ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT

Guillermo López-García, José M. Jerez, and Francisco J. Veredas

Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática,
Universidad de Málaga, Málaga (Spain)
{guilopgar,jmjerez,franveredas}@uma.es

Abstract. This working notes paper presents our contribution to the CLEF eHealth 2020 Task 1. Our team has participated in the CodiEsp-D subtask, the first shared task consisted in the automatic clinical coding of medical cases in Spanish, annotated with ICD-10-CM codes. We tackled the task as a multi-label classification problem using BERT model [4]. With the aim of leveraging all the language modeling capacities of the deep bidirectional encoder architecture of BERT, we developed a tailored approach to annotate short fragments of text extracted from the long clinical cases present in the CodiEsp corpus and use them as input to the model. Two publicly available Spanish versions of BERT, namely BETO [3] and BERT-SciELO [1], were fine-tuned on the CodiEsp-D corpus extended by a set of abstracts annotated with ICD-10 codes, following our fragment-based classification approach. BERT-SciELO, a BERT-Base model pre-trained from scratch on an unlabeled corpus of biomedical articles in Spanish, achieved the best results among our three submitted systems, obtaining a final Mean Average Precision (MAP) metric score of 0.482 on the evaluation set.

Keywords: Clinical coding · Spanish clinical cases · BERT · Text classification · Transfer learning · Clinical NLP

1 Introduction

The increasingly adoption of electronic health records (EHRs) as a key component in many hospital information systems across the globe has posed a series of questions to the scientific community that remain partially unresolved. One of the main issues is how to effectively leverage the information stored in the system to improve patient care. EHRs store heterogeneous data in a wide variety of formats, including *free-text* documents like clinical notes [20]. These medical textual representations contain crucial patient information related to diagnosis, treatments, or procedures. However, their unstructured nature makes it specially challenging to extract the relevant medical concepts from the data.

Automatic clinical coding is the task of transforming unstructured clinical text into a structured format, following standard coding terminologies and using computational methods [22]. The produced structured data can be subsequently used for medical billing, conducting epidemiological studies, exchanging information between medical institutions, performing statistical analysis, and many other purposes, with the additional advantage of not requiring any human intervention throughout the coding process. Given the importance of storing natural language descriptions of medical cases in modern EHR systems, automatic clinical coding constitutes an essential task in the process of extracting valuable information from EHR data, improving many aspects of clinical care.

Historically, natural language processing (NLP) techniques have been applied to the problem of clinical coding [15, 18, 17, 8]. However, most of the previous works only focus on English text, as the availability of corpora annotated with clinical coding information and additional linguistic resources in languages other than English is scarce. With the intention of overcoming this issue, over the past four years, the CLEF eHealth Lab has organised a series of clinical coding shared tasks on non-English or multilingual corpora. Concretely, CLEF eHealth 2016 Task 2 focused on the assignment of International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes to French free-text death certificates, using CépiDC corpus [12]. In 2017, the CLEF eHealth Task 1 continued with automatic coding of death certificates, but turning the challenge into a bilingual task, using both CépiDC (French) and CDC (English) annotated corpora [11]. One year later, the CLEF eHealth 2018 Task 1 explored a multilingual clinical coding challenge, with the assignment of ICD-10 codes to death reports in French, Hungarian and Italian [13]. Finally, last year, the CLEF eHealth 2019 Task 1 again focused on the assignment of ICD-10 codes, but instead of using death reports, non-technical summaries of animal experiments in German were employed [14].

This year, CLEF eHealth 2020 Task 1 corresponds to the CodiEsp track [10], the first shared task consisted in the automatic coding of clinical cases in Spanish, using the Spanish version of ICD-10 (CIE-10) and the CodiEsp corpus, a synthetic corpus of $1K$ clinical cases in Spanish manually curated by the organisers of the task. The CodiEsp track is composed of three different subtasks: CodiEsp Diagnosis (CodiEsp-D) subtask, CodiEsp Procedure (CodiEsp-P) subtask and Explainable AI (CodiEsp-X) subtask. Given a free-text clinical case in Spanish, CodiEsp-D subtask requires assigning a set of diagnosis codes—ICD-10-CM or CIE-10 *Diagnóstico* in Spanish—to the medical document, whereas in CodiEsp-P subtask procedures codes—ICD-10-PCS or CIE-10 *Procedimiento*—are predicted for each clinical case. On the other hand, systems participating in CodiEsp-X subtask are required to predict both diagnosis and procedures codes, but also to provide references in the text justifying the coding predictions.

In this work, we present our contribution to the CLEF eHealth 2020 [5], where our team has participated in the CodiEsp-D subtask. We have tackled the problem as a multi-label text classification task using BERT [4], a contextualized neural language model that achieved state-of-the-art results on eleven distinct

NLP tasks, and was employed by the best performing team in the CLEF eHealth 2019 Task 1 [19]. Since BERT was specially designed to process short fragments of text—in contrast with the long clinical notes present in the CodiEsp corpus—, we have developed a tailored approach to turn the text classification task into a short segments text classification problem, in order to leverage all the predictive capabilities of the BERT model when applied to the CodiEsp corpus. In this way, each medical document from the corpus was split into short fragments of text. Then, using the information available for the CodiEsp-X subtask, we annotated each short fragment with ICD-10-CM codes information, and used the annotated segments as input to the model. Once the probabilities for individual codes were predicted by the model on every fragment of a document, a maximum probability criterion was used to obtain the probability of each ICD-10 diagnosis code at the document level. Finally, for every document, a list of codes ordered by confidence or relevance was produced, which was used by the organisers to evaluate the performance of the participating systems. For reproducibility purposes, all the code generated to implement our approach is publicly available at `https://github.com/guilopgar/CLEF-2020-CodiEsp`.

The rest of the paper is organised as follows. In Section 2, a short description of the CodiEsp corpus is given, as well as the details of our proposed strategy to tackle the CodiEsp-D subtask are described. We analyze the obtained results in Section 3, whereas Section 4 presents some conclusions and perspectives for future work.

2 Materials and Methods

2.1 Corpora

According to the task organisers[1], the CodiEsp corpus comprises $1K$ clinical case studies covering a broad diversity of medical subjects, such as pneumology, urology, cardiology or oncology. The entire corpus was randomly split into three different subsets, the training (500 documents), development (250 documents) and test (250 documents) sets. Also, jointly with the test set, a background set of 2751 clinical cases was released. The latter had the intention of guaranteeing that teams participating in the CodiEsp track could not do manual corrections, since submissions to the task had to include predictions for the 3001 medical documents present in both the background and test sets, but they were only evaluated on the test set.

For the CodiEsp-D subtask, annotation tables containing the assignment of ICD-10-CM codes to the medical documents in the corpus were also available. Furthermore, along with the CodiEsp corpus, a set of additional documents was provided[2]. The supplementary documents correspond to Spanish abstracts obtained from LILACS[3] and IBECS[4] biomedical literature databases, which were

[1] `https://temu.bsc.es/codiesp/index.php/category/data/`
[2] `https://zenodo.org/record/3606662#.XvxBT59fg8o`
[3] `https://lilacs.bvsalud.org/es/`
[4] `http://ibecs.isciii.es/`

annotated with ICD-10 codes information. From the whole set of abstracts, we only selected those texts annotated with ICD-10-CM codes present in the list of valid codes for the CodiEsp track supplied by the organisers[5].

Table 1 contains a basic description of the CodiEsp-D corpus (training, development and test subsets) as well as the additional abstracts annotated with diagnosis codes information. As it is shown in the table, CodiEsp-D is a considerably challenging task, given the limited number of clinical cases (1000) and the large number of unique codes (2557) present in the corpus, including 363 codes that are only present in the test subset. Additionally, the number of codes annotations is also scarce, resulting in a highly imbalanced multi-label classification problem, where for each code, the number of negative samples clearly surpasses the number of positive cases, i.e. number of documents annotated with a certain code. For these reasons, we decided to expand the CodiEsp-D corpus using the additional set of abstracts. Considering that most of the diagnosis codes (2153 out of 2984) included in the abstracts corpus are not present in the CodiEsp-D annotations, we experimented with two distinct ways of expanding the training and development CodiEsp-D corpora: either using all available abstracts, or, alternatively, solely using the abstracts annotated with ICD-10-CM codes contained in the training and development sets, leading to a reduced version of the abstracts corpus comprising 115457 documents and 160652 codes annotations, from which 733 are unique ICD-10 codes.

Table 1. Summary of CodiEsp-D and additional abstracts corpora annotated with ICD-10-CM codes.

	Training	**Development**	**Test**	**Abstracts**
Documents	500	250	250	170120
Total ICD Codes	5639	2677	2842	403856
Avg. ICD codes per doc.	11.278	10.708	11.368	2.374
Unique ICD codes	1767	1158	1143	2984
Avg. docs. per ICD code	3.191	2.312	2.486	135.340
Unique unseen ICD codes	-	427	363	2153

2.2 Classification system

We have tackled the CodiEsp-D challenge using BERT model [4]. BERT is a contextual language representation model, based on the encoder part of the Transformer architecture [24], designed to extract deep bidirectional contextual representations both at the token and sentence level. The model can be pretrained on an unlabeled corpus in an unsupervised manner using two language modeling objectives, namely next sentence prediction and masked language modeling. Unlike other contextual language models, BERT can be transferred to be

[5] `https://zenodo.org/record/3706838#.XvxD3J9fg8o`

used in a downstream task, by fine-tuning the whole architecture in a supervised way, then adapting all its pretrained weights to solve a specific task.

BERT has gained a lot of attention in the NLP community, including in biomedical and clinical domains, where BERT-based approaches have obtained state-of-the-art results in a wide range of tasks [7, 21]. In this work, we have experimented with two Spanish versions of the BERT-Base architecture: BETO [3] and BERT-SciELO [1] models. While both models were pretrained from scratch on unlabeled Spanish corpora, BETO was trained on a compilation of general domain texts[6], whereas BERT-SciELO was pre-trained on a corpus of biomedical articles retrieved from SciELO[7].

In contrast to sequential models such as recurrent neural networks (RNNs), for computational feasibility reasons, Transformer architectures cannot deal with long input sequences of variable length, since for the self-attention layers the complexity is quadratic on the length of the input sequence [24]. For instance, in the original implementation of BERT—which uses WordPiece [6] tokenization to subdivide each input token into further sub-token units—, the maximum input sub-tokens sequence length is 512. This constitutes an important limitation when dealing with document or long-text classification tasks like CodiEsp-D, where the sub-tokens sequence size of many clinical cases is clearly above the maximum length supported by BERT.

Over the last year, a few works have already explored different strategies to overcome this limitation. The most straightforward approach is to use a text truncation method, like the one adopted by the CLEF eHealth 2019 Task 1 best performing team [19], which simply consisted in using the first 510[8] sub-tokens of each document as input to the model. In [23], additional truncation strategies were considered, and the best performing method was to select the first 128 and the last 382 sub-tokens from each input text. Authors hypothesized that the most relevant information in a document appears at the beginning and at the end of it, which may be the case for the texts present in the corpora analyzed in that work, specifically the movie review IMDb corpus and the Chinese Sogou news articles dataset. However, in the CodiEsp-D corpus, clinical information relevant to solve the task may be spread anywhere within the clinical cases, hence eliminating parts of the documents may not be the most appropriate strategy for our needs.

Another recent work explored a different approach that does not make use of any truncation method to adapt BERT to solve a text classification task. In this way, authors in [16] proposed to segment the input texts into smaller parts, and then fine-tune the BERT model on a segment-level supervised task, assigning to each segment the labels associated with the entire document where the segment comes from. If we applied the same strategy to solve the CodiEsp-D multi-label classification task, we would annotate all the segments obtained from a single

[6] `https://github.com/josecannete/spanish-corpora`

[7] `https://www.scielo.org/es/`

[8] Since BERT always adds two special tokens (`[CLS]` and `[SEP]`) at the first and last positions, respectively, of an input sequence.

clinical case with the same ICD-10-CM codes present in the complete document. This constitutes a problematic situation, as many fragments would be annotated with diagnosis codes which are not represented within the fragment, but appear in other segments from the same clinical document.

With the aim of adopting a strategy to adjust BERT to the distinctive features of the CodiEsp-D subtask, we have developed a three-phases custom approach that transforms the multi-label long-text classification task into a multi-label short-fragment classification problem. Unlike [16], using the annotations available for CodiEsp-X named-entity recognition (NER) subtask, for each fragment we only assign the labels occurring within the specific fragment, avoiding misleading the model by assigning codes to a fragment that are present in other parts of the document. In the next paragraphs, the three stages of the developed approach are described.

Splitting each clinical case into fragments. As indicated before, BERT supports input sequence lengths up to a maximum value of N ($N = 512$ in the original implementation). For this reason, for every clinical case in the CodiEsp-D corpus, after performing WordPiece tokenization, we split the resulting sub-tokens sequence $w = (w_1, w_2, ..., w_k)$, of length k, into a sequence of $m = \lceil k/(N-2) \rceil$ contiguous sub-token fragments $f = (f_1, f_2, ..., f_m) = ((w_1, ..., w_{N-2}), (w_{(N-2)+1}, ..., w_{2*(N-2)}), ..., (w_{(m-1)*(N-2)+1}, ..., w_k))$, in which the $m-1$ first fragments $f_1, .., f_{m-1}$ have a length of $N-2$ sub-tokens while the last fragment f_m contains the remaining final sub-tokens of the document (considering that the tokens `[CLS]` and `[SEP]` have to be later added at the beginning and the end of each fragment, respectively, to get sequences of size N sub-tokens that will constitute the input to BERT).

Annotating text fragments with ICD-10-CM codes. Annotations available for the CodiEsp-D subtask contain the assignment of a set of diagnosis codes to each clinical document (see Fig. 1A). On the other hand, in CodiEsp-X subtask—which uses the same CodiEsp corpus as the CodiEsp-D subtask—, codes annotations include an extra field indicating the reference in the text that explains the coding assignment (see Fig. 1B). In this way, using the information provided for the CodiEsp-X subtask, we managed to annotate each of the fragments—resulting from the splitting procedure—exclusively with those ICD-10-CM codes annotations whose text references were contained within the fragment. Thus, given a certain fragment f_l and the set of its associated diagnosis codes C_{f_l}, as well as a CodiEsp-X annotation a_i consisted in a code c_i and a text reference spanning a group of λ continuous or discontinuous sub-tokens[9] $t_i = (w_{i1}, w_{i2}, ..., w_{i\lambda})$, we annotated f_l with c_i when any of the sub-tokens of t_i was contained inside the limits of f_l, i.e. $c_i \in C_{f_l} \iff \exists w_{ij} \in t_i \mid w_{ij_s} \geq f_{l_s} \wedge w_{ij_e} \leq f_{l_e}$, where w_{ij_s} and w_{ij_e} stand for the starting and ending character positions of the sub-token w_{ij}, respectively, whereas f_{l_s} and f_{l_e} represent the

[9] Considering that annotations text references were also tokenized using WordPiece.

starting and ending character positions of the fragment f_l. To illustrate the annotation process, in Fig. 2, we show the annotations generated for the fragments extracted from the `S1139-76322012000400011-1` CodiEsp document using the information available for the same document in the CodiEsp-X subtask (see Fig. 1B).

Obtaining codes probabilities at document level. Using the sub-tokens fragments from the clinical cases in the CodiEsp corpus annotated with ICD-10-CM codes, we trained the BERT model on a fragment-level classification problem. Since many distinct codes may be assigned to the same medical document, we treated the task as a multi-label classification problem. To fine-tune the whole architecture of BERT on a supervised learning task at text-level, the representation produced by the model for the initial `[CLS]` token was fed into a final output layer for classification. As we were dealing with a multi-label task—which is equivalent to multiple independent binary classification tasks—, we used the sigmoid activation function and D output units, with D representing the number of unique diagnosis codes occurring in the texts used to train the model. Hence, given a text fragment as input to the model, the produced output vector could be interpreted as the probability of each code to occur within the input fragment. However, CodiEsp-D task was a document classification problem, and the evaluation of the participating systems was performed at document level. Accordingly, using a maximum probability criterion, we post-processed the model predicted probabilities to obtain codes probabilities at document level. Thus, given a sequence f containing all fragments generated from a single document d as input, the model produces the output probability matrix $M \in \mathbb{R}^{|f| \times D}$. Then, selecting the maximum probability value across every column of M, a final vector $p \in \mathbb{R}^D$ of codes probabilities at document level is generated, which contains the probability of each code to appear in d. Using this method, we were able to produce, for each clinical case, a list of D distinct codes sorted in descending order according to their probability values predicted by the model, which was finally used by the organisers to evaluate the performance of the classification system.

It should be noted that, in the case of the additional abstracts corpus (see Section 2.1), the available codes annotations did not contain any extra field indicating the text reference that supports the coding assignment. Therefore, our fragment-based custom approach could not be applied to the abstracts corpus, but only to the CodiEsp-D corpus enriched with the information available for the CodiEsp-X subtask. For this reason, exclusively the abstracts that, after performing WordPiece tokenization, contain a maximum number of $N - 2$ sub-tokens were used to expand the CodiEsp-D corpus.

2.3 Experiments

We experimented with two Spanish versions of the BERT-Base model, namely BETO and BERT-SciELO. Since the vocabulary of the BERT-SciELO model

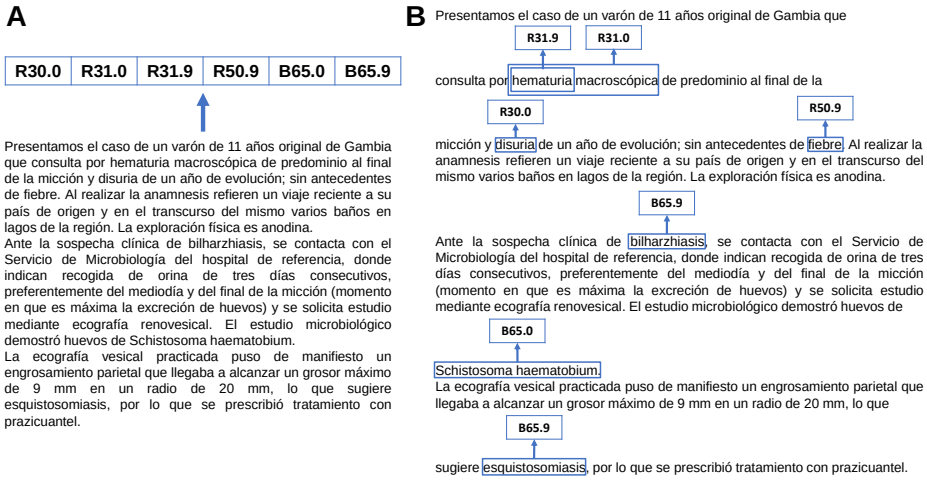

Fig. 1. Illustration of annotation formats using the `S1139-76322012000400011-1` clinical case from the CodiEsp training corpus. **A** Codes annotations available for the CodiEsp-D subtask. **B** Diagnosis codes annotations available for the CodiEsp-X subtask.

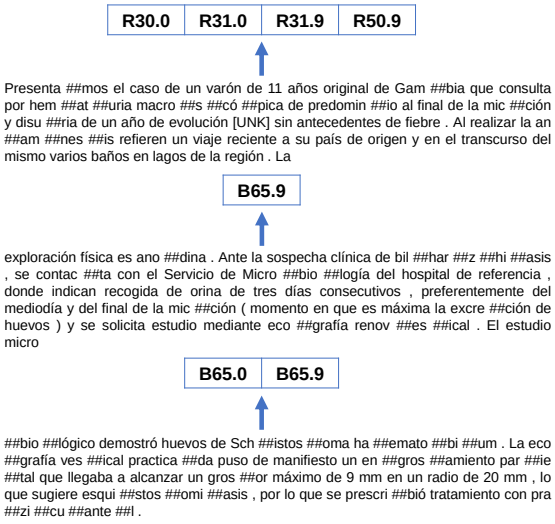

Fig. 2. Illustration of the text fragments ICD-10-CM codes annotations obtained after applying the first two phases of our developed approach to the `S1139-76322012000400011-1` clinical case from the CodiEsp training corpus. To generate the sub-tokens fragments, we used the WorPiece tokenizer of the BETO model, setting a maximum fragment length of $N - 2 = 78$ sub-tokens.

did not include any punctuation character, we used a pre-processed version of the expanded CodiEsp-D corpus (together with the additional abstracts) in which punctuation marks were substituted by spaces. In the case of BETO, the raw text from the documents was employed, as punctuation marks were contained in the vocabulary of the model. Regarding the hardware resources employed, all experiments were executed on a single GeForce GTX 1080 Ti 11 GB GPU. Given the limitations imposed by the hardware, we used a maximum input sequence length of $N = 230$ for the BERT-SciELO and a value of $N = 275$ for the BETO model, with both models having $\sim 184M$ trainable weights. Finally, with respect to the hyperparameters of the models, for fine-tuning, we used RAdam [9] with learning rate of 3×10^{-5}, a batch size of 16 and the number of epochs were experimentally determined on the CodiEsp-D development set using early-stopping, with an upper limit of 40 epochs.

3 Results

In this section, we present the results obtained by our team, ICB-UMA, at the CodiEsp-D subtask. Task organisers allowed each participating team to submit up to 5 runs of their classification systems. We submitted three distinct runs. The first two submissions (ICB-UMA-run1 and ICB-UMA-run2) corresponded to the BERT-SciELO model fine-tuned on the CodiEsp-D training and development corpora expanded using additional abstracts annotated with ICD-10-CM codes present in the training and development sets. Thus, the extended corpus contained 2194 distinct diagnosis codes—the number of unique ICD-10-CM codes present in the training and development corpora (see Table 1)—, having a final BERT-SciELO model with an output classification layer of $D = 2194$ units. Since exclusively the abstracts containing a maximum number of $230 - 2$ WordPiece sub-tokens were considered for the BERT-SciELO model (see Section 2.2), a final set of 66620 abstracts with 92239 codes annotations were employed to expand the training and development corpora. Submission ICB-UMA-run2 used a dropout probability of 0.1 on the output layer of BERT-SciELO model, whereas in submission ICB-UMA-run1 no dropout was used on the output layer. For its part, in submission ICB-UMA-run3 the BETO model was fine-tuned on the CodiEsp-D training and development sets extended using all available abstracts with a maximum number of $275 - 2$ WordPiece sub-tokens, which comprised a total of 87871 documents and 208076 diagnosis codes annotation, from which 2204 were unique ICD-10-CM codes, i.e. codes that were not present in the training and development corpora. Finally, for submission ICB-UMA-run3 the output layer of the BETO model had $D = 4398$ units, with no dropout on the final layer. Apart from these three runs, we also experimented with the BERT-SciELO model fine-tuned on the CodiEsp-D corpus expanded using all available abstracts, as well as with the BETO model fine-tuned on the CodiEsp-D corpus extended using additional abstracts annotated with training and development diagnosis codes. However, these two approaches obtained worse results on the

development set, and we submitted the three strategies that best performed on the development corpus.

Table 2 and Table 3 show the predictive performance of our three different submitted runs on the CodiEsp test corpus, as a result of the evaluation performed by the CodiEsp track organisers. Namely, Table 2 presents the results obtained according to the official evaluation metric of the CodiEsp-D subtask, i.e. the Mean Average Precision (MAP). The second column of the table contains the MAP values calculated considering all codes present in the CodiEsp test subset, whereas the results shown in the third column (*MAP codes*) were computed taking into account only the predictions for the test codes that were also present in the training and development subsets. Finally, the fourth column (*MAP30*) measures MAP-at-k metric (MAP@k), with k equals 30, while the last column (*MAP30 codes*) measures MAP@30 considering only the test codes that were present in the training and development corpora (as in the third column of the table). According to the results observed in Table 2, the BERT-SciELO-based model outperformed the BETO-based classifier on the CodiEsp-D subtask, since both the ICB-UMA-run1 and ICB-UMA-run2 systems achieved higher values for all analyzed metrics than the ICB-UMA-run3 submitted system. The best performance is obtained by the BERT-SciELO model when no dropout is used on the output layer, as the ICB-UMA-run1 results slightly surpassed the performance of the ICB-UMA-run2 system across the four examined evaluation metrics. To summarize, we can say that, according to our obtained results, the BERT-SciELO model outperformed the BETO on the CodiEsp-D predictive task. Therefore, pre-training the BERT-Base architecture from scratch on an unlabeled corpus of Spanish biomedical articles, rather than using a general domain Spanish corpus, leads to a better performance of the model on a clinical coding task in the context of Spanish medical narrative. The results obtained in this work for the CodiEsp-D subtask support the hypothesis already explored in previous works [2, 21], claiming that a clinical domain-specific BERT yields superior performance on medical classification problems than a general domain version of the model.

Table 2. Classification performance of each submitted run assessed using MAP, the official CodiEsp-D subtask evaluation metric.

Submission	**MAP**	**MAP codes**	**MAP30**	**MAP30 codes**
ICB-UMA-run1	0.482	0.567	0.460	0.542
ICB-UMA-run2	0.471	0.554	0.449	0.529
ICB-UMA-run3	0.455	0.536	0.430	0.509

On the other hand, to perform a larger analysis of the results, the task organisers evaluated the performance of the systems according to a set of additional metrics, other than the official ones. In this way, in Table 3, the second, third and fourth columns show the calculated values using precision (P), recall (R) and the F_1 score ($F1$) metrics, respectively, considering all codes present in the

CodiEsp-D test set, while the fifth (*P codes*), sixth (*R codes*) and seventh (*F1 codes*) columns contain the results computed using the same three metrics but taking into consideration only the codes present in the training and development subsets. Finally, in the last three columns (*P cat*, *R cat* and *F1 cat*), the previous metrics are used to evaluate the submitted predictions at the hierarchical category level of the ICD-10-CM codes contained in the test set. As it can be observed from Table 3, for the precision and consequently for the F_1 score, our three submitted systems obtained extremely poor values, though for the recall the computed values are abnormally high. The reason for this is that, with the aim of maximizing the score obtained for the official evaluation metric, i.e. MAP, for each test document we submitted all codes considered by the model—2194 codes in the case of ICB-UMA-run1 and ICB-UMA-run2 and 4398 codes for the ICB-UMA-run3 submission—ordered by their predicted probability of occurrence (see Section 2.2). If we had optimized precision, recall and F1 score metrics instead of MAP, in place of submitting all codes, a decision threshold would have been defined to select solely a subset of the codes according to their predicted probabilities.

Table 3. Classification performance of each submitted run evaluated according to additional metrics.

Submission	**P**	**R**	**F1**	**P codes**	**R codes**	**F1 codes**	**P cat**	**R cat**	**F1 cat**
ICB-UMA-run1	0.004	0.858	0.009	0.004	1.0	0.009	0.010	0.968	0.021
ICB-UMA-run2	0.004	0.858	0.009	0.004	1.0	0.009	0.010	0.968	0.021
ICB-UMA-run3	0.002	0.897	0.005	0.004	1.0	0.009	0.008	0.987	0.016

4 Conclusion

In this paper, we present our contribution to the CodiEsp-D subtask from the CodiEsp track [10] of CLEF eHealth 2020 [5]. The shared task proposes the automatic assignment of ICD-10-CM codes to Spanish clinical cases. The scarce number of medical cases present in the CodiEsp corpus in combination with the large quantity of unique diagnosis codes used to annotate the documents, make CodiEsp-D a considerably challenging task.

We have tackled the challenge as a multi-label text classification task using BERT model [4]. A fragment-based classification approach was developed in order to take advantage of all the predictive capacity of BERT when receiving the long-text clinical cases contained in the CodiEsp corpus as input to the model. Our strategy consisted in using the available annotations for the CodiEsp-X NER subtask to turn the CodiEsp-D multi-label document classification task into a multi-label short-fragment classification problem. We experimented with two publicly available Spanish versions of the BERT model, fine-tuned on the CodiEsp-D corpus expanded using a set of available abstracts annotated with

ICD-10-CM codes. BERT-SciELO [1], a BERT-Base architecture pre-trained on a corpus of biomedical articles in Spanish, yielded the best performance among our three submitted systems, obtaining a MAP score of 0.482 on the evaluation set. The obtained results in this work reinforced the idea that a medical domain version of BERT achieves higher performance on clinical classification tasks than nonspecific domain versions of the model.

In future works, we will try to enhance the developed fragment-based classification strategy to further improve the obtained results on the CodiEsp-D subtask. For instance, when splitting each clinical case into fragments, we could perform the text segmentation at the sentence level, producing fragments comprising a sequence of sentences with a complete semantic meaning. On the other hand, given the superior performance observed from the BERT-SciELO model, it is worth investigating whether alternative clinical-specific versions of BERT pre-trained on Spanish medical corpora more similar to the CodiEsp corpus could increase the results further. Additionally, due to the widespread adoption of BERT in multi-lingual setups across domains, we could also explore the pre-training of the model on a multi-lingual medical corpus. This would permit the creation of enormous medical corpora comprising clinical documents written in many different languages. Because of the sub-word vocabulary employed by BERT and the common etymology of numerous medical terms across distinct languages, multi-lingual clinical corpora could serve as a valuable source of data to pre-train BERT models. The resulting BERT's deep bidirectional architecture could leverage its language modeling capabilities to produce effective contextual representations that could be used in applications within a vast number of medical NLP information-extraction problems.

Acknowledgments. This work was partially supported by the project TIN2017-88728-C2-1-R, MINECO, Plan Nacional de I+D+I, and I Plan Propio de Investigación y Transferencia of the Universidad de Málaga.

References

1. Akhtyamova, L., Martínez, P., Verspoor, K., Cardiff, J.: Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives. Preprint (Version 1) available at Research Square (2020). https://doi.org/10.21203/rs.2.22697/v1
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/W19-1909
3. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: to appear in PML4DC at ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
5. 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., andNicola Ferro, L.C. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science, vol. 12260 (2020)
6. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
7. 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. Bioinformatics **36**(4), 1234–1240 (2019). https://doi.org/10.1093/bioinformatics/btz682
8. Li, M., Fei, Z., Zeng, M., Wu, F., Li, Y., Pan, Y., Wang, J.: Automated ICD-9 Coding via A Deep Learning Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics **16**(4), 1193–1202 (2019)
9. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the Variance of the Adaptive Learning Rate and Beyond (2019)
10. 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)
11. Névéol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Zweigenbaum, P.: CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: Proc of CLEF eHealth Evaluation lab. Dublin, Ireland (2017)
12. Névéol, A., Grouin, C., Cohen, K.B., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Proc of CLEF eHealth Evaluation lab. Evora, Portugal (2016)
13. 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: Proc of CLEF eHealth Evaluation lab. Avignon, France (2018)
14. Neves, M.L., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the CLEF eHealth Evaluation Lab 2019. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. Lecture Notes in Computer Science, vol. 11696. Springer, Cham (2019)
15. 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). https://doi.org/10.1197/jamia.M2077

16. Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N.: Hierarchical Transformers for Long Document Classification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 838–844 (2019). https://doi.org/10.1109/ASRU46091.2019.9003958
17. Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., Elhadad, N.: Diagnosis code assignment: models and evaluation metrics. Journal of the American Medical Informatics Association **21**(2), 231–237 (2013). https://doi.org/10.1136/amiajnl-2013-002159
18. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B., Duch, W.: A Shared Task Involving Multi-Label Classification of Clinical Free Text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. p. 97–104. BioNLP '07, Association for Computational Linguistics, USA (2007)
19. Sänger, M., Weber, L., Kittner, M., Leser, U.: Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task 1. In: CLEF (Working Notes) (2019)
20. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics **22**(5), 1589–1604 (2018)
21. Si, Y., Wang, J., Xu, H., Roberts, K.: Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association **26**(11), 1297–1304 (2019). https://doi.org/10.1093/jamia/ocz096
22. Stanfill, M.H., Williams, M., Fenton, S.H., Jenders, R.A., Hersh, W.R.: A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association **17**(6), 646–651 (2010). https://doi.org/10.1136/jamia.2009.001024
23. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to Fine-Tune BERT for Text Classification? In: Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, vol. 11856. Springer, Cham (2019)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)