VICOMTECH at MedProcNER 2023: Transformers-based Sequence-labelling and Cross-encoding for Entity Detection and Normalisation in Spanish Clinical Texts

Elena Zotova[1,2,†], Aitor García-Pablos[1,†], Montse Cuadros[1,†] and German Rigau[2,3]

[1]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain

[2]Department of Languages and Computer Systems. University of the Basque Country (UPV-EHU), Paseo Manuel de Lardizabal, 1, Donostia/San-Sebastián, 20018, Spain

[3]HiTZ Basque Center for Language Technologies, Paseo Manuel de Lardizabal, 1, Donostia/San-Sebastián, 20018, Spain

Abstract

This paper describes the participation of the Vicomtech NLP team in the MedProcNER 2023 shared task about detecting mentions of procedures in clinical texts written in Spanish and normalising them to SNOMED CT codes. We participate in each of the three tasks, combining multiple approaches and strategies. For Task 1 (NER) we use a Transformer-based model to perform sequence labelling. For Task 2 (Normalisation) we use Semantic Text Search approaches to relate entity mentions to their codes. The solution for Task 3 (Indexing) is built on top of the two first tasks. For Task 1 our system obtained 77.96% of F1-score. Our approaches for Task 2 and Task 3 achieved the highest F1 scores in the official evaluation results—57.07% and 62.42%, respectively.

Keywords

Named Entity Recognition, Entity Linking, Entity Normalisation, Clinical Coding, Document Indexing, SNOMED CT

1. Introduction

This paper describes Vicomtech's participation in MedProcNER 2023 shared task [1], which is part of the BioASQ Workshop in the CLEF 2023 conference [2]. This challenge is focused on the detection, normalisation and indexing of clinical procedures in clinical documents in Spanish. It is split in three tasks:

- Task 1. Clinical Procedure Recognition. In this subtask, participants are challenged to automatically detect mentions of clinical procedures in clinical reports in Spanish. Using

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

*Corresponding author.

†These authors contributed equally.

ezotova@vicomtech.org (E. Zotova); agarciap@vicomtech.org (A. García-Pablos); mcuadros@vicomtech.org (M. Cuadros); german.rigau@ehu.eus (G. Rigau)

0000-0002-8350-1331 (E. Zotova); 0000-0001-9882-7521 (A. García-Pablos); 0000-0002-3620-1053 (M. Cuadros); 0000-0003-1119-0930 (G. Rigau)

the MedProcNER corpus as training data, they must build systems capable of retrieving the start and end position of clinical procedures mentioned in the text.

- Task 2. Clinical Procedure Normalisation. The challenge of this subtask is to automatically normalise mentions of clinical procedures in published clinical reports in Spanish. The proposed systems should assign SNOMED CT codes to the mentions retrieved in Task 1.
- Task 3. Clinical Procedure-based Document Indexing. The subtask aims to automatically assign clinical procedure codes to the full clinical case report texts. Using the MedProcNER corpus as training data, participants must create systems that can assign SNOMED CT codes to the full case report so that they can be indexed.

We refer the reader to the overview article [1] and the competition's official website[1] for detailed information about MedProcNER 2023. Vicomtech's NLP team has implemented several tools to address the different stages of the task incrementally: entity mention detection has been addressed with a Transformer-based sequence labelling system; the entity normalisation has been tackled with Semantic Text Similarity (STS) techniques; finally, both methods were applied to resolve the clinical indexing problem.

This paper is organised as follows. Section 2 describes the annotated corpus provided by the organisers, the custom train and validation split used for the experiments, the SNOMED CT gazetteer database and the method for the terminology enrichment. Section 3 presents the systems implemented to tackle Task 1. Section 4 describes the approach for Task 2 (normalisation). Section 5 briefly explains how we treated the task of clinical indexing. Section 6 shows the challenge's official results and discusses the presented systems' performance.

2. Data Description

The official dataset consists of 1,000 manually annotated text clinical reports in Spanish from which 750 documents are prepared for training purposes and 250 documents are reserved for the participants' systems' testing. Clinical case reports are a type of textual genre in medicine that describe a patient's medical history, symptoms, diagnosis, and treatment in detail. Task 1 annotations consist of 11,065 spans of entity mentions with their corresponding label PROCEDIMIENTO (*procedure*), Task 2 train set is annotated with 4,857 SNOMED CT codes, and Task 3 contains 250 documents, where each document is annotated with a set of SNOMED CT codes.

The majority of the entities for the normalisation task are single code annotations; nevertheless, about 2.6% (125) of all spans are annotated with a composite code, formed with two or more SNOMED CT IDs, concatenated with a "+" sign. We also check if all annotated codes correspond to the procedure semantic tag in SNOMED CT gazetteer. 4,602 codes from the train set (94.75%) have *procedure* tag; there are also 255 items with the other tags; for instance, 101 codes are labelled as *regime/therapy* in SNOMED CT, 51 code is marked as CODE_NOT_IN_DICT, 25 codes have the tag *physical object*, 24 codes are tagged as *product*, 12 codes are tagged as *substance*, etc.

To perform the experiments, we randomly split the documents into training (90%) and validation (10%) sets. We can see the details of the split in Table 1.

[1] https://temu.bsc.es/medprocner/

Table 1
Training dataset for tasks 1 and 2

	Documents	PROCEDIMIENTO	Unique SNOMED CT codes
Train	675	4,346	1,538
Validation	75	511	291

2.1. SNOMED CT Dictionary Enrichment

The shared task organisers provided a TSV file containing the SNOMED CT codes, their definitions, and the training and validation data. The SNOMED CT taxonomy contains 242,228 entries of 130,219 unique concepts, which means that some of the concepts have various synonyms (up to 32 entries with the same code). These codes must be assigned to each entity in the input texts as part of Task 2. As a data pre-processing step, we have extended the provided dictionary entries using the manually labelled terms from the training set of our train-validation split. This adds 2,697 unique terms and synonyms, so the final number of entries in the SNOMED CT taxonomy becomes 244,924.

If a complex code occurs, we treat it as a single atomic code. We refer to complex code when an entity is assigned multiple SNOMED CT codes. Example:

> Al examen físico las mucosas son húmedas y normocoloreadas, la **auscultación cardio-respiratoria** es normal [...] (*On physical examination the mucous membranes are moist and normal-coloured, cardio-respiratory auscultation is normal [...])*

This phrase is annotated with composite SNOMED CT code 449263002+449264008 "auscultación del corazón" + "auscultación del tracto respiratorio inferior" (*"auscultation of the heart" + "auscultation of the lower respiratory tract"*), which means that both codes occur in the marked span. In the enriched SNOMED CT taxonomy it is defined as 449263002+449264008 "auscultación cardio-respiratoria".

3. Task 1: Clinical Procedure Recognition

In the task 1, participants are requested to automatically detect mentions of clinical procedures in the provided clinical reports. In other words, it is a Named Entity Recognition task. We have faced the task as a regular sequence labelling task, using IOB-tagging to emit one of B-TAG, I-TAG or O, where the only possible TAG type is "PROCEDIMIENTO".

The sequence labelling is performed by a Transformer-based model, which encodes each input token into its contextual word embedding. These word embeddings pass through a classification head that projects the word embedding into the output label space. We have experimented with several Transformer models.

Since the MedProcNER documents are generally too long to fit in one piece into a Transformer model, we have applied the sliding windows technique, as described elsewhere [3]. In a few words, we surround each window with a number of context tokens. These tokens are ignored when rebuilding the original document; they provide valuable information to resolve the central

Table 2
Corpus for Longformer-es model

	Paragraphs	Max Tokens/Paragraph	Mean Token/Paragraph
Train	3,432	711	79.96
Validation	389	533	78.96

Table 3
Results of the NER models on our validation set.

System	P	R	F1
xlm-roberta-large	**0.7872**	0.7514	0.7689
roberta-bio-es	0.7649	**0.7755**	**0.7702**
longformer-bne-es	0.7402	0.7560	0.7480

window by avoiding hard, meaningless segmentation cuts. We have applied this sliding windows technique with the xlm-roberta-large[2] model and with the roberta-base-biomedical-es from the BSC[3] as they are listed in the HuggingFace model hub [4].

As an alternative to the sliding windows approach, we have tried a longformer model, which allows a large enough sequence-length encoding up to 4096 to accommodate any MedProcNER clinical procedures in one shot. We select longformer-base-4096-bne-es[4] which is the longformer version of the RoBERTa model for the Spanish language [5]. It allows us to process larger contexts as input without additional aggregation strategies. We split the clinical reports documents into paragraphs following the line-break characters and tokenize them into words using SpaCy[5] to obtain token-label pairs. The resulting corpus is shown in Table 2, we get 3,821 paragraphs where the longest one is 711 tokens. Transformer encoding tokens do not correspond to the grammatical word and punctuation tokens, so we assume that the maximum sequence length of 2,048 will be enough to encode all the paragraphs.

Table 3 shows the models' performance where the best-performing model is roberta-bio-es, trained on biomedical domain corpora, the second best model is xlm-roberta-large which shows close to the best performance due to its size and parameters number. The longformer strategy is not the best, which might be because of the size of the training corpus reduces with the paragraph split method.

4. Task 2: Clinical Procedure Normalization

Our primary approach is Semantic Text Similarity (STS) techniques. STS determines how similar two pieces of text are by measuring their degree of semantic closeness. Semantic search is based on STS, allowing retrieval of relevant text results beyond mere lexical matching. The

[2]https://huggingface.co/xlm-roberta-large
[3]https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-es
[4]https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es
[5]https://spacy.io/

main concepts of semantic search are query, collection of documents, and degree of relevance between a query and retrieved documents. There are different methods of measuring the degree of relevance and relatedness of two pieces of text—cosine distance, inner product, etc.

As a baseline model for resolving the normalisation problem, we implement similar document retrieval with the BM25 ranking function [6]. This function ranks a set of documents based on the query terms appearing in each document, regardless of their proximity, and it works on the concept of bag-of-words and TF-IDF. We use it as the simplest statistical method.

4.1. Transformer-based Semantic Search

The semantic search involves embedding all entries (sentences, documents, or, in this case, taxonomy descriptions) into a vector space. At search time, the query, represented in this task by the detected entity mention, is also embedded into the same vector space. This allows a direct comparison of vectors using distance. Nowadays, the most extended method to encode text is to use a pre-trained Transformer model [7] to obtain the corresponding embeddings (multidimensional vectors) and compute the similarity score using a similarity metric (e.g., in this case, it is the inner product of the normalised vectors).

We encode the entity words and SNOMED CT code descriptions with a SapBERT-XLMR-large model [8]. This model is pretrained with UMLS database [9] using XLM-RoBERTa-large as the base model. We find injecting UMLS knowledge of multilingual clinical terminology into a pre-trained language model especially helpful for the normalisation task; an embedding dimension of 1024 is enough to encode all the terminology and corpus entities without truncation. [CLS] token of the transformer's architecture is used for the vector representation of a text.

Next, each corpus entity's closest candidate from the SNOMED CT database is retrieved. The code of the most similar taxonomy entry is used as the predicted code for each given entity.

4.2. Cross-Encoders

We also experiment with a cross-encoder model [10] training. Cross-encoders handle sentence pair scoring and classification tasks [11]. They have been proven successful in the clinical domain also [12]. In contrast to an unsupervised semantic similarity function, the cross-encoder is trained by encoding both sentences simultaneously and produces a value between 0 and 1 that indicates the similarity or relatedness of the input sentence pair (see Figure 1b). Cross-encoders are trained using a set of text pairs labelled as similar/related (i.e., positive) or dissimilar/unrelated (negative).

We have used the same train and evaluation split from the dataset to retrieve positive examples from the SNOMED CT file. For each entity labelled in the corpus, we have created pairs with the entity's text and the SNOMED CT description corresponding to the entity's SNOMED CT code. To obtain negative examples, we have used negative sampling by choosing pairs that are not related. We implement three types of corpus preparation with negative sampling.

- **NS1**. Semantic search with SapBERT-XLMR-large model [8]. We take the first 64 candidates, and for each of them, we assign the value 1.0 to the query-candidate pair if the query has the same code as a candidate, and the value 0.0 if the code is different. The column SNOMED CT Description in Table 4 contains ten pairs for a query "exploración

Table 4
Examples for training the cross-encoder created with negative sampling of the query "exploración ginecológica" (*gynecologic examination*), SNOMED CT code 83607001. The label 1 indicates that the retrieved entry matches the code assigned in the training data to "exploración ginecológica". The UMLS Synonym column shows the definitions of the term in the clinical taxonomies included in UMLS

Label	SNOMED CT Description
1	examen ginecológico
0	exploración del aparato genital femenino
0	examen ginecológico endoscópico
0	endoscopia ginecológica
0	examen ginecológico de rutina
0	exploración de vagina
0	examen vaginal
0	exploración del aparato genitourinario
0	incisión y exploración de la vagina
0	pruebas uroginecológicas
Label	**UMLS Synonym**
1	gynecologic examination
1	gynaecologic examination
1	female genital examination
1	examination of female genitals
1	gynecologic examination (procedure)
1	examen ginecológico (procedimiento)

ginecológica" (SNOMED CT code 83607001). We hypothesise that labelled semantically similar pairs will help to discriminate the correct term.

- **NS2**. Semantic search with SapBERT-XLMR-large model [8] adding UMLS synonyms. We enrich the negative sampling corpus with definitions from different clinical taxonomies presented in UMLS, such as ICD-10, CUI, etc. We use the ClinIDMap mapping tool [13] to get new synonyms for all corpus codes. The number of positive pairs increases more than three times. The column UMLS Synonym shows some examples for the same query in Table 5.
- **NS3**. We get all composite codes from the train set and make composite descriptions concatenating the terms of these codes from SNOMED CT. With this, the positive pair will be:

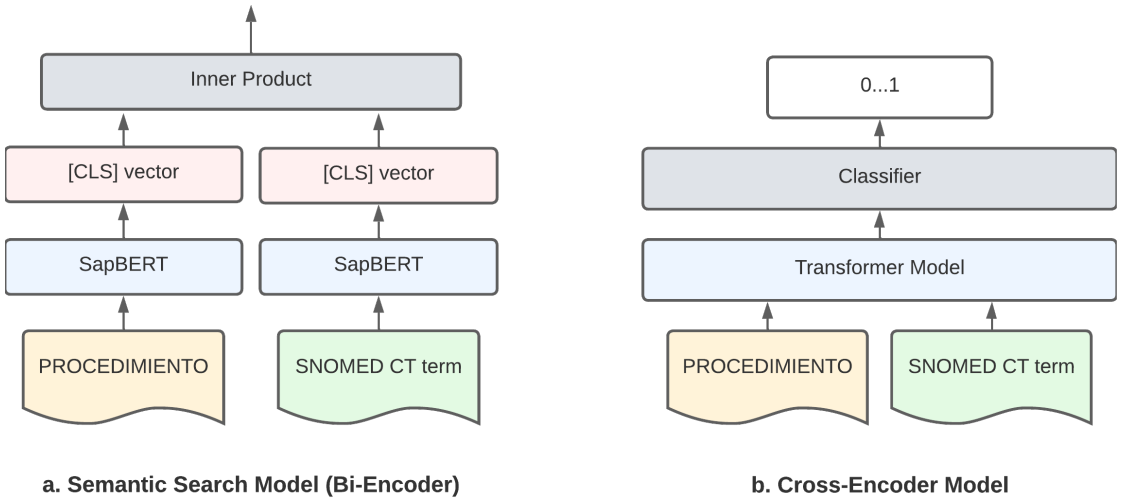
  > corpus span: auscultación cardio-respiratoria (*cardio-respiratory auscultation*)
  > SNOMED CT description: auscultación del corazón; auscultación del tracto respiratorio inferior (*auscultation of the heart; auscultation of the lower respiratory tract*)

  Negative examples are obtained with the same method as described above. We add the composite code examples to the NS2 corpus.

Table 5 describes the datasets obtained with the methods of negative sampling. We have

Table 5
Size of training and development corpus for cross-encoder training.

Method	Train			Validation		
	Total	Negative	Positive	Total	Negative	Positive
NS1	173,430	168,565	4,865	25,771	24,957	814
NS2	186,179	168,565	17,614	27,759	24,957	2,802
NS3	186,294	168,578	17,716	27,776	24,955	2,821



a. Semantic Search Model (Bi-Encoder) **b. Cross-Encoder Model**

Figure 1: Semantic models. Bi-encoder encodes two text pieces separately and measures their relatedness with semantic similarity function. Cross-encoder model is trained with pairs of text and produces a score between 0 and 1.

trained three cross-encoder models with these datasets on top of the model RoBERTa base pre-trained with data from the National Library of Spain (BNE) [5] as it performed well on Spanish texts. We use the Sentence-Transformers [10] framework to train the models. The best checkpoint within 20 training epochs was chosen according to the macro F1-score calculated on the validation dataset—F1=0.6843 for NS1 method, F1=0.9118 for NS2 method and F1=0.9046 for NS3 method.

4.3. Approaches to SNOMED CT code prediction

Based on the described semantic text search techniques and models, we have experimented with three approaches to predict the correct set of codes given to an entity. We also implement the search in SNOMED CT database only and in the enriched SNOMED CT+train database. We also try to filter SNOMED CT database to *procedure* semantic tag only, but, as was commented in Section 2, some of the annotated codes are not procedures, the performance drops, so we have not considered it.

Semantic Search (SS) We use two types of semantic search: (1) with BM25 ranking and (2) transformer-based semantic search. For the second type we encode both the entity words and SNOMED CT with SapBERT model and retrieve the closest candidate from the SNOMED CT using the cosine similarity function (see Figure 1a). The code of the most similar taxonomy entry is used as the predicted code for each given entity.

Semantic Search and Rerank (SS-R) With this approach the prediction of the SNOMED CT codes consists on the following steps:

1. Retrieve 64 candidates from SNOMED CT with the semantic search or with BM25 approaches described above.
2. Rerank the retrieved documents with the cross-encoder model described in Section 4.2. The cross-encoder model re-scores the candidates retrieved in the first step.
3. Get the candidate with the highest score from the cross-encoder and pick its corresponding SNOMED CT code for the entity.

Semantic Search and Conditional Post-Processing (SS-C) Examining the top most similar items retrieved from semantic search, we observe that sometimes the correct answer occurs at the second position, increasing the evaluation score of Accuracy@K where K=2 by 5 points. Low scores in the first position suggest the retrieved term is incorrect, and the second following may be correct. We experiment with the similarity score threshold on the development set and select the threshold with slightly higher accuracy (see SS-C in Table 6). If the score of the first retrieved is less than the threshold, we choose the second item.

All semantic similarity experiments are implemented using FAISS [14] and the reranking process is implemented in Sentence-Transformers [10] framework. We used the normalised inner product to calculate the similarity metric. The performance is calculated over the Task 2 validation set. As depicted in Table 6, the best-performing system is the semantic search with SapBERT model over SNOMED CT dictionary enriched with the entities from the training corpus. The performance of any semantic search system depends not only on the semantic relatedness but also on lexical similarity, which is essential for a short text search like in this case. 112 unique entities from the validation set precisely match the entities in the training set but not in SNOMED CT (“oligoelementos”, “nutrición parenteral”, ‘coronariografía”, “acto quirúrgico”, “fluidoterapia”, “diagnóstico anatomopatológico” etc), making predictions easy. For this reason, we have selected this system for our submission to the task competition. To test our ideas, we also submit the cross-encoder models on top of our best NER model (see Table 3).

Since we were limited to submit only five runs, we decided not to use BM25 method due to its lower scores. But it should be noted that the cross-encoder model helps to re-rank the retrieved with BM25 system, BM25+cross-encoder performs 5 points better than BM25 only.

The sampling methods for cross-encoder training show very similar behaviour. As for the transformer-based semantic search and cross-encoder, it does not help in the overall evaluation. We manually examine the errors of the cross-encoder and see that in some cases it rescores incorrect predictions of the semantic search system to the correct ones. So, it could be used in an ensemble with the SS system. For instance, corpus entity “anatomía patológica del lavado”

Table 6
Experiment Results

System	Database	Accuracy@1	F1 macro
SS			
BM25	SNOMED CT	0.1996	0.0965
BM25	SNOMED CT+train	0.4149	0.2152
SapBERT	SNOMED CT	0.4344	0.2554
SapBERT	SNOMED CT+train	**0.6810**	**0.4095**
SS-R			
BM25+NS1	SNOMED CT+train	0.4775	0.2579
BM25+NS2	SNOMED CT+train	0.4775	0.2604
BM25+NS3	SNOMED CT+train	0.4716	0.2604
SS+NS1	SNOMED CT+train	**0.6458**	0.3703
SS+NS2	SNOMED CT+train	0.6438	**0.3814**
SS+NS3	SNOMED CT+train	0.6360	0.3646
SS-C	Threshold		
	14	0.6810	0.4095
	16	**0.6830**	**0.4113**
	18	**0.6830**	0.4104
	20	0.6810	0.4071
	22	0.6732	0.3939
	24	0.6810	0.4017
	26	0.6693	0.3892
	28	0.6556	0.3738
	30	0.6478	0.3631

(*pathological anatomy of lavage*) is normalised to the code 67889009 "lavado" (*lavage*), while reranking model assigns the correct term "anatomía patológica" (*pathological anatomy*).

5. Task 3: Clinical Procedure-based Document Indexing

Our approach for the MedProcNER Task 3 is directly based on the previous tasks. The Task 1 and 2 detect and normalise mentions to clinical procedures, obtaining the exact span in which the occur, together with their SNOMED CT code. For the Task 3 we just gather the codes for each document, retaining the set of unique codes per document, and using that as the outcome for the Task 3.

6. Results and Discussion

This section describes the official results obtained by our submitted systems in the MedProcNER 2023 Shared Task. At the time of this writing, the results from all the other participants have not been disclosed by the organisers, and the only information we have to compare our systems

Table 7
Test results, provided by the organisers; the bold font refers to the best scores.

System	P	R	F1
Task 1			
Run 1: xlm-roberta-large	**0.8054**	0.7535	0.7786
Run 2: roberta-bio-es	0.7679	**0.7629**	0.7653
Run 3: longformer-bne-es	0.7478	0.7588	0.7533
Best			**0.7985**
Task 2			
Run 1: xlm-roberta-large-SS	**0.5902**	0.5525	**0.5707**
Run 2: roberta-bio-es-SS	0.5665	**0.5627**	0.5646
Run 3: roberta-bio-es-SS-C	0.5662	0.5625	0.5643
Run 4: roberta-bio-es-SS-R-NS2	0.5248	0.5213	0.5230
Run 5: longformer-bne-es-SS	0.5498	0.5580	0.5539
Best			**0.5707**
Task 3			
Run 1: roberta-bio-es-SS	0.6182	0.6295	0.6238
Run 2: roberta-bio-es-SS-R-NS2	0.5885	0.5917	0.5901
Run 3: longformer-bne-es-SS	0.6039	0.6288	0.6161
Run 4: xlm-roberta-large-SS	**0.6371**	0.6109	0.6239
Run 5: roberta-bio-es-SS-C	0.6190	**0.6295**	**0.6242**
Best			**0.6242**

is the best score for each task. The results are shown in Table 7, where the scores provided by the organisers are marked in bold. According to this, in the Task 1 our best system is 2 points below the best scoring participant, while our best system for Task 2 and Task 3 has obtained the highest score.

Our models exhibit a similar behaviour on our custom validation set and in the official test set. The precision of xlm-roberta-large model is notably higher than in the other models both in the validation and test set, which might affect the performance and robustness of the model and might be helpful in situations when the precision metric is more important than recall.

We examined the errors in the development set and concluded that it is difficult for the models to distinguish between semanticaly or lexically close corpus entity and its code definition. For instance, corpus entity "Gammagrafía ósea con Tc99m-MDP" (*Bone scintigraphy with Tc99m-MDP*) is manually annotated as code 418832007 which has the definition "gammagrafía ósea de cuerpo entero" (*whole body bone scintigraphy*). The system assigns to the entity code 425559005 with the definition "resonancia magnética nuclear de hueso" (*bone magnetic resonance imaging*). We suppose that the model calculates high relatedness because of the words "ósea" and "hueso" (which both have meaning *bone*). Also the word "gammagrafía" is related to nuclear medicine, which is semantically close to "resonancia magnética nuclear".

There are also some other challenging points. For instance, a type of error is related to the strict match of the entity and SNOMED CT definition. There are cases where the corpus word

is equal to the SNOMED CT definition but they are annotated with different codes. This is an example from the corpus:

> Pruebas complementarias: **hemograma** con ligera leucocitosis sin desviación de fórmula leucocitaria siendo el resto normal, bioquímica y coagulación normales. (*Complementary tests:* ***hemogram*** *with slight leukocytosis without deviation of the leukocyte formula, the rest being normal, normal biochemistry and coagulation.*)

In this context, the entity "hemograma" (*blood count*) is manually annotated with code 26604007, definition "recuento sanguíneo completo" (*complete blood count*). The predicted code is 43789009 with definition "hemograma" (*blood count*). The code 43789009 has two definitions in SNOMED CT: "hemograma completo sin fórmula diferencial" (*complete blood count without differential formula*) and "hemograma" (*blood count*). We rely on the similarity model in the whole database, and it always matches two identical words if they occur. In this case, the possible solution might be additional word sense disambiguation using the context of the corpus entity.

Composite codes are difficult to predict also, for instance, code 363679005+182770003 "estudios de imagen y preanestésicos" (*imaging and pre-anaesthetic studies*) is composed of two definition—"procedimiento de estudio por imágenes" (*imaging procedure*) and "evaluación preanestésica" (*pre-anaesthetic assessment*). The predicted code is 182770003, which is only one part of the code. A possible method to tackle this problem could be a specific classifier to distinguish between simple and composite codes. Further, as a composite code may consist of more than two codes, in case of using a similarity search approach it would be difficult to guess the number of top most-similar codes to select.

Again, it must be noted that the results of each task depend on the results from the previous tasks. Any error in the Task 1 impacts the results for the Task 2 and Task 3, and that must be taken into account when examining the overall results for the tasks.

7. Conclusions

In this paper we have described our participation in the MedProcNER 2023 Shared Task. We presented three runs for Task 1 (NER), which requires finding mentions of procedure entities in the provided clinical texts. For this first task we have trained several sequence labelling models based on multilingual and Spanish pre-trained Transformer models. For Task 2 (Normalisation), which requires assigning specific SNOMED CT codes to each detected entity, we have implemented a system based on Semantic Text Similarity and cross-encoders. Our approach for Task 3 (document indexing) is directly based on the systems for the previous two tasks; we detect procedure entities, normalise them, and then get a set of unique codes for each document. Our submissions for Task 2 and 3 have achieved the highest scores for the competition.

References

[1] 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 Detec-

tion and Entity Linking at BioASQ 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[2] 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.
[3] A. García-Pablos, N. Perez, M. Cuadros, Vicomtech at CANTEMIST 2020, 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. 489–498.
[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, HuggingFace's Transformers: State-of-the-art Natural Language Processing, arXiv:1910.03771 (2019) 1–11.
[5] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento del Lenguaje Natural 68 (2022).
[6] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive, in: Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998, volume 500-242 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1998, pp. 199–210.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.
[8] 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.
[9] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267–270.
[10] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992.
[11] R. Nogueira, K. Cho, Passage Re-ranking with BERT, ArXiv abs/1901.04085 (2019).
[12] A. Rahimi, T. Baldwin, K. Verspoor, WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5957–5962.
[13] E. Zotova, M. Cuadros, G. Rigau, ClinIDMap: Towards a Clinical IDs Mapping for Data Interoperability, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3661–3669.

[14] J. Johnson, M. Douze, H. Jégou, Billion-scale Similarity Search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.