word class predictions. To do this, we use the aggregation capabilities of the Hugging Face token classification pipeline[9], setting the aggregation strategy to 'first'. This groups the tokens of a word, while preserving word boundaries. For instance, *'resonancia magnética'* may be tokenized as *'resonancia'*, *'magnét'* and *'##ica'*. These three tokens may be classified as *B-PROCEDIMIENTO*, *I-PROCEDIMIENTO*, and *O*. The chosen aggregation strategy would override the classification of the last token in order to preserve word boundaries when merging. As a result, *'resonancia magnética'* (and not *'resonancia magnét'*) is recognized as a clinical procedure mention[10]. This approach might be beneficial for models pre-trained exclusively on an English corpus since it is more likely that they misclassify suffixes of out-of-vocabulary words. Note this could only be applied to word-based models, where there is the notion of a word boundary. For the RoBERTa Biomedical-clinical model and the BioM-ALBERT-Large model, we include an additional postprocessing step - removing relatively rare leading and trailing whitespace characters and unclosed parenthesis occurrences which are a result of the preprocessing, e.g. ' PAAF)' is transformed into 'PAAF'.

4.4. Subtask 2 - Entity Linking

For the task of Entity Linking, we construct a knowledge base (KB) using the MedProcNER train set and gazetteer described in Section 3.1, which contains procedure SNOMED-CT codes from the task gazetteer and all their aliases from the train set. For all entries in the KB, we create representations using the SapBERT XLMR-large model [4]. The SapBERT XLMR-large model was pre-trained on UMLS data for all available languages including Spanish using a

[9] Hugging Face token classification pipeline
[10] Another example of using the 'first' strategy from the Hugging Face documentation.

self-alignment objective so that phrases that represent a concept in different languages are grouped closely in the embedding space. Spanish is the second most common language in UMLS 2020AB and the dataset used to pre-train the model consists of 10.7% Spanish terms. The model achieves 56.4% F1 score on the Spanish version of the WikiMed dataset [4].

We create an entity representation for the input span and use cosine similarity search in the knowledge base. The system selects the entity with the highest cosine similarity. The implementation uses the NeMO library [11] adapted to use the SapBERT XLMR-large model. The NeMO library implements the approach for entity linking described by Liu et al [11]. It uses models pre-trained using the self-alignment objective like SapBERT and generates linking candidates using cosine similarity search. Future enhancements of the system can support predicting the NIL object (i.e. NO_CODE), or multiple predictions.

5. Experiments and Results

5.1. Train and Validation Datasets

To form the training and validation sets for fine-tuning, we first split all of the clinical case texts in the MedProcNER train dataset, then 80% of the sentences are randomly selected for training, and the remaining 20% are used for validation.

5.2. Evaluation Metrics

We used micro-averaged precision, recall, and F1-score as metrics for both subtasks.

Since no official evaluation library was provided, we modified the DisTEMIST evaluation script by replacing the gazetteer with the MedProcNER gazetteer, and updating the entity type to 'PROCEDIMIENTO' [6].

5.3. Subtask 1 - Named Entity Recognition

5.3.1. NER Model Performance

Table 4
Subtask 1 results on the validation and test sets.

Model	Val Precision	Val Recall	Val F1	Test Precision	Test Recall	Test F1
Pre-trained Spanish RoBERTa	0.6944	**0.6887**	**0.6915**	**0.7165**	**0.7143**	**0.7154**
CLIN-X-ES	**0.6985**	0.6841	0.6912	0.7047	0.6916	0.6981
BioM-BERT Large	0.6581	0.6317	0.6447	0.6894	0.6599	0.6743
BioM-ALBERT Large + Adapter	0.6703	0.6098	0.6387	0.6928	0.6264	0.6580

The results of the different classifiers on the validation set and the test set are presented in Table 4. Being a monolingual model, highly specialized in the clinical context, the Spanish Biomedical-Clinical RoBERTa outperforms the rest of the models on both datasets, in terms of

[11] NeMO library

F1. The multilingual CLIN-X-ES shows competitive performance as well. Interestingly, despite being trained exclusively on English texts, both BioM models perform relatively well on the test set, with BioM-BERT Large having a slight lead, perhaps because it is trained on a more diverse dataset than the BioM-ALBERT. Adam optimizer parameter values (ϵ, β_1, and β_2) are the same for all models - the HuggingFace trainer defaults - 1e-8, 0.9, 0.999 respectively.

5.3.2. Effect of Further Pre-training

We further pre-trained the Spanish RoBERTa-Biomedical-Clinical for two epochs on the Gazetteer terms dataset. This led to a stable performance improvement, as it conditioned the model to the target clinical procedure terms that may appear in the clinical case texts. Table 5 compares the performance on the validation set, with and without pre-training.

Table 5
Comparison of the performance of Spanish RoBERTa-Biomedical-Clinical on the validation set, with and without additional pre-training.

Model	Val Precision	Val Recall	Val F1
Spanish RoBERTa	0.688	0.6755	0.6817
Pre-trained Spanish RoBERTa	**0.6944**	**0.6887**	**0.6915**

5.3.3. Effect of Preprocessing

Preprocessing the clinical report texts by removing special characters, and replacing numbers with the 'NUMBER' literal, has a positive impact on the performance of the RoBERTa-Biomedical-Clinical model, allowing it to better generalize, ignoring dates, patient age, and other numeric mentions that, in the context of a clinical procedure, only add noise. On the other hand, Roman numerals are preserved, as they are part of some disease names (e.g. diabetes tipo II). Table 6 shows that preprocessing improves recall without significantly affecting precision.

Table 6
Effect of preprocessing on the performance of Spanish RoBERTa-Biomedical-Clinical on the validation set.

Model	Val Precision	Val Recall	Val F1
Spanish RoBERTa	**0.6897**	0.6513	0.6699
Spanish RoBERTa + Preprocessing	0.6880	**0.6755**	**0.6817**

5.3.4. Language Adapter Applicability

In an attempt to allow BioM models, which are trained exclusively on English data, to adapt to the specifics of the Spanish language without causing catastrophic forgetting of their representations of biomedical terms, we pre-trained a language adapter for the BioM-ALBERT model. Adapters [27] offer a less computationally expensive alternative to traditional full pre-training and fine-tuning. It involves introducing a small number of new parameters (relative to the size of the original model), and training just the new ones, while keeping the rest of the model frozen.

Table 7
Adapted BioM-ALBERT-Large performance compared to base model on the validation set.

Model	Val Precision	Val Recall	Val F1
BioM-ALBERT-Large	0.6566	**0.6135**	0.6343
BioM-ALBERT-Large + Adapter	**0.6703**	0.6098	**0.6387**

Adapters have successfully been applied to biomedical NLP tasks [28]. Our adapter has the MAD-X language adapter architecture and was trained on the Custom Spanish Medical Procedure dataset with the masked language modeling objective for 1 epoch. Although this did not lead to a significant performance improvement (only affecting precision), it might be an approach worth exploring on a larger dataset, and training for more epochs. Table 7 compares the performance of adapted, and base BioM-ALBERT-Large models on the validation set.

Table 8
Hyperparameter values used during fine-tuning of the NER classifier models. RoBERTa values are identical to those used in the original paper [3]. CLIN-X-ES values are the same as the RoBERTa ones, apart from the batch size, as the model had to fit in memory. BioM-BERT-Large values are a result of a hyperparameter search using Tune [29]. BioM-ALBERT-Large values are from the original paper [25].

Model	Learning Rate	Weight Decay	Warmup Ratio	Batch Size	Gradient Accumulation Steps	Epochs
Spanish RoBERTa	5e-5	0	0	32	2	20
CLIN-X-ES	5e-5	0	0	16	2	20
BioM-BERT-Large	2e-5	0.01	0	16	0	3
BioM-ALBERT-Large	3e-5	0	0	16	0	4

5.4. Subtask 2 - Entity Linking

For the entity linking task, we perform an experiment to compare three different SapBERT flavors for generating the mention representation. We report the accuracy at @1, @5, @10, and @25. In order to test the Entity Linking independently, we take the annotated train set for subtask 1 and predict the SNOMED-CT code for each mention. The knowledge base used for this validation consists of the procedures from the gazetteer (Section 3.3), UMLS (Section 3.1), and the mentions from the train set which are not part of the validation set.

The results of the experiment for SapBERT [11], SapBERT XLMR [4], and SapBERT XLMR-large [4] are shown in table 9.

The original SapBERT has the lowest score, explained by the fact that it was trained only with English terms on PubMed texts, while the cross-lingual versions are trained on UMLS terms.

The best result @1 is achieved using SapBERT XLMR and SapBERT XLMR-large scores just 0.93% less than the smaller model which is not statistically significant (Chi-square = 0.0871, p=0.7678>0.05 for the 968 validation set samples). For the purposes of the final system, we decided to use the large model since it showed better results in the original paper [4].

The SapBERT XLMR-large model shows an accuracy @25 of approximately 0.84, so future enhancements of the system can use it as a candidate generator and use a separate ranking function to select the best candidate and increase the likelihood of finding the correct one.

Table 9
Subtask 2 SapBERT model comparison results.

Model	Acc@1	Acc@5	Acc@10	Acc@25
SapBERT	0.5805	0.6921	0.7128	0.7613
SapBERT XLMR	**0.6673**	0.7747	0.8016	0.8305
SapBERT XLMR-large	0.6580	**0.7851**	**0.8099**	**0.8388**

5.5. Overall System

We perform a comparison of the different NER models in the end-to-end system performance. We measure F1 for both the NER and EL tasks. We use two different dataset scenarios for comparison:

- Using a split of the train dataset - 80% of data used for training, 20% used for validation testing;
- Using the full train dataset for training and the test set for testing;

The results of the experiment on the validation dataset are shown in table 10, and table 11 shows the results from the test dataset evaluated by the organizers [2]. The highest score on the validation NER model is achieved using the BioM-BERT-Large model, while the highest EL score is achieved using the CLIN-X-ES + SapBERT XLMR-Large model. Overall, most of the models produce very close results and the top-scoring models change when evaluating on the test set. The best-performing model for both NER and EL on the test set is the Pre-trained Spanish RoBERTa which scores 0.71 on NER and 0.53 on the EL task.

The reason why the entity linking model shows lower results on the validation set may be explained by the fact that the KB it uses has all the validation mentions explicitly removed to account for a worst-case scenario. A more detailed comparison may be performed once the test set annotations become available.

Table 10
End-to-end comparison results on the validation dataset.

Model NER	Model EL	Validation F1 NER	Validation F1 EL
BioM-BERT-Large	SapBERT XLMR-Large	**0.7055**	0.4514
BioM-BERT-Large + Preprocessing	SapBERT XLMR-Large	0.6576	0.4644
CLIN-X-ES	SapBERT XLMR-Large	0.6912	**0.4845**
Pre-trained Spanish RoBERTa	SapBERT XLMR-Large	0.6915	0.4838
BioM-ALBERT Large + Adapter	SapBERT XLMR-Large	0.6387	0.4402

Table 11
End-to-end comparison results on the test dataset.

Model NER	Model EL	Test F1 NER	Test F1 EL
BioM-BERT-Large	SapBERT XLMR-Large	0.6769	0.5293
BioM-BERT-Large + Preprocessing	SapBERT XLMR-Large	0.6743	0.5216
CLIN-X-ES	SapBERT XLMR-Large	0.6981	0.5283
Pre-trained Spanish RoBERTa	SapBERT XLMR-Large	**0.7154**	**0.5369**
BioM-ALBERT Large + Adapter	SapBERT XLMR-Large	0.6580	0.51870

5.6. Error Analysis

5.6.1. Subtask 1 - Named Entity Recognition

In this section, we look into the errors of our best NER model (a further pre-trained Spanish RoBERTa-Biomedical Clinical) on the validation set. The following types of errors are identified:

- **Redundant context** - for short mentions, instead of only identifying the mention itself, the model outputs a phrase containing the mention, thus adding unnecessary context. For example, in the sentence *'Reacción de Mantoux negativa.'* our model outputs the phrase *'Reacción de Mantoux'* instead of the ground truth - *'Mantoux'*. This is also the case when multiple separate mentions are listed as one, e.g. *'Mantoux, electrocardiograma (ECG), radiografía de tórax...'* are identified as a single mention, instead of 3 separate ones.
- **Partitioned mentions** - parts of a single true procedure mention are identified as separate mentions. For example, in the sentence *'El paciente se intervino quirúrgicamente, realizándose una cistoprostatectomía radical con linfadenectomía más derivación tipo Indiana y cierre del defecto de pared abdominal con fascia lata.'* our model predicts two separate mentions *'cistoprostatectomía radical'* and *'linfadenectomía'*, instead of the truth - *'cistoprostatectomía radical con linfadenectomía'*. Overall, the model has a difficulty identifying long clinical procedure mentions, such as surgery descriptions.
- **Nested mentions are not identified** - this is an inherent problem of the token classifier, where each token is tagged with *only* one of {B, I, O}. Consequently, a mention that is part of another mention, is not identified. For instance, the following mention *'tomografía axial computarizada (TAC) abdómino-pélvica'* contains a nested mention *'TAC'* that is not recognised by our model. Since about 2% of the training data consists of nested mentions, the model is not able to recognize them correctly.

5.6.2. Subtask 2 - Entity Linking

As the linking module uses the results from the NER module, successful linking is limited by the correctly identified mentions. Several factors contribute to entity-linking errors:

- Zero or multiple codes matching the mention - the model supports the prediction of a single entity from the knowledge base, so mentions that should be linked with zero or multiple codes will not be predicted correctly. The validation set contains only 30 mentions which correspond to NO_CODE or multiple codes, accounting for approximately 3% of the validation set errors.

- Missing entities in the KB cannot be predicted by the model. About 5.5% of the mentions in the validation set do not have a record in the validation KB for the same entity code.

A common error is suggesting an entity that is close to the target entity like a parent or entity term which has common (sub)-words, as shown in table 12. For example, when the query is "somatometría *(somatometry)*", the system predicts the code for "medición de somatomedina *(somatomedin measurement)*" due to the common sub-word "somato".

Table 12
Examples of entity linking errors.

Query	Predicted code	True code	Predicted text	True text
estudio radiográfico *(radiographic study)*	363679005	363680008	procedimiento de diagnóstico por imágenes *(diagnostic imaging procedure)*	procedimiento radiográfico *(radiographic procedure)*
somatometría *(somatometry)*	104936007	54709006	medición de somatomedina *(somatomedin measurement)*	medición corporal *(body measurement)*
antibiótico *(antibiotic)*	58427002	281789004	medición de antibiótico *(antibiotic measurement)*	tratamiento con antibióticos *(antibiotic treatment)*

6. Conclusion

The proposed system shows satisfactory results on the named entity recognition, and entity linking subtasks of the MedProcNER challenge, exploring various approaches, with a focus on the impact of further pre-training and adapting on the performance of both monolingual (Spanish or English), and multilingual models. The system was evaluated on the test set and scored 0.71 F1 on the NER task and 0.53 F1 on the entity linking task. By comparison, the best system in the competition has shown an F1 score of 0.7985 on the NER task and 0.5707 on the entity linking task.

The conducted experiments suggest that further conditioning a specialized monolingual Spanish model via training on a focused dataset, such as one featuring gazetteer terms, leads to a stable performance improvement on the NER task. Large cross-lingual models proved to be an optimal choice for entity linking, achieving competitive results.

As further work, additional training data could be generated by replacing procedure mentions with synonyms. Also, introducing machine translation and alignment as a step in the NER pipeline could be beneficial, as it would allow utilizing state-of-the-art monolingual transformer models that are trained on English biomedical-clinical datasets. Last, but not least, it would be worth looking into alternative strategies for constructing entity mentions from token classification results, as the approach we employed, although simple, might be limiting in some cases (e.g. longer mentions). For the entity linking task further work can explore using the SapBERT XLMR-Large model to generate candidates and training a model to rank them. Also, exploring how to take advantage of the SNOMED-CT structure and entity relationships can be useful to provide additional information for linking.

Acknowledgments

This research is partially funded by Project UNITe BG05M2OP001-1.001-0004 funded by the OP "Science and Education for Smart Growth", co-funded by the EU through the ESI Funds, and partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[2] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[3] 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, 2021.

[4] 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.

[5] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022, pp. 179–203.

[6] V. Moscato, M. Postiglione, G. Sperlì, Biomedical spanish language models for entity recognition and linking at bioasq distemist, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022, pp. 315–324.

[7] Y. Xiong, Y. Huang, Q. Chen, X. Wang, B. Tang, A joint model for medical named entity recognition and normalization, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), 2020, pp. 499–504.

[8] Y. Xiong, Y. Shen, Y. Huang, S. Chen, B. Tang, X. Wang, Q. Chen, J. Yan, Y. Zhou, A deep learning-based system for pharmaconer, in: Proceedings of the 5th Workshop on

BioNLP Open Shared Tasks, Association for Computational Linguistics, 2019, pp. 33–37. URL: https://aclanthology.org/D19-5706. doi:10.18653/v1/D19-5706.

[9] 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.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[11] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238.

[12] F. Borchert, M.-P. Schapranow, Hpi-dhc @ bioasq distemist: Spanish biomedical entity linking with pre-trained transformers and cross-lingual candidate retrieval, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022, pp. 244–258.

[13] B. F. L. M.-R. Matic Bernik, Robert Tovornik, DiagÑoza: a natural language processing tool for automatic annotation of clinical free text with snomed-ct, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022, pp. 235–243.

[14] L. Lange, X. Dai, H. Adel, J. Strötgen, NLNDE at CANTEMIST: neural sequence labeling and parsing approaches for clinical concept extraction, CoRR abs/2010.12322 (2020).

[15] 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.

[16] Q. C. X. W.-Y. N. Ying Xiong, Yuanhang Huang, B. Tang, A joint model for medical named entity recognition and normalization, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020, pp. 499–504.

[17] M. C. Aitor García-Pablos, Naiara Perez, Vicomtech at cantemist 2020, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020, pp. 489–498.

[18] S. L. López, E. F. Maduell, L. G. Sánchez, M. Krallinger, MedProcNER/ProcTEMIST Corpus: Gold Standard annotations for Clinical Procedures Information Extraction, 2023. URL: https://doi.org/10.5281/zenodo.7929830. doi:10.5281/zenodo.7929830.

[19] International Health Terminology Standards Development Organisation (IHTSDO), SNOMED Clinical Terms (SNOMED CT), [2022]. URL: [https://confluence.ihtsdotools.org/display/RMT/October+2022+Spanish+edition+release].

[20] 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.

[21] V. Krishnan, V. Ganapathy, Named entity recognition, 2005.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[23] L. Lange, H. Adel, J. Strötgen, D. Klakow, iCLIN-x/i: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain, Bioinformatics 38 (2022) 3267–3274. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtac297. doi:10.1093/bioinformatics/btac297.

[24] M. Villegas, A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, M. Krallinger, The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations, in: LREC MultilingualBIO: Multilingual Biomedical Text Processing, ELRA, 2018, pp. 32–39.

[25] S. Alrowili, V. Shanker, BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 221–227. URL: https://www.aclweb.org/anthology/2021.bionlp-1.24.

[26] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7654–7673. URL: https://aclanthology.org/2020.emnlp-main.617. doi:10.18653/v1/2020.emnlp-main.617.

[27] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, Adapterfusion: Non-destructive task composition for transfer learning, 2021. arXiv:2005.00247.

[28] J. Papaioannou, P. Grundmann, B. van Aken, A. Samaras, I. Kyparissidis, G. Giannakoulas, F. Gers, A. Loeser, Cross-lingual knowledge transfer for clinical phenotyping, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 900–909. URL: https://aclanthology.org/2022.lrec-1.95.

[29] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv:1807.05118 (2018).