FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding

Nuria García-Santa[1] and Kendrick Cetina[1]

Fujitsu Laboratories of Europe (FLE), Pozuelo de Alarcón (Madrid) 28224, Spain
{nuria.garcia, kendrick.cetina}@uk.fujitsu.com
http://www.fujitsu.com/emea/about/fle/

Abstract. In Healthcare domain, several documents are provided in a narrative way, following textual unstructured formats. This is the case of the discharge summaries, which are clinical texts where physicians describe the conditions of the patients with natural language, making the automated processing of such texts hard and challenging. The objective of the tasks of the 2020 CLEF eHealth for Multilingual Information Extraction is to develop solutions to automatically annotate Spanish clinical texts with codes from the International Classification of Diseases, 10th version (ICD-10). In the present paper, we show our approach which is based on Named Entity Recognition (NER) to detect the diagnoses and procedures, and semantic linking against a Knowledge Graph to extract the ICD-10 codes. Besides, we exploit text augmentation techniques to generate synthetic input samples and we use BERT pre-trained models and architecture to train the NERs.

Keywords: CLEF eHealth · Clinical Encoding · Text Mining · Semantic Knowledge · Named Entity Recognition (NER)

1 Introduction

Automated Clinical Encoding covers multiple computer-assisted techniques to extract valuable knowledge within clinical documents written in natural language and transform such knowledge to structured information. Clinical documents usually include medical entities that correspond to diagnoses, procedures, symptoms, drugs, etc., but the use of narrative and informal language is challenging for the automatic processing of this information. Among Automated Clinical Encoding tasks, a popular one is the assisted assignment of codes to the clinical documents from standard medical classifications, such as the International Classification of Diseases (ICD) [7]. Traditionally, this code assignment is done manually by healthcare professionals. Therefore, the main objective of automated approaches is to support clinicians in their daily activities by helping them save time and resources.

The CLEF eHealth challenge on Multilingual Information Extraction of this year is focused on this kind of techniques [6][14]. The three sub-tasks of this 2020 challenge are based on the automated code assignment of Spanish clinical documents for diagnoses and procedures of the International Classification of Diseases (ICD).

In previous years, the challenges worked in the same line of research. For CLEF eHealth 2017 challenge, participants provided solutions to extract ICD codes (10th version , ICD-10) in death certificates for English and French [17], and, in 2018, for French, Italian, and Hungarian [18]. In the CLEF eHealth challenge of 2019, the shared task was focused on the automatic detection of ICD-10 codes for German Non-technical summaries (NTSs), which are short descriptions of planned animal experiments [19].

In these past challenges, the best approaches were mainly based on neural network architectures. In the best approach of 2017, the authors provided sequence-to-sequence deep learning models based on Recurrent Neural Networks (RNNs) [12]. In 2018, the best solution followed a similar way, proposing a machine learning sequence-to-sequence neural model to map input text snippets with the output ICD-10 codes [2] . And, in 2019, the best two approaches developed different neural network designs, such as CNNs and Attention models [1], or logistic regression classifiers [21], but both used multilingual BERT [4].

In the literature, a wide range of approaches have been published, since semantic-based or rule-based solutions to machine learning proposals. Several examples of semantic approaches are works such as Pakhomov et al. [20], where the authors presented a system that relies on a Knowledge Base obtained by manually coded data, collected over 10 years, or García-Santa et. al [5], that developed a solution to return automatically the k-top ICD codes associated to a clinical text through exploitation of enriched Knowledge Graphs and heuristic rules. In machine learning research, Mullenbach et al. [16] proposed a method called Convolutional Attention for Multi-Label classification (CAML) that is based on a CNN and a per-label attention mechanism, and, it includes explanations of the code assignments. Baumel et al. [3] described a Hierarchical Attention bidirectional Gated Recurrent Unit (HA-GRU) to identify the relevant sentences for each code. In this approach, the authors compared the results with an SVM-based one-vs-all model, a continuous bag-of-words (CBOW) model, and a CNN.

To address the CLEF eHealth 2020 challenge, our FLE team has developed a solution focused on Named Entity Recognition (NER) and semantic-based approaches exploiting Knowledge Graphs. Our Knowledge Graph has been enriched with the annotations coming from training and validation sets provided by the organizers of the challenge. Besides, we extended the input datasets for the NER models by creating synthetic samples with text augmentation techniques over the train/validation sets and we also used BERT pre-trained models and architecture for the NER training.

2 Material and Methods

2.1 Problem definition

Current CLEF eHealth 2020 task deals with multilingual Information Extraction (IE). Concretely, this year, the challenge is focused on automatic coding of Spanish clinical textual documents to the International Classification of Diseases [7], version 10 (ICD-10)[1], in its Spanish distribution (CIE-10)[2]. This is the first community task oriented exclusively to the automatic coding of clinical cases in Spanish [14].

The challenge is divided in three sub-tasks:

- Task 1: Automatic code assignment of Spanish texts to CIE10-CM, i.e. to specific diagnoses of the standard.
- Task 2: Automatic code assignment of Spanish texts to CIE10-PCS, i.e. to specific procedures of the standard.
- Task 3: Addition of explainability references for the two aforementioned tasks. It requires a joint automatic code assignment of diagnoses and procedures, including the positions of the key entities that justify such code assignment.

The output has to be a list of ICD-10 codes for each text document. In the first two sub-tasks, this list must be arranged in descending order, based on the relevance of the code to the corresponding document. In the last sub-task, order of relevance is not required but a joint list of codes has to be presented for diagnoses and procedures, specifying the position of related entities in the text documents.

Our team has participated in the three sub-tasks through a multi-task approach. In our proposal, the core and main techniques have been shared and reused to address the three sub-tasks in a unified way.

2.2 Datasets and Resources

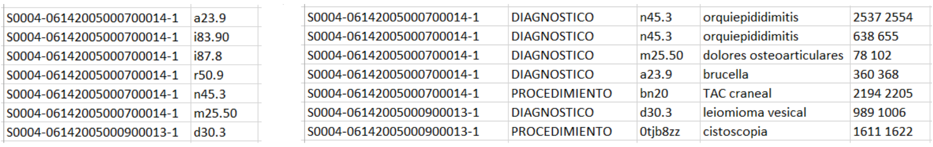
For all the sub-tasks, a synthetic corpus of 1000 clinical case studies has been published. The dataset was manually annotated by clinical professionals. In the official source of the challenge it is specified that the dataset comprises 16,504 sentences and 396,988 words, with an average of 396.2 words per clinical case. This corpus is freely accessible[3]. There are separate directories for train, dev and test datasets. The train set has 500 clinical cases, the dev set has 250, and the test set with gold standard annotations has 250 clinical cases. In addition, organizers shared a background set without annotations of 2,751 clinical cases. Besides the texts of the clinical cases, in the corpus of train and dev set, tab-separated files for each sub-task are included. These files include the annotations associated to each clinical case. In figure 1, an excerpt sample of these files is shown for task 1 (CIE10-CM) and task 3 (Explainability).

The training dataset of the task 3 (Explainability) has 9,211 annotated codes, of which, 2,392 are unique. Taking into account that CIE10-CM reports a number of 71,486 diagnoses and CIE10-PCS has a number of 87,170 procedures, we have a total of 158,656 codes [15]. This quantity is very far from the unique number of annotated codes in the dataset, which means that we are losing a wide spectrum of potential codes of assignment. This makes it more difficult to provide scalable systems in supervised learning approaches. Besides, the number of the code annotations in the dataset is very unbalanced. If this issue is not addressed, it could increase classification biases to most frequent codes. Figure 2 shows the forty most frequent codes in the dataset.

[1] https://www.who.int/classifications/icd/en/

[2] https://eciemaps.mscbs.gob.es/ecieMaps/browser/metabuscador.html

[3] https://zenodo.org/record/3837305#.XtTwHjozYgx



S0004-06142005000700014-1	a23.9
S0004-06142005000700014-1	i83.90
S0004-06142005000700014-1	i87.8
S0004-06142005000700014-1	r50.9
S0004-06142005000700014-1	n45.3
S0004-06142005000700014-1	m25.50
S0004-06142005000900013-1	d30.3

S0004-06142005000700014-1	DIAGNOSTICO	n45.3	orquiepididimitis	2537 2554
S0004-06142005000700014-1	DIAGNOSTICO	n45.3	orquiepididimitis	638 655
S0004-06142005000700014-1	DIAGNOSTICO	m25.50	dolores osteoarticulares	78 102
S0004-06142005000700014-1	DIAGNOSTICO	a23.9	brucella	360 368
S0004-06142005000700014-1	PROCEDIMIENTO	bn20	TAC craneal	2194 2205
S0004-06142005000900013-1	DIAGNOSTICO	d30.3	leiomioma vesical	989 1006
S0004-06142005000900013-1	PROCEDIMIENTO	0tjb8zz	cistoscopia	1611 1622

Fig. 1: Excerpt sample of the training datasets. In the left, for the task 1 (two columns; clinical case ID and CIE10-CM code). In the right, for the task 3 (four columns; clinical case ID, entity label type, CIE-10 code, entity label and position in text).

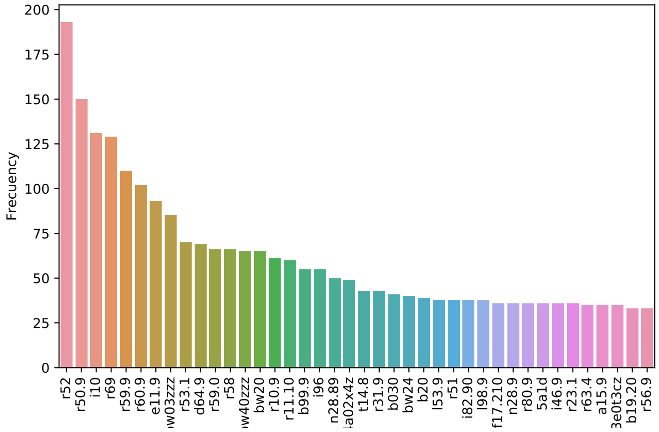


Fig. 2: Top 40 most frequent codes in the training dataset of the task 3.

Aside from these core datasets of the challenge, we also tried the additional Spanish abstracts provided by the organizers[4]. These abstracts were a total of 176,294 texts and they were annotated automatically [13]. After several tests, we decided to discard the use of this resource in final versions because it increased the noise of our annotations and the performance was affected negatively.

External sources such as PubMed [11] and MIMIC-III [8] have been used to support Named Entity Recognition (NER) tasks. Annotated samples of diagnoses and procedures entities mentioned in medical literature from PubMed (through PubTator FTP service [22]) and in clinical notes from MIMIC-III databases are exploited to train NER models. We translated those annotated datasets from English to Spanish. In this NER task, the pre-trained language model of Multi-lingual BERT[5] [4] was also used.

Other linguistic resource that we used is the NegEx-MES tool[6] to detect entities negated in Spanish texts. This resource was exploited in several of our final versions for post-processing steps.

[4] https://zenodo.org/record/3606662#.XtUVdDozYgx

[5] https://github.com/google-research/bert/blob/master/multilingual.md

[6] https://github.com/PlanTL-SANIDAD/NegEx-MES

And, finally, we used the lists of valid codes for Spanish ICD-10 [15], for diagnoses and procedures (CIE-10).

2.3 Automated Clinical Encoding Methodology

The main workflow and steps of our system are depicted in figure 3. We followed an approach based on named entity detection from clinical texts and later Knowledge Graph (KG) entity linking to CIE-10 codes. We also tested an approach based on text classification through simple Convolutional Neural Networks (CNNs). However, performance decreased and the system was less scalable because of imbalance nature and low coverage of datasets, as we explained in previous section. Because of these reasons, we discarded a full Machine Learning approach in our final system, instead, we developed a combination of Machine Learning for Named Entity Recognition (NER) and semantic-based approach for CIE-10 entity linking through Knowledge Graph (KG) construction.

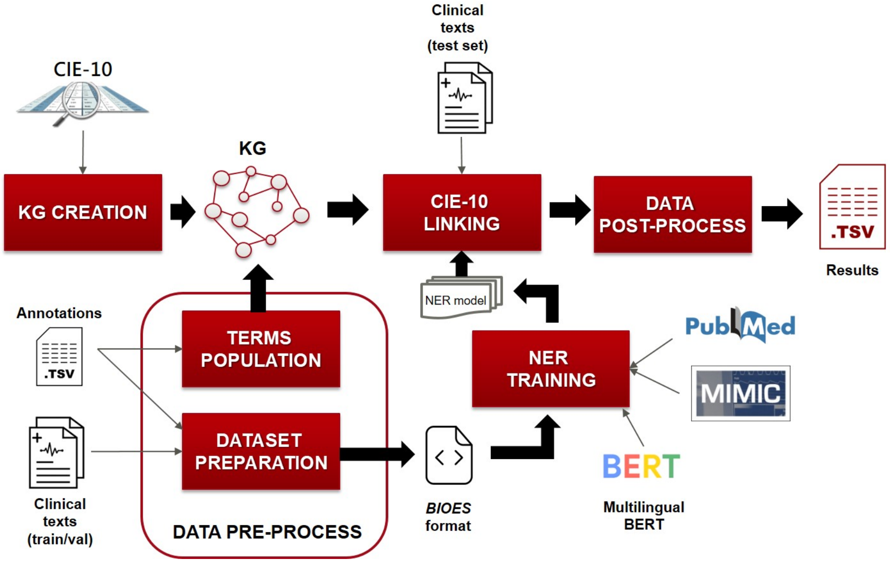


Fig. 3: Workflow of our system

The first steps include the activities of KG Creation and Data Pre-process. For the KG Creation, we take the lists of valid codes in CIE-10 and we implement a Knowledge Graph with that information. This data contains nodes with the CIE-10 code, the description in Spanish, the description in English and hierarchical relations between the CIE-10 nodes. In this implementation we use the framework of Neo4j[7]. In Data

[7] https://neo4j.com/

Pre-process, we perform Terms Population to the KG. These terms come from the samples annotated in the training and validation sets provided by the organizers. Taking the datasets annotated in the task 3, we get the CIE-10 codes and its related term mentions and we create a new array attribute for that CIE-10 node in the KG, adding all the different ways to name a diagnosis or procedure (term mentions). We also perform Dataset Preparation, cleaning the clinical texts of special characters and encoding issues, and solving several errors in the annotated samples regarding wrong labels (e.g. a diagnosis with the type label of procedure or vice versa), and codes that belonged to other standards. In this step, we also adapt the format of input annotated samples and clinical texts to the *BIOES* format or *IOBES*[8] in order to be used later for training the NER.

In the intermediate step, we proceed to train a Named Entity Recognition (NER) model. We use the pre-trained language model of Multilingual BERT to initialize the neural network. We have performed different trainings depending on variations in the input annotated samples. Below, we point out the different settings (all samples follow the *BIOES* format):

- Baseline: Input annotated samples from train and dev sets provided by organizers.
- Baseline + Abstracts: Previous samples + annotated samples from the additional Spanish abstracts resource.
- Baseline + MIMIC-III: Baseline samples + annotated samples from MIMIC-III dataset.
- Baseline + Text Augmentation: Baseline samples + annotated samples augmented with Fujitsu's proprietary technology.

For the final versions, we carried out the setting of *'Baseline + Text Augmentation'* because we achieved the best results as shown in the evaluation section. We trained two different NER models for each type of label; one for diagnoses recognition, and the other for procedures recognition. In the neural network, we follow the BERT architecture. Bidirectional Encoder Representations from Transformers (BERT) [4] is a bidirectional transformer encoder whose main features are multi-headed self attention, multi-layer feed forward and positional embeddings. We fine-tuned the Multilingual BERT for our supervised NER model.

Next, in the CIE-10 Linking process, the previous models were run over the test sets to extract all named entities within the clinical texts. In this way, we are able to obtain the named entity, its label (diagnosis or procedure) and its position in the text. Once we have detected the named entities, we perform the linking algorithm to extract the correspondent CIE-10 codes. For this activity, we use Levenshtein string similarity distance [10]. We compare the named entities against the terms and descriptions of the KG, getting the most suitable CIE-10 codes.

In the final step, we develop Data Post-process methods to create the definitive results. For the task 1 and the task 2 we apply frequency-based techniques to sort the results. We assume the most frequent named entities in a text are the most relevant ones. The CIE-10 codes of such named entities would be in the first positions of the list of a clinical text. However, in those cases when the frequency is the same, we apply reordering based on the text position of named entities. We give more relevance

[8] https://donovanong.github.io/ner/tagging-scheme-for-ner.html

to named entities nearer to the end of the text, where conclusions and the diagnostics usually take place. We assign lower relevance to the entities located at the beginning of the text, where the antecedents are usually exposed. For all these tasks, we created a version of results where we removed the entities negated with NegEx-MES tool. We also created other versions where we analyzed overlapping of named entities in different text positions to normalize the CIE-10 code of the longest named entity.

3 Evaluation

In this section we first describe the environment and the tools that we used. Then, we describe the experiments carried out to select the tools and the methods used. And finally, we present and discuss the CLEF eHealth performance evaluation of our results.

3.1 Environment setup

All the NER models described here are obtained by fine-tuning a BERT-Based Multilingual cased architecture with 110 million trainable parameters. The weights are available on GitHub[9]. It supports 104 languages (Spanish included). We chose this architecture empirically from the experience we have on training biomedical domain models. We used the script `run_ner.py`, which is available on GitHub[10] to fine-tune our NER models. This script was also previously used to train BioBERT[9].

For text augmentation of the NER input datasets, we trained a text generation model with Fujitsu's proprietary technology based on decentralized learning. We used a subset of MIMIC-III [8] and PubMed [11] databases. We selected the first 10% of each database and we created sequences of sizes 40 and 50 for MIMIC-III and PubMed respectively. We trained a total of 4,113,665 parameters with batch size of 128 for 250 epochs. The training time for this model was 16.6 hours. This is the main version of the Fujitsu Text Generation model, but we generated a second analogue model with the training data provided by CLEF eHealth challenge. In our experiments, we tested the performance of both versions.

After the prediction of the NER and the linking process to the CIE-10 codes with the help of the Knowledge Graph, we performed 4 different methods of post-processing of the results. We presented 4 versions of our results, one for each type of post-processing method applied. Table 1 presents the identifiers of each version, and then we describe the methods of post-processing followed.

- *No position overlap:* After locating the position of each entity found in the text, we remove entities with overlapping positions and we only keep the longest entity. For instance, if we detect the entities “hipertensión ocular” and “hipertensión”, the second entity is a substring of the first, that is, it is in the same position of the word “hipertensión” from the first entity. In this case, we only keep the first entity, which is the longest one.

[9] https://github.com/google-research/bert

[10] https://github.com/dmis-lab/biobert

Table 1: Identifiers of the results by task presented.

Results Identifiers			Description
Task 1	Task 2	Task 3	
CodiEspD_v1	CodiEspP_v1	CodiEspX_v1	No position overlap.
CodiEspD_v2	CodiEspP_v2	CodiEspX_v2	No position overlap. Denied entities removed.
CodiEspD_v3	CodiEspP_v3	CodiEspX_v3	No word overlap.
CodiEspD_v4	CodiEspP_v4	CodiEspX_v4	No word overlap. Denied entities removed.

- *No position overlap. Denied entities removed:* This follows the same approach as above. We remove overlapped entities, but we add an extra step by using the NegEx-MES tool to find instances of negated entities. This includes instances where a text states that a patient does not present a disease.
- *No word overlap:* It is similar to *No position overlap* method but here, on top of the position-level overlap, we also consider the word-level overlap. We keep the code of the longest entity when word overlap exists, regardless of the position of the entity in the text. Following the previous example, if in a text the entity “hipertensión ocular” exists while in other part of the same text the entity “hipertensión” also exists, then we associate the code of “hipertensión ocular” to the entity “hipertensión”.
- *No word overlap. Denied entities removed:* This result is obtained by applying the NegEx-MES tool to the *No word overlap* process.

3.2 Experiments and Results

The performance of each experiment in the building of the NER model is shown in the Table 2.

Table 2: Results of the experiments performed to build our Named Entity Recognition model.

Experiment Identifier	**Train Data**	**Test Data**	**Learned Entity**	**F1-Score**
Baseline	Data-1	Val Data-1	diagnostico	56.82
Abstracts	Data-1 + Abstracts	Val Data-1	diagnostico	24.05
MIMIC-III	Data-1 + MIMIC-III	Val Data-1	diagnostico	57.10
Fujitsu Augmentation	Data-1 + Augmentation	Val Data-1	diagnostico	58.03
Fujitsu Augmentation fine-tune	Data-1 + Augmentation_extended	Val Data-1	diagnostico	58.40
Final T1	Data-2 + Augmentation_extended	Val Data-2	diagnostico	**72.45**
Final T2	Data-2 + Augmentation_extended	Val Data-2	procedimiento	**76.70**

- *Baseline:* Firstly, we obtain a baseline performance by fine tuning BERT for the entity "diagnostico". This baseline is trained with the first version of the dataset provided by the organizers (Data-1) and it is tested with the validation set (Val Data-1). We achieve an F1-Score of 56.82 with this experiment. Based on this result, we are going to proceed to adjusting training parameters.
- *Abstracts:* Besides training and validation sets, organizers provided extra text annotations from literature abstracts. So, in this experiment, we trained an NER with the initial training set and the data from the abstracts. The performance decreased, so we discarded the abstract data for the following training iterations. We believe the data imbalance and noise introduced to the training by the abstracts are responsible of this low performance.
- *MIMIC-III:* Similar to the previous experiment, we trained an NER with the initial training set plus data from MIMIC-III dataset, taking care not to outnumber the data points of CLEF eHealth data. This experiment resulted in a better performance over the baseline. Therefore, we thought that data augmentation would be a useful tool to achieve higher performance.
- *Fujitsu Augmentation:* We used the Text Generation model trained with Fujitsu's proprietary Decentralized Learning technology. This Text Generation model was trained with MIMIC-III and PubMed data. This model receives a seed text as input and it generates text similar to the medical domain text. In our experiments, the seed text is the word or the set of words that conform to an entity. By using our Text Generation model, we duplicate the number of samples in the initial train set. And by the randomness of the generated text, we add robustness to the final NER model. In this experiment we achieve an F1-Score of 58.03.
- *Fujitsu Augmentation fine-tune:* In this experiment we fine-tune the Text Generation model with data from CLEF eHealth challenge. This means the Text Generation model learns from MIMIC-III, PubMed and CLEF eHealth data. Similar to the previous experiment, we duplicate the training samples and we obtain an NER with a performance of 58.40. This is the final methodology used to train the NER model in our system.

Using the method of experiment "Fujitsu Augmentation fine-tune" we trained two NER models with the latest released datasets (Data-2) from the task 3 for type "diagnostico" (Final T1) and for type "procedimiento" (Final T2). With this method and the updated dataset we go from 56.82 to 72.45 F1-Score for the entity "diagnostico". And these NER models are the ones used alongside the linking algorithm with our Knowledge Graph to obtain our results over the test set.

After post-processing the data with the 4 methods described in Section 3.1, we submitted the results for evaluation. Table 3 shows the performance obtained after the evaluation carried out by the organizers of the CLEF eHealth challenge.

From the evaluation tables we see that for task 1, $V1$ achieves the highest Mean Average Precision (MAP) and F1-Score. $V3$ and $V4$ achieved the highest MAP and F1-Score for task 2 of the challenge and $V1$ was the highest performance approach for task 3. We associate the lower results in task 2 to issues with bigger specificity regarding body parts and concrete parameters in CIE10-PCS. There are cases where our approach

Table 3: Performance results provided by organizers of the CLEF eHealth challenge for the four versions of our results over the test set.

File	MAP	MAP codes	MAP30	MAP30 codes	P	R	F1	P codes	R codes	F1 codes	P cat	R cat	F1 cat
CodiEspD_v1	**0.519**	0.598	0.519	0.597	0.732	0.633	**0.679**	0.767	0.699	0.731	0.802	0.734	0.766
CodiEspD_v2	0.481	0.553	0.48	0.553	0.733	0.588	0.652	0.768	0.646	0.702	0.804	0.687	0.741
CodiEspD_v3	0.501	0.576	0.501	0.576	0.74	0.604	0.665	0.775	0.665	0.716	0.807	0.714	0.758
CodiEspD_v4	0.46	0.528	0.46	0.528	0.739	0.556	0.635	0.774	0.61	0.682	0.809	0.662	0.728
CodiEspP_v1	0.434	0.515	0.433	0.514	0.587	0.448	0.508	0.627	0.539	0.58	0.665	0.468	0.549
CodiEspP_v2	0.433	0.513	0.432	0.512	0.587	0.446	0.507	0.626	0.537	0.578	0.665	0.465	0.548
CodiEspP_v3	**0.443**	0.525	0.443	0.525	0.643	0.428	**0.514**	0.692	0.514	0.59	0.687	0.462	0.552
CodiEspP_v4	**0.440**	0.52	0.440	0.52	0.642	0.424	**0.511**	0.692	0.51	0.587	0.687	0.458	0.55
CodiEspX_v1	-	-	-	-	0.669	0.562	**0.611**	0.704	0.634	0.667	-	-	-
CodiEspX_v2	-	-	-	-	0.667	0.527	0.589	0.702	0.592	0.642	-	-	-
CodiEspX_v3	-	-	-	-	0.687	0.537	0.603	0.725	0.604	0.659	-	-	-
CodiEspX_v4	-	-	-	-	0.685	0.505	0.581	0.722	0.566	0.635	-	-	-

is missing the specifications of the procedure. In those cases, we are retrieving wrong code annotations, what decreases the precision and recall.

We can highlight the versions of post-processing that did not remove negated entities as the highest achieving approaches. We conclude, after analysis over the ground-truth, that negated entities are contemplated as part of the expected results. In the case of task 1, out of the 2841 negated entities that we removed from our results, 2059 appear in the ground-truth. And in case of task 2, the ground-truth contains 61 appearances from the 58 negated entities removed from our results. This is the reason why versions of our results that do not remove negated entities ($V1$ and $V3$) outperform the versions that remove negated entities ($V2$ and $V4$) across all the tasks. And due to the small number of negated entities we found in task 2, the performance differences between versions is negligible.

4 Conclusions

In this paper we present the methods used in the CLEF eHealth 2020 Challenge for Automated Clinical Encoding. This challenge consisted of 3 tasks for CIE-10 code assignment of texts written in Spanish for diagnoses (task 1) and procedures (task 2), and identifying the positioning of those entities in the texts (task 3).

We followed an approach composed of a Knowledge Graph created with the CIE-10 standard and the training data annotations. Then we fine-tuned a multilingual BERT-based network for Named Entity Recognition to predict entities in the clinical texts. We used data augmentation through Fujitsu proprietary Text Generation Model, trained with Decentralized Learning from the datasets MIMIC-III and PubMed, to create synthetic samples for the NER training. With the output of our NER models and the Knowledge Graph, we developed a linking algorithm to assign a CIE-10 code to predicted entities in the texts. Finally, we post-processed the output of our linking algorithm to remove negated entities using the NegEx-MES tool. We also analyzed the output to provide 4 different versions of our results taking into account the overlapping words and positions of predicted entities.

Our approach achieves F1-Scores of 0.67, 0.51 and 0.61 for tasks 1, 2 and 3 respectively. The versions of our results that achieve higher performance are the ones that do not remove negated entities, named CodiEspD_v1, CodiEspP_v3 and CodiEspX_v1.

References

1. Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K.A., Wixted, M.K.: MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_67.pdf
2. Atutxa, A., Casillas, A., Ezeiza, N., Fresno, V., Goenaga, I., Gojenola, K., Martínez, R., Anchordoqui, M.O., Perez-de-Viñaspre, O.: IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence Approach. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/paper_167.pdf
3. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., Elhadad, N.: Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment. In: The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Workshops, vol. WS-18, pp. 409–416. AAAI Press (2018), https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16881
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. García-Santa, N., San-Miguel, B., Ugai, T.: The Magic of Semantic Enrichment and NLP for Medical Coding. In: The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers. Lecture Notes in Computer Science, vol. 11762, pp. 58–63. Springer (2019). https://doi.org/10.1007/978-3-030-32327-1_12
6. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez Gonzales, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth Evaluation Lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . LNCS Volume number: 12260 (2020)
7. ICD, W.: 10: International statistical classification of diseases and related health problems. World Health Organization, Geneva (1992)
8. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**, 160035 (2016). https://doi.org/10.13026/C2XW26
9. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. **36**(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682
10. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)
11. Lindberg, D.: Internet access to the National Library of Medicine. Effective clinical practice: ECP **3**(5), 256 (2000)

12. Miftahutdinov, Z., Tutubalina, E.: KFU at CLEF ehealth 2017 task 1: ICD-10 coding of english death certificates with recurrent neural networks. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1866/paper_64.pdf
13. Miranda, A., Rana, A., Krallinger, M.: Abstracts from Lilacs and Ibecs with ICD10 codes (Jan 2020). https://doi.org/10.5281/zenodo.3606626, https://doi.org/10.5281/zenodo.3606626, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
14. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)
15. Miranda-Escalada, A., Krallinger, M.: CodiEsp codes: list of valid CIE10 codes for the CodiEsp task (Jan 2020), https://doi.org/10.5281/zenodo.3706838, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
16. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable Prediction of Medical Codes from Clinical Text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). pp. 1101–1111. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-1100
17. Névéol, A., Robert, A., Anderson, R., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1866/invited_paper_6.pdf
18. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., Ramadier, L., Rey, G., Zweigenbaum, P.: CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/invited_paper_18.pdf
19. Neves, M.L., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the CLEF eHealth 2019 Multilingual Information Extraction. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_251.pdf
20. Pakhomov, S.V., Buntrock, J.D., Chute, C.G.: Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association **13**(5), 516–525 (2006)
21. Sänger, M., Weber, L., Kittner, M., Leser, U.: Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task 1. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_81.pdf
22. Wei, C.H., Allot, A., Leaman, R., Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic acids research **47**(W1), W587–W593 (2019)