ICB-UMA at CANTEMIST 2020: Automatic ICD-O Coding in Spanish with BERT

Guillermo López-García[a], José M. Jerez[a], Nuria Ribelles[b], Emilio Alba[b] and Francisco J. Veredas[a]

[a] *Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática, Universidad de Málaga, Málaga, Spain*
[b] *Unidad de Gestión Clínica Intercentros de Oncología, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain*

Abstract

This working notes paper presents our contribution to the CANTEMIST track. Our team has participated in the CANTEMIST-CODING subtask, the first shared task consisted in the automatic assignment of ICD-O-3 codes to Spanish oncology clinical cases. We addressed the task as a multi-label text classification problem using BERT model [1]. In order to leverage all the language modelling capabilities of the BERT architecture when applied to the CANTEMIST corpus, we have used an enhanced version of our fragment-based classification approach initially developed to tackle the CodiEsp-D task [2]. Hence, applying the improved version of our fragment-based strategy to the CANTEMIST corpus, we produced short fragments of text comprising a sequence of sentences from the long clinical documents present in the oncology corpus, and used them as input to the model. Two different versions of the BERT-Base model, namely the Multilingual BERT [3] and the BERT-SciELO [4] models, were fine-tuned on the CANTEMIST-CODING annotated corpus. The Multilingual BERT model further pre-trained on an unlabeled Spanish corpus of oncology clinical cases retrieved from Galén [5], achieved the highest classification performance among our five submitted systems, obtaining a final Mean Average Precision (MAP) score of 0.847 on the evaluation set.

Keywords

Clinical NLP, BERT, Spanish oncology clinical cases, Automatic clinical coding, Transfer learning, Text classification

1. Introduction

There is a growing interest in processing clinical documents using text mining and Natural Language Processing (NLP) techniques, giving rise to the birth of a new scientific subdiscipline situated at the intersection between Medicine, Linguistics and Computer Science, namely Clinical NLP [6]. One of the most active areas of research in Clinical NLP is the development of tools that perform automatic clinical coding, i.e. the task of autonomously extracting valuable structured information contained in the unstructured medical notes, following standardised coding terminologies [7].

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: guilopgar@uma.es (G. López-García); jja@lcc.uma.es (J.M. Jerez); ealbac@uma.es (E. Alba); franveredas@uma.es (F.J. Veredas)
ORCID: 0000-0001-5903-1483 (G. López-García); 0000-0002-7858-2966 (J.M. Jerez); 0000-0003-0739-2505 (F.J. Veredas)

CEUR Workshop Proceedings (CEUR-WS.org)

Historically, Clinical NLP researchers have focused mainly on English text, generating and exploiting clinical coding resources in the English language [8, 9, 10]. With 483 million native speakers [11], there exists a noteworthy interest in processing medical documents in Spanish. However, the lack of clinical linguistic resources for non-English languages makes it specially arduous to develop tools tailored to the Spanish clinical documents.

With the aim of promoting the application of Clinical NLP techniques to Spanish medical texts, the CANcer TExt MIning Shared Task (CANTEMIST) [12] has been organised in the context of the Iberian Languages Evaluation Forum (IberLEF 2020). CANTEMIST is the first shared task consisted in the automatic clinical coding of oncology medical cases written in Spanish. The CANTEMIST corpus comprises $1.3K$ clinical cases manually annotated by experts in oncology using the Spanish version (eCIE-O-3.1) of the International Classification of Diseases for Oncology (ICD-O-3) codes. The CANTEMIST track is composed of three distinct subtasks: CANTEMIST-NER, CANTEMIST-NORM and CANTEMIST-CODING. CANTEMIST-NER subtrack requires identifying tumour morphology mentions contained in a free-text clinical case, whereas in CANTEMIST-NORM subtask tumour morphology mentions must be both identified and normalised by assigning their corresponding ICD-O-3 codes. On the other hand, CANTEMIST-CODING task requires assigning a set of ICD-0-3 codes to each medical document in the corpus.

In this work, we present our contribution to the IberLEF 2020, where our team has participated in the CANTEMIST-CODING subtask. We have addressed the task as a multi-label text classification problem using BERT [1], a Transformer-based [13] language model that achieved state-of-the-art results on several different NLP tasks. BERT was initially designed to receive short fragments of text as input to the model, as opposed to the long oncology texts present in the CANTEMIST corpus. In order to leverage all the language modelling capabilities of the BERT architecture when applied to the CANTEMIST corpus, we have employed an improved version of our fragment-based classification approach originally developed for the CodiEsp-D task [2]. Hence, using the annotations available for the CANTEMIST-NORM subtask, we turned the CANTEMIST-CODING multi-label document classification problem into a multi-label short-fragment classification task, producing annotated short fragments of text with a full semantic meaning. Furthermore, we experimented with two different versions of the BERT-Base model: the Multilingual version [3], and the BERT-SciELO [4] model. Besides, different alternatives were explored to adapt the two models to the clinical domain, by further pre-training their weights on medical corpora before fine-tuning the models on the CANTEMIST corpus.

For reproducibility purposes, the implementation of our approach is publicly available at https://github.com/guilopgar/CANTEMIST-2020.

2. Materials and Methods

2.1. Corpora

2.1.1. Clinical corpora

In this work, we experimented with two different BERT-Base models, namely Multilingual BERT and BERT-SciELO. Multilingual BERT was pre-trained on an extensive multilingual

Table 1
Summary of the Galén oncology and MIMIC-III discharge summaries unlabeled corpora. The number of tokens and the number of characters were obtained using the Linux wc -w and wc -c commands, respectively.

Corpus	Documents	Tokens	Characters
Galén oncology	30.9K	64.4M	437.6M
MIMIC-III discharge summaries	57.4K	81M	534.7M

general domain corpus comprising texts from 104 different languages [3], including Spanish. On the other hand, BERT-SciELO model was pre-trained on a corpus of biomedical articles in Spanish [4]. With the aim of adapting both models to the distinctive features of the Spanish clinical texts domain, we decided to further pre-train the models on a corpus of de-identified medical texts in Spanish retrieved from the Galén Oncology Information System [5]. The corpus comprises $30.9K$ unlabeled documents containing oncology clinical notes written by physicians from the Oncology Departments of the Hospital Universitario Virgen de la Victoria (HUVV) and the Hospital Regional Universitario (HRU) in Málaga, Spain.

Moreover, in order to exploit the cross-lingual features extracted by the Multilingual BERT model, in addition to the corpus of medical texts retrieved from Galén, we also used a clinical corpus in English to pre-train the Multilingual BERT model. In this way, we joined the Galén oncology documents and the discharge summaries from the MIMIC-III database [14] in a single bilingual clinical corpus, used to perform the unsupervised pre-training of the multilingual model. From the multiple categories of documents stored in the MIMIC-III database, e.g. radiology, nursing, nutrition and pharmacy reports, we only selected the discharge summaries texts, given the high similarity between the medical texts from Galén and the content of the discharge summaries. In Table 1, a brief description of both the Galén oncology corpus and the discharge summaries corpus retrieved from the MIMIC-III database is given.

2.1.2. CANTEMIST corpus

The CANTEMIST corpus contains $1.3K$ clinical cases manually curated by oncology experts, covering a wide variety of cancer types. For the CANTEMIST-CODING subtask, the documents from the corpus were annotated with ICD-O-3 codes. The entire corpus was divided into four distinct subsets of annotated texts, the training (501 documents), development-1 (249 documents), development-2 (250 documents) and test (300 documents) sets. Teams participating in the CANTEMIST-CODING subtrack were evaluated on the test set.

In Table 2, a basic description of the CANTEMIST-CODING corpus is given. As it can be seen from the table, the number of codes annotations is scarce considering the limited number of documents contained in the corpus, having a low average number of texts where each ICD-O-3 code is present. This results in an imbalanced multi-label classification problem, in which for each code, the annotated texts (positive samples) are clearly outnumbered by the documents where the code is not present (negative samples).

Table 2
Summary of the CANTEMIST-CODING annotated corpus.

	Training	Development-1	Development-2	Test
Documents	501	249	250	300
Total ICD-O codes	2756	1385	1279	1599
Avg. ICD-O codes per doc.	5.501	5.562	5.116	5.330
Unique ICD-O codes	493	338	334	386
Avg. docs. per ICD-O code	5.590	4.098	3.829	4.142
Unique unseen ICD-O codes	-	130	120	107

2.2. Classification system

We have tackled the CANTEMIST-CODING challenge using BERT model [1]. One of the characteristic features of BERT is that its Transformer-based architecture is designed to process an input sequence of WordPiece [15] sub-tokens with a limited length N ($N = 512$ in the original implementation of BERT). This entails an important constraint when dealing with long-document classification tasks such as CANTEMIST-CODING, in which most of the clinical cases exhibit a WordPiece sub-tokens sequence length high above the maximum length supported by BERT.

In order to overcome this limitation, we have applied our fragment-based classification approach initially developed to address the CodiEsp-D task [2]. Given the high correspondence between CodiEsp-D and CANTEMIST-CODING subtasks, the three-phases custom approach was applied in a straightforward manner. In this way, we firstly split each clinical document of the CANTEMIST corpus into short fragments of text. Subsequently, using the ICD-O-3 codes annotations available for the CANTEMIST-NORM subtask, we annotated each fragment with the oncology codes exclusively occurring within the fragment. Then, we used the annotated fragments to perform the supervised fine-tuning of the BERT model on a fragment-level multi-label classification task. Finally, since the evaluation of the CANTEMIST-CODING participating systems was performed at document level, we post-processed the probabilities predicted by the model at fragment level using a maximum probability criterion, producing a list of codes ordered by relevance for each clinical document [2].

Nevertheless, in this work, we have not directly applied the first phase of the fragment-based classification approach described above. Instead, we have modified the first of the three-phases forming the fragment-based strategy, performing the text segmentation at the sentence level. Concretely, during the splitting phase, for each clinical case, the text was firstly split into sentences using the SPACCC Sentence Splitter tool [16]. Then, the WordPiece tokenization was performed on each sentence, producing a sequence of sub-tokens $s_i = (w_{i1}, ..., w_{1k})$ for every sentence, with length k. For each sentence s_i, if $k > N - 2$ (considering that BERT always adds the sub-tokens `[CLS]` and `[SEP]` at the start and end positions, respectively, of an input sequence), we split s_i into $\lceil k/(N-2) \rceil$ further sentences, ensuring that each finally produced sentence had a maximum length of $N - 2$ sub-tokens. Hence, a final sequence $s = (s_1, ..., s_m) = ((w_{11}, ..., w_{1\alpha}), ..., (w_{m1}, ..., w_{m\omega}))$ of m sub-tokens sentences was generated. Lastly, we split the sequence s into a sequence of fragments of contiguous sentences $f =$

$(f_1, f_2, ..., f_l) = ((s_1, ..., s_\lambda), (s_{\lambda+1}, ..., s_\beta)..., (s_\sigma, ..., s_m))$ using a simple greedy strategy: each fragment f_i contained the maximum number of adjacent sentences such that $\sum_{s_j \in f_i} |s_j| \leq N - 2$. Using this enhanced version of the splitting phase, along with the other two stages of the original version of our fragment-based classification approach [2], we could produce annotated short fragments of text comprising a sequence of sentences with full semantic meaning.

2.3. Experiments

We used two different versions of the BERT-Base model, namely the Multilingual BERT and the BERT-SciELO models. To perform the unsupervised pre-training of both models on the clinical corpora described in Section 2.1.1, we made use of the original TensorFlow implementation of BERT [17]. As the vocabulary of the BERT-SciELO model does not account for any punctuation character, we performed a pre-processing procedure consisted in substituting all punctuation marks contained in the Galén oncology corpus (see Section 2.1.1) by spaces, and used the pre-preprocessed version of the corpus to pre-train the weights of the model. Once pre-trained, the model was fine-tuned on the CANTEMIST-CODING task, applying the same pre-processing procedure to the CANTEMIST corpus before training the architecture. In the case of the Multilingual BERT model, since punctuation marks are considered in its vocabulary, the raw texts of both the clinical corpora and the CANTEMIST corpus were employed to pre-train and then fine-tune the architecture, respectively. Regarding the models hyper-parameters, for fine-tuning, we used a maximum input sequence length of $N = 100$ for the Multilingual BERT and a value of $N = 72$ for the BERT-SciELO model; for both models, we used RAdam [18] with a learning rate of 3×10^{-5}, a batch size of 16 and the number of epochs were experimentally determined on the CANTEMIST-CODING development-2 subset using early-stopping, with an upper limit of 40 epochs. Finally, with respect to the hardware resources, all experiments were executed on a single GeForce GTX 1080 Ti 11 GB GPU.

3. Results

In this section, we describe the results obtained by our team, ICB-UMA, at the CANTEMIST-CODING task. We submitted five different runs of our fragment-based classification system. The first (ICB-UMA run1) and the fourth (ICB-UMA run4) submissions corresponded to the original Multilingual BERT and BERT-SciELO models, respectively, fine-tuned on the CANTEMIST-CODING training, development-1 and development-2 corpora. Submissions ICB-UMA run2 and ICB-UMA run 5 contained the codes predicted by the Multilingual BERT and the BERT-SciELO models, respectively, further pre-trained on the Galén oncology corpus (see section 2.1.1) and then fine-tuned on the CANTEMIST-CODING corpus. Finally, submission ICB-UMA run3 corresponded to the Multilingual BERT model further pre-trained on the bilingual clinical corpus comprising the texts from the Galeń oncology corpus and the discharge summaries from the MIMIC-III database, and subsequently fine-tuned on the CANTEMIST-CODING corpus. To fine-tune the weights of each of the submitted BERT models, the representation generated by BERT for the initial `[CLS]` sub-token was fed into an output multi-label classification layer of 743—the number of unique codes present in the training, development-1 and development-2 CANTEMIST-CODING corpora (see Table 2)—units.

Table 3
Predictive performance of each submitted system assessed using MAP, the main evaluation metric of the CANTEMIST-CODING subtask.

Submission	MAP	MAP No-Code
ICB-UMA run1	0.821	0.794
ICB-UMA run2	0.847	0.821
ICB-UMA run3	0.837	0.813
ICB-UMA run4	0.800	0.769
ICB-UMA run5	0.812	0.784

Table 4
Predictive performance of each submitted system assessed according to additional evaluation metrics.

Submission	P	R	F1	P No-Code	R No-Code	F1 No-Code
ICB-UMA run1	0.007	0.928	0.013	0.006	0.914	0.011
ICB-UMA run2	0.007	0.928	0.013	0.006	0.914	0.011
ICB-UMA run3	0.007	0.928	0.013	0.006	0.914	0.011
ICB-UMA run4	0.007	0.928	0.013	0.006	0.914	0.011
ICB-UMA run5	0.007	0.928	0.013	0.006	0.914	0.011

In Table 3 and Table 4, we show the classification performance of our five submitted runs on the CANTEMIST-CODING test corpus. In particular, Table 3 describes the results obtained according to the main evaluation metric of the CANTEMIST-CODING subtask, i.e. the Mean Average Precision (MAP) [19]. The second column of the table shows the MAP values computed considering all codes contained in the CANTEMIST-CODING test annotated corpus, whereas the values presented in the last column (*MAP No-Code*) were calculated without considering the overrepresented metastasis ICD-O-3 code (8000/6). According to the results observed in Table 3, the Multilingual version of BERT outperformed the BERT-SciELO model, as ICB-UMA run1, ICB-UMA run2 and ICB-UMA run3 systems achieved higher values than ICB-UMA run4 and ICB-UMA run5 systems for the two analysed metrics in the table. The best performance is obtained by the Multilingual BERT model pre-trained on the Galén oncology corpus (ICB-UMA run2), followed by the same model pre-trained on the bilingual medical corpus (ICB-UMA run3). If we compare ICB-UMA run4 and ICB-UMA run5 systems, we can see that the BERT-SciELO model further pre-trained on the Galén oncology corpus (ICB-UMA run5) outperformed the original version of the BERT-SciELO model (ICB-UMA run4). Thus, the obtained results in this work demonstrate that, both the Multilingual BERT and the BERT-SciELO models, when adapted to the Spanish clinical texts domain by means of further pre-training their weights on an unlabeled medical corpus in Spanish, outperformed the original version of the models on the CANTEMIST-CODING task.

On the other hand, to perform a more extensive analysis of the obtained results, the organisers of the CANTEMIST track evaluated the classification performance of the submitted systems according to a set of additional metrics. Hence, in Table 4, the second, third and fourth columns present the computed values using precision (*P*), recall (*R*) and the F-score (*F1*) metrics, respectively, taking into consideration all codes contained in the CANTEMIST-CODING test subset, while the last three columns (*P No-Code*, *R No-Code* and *F1 No-Code*) show the results calculated using the same three metrics but without considering the 8000/6 ICD-O-3 code. As it can be seen from Table 4, our five submitted systems obtained really poor values for both precision and F-score metrics, while for the recall metric the obtained values were unusually high. The reason is that, with the goal of maximising the score obtained for the main evaluation MAP metric, as performed in [2], for each clinical case from the test corpus, we submitted all ICD-O-3 codes considered by the classification system—743, the number of units of the output classification layer of the models—sorted by their predicted probability of occurrence in descending order. On the contrary, if we had maximised precision, recall and F-score metrics, instead of submitting all considered codes, we would have defined a classification threshold to select only a subset of the codes according to their predicted probabilities.

4. Conclusion

In this paper, we present our contribution to the CANTEMIST-CODING subtask from the CANTEMIST track [12], in the context of the IberLEF 2020. This shared task consisted in the automatic assignment of ICD-O-3 codes to oncology clinical cases written in Spanish. We have addressed the task as a multi-label text classification problem using BERT model [1]. With the goal of adapting BERT to the distinctive features of the CANTEMIST-CODING corpus, we have applied an improved version of our fragment-based classification approach initially developed for the CodiEsp-D task [2]. In this way, using the available information for the CANTEMIST-NORM subtask, we converted the CANTEMIST-CODING multi-label long-text classification task into a multi-label short-text classification problem, generating short fragments of text with full semantic meaning annotated with ICD-O-3 codes. We experimented with two different versions of the BERT-Base model, namely the Multilingual BERT [3] and the BERT-SciELO [4] models. The best classification performance among our five submitted systems was achieved by the Multilingual BERT model further pre-trained on a medical corpus of Spanish oncology clinical cases, obtaining a MAP score of 0.847 on the evaluation set. Besides, both Multilingual BERT and BERT-SciELO models further pre-trained on a medical corpus outperformed the original versions of the models on the CANTEMIST-CODING subtask, reinforcing the idea that a clinical domain version of BERT achieves higher performance on medical classification tasks that a non-clinical domain version of the model.

In future works, given the promising results obtained by the Multilingual BERT model when applied to the CANTEMIST corpus, we will tackle other Spanish medical text classification tasks, such as CodiEsp-D subtask [2], using the Multilingual BERT fragment-based classification approach developed in this work. Furthermore, we will investigate whether further pre-training the model architecture using not only Spanish and English medical texts, but also French, Italian or German clinical documents, could leverage all the cross-lingual features extracted by the Multilingual BERT model and improve the results presented in this work for a Spanish oncology text classification task.

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] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[2] G. López-García, J. M. Jerez, F. J. Veredas, ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT, in: Working Notes of Conference and Labs of the Evaluation Forum (CLEF), CEUR Workshop Proceedings, 2020.

[3] Google Research, Multilingual BERT, 2019. URL: https://github.com/google-research/bert/blob/master/multilingual.md.

[4] L. Akhtyamova, P. Martínez, K. Verspoor, J. Cardiff, Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives, Preprint (Version 1) available at Research Square (2020). doi:10.21203/rs.2.22697/v1.

[5] N. Ribelles, J. M. Jerez, D. Urda, J. L. Subirats, A. Márquez, C. Quero, E. Torres, L. Franco, E. Alba, Galén: Sistema de Información para la gestión y coordinación de procesos en un servicio de Oncología, FeSALUD 6 (2010).

[6] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, T. Botsis, Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review, Journal of Biomedical Informatics 73 (2017) 14–29. doi:10.1016/j.jbi.2017.07.012.

[7] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, W. R. Hersh, A systematic literature review of automated clinical coding and classification systems, Journal of the American Medical Informatics Association 17 (2010) 646–651. doi:10.1136/jamia.2009.001024.

[8] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, W. Duch, 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, BioNLP '07, Association for Computational Linguistics, USA, 2007, p. 97–104.

[9] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, Journal of the American Medical Informatics Association 21 (2013) 231–237. doi:10.1136/amiajnl-2013-002159.

[10] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, J. Wang, Automated ICD-9 Coding via A Deep Learning Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics 16 (2019) 1193–1202.

[11] D. F. Vítores, El español: una lengua viva. Informe 2019. Instituto Cervantes, 2019. URL: https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2019.pdf.

[12] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normal-

ization and clinical coding: Overview of the CANTEMIST track for cancer text mining in Spanish, Corpus, Guidelines, Methods and Results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. 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 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[14] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016).

[15] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351.

[16] Plan TL-Sanidad, The Sentence Splitter (SS) for Clinical Cases Written in Spanish, 2019. doi:10.5281/zenodo.2586995.

[17] Google Research, BERT, 2019. URL: https://github.com/google-research/bert.

[18] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the Variance of the Adaptive Learning Rate and Beyond (2019). arXiv:1908.03265.

[19] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, USA, 2008.