A deep learning model for the analysis of medical reports in ICD-10 clinical coding task

Marco Polignano
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
marco.polignano@uniba.it

Pierpaolo Basile
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
pierpaolo.basile@uniba.it

Marco de Gemmis
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
marco.degemmis@uniba.it

Pasquale Lops
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
pasquale.lops@uniba.it

Giovanni Semeraro
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
giovanni.semeraro@uniba.it

Abstract

English. The practice of assigning a uniquely identifiable and easily traceable code to pathology from medical diagnoses is an added value to the current modality of archiving health data collected to build the clinical history of each of us. Unfortunately, the enormous amount of possible pathologies and medical conditions has led to the realization of extremely wide international codifications that are difficult to consult even for a human being. This difficulty makes the practice of annotation of diagnoses with ICD-10 codes very cumbersome and rarely performed. In order to support this operation, a classification model was proposed, able to analyze medical diagnoses written in natural language and automatically assign one or more international reference codes. The model has been evaluated on a dataset released in the Spanish language for the eHealth challenge (CodiEsp) of the international conference CLEF 2020, but it could be extended to any language with latin characters. We proposed a model based on a two-step classification process based on BERT and BiLSTM. Although still far from an accuracy sufficient to do without a licensed physician opinion, the results obtained show the feasibility of the task and are a starting point for future studies in this direction.

Italian. La pratica di assegnare un codice univocamente identificabile e facilmente riconducibile ad una patologia a partire da diagnosi mediche e un valore aggiunto alla attuale modalità di archiviazione dei dati sanitari raccolti per costruire la storia clinica di ciascuno di noi. Purtroppo però, lenorme numero di possibili patologie e condizioni mediche ha portato alla realizzazione di codifiche internazionali estremamente ampie e di difficile consultazione anche per un essere umano. Tale difficolt rende la pratica di annotazione delle diagnosi con i codici ICD-10 molto complessa e raramente svolta. Col fine di supportare tale operazione si è proposto un modello di classificazione, in grado di analizzare le diagnosi mediche scritte in linguaggio naturale ed assegnarle automaticamente uno o più codici internazionali di riferimento. Il modello è stato valutato su un dataset rilasciato in lingua Spagnola per la challenge (CodiEsp) di eHealth della conferenza internazionale CLEF 2020 ma è di semplice estensione su qualsiasi lingua con caratteri latini. Abbiamo proposto un modello basato su due passi di classificazione e basati sullutilizzo di BERT e delle BiLSTM. I risultati ottenuti, seppur ancora lontani da una accuratezza sufficiente per far a meno di un parere di un medico esperto, mostrano la fattibilità del task e si pongono come punto di partenza per futuri studi in tale direzione.

1 Introduction

In many of the existing systems for storing patient clinical data, the medical report of the doctor is stored in the form of textual data. Only in a few cases, one or more identification codes are assigned to each of the diagnosed conditions. The process of assigning a unique code to pathologies, symptoms, clinical situations, and drugs is commonly referred to as Clinical Coding. Among the most widely used coding systems, we can find the tenth version of the international medical glossary published by WHO (World Health Organization), commonly known as ICD-10. It contains codes for diseases, signs, and symptoms, abnormal results, complaints, social situations, and external causes of injury or illness. The United States uses its national variant of the ICD-10 called the ICD-10 Clinical Modification (ICD-10-CM). A procedure classification called the ICD-10 Procedure Coding System (ICD-10-PCS) has also been developed for the acquisition of hospitalization procedures. There are over 70,000 ICD-10-PCS procedure codes and over 69,000 ICD-10-CM diagnosis codes, compared to about 3,800 procedure codes and about 14,000 diagnosis codes in the previous ICD-9-CM. The use of an international classification to annotate medical diagnoses makes the health system interoperable between different countries. Among the many possibilities of using ICD codes, a doctor of any nationality could thus be able to read, analyze, and use the medical history of a patient even if of a different nationality. In addition, diagnostic patterns used by clinicians could be identified to improve automatic disease prediction strategies and provide automatic specialist support for decision making. These observations strongly support the need for automated systems to support clinicians to perform this task quickly and without human intervention. The contribution of our work is a novel model for ICD-10 coses annotation We proposed a model based on Bi-LSTM and BERT to assign one or more ICD-10 codes to the medical diagnosis. Specifically, we have designed our approach as a two-step process. In the first one, we use a BERT-based classifier to select from the dataset only a subset of sentences that could be candidates for the annotation step. In particular, those phrases are them that could e annotated with one or more codes. The phrases left were generally speaking expressions. In the second step, we used a Bi-LSTM model to analyze the candidate sentences and assign one or more codes to them. The results are encouraging and a good starting point for further investigations. The rest of the paper is structured as follows. We start analyzing related works, and we go through the description of the model and the dates. Finally, we analyze the results obtained, and we expose our consideration of the task in the conclusion section.

2 Related Work

The scientific community has long addressed the task of analyzing medical diagnoses in order to assign a unique code for each pathology. In particular, since 1973, the first corpus and state of the art analysis of the clinical coding task have been released. In 1999, Chapman stated that a system based on algorithms would be better than a human being to perform this task. If we think about the high number of codes in the ICD-10 glossary (about 70000), it immediately comes to mind that even a human being cannot be very accurate in the assignment, as long as he must know perfectly each of the countless codes. In 2006 Kukafka et al. affirm that the algorithmic path to be followed for the resolution of the task was through the NLP. It is, in fact the most suitable for the resolution of the task, currently. To get automatic systems able to afford the task efficiently, it is necessary to wait until 2017. Miftahutdinov (Miftahutdinov and Tutubalina, 2017) at CLEF eHealth 2017 uses an LSTM on a TF-IDF representation of the text to identify the most suitable ICD-10 code for the input sequence. This allows to obtain an F1 score equal to 0.85, considering a classification on 1,256 distinct classes. In 2018, Atutxa et al. (Atutxa et al., 2018), proposed a three-level sequence-to-sequence neural network-based approach. The first neural network tries to assign one set of ICD-10 codes to the whole document, then they are refined to assign one set of codes to the line, and finally one specific code. This strategy allowed the model to obtain an F1 score between 0.7086 and 0.9610, depending on the language of the dataset on which the system has been evaluated. At CLEF eHealth 2019, the best system was proposed by Sänger et al. (Sänger et al., 2019), obtaining an F1 score of 0.80. The proposed model utilized a multilingual BERT (Devlin et al., 2019) text encoding model, fine-tuned on additional training data of German clinical trials also annotated with ICD-10 codes. The model is

extended by a single output layer to produce probabilities for specific ICD-10 codes. Considering the successful models presented as state of the art, we decided to use a machine learning approach that combines CNNs, Bidirectional LSTMs, Attention Layers, and BERT.

3 Model and dataset

The CodiEsp evaluation track proposed by CLEF 2020 (Stanfill et al., 2010; Goeuriot et al., 2020; Miranda-Escalada et al., 2020) was structured as three sub-tracks about the analysis of clinical reports: CodiEsp-D requires automatic ICD10-CM [CIE10 Diagnóstico] code assignment; CodiEsp-P requires automatic ICD10-PCS [CIE10 Procedimiento] code assignment; CodiEsp-X requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The correctness of the provided reference is assessed in this subtrack, in addition to the code prediction. We decided to create a model based on deep learning able to deal with the first two subtasks, without performing an operation of reference span detection as required in the third task. An high level architecture of the model is described in Fig. 1. Specifically, we decided to use a BERT-based classifier to perform a pre-filtering operation in order to select a subset of sentences possibly referring to a clinical state. Later, the candidate sentences are submitted to a classifier based on BiLSTM, CNN, and self-attention to assign them one or more clinical codes. As pre-trained BERT model, we decided to use BETO (Caete et al., 2020), a Spanish pre-trained version of BERT. The authors trained BETO using 12 self-attention layers with 16 attention-heads each and 1,024 as hidden size. They used all the data from Wikipedia and all of the sources of the OPUS Project (Tiedemann, 2012), having the text in Spanish. We decided not to use the multilingual version of BERT because it has been shown that a version trained on the native language performs much better in many NLP tasks (Polignano et al., 2019b; Polignano et al., 2019c). The sentences classified as possible references of clinical codes are consequently passed to the second part of our model. First of all, the sentences are encoded into word embeddings. In this step, we decided to use a FastText embedding strategy (Bojanowski et al., 2017), which proved to be more effective than GLoVE (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) when many domain-specific words occur in the dataset (Polignano et al., 2019a). For our final configuration of the model, we chose the one released by José Cañete[1] made of 300 dimensions, trained on the Spanish Unannotated Corpora[2] containing more than 3 billion words. The block of the model that uses BiLSTM, CNN and self attention has been already proposed by the authors of this contribution for the emotions classification task. More details about it can be found in (Polignano et al., 2019a). As model parameters of the BiLSTM architecture we decided to set the value of hidden units to 64 and the internal dropout value to 0.3. We have also decided to vary the function of activation used by the net, setting it to the hyperbolic tangent function (tanh). A level of self-attention is added following the LSTM. Consequently, we applied the CNN layer on the result of the attention algorithm. In detail we apply a 1D Convolutional network with 64 filters and 5x5 kernel. We used ReLu as activation function, that unlike the hyperbolic tangent is faster to calculate. On the top of the CNN layer, we added a Max Pooling function for subsampling the values obtained, reducing the computational load and, the number of parameters of the model. The hidden model obtained until this step has been merged with the output of the previous Bi-LSTM. After that, we used a max-pooling layer for 'flattening' the results and reduce the model parameters. Finally, another dense layer with a soft-max activation function has been applied for estimating the probability distribution of each clinical code available in the dataset. Further details about the model can be found in (Polignano et al., 2019a; Polignano et al., 2020) and the source code of the model is publicly available on GitHub[3].

4 Experimental evaluation

The CodiEsp corpora contains manually annotated clinical reports, written in Spanish, with corresponding clinical codes. The training set contains 500 clinical cases, while the development and the test set provide 250 clinical cases each. The CodiEsp corpus format is plain text with UTF8 encoding, where each clinical case is stored in a single file whose name is the clinical unique case identifier. The final collection of the

[1] https://github.com/dccuchile/spanish-word-embeddings
[2] http://crscardellino.github.io/SBWCE/
[3] https://github.com/marcopoli/CODIESP-10

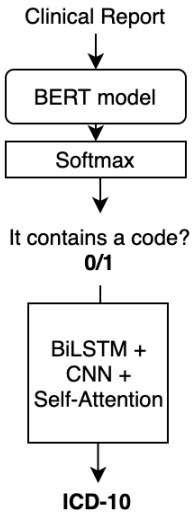

Figure 1: General design of the proposed model.

1,000 clinical cases of the corpus contains 16,504 sentences, with 16.5 sentences per clinical case on average. It contains 396,988 words, with 396.2 words per clinical report on average. The final architecture of the model, previously proposed was obtained after conducting several experiments on 20% of the training dataset released for the Codiesp-D subtask. For each task we used as classification labels only them with at least one example set in the training set. In particular we used 1788 codes for the Codiesp-D and 546 codes for the Codiesp-P task. For lacking of space, we are going to report int his contribution only the most relevant. As first experiment we trained different classification models with the purpose of directly classify the single medical reports with one ICD-10 code. In particular we developed the following models: *LSTM, BiLSTM, CNN, CNN + Self Attention, BERT, BiLSTM + CNN, BiLSTM + CNN + Self Attention,(Pre-filtering) BERT - (Classification) BiLSTM + CNN + Self Attention*.

Analyzing the results in Tab. 1, it is possible to notice that considering the F1 score as a metric, models based on deep learning approaches with LSTM and CNN strategy are able to obtain very similar results. It is evident that the differences between these methodologies are minimal and that generally, a combination of them improves the performance. Starting from an F1 measure of 0.09789 obtained using a biLSTM layer, a score of 0.10410 is reached when combining the biLSTM, CNN, and self-attention layers. The BERT-based

Table 1: Results obtained holding the FastText SUC word embedding, and varying the model.

	Macro-P	**Macro-R**	**Macro-F1**
LSTM	0.09357	0.09903	0.08845
BiLSTM	0.10354	0.10909	0.09789
BiLSTM + CNN	0.09995	0.11552	0.09831
BiLSTM + CNN + SelfAtt.	**0.10629**	**0.11887**	**0.10410**
CNN	0.09511	0.10095	0.09100
CNN + SelfAtt.	0.09279	0.09484	0.08706
BERT (BETO)	0.10381	0.10821	0.10294

classification model requires particular attention. It succeeds, in fact, to obtain a score of F1 comparable to that obtained by the model that combines the single techniques (BiLSTM+CNN+Self.Att). Thus, we considered good candidates for the final classification model, both BERT and the BiLSTM+CNN+Self.Att. models.The second step of experimentation was to understand if using a classifier that performs the task of pre-filtering diagnoses not containing an ICD-10 code could help the classification performance. Tab. 2. shows how we decided to test the combinations of the two models previously chosen as candidates with a pre-filtering approach followed by a classification approach. Observing the results in terms of F1 measure, it is possible to observe that the model using BERT in the first phase of selection and BiLST + CNN + Self.Att. in the phase of choice of the final codes, is the one that leads to better results. We were thus able to increase the F1 score by about 0.03 points compared to the previous step, reaching the value of 0.13632. Finally, we decided to use a threshold on the result of the last layer of the here proposed model (dense layer with a softmax function) in order to extract more than one label for each medical report. We performed different experiment in order to decide this threshold and finally we used a value of 0.10 able, from our evaluation, to maximize the F1 score of the model, reaching a score of 0.16011. The model we implemented, has been used for participating at both CodiEsp subtasks, i.e., CodiEsp-D and CodiEsp-P.

5 Conclusion

The ability to automatically annotate medical reports with international codes is an open and relevant research challenge for future technologies of global medical data sharing. In this work, we have

Table 2: Results obtained holding the FastText SUC word embedding, the models (BERT, BiLSTM + CNN + SelfAttention) and varying their combinations for pre-filtering and classification.

	Macro-P	**Macro-R**	**Macro-F1**
(Pre-filtering) BiLSTM + CNN + SelfAtt. – (Classification) BiLSTM + CNN + SelfAtt.	0.13241	0.10934	0.11534
(Pre-filtering) BiLSTM + CNN + SelfAtt. – (Classification) BERT (BETO multi-class)	0.09180	0.08871	0.10022
(Pre-filtering) BERT (BETO) – (Classification) BERT (BETO multi-class)	0.09704	0.10092	0.11734
(Pre-filtering) BERT (BETO) – (Classification) BiLSTM + CNN + SelfAtt.	**0.13823**	**0.12053**	**0.13632**

proposed a model based on the state of the art technologies to support the doctors during this time-consuming task. For example we can imagine the use of our model in a system that can suggest to the doctor a set of possible codes to choose from to note down the medical diagnosis. Such a suggestion could make this annotation process less complex for any human annotator. It is important to keep in mind that the ICD-10 code contains more than 70,000 codes and their number is constantly increasing. During the competition it was shown that the large number of possible labels was also the most difficult problem to deal with even for automatic systems. In this regard, in fact, many systems based on machine learning have found many more difficulties than hybrid systems able to select through linguistic rules a subset of possible assignable codes. This observation will be used by us in the future to improve the model proposed here. Currently, the results obtained are encouraging and confirm the possibility of tackling it with current technologies. Nevertheless, our system cannot obtain a score of reliability, such as acting independently, and the opinion of a human coder is still essential. The system proposed here has not won the challenge at CLEF 2020 but is still a good starting point for further studies that want to use those technologies to resolve the clinical coding task.

Acknowledgment

This work is funded by project "DECiSION" codice raggruppamento: BQS5153, under the Apulian INNONETWORK programme, Italy.

References

Aitziber Atutxa, Arantza Casillas, Nerea Ezeiza, Víctor Fresno, Iakes Goenaga, Koldo Gojenola, Raquel Martínez, Maite Oronoz Anchordoqui, and Olatz Perez-de Viñaspre. 2018. Ixamed at clef ehealth 2018 task 1: Icd10 coding with a sequence-to-sequence approach. In *CLEF (Working Notes)*, page 1.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jos Caete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Prez. 2020. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Saez Gonzales, Marco Viviani, and Chenchen Xu. 2020. Overview of the CLEF eHealth evaluation lab 2020. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurlie Nvol, and Linda Cappellato andNicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)* , LNCS Volume number: 12260.

Zulfat Miftahutdinov and Elena Tutubalina. 2017. Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. In *CLEF (Working Notes)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estap, and Martin Krallinger. 2020. 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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019a. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019b. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CLiC-it*.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019c. AlBERTo: Modeling Italian Social Media Language with BERT. *Italian Journal of Computational Linguistics - IJCOL*, -2, n.2.

Marco Polignano, Vincenzo Suriano, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2020. A study of machine learning models for clinical coding of medical reports at codiesp 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurlie Nvol, editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. CEUR.org.

Mario Sänger, Leon Weber, Madeleine Kittner, and Ulf Leser. 2019. Classifying german animal experiment summaries with multi-lingual bert at clef ehealth 2019 task 1. In *CLEF (Working Notes)*.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.