Text Augmentation Techniques for Clinical Case Classification

Anaïs Ollagnier[1 [0000−0002−4349−5678]] and Hywel Williams[1 [0000−0002−5927−3367]]

Computer Science, University of Exeter, Exeter EX4 4QE, UK
{a.ollagnier,h.t.p.williams}@exeter.ac.uk

Abstract. Clinical coding consists in the transformation (or classification) of patient record information into a structured or coded format using internationally recognized class codes. Coding accuracy is an ongoing challenge which has led to the organization of challenges and shared tasks to evaluate AI-enhanced, computer-assisted coding systems. In this paper we present our contribution at CodiEsp: Clinical Case Coding Task (CLEF eHealth 2020) on the automatic assignment of clinical coding (diagnosis and procedures) to clinical cases in Spanish. We approach the task as multi-label classification problem and leverage the powerful language model: Multilingual BERT (M-BERT) to represent the clinical cases and design various deep learning architectures based on a Convolutional Neural Network and a Long Short-Term Memory Network (CNN-LSTM) classifier. To handle the class-imbalance problem, we present other models based on data augmentation techniques (i.e. word-level transformations and text generation methods) for synthesizing labeled-data. Models based on data augmentation pipelines obtain the best results, measured by the F1-score, in comparison to the other proposed models for both tasks. The pipeline based on the word-level transformations obtains the best F1-score (0.143) for the CodiEsp-D task, while the data augmentation technique using the text generation method achieves the best F1-score (0.216) for the CodiEsp-P task.

Keywords: Medical text classification · Data augmentation · Text generation.

1 Introduction

The International Classification of Diseases (ICD) is a health care classification system which provides standardized codes for reporting diseases and health conditions[1]. ICD codes are widely used in *Electronic Medical Records* (EMR) to describe a patient's diagnosis or treatment. In current practice medical coders

[1] https://www.who.int/classifications/icd/factsheet/en/ Date of access: 15th June 2020.

review a physician's clinical diagnosis (almost always recorded as free text) then manually assign ICD codes according to coding guidelines. While the process of standardizing EMR is important for making clinical and financial decisions manual ICD coding is expensive, time-consuming and prone to error [12,3]. Considering these constraints, automated ICD coding has become an important line of research in the Artificial Intelligence community. Traditional machine learning and deep learning techniques have been applied successfully in this context and show promising results [13,8,16,10]. However, developing an accurate computational system to support automated ICD coding is still a challenging task. The idiosyncrasies of medical language, the scarcity of hospitals using EMR and the class-imbalance problem in training datasets are among the persistent challenges [4]. Issues related to clinical coding have led to the organization of challenges and shared tasks aiming to evaluate automated clinical coding systems such as the CLEF eHealth Evaluation Lab. The CLEF eHealth[2] [7], established in 2012 as part of the Conference and Labs of the Evaluation Forum (CLEF), is a workshop offering evaluation labs (datasets, evaluation frameworks, and events) in the medical and biomedical domain on different tracks such as information extraction, information management and information retrieval in a mono- and multilingual setting. During the CLEF eHealth 2020 the Clinical Case Coding in Spanish Shared Task[3] (CodiEsp) [11] was introduced with the aim to evaluate systems devoted to the automatic assignment of ICD codes to EMR in Spanish. This task includes three sub-tasks: (1) *Codiesp Diagnosis Coding* (CodiEsp-D) which consists of automatically assigning ICD10-Clinical Modification codes to clinical cases in Spanish; (2) *Codiesp Procedure Coding* (CodiEsp-P) which focuses on assigning ICD10-Procedure codes to clinical cases in Spanish; (3) *Codiesp Explainable Artificial Intelligence* (CodiEsp-X) which evaluates the explainability/interpretability of the proposed systems (i.e. request to return the text spans supporting the ICD10 code assignment).

This paper presents our contribution at the CLEF eHealth CodiEsp 2020 CodiEsp-D and CodiEsp-P sub-tasks. In total five models were submitted during the official evaluation, all based on a Convolutional Neural Network and Long Short-Term Memory Network (CNN-LSTM) classifier. Multilingual BERT (M-BERT) achieved the best performances in the CLEF eHealth 2019 Multilingual Information Extraction task [15], hence we proposed to leverage the M-BERT pre-trained model as a part of various deep learning architectures. Then, in order to handle the class-imbalance problem, we designed data augmentation pipelines exploring word-level transformation and a text generation method for synthesizing labeled-data. To compare all the proposed systems we carried out empirical comparisons against a standard CNN architecture, used here as a baseline.

[2] `https://clefehealth.imag.fr/` Date of access: 18th June 2020.
[3] `https://temu.bsc.es/codiesp/` Date of access: 18th June 2020.

2 Data

The Codiesp corpus[4] consists of a set of 1000 clinical cases manually annotated by clinical coding professionals [5]. Documents were coded with clinical diagnosis and procedure codes from the Spanish official version of ICD10-Clinical Modification and ICD10-Procedure. The released corpus has around 16,504 sentences and 396,988 words, with an average of 396.2 words per clinical case. The corpus has been randomly sampled into three subsets: the training set (500 clinical cases), the development set and the test sets (250 clinical cases each). Each subset provides clinical cases in plain text format stored as single files (each filename corresponds to an unique clinical case identifier) and a tab-separated file with either ICD10-Diagnostico (equivalent to ICD10-CM) or ICD10-Procedimiento (equivalent to ICD10-PCS) code assignments according to the target task. Table 1 summarises the top-5 most frequent ICD10-Diagnostico and ICD10-Procedimiento codes from the training and development datasets for both tasks.

Table 1. Top-5 most frequent ICD10-Diagnostico and ICD10-Procedimiento codes.

CodiEsp-D		**CodiEsp-P**	
R52	118 (15.73%)	bw03zzz	74 (9.87%)
R69	106 (14.13%)	bw40zzz	61 (8.13%)
R50.9	99 (13.2%)	bw20	56 (7.47%)
i10	81 (10.8%)	bw24	36 (4.8%)
R60.9	70 (9.33%)	4a02x4z	34 (4.53%)

As we can observe in Table 1, the datasets provided are high imbalanced (i.e. there is high disparity between classes), with 15.73% and 9.87% respectively as the highest frequency rates for the Codiesp-D and Codiesp-P tasks. In total, 10,711 codes were assigned for both tasks, of which, 1819 are unique in the Codiesp-D datasets and 608 in the Codiesp-P datasets. The proportion of rare classes (i.e. classes with only one observation) is also high, 1022 classes (i.e. 56.18%) in the codiesp-D datasets and 393 (i.e. 64.64%) in the Codiesp-P datasets. These findings led us to investigate data augmentation techniques which have shown promise in scarce labeled data situations [17,2].

Moreover, to expand the training and development corpora, the organizers have also released several additional data resources[6] including medical literature abstracts (i.e. abstracts from Lilacs and Ibecs with ICD10 codes), linguistic

[4] Codiesp corpus available online: `https://zenodo.org/record/3837305#.XvsEN5bTVhF` Date of access: 30th June 2020.

[5] Information about annotation guidelines: `https://zenodo.org/record/3632523#.Xvw2N5bTU5m` Date of access: 1st July 2020.

[6] `https://temu.bsc.es/codiesp/index.php/2019/09/19/resources/` Date of access: 30th June 2020.

resources, gazetteers and a machine-translated version from English of Codiesp corpus clinical cases.

3 System architectures

Empirical studies conducted on the development sets for each task[7] found best performance using a CNN-LSTM classifier. Figure 1 details the architecture used and the shared parameters for both tasks.The model takes as input a time-ordered sequence of tokens (words) of arbitrary length (truncated to 396 words which corresponds here the averaged number of words per document, and then padded with zero vectors) and outputs a document-level prediction. After the embedding layer, the layer corresponding to the CNN classifier (one-headed) is introduced using a configuration of 100 parallel feature maps and a kernel size of 3. Immediately afterwards, a LSTM layer is added (set to 100 internal units). Then a dense layer of 64 nodes with ReLu is inserted. Finally an output layer is used with one node containing softmax function. The models have been trained using the Adam optimizer, with a learning rate of 0.001 and a batch size fixed to 32 for both tasks.

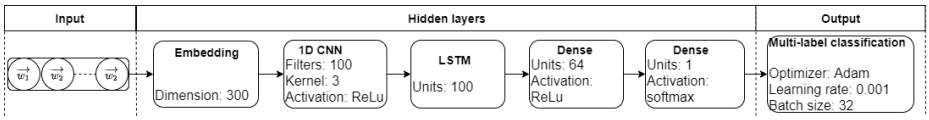

Fig. 1. System architecture details.

4 Proposed Methods

All proposed methods were trained and tested using the Spanish version of the released corpora. Concerning the preprocessing steps, clinical cases were converted to lowercase and stop-words were removed. After the tokenization process, all tokens based only on non-alphanumeric characters and all short tokens (with < 3 characters) were also deleted. In total five models were submitted to the official evaluation, we provide below a detailed description of each of them:

- **CNN-LSTM**: this default approach (used here as a baseline) is based on the architecture presented in section 3.
- **M-BERT**: this approach is based on the BERT language model [5]. Briefly, BERT, which is based on a transformer architecture, is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context. Several pre-trained language models

[7] Evaluations (not reported here) were conducted on LSTM, BiLSTM, BiGRU, CNN and CNN-LSTM using the same architecture as presented here.

(PTM) have been built from this text encoding model which has previously been successfully applied to various biomedical NLP tasks [9]. In the CLEF eHealth 2019 Multilingual Information Extraction task, models relying on BERT and its variants (BioBERT and M-BERT) obtained the best results [1,6,15]. Here, we propose to explore a *Sequential Transfer Learning*-based technique (STL) using M-BERT[8] . In a STL scenario the source and target tasks are different and training is performed in sequence. Typically, STL consists of two stages: a *pre-training phase* in which general representations are learned on a source task or domain, and an *adaptation phase* during which the learned knowledge is transferred to the target task or domain [14]. In the proposed models the pre-trained language representation (i.e. M-BERT) is introduced during the *pre-training phase* and then the CNN-LSTM architecture (c.f. section 3) is applied during the *adaptation phase* to fine-tune models to the target task.

- **WordNet**: this approach explores a traditional textual data augmentation technique consisting of a word-level transformation: synonym replacement. Introduced in [17], the application of this kind of local change was shown to improve performance on text classification tasks, especially for small training datasets. The process of synonym replacement is implemented as a preprocessing step in a data generator pipeline (this pipeline generates batches of tensor data with real-time data augmentation). For each batch, 10% of documents' words (randomly selected) are substituted by WordNet's synonyms[9] (except stopwords). Finally, edited documents are used to feed models relying on a CNN-LSTM architecture. Below is an example of synonym replacement on a clinical case sample.

  - original: Paciente de 50 años con antecedente de litiasis **renal** de repetición que consultó por hematuria recidivante y **sensación** de malestar. El estudio citológico seriado de orinas demostró la **presencia** de células atípicas sospechosas de malignidad.
  - edited: Paciente de 50 años con antecedente de litiasis **nefrítico** de repetición que consultó por hematuria recidivante y **percepción** de malestar. El estudio citológico seriado de orinas demostró la **apariencia** de células atípicas sospechosas de malignidad.

- **WordNet M-BERT**: based on the two previous approaches: **WordNet** and **M-BERT**, we explore the combination of a word-level data augmentation technique and the M-BERT pre-trained language model representation. Also implemented as a preprocessing step of the data generator pipeline, synonym replacement is based on the same setup as in the original approach i.e. 10% of documents' words are substituted. Then, as introduced in the M-BERT approach, models are trained as a part of a STL scenario.
- **TEXT GEN**: in this approach we propose to explore a novel data augmentation technique based on a text generation method. This strategy was recently

[8] BioBERT trained from the original BERT pre-trained model and medical resources in English can't be applied to clinical cases in Spanish

[9] Synonym replacement is performed using the python library NLPAug[10]

introduced for synthesizing labeled data to improve text classification tasks. Approaches leveraging text generation have shown promise, outperforming state-of-the-art techniques for data augmentation, specifically for handling scarce data situations [2]. The proposed data augmentation pipeline consists in two stages: a *pre-training phase* in which a language model is learned from the given training sets and a *generative phase* during which the pre-trained language model is used to generate artificial data. In detail, the pre-trained language model is built using an n-gram modeling approach which estimates n-gram distribution probabilities learned from a given corpus. The language models are trained for both tasks using the CNN-LSTM architecture presented in section 3. During the *generative phase* the appropriate pre-trained language model is introduced to generate artificial data as a preprocessing step in the data generator pipeline. 30% of each document is altered for each mini batch. Formally, each document is split into sentences then 30% of the sentences are replaced by synthesized data. To synthesize new data 30% of the beginning of a given sentence is used as a seed then extended according to the average length of sentences in the corpus (set to 20 words). Below is an example of a synthesized sentence using the pre-trained language model learned from the CodiEsp-D training set.

- original: **Analytical analysis showed** hydroxyvitamin lion ponesium sodium.
- edited: **Analytical analysis showed** sequence made transoperative urine outpatient image microbiological markers flap immunohistochemical intravenous dorsolumbar remained level signs 1788 partially establishing transplantation.

5 Experimental Results on the Test sets

Models were trained on a workstation with a 36-core CPU and an AMD FirePro W2100 GPU. Systems were evaluated according to the following metrics: Mean Average Precision (MAP), MAP@30, precision, recall and F1-score. For experimental purposes two versions of the F1-score metric are computed: the F1-score measure which considers the full code for both tasks and the F1-score CAT which considers only the first three digits of ICD10-Clinical Modification codes (e.g. codes R20.1 and R20.9 are mapped to R20) and the first four digits of ICD10-Procedure codes (e.g. the code bw40zzz is mapped to bw40). Table 2 summarizes the results obtained for both the CodiEsp-D task and the CodiEsp-P task on the test sets. For readability purposes only the MAP, the F1-score and the F1-score CAT are reported. Due to lack of time, not all models for each task were proposed at the official evaluation. However we performed the missing evaluations using the evaluation library released by the organizers[11], results are presented in italic font.

[11] Evaluation library: `https://temu.bsc.es/codiesp/index.php/2019/09/19/evaluation-library/`. This library only allows to compute the MAP. Date of access: 13th July 2020.

Table 2. Official Results of the Clinical Case Coding in Spanish Shared Task (eHealth CLEF 2020).

Model	CodiEsp-D			CodiEsp-P		
	MAP	F1-score	F1-score CAT	MAP	F1-score	F1-score CAT
Baseline	0.0.76	0.114	**0.166**	0.123	0.114	0.12
M-BERT	0.081	0.13	0.151	0.123	0.124	0.129
WN M-BERT	0.078	0.136	0.153	**0.125**	0.117	0.121
WN	**0.082**	**0.143**	0.165	*0.132*	-	-
TEXT GEN	*0.071*	-	-	0.121	**0.145**	**0.216**

As we can observe the results obtained depend on both the tasks and the models used. For the CodiEsp-D task the Wordnet model (WN) achieves the best MAP, followed closely by the model based on the pre-trained language model M-BERT. For the F1-score, the WN model also obtains the best performance against the other proposed approaches (+0.029 from the baseline). For the F1-score CAT, the baseline is first-ranked, slightly outperforming the WN model (−0.001 from the baseline). For the CodiEsp-P task the best MAP is obtained using the combination of M-BERT and the data augmentation technique based on synonym replacement (WN M-BERT) while the TEXT GEN model achieves the best performance for both F1-scores (+0.031 for the F1-score and +0.096 for the F1-score CAT, measured relative to the baseline). Concerning the unofficial results (in italic font), the model WN is first-ranked on the CodiEsp-P task while the TEXT GEN model is the less efficient for the CodiEsp-D task.

In the overall evaluation, the use of a STL-based architecture combined with a pre-trained model has shown its efficiency, outperforming the baseline on both tasks on the majority of evaluation metrics. Concerning data augmentation techniques, despite missing evaluations, the proposed techniques produced strong results in comparison with the other models, outperforming both the baseline and the models based on M-BERT on both tasks on the majority of evaluation metrics.

6 Conclusion

In this paper we presented our contribution to the CLEF eHealth CodiEsp 2020 CodiEsp-D and CodiEsp-P sub-tasks. In total we proposed five models during the official evaluation in which we explored both the powerful language model: Multilingual BERT (M-BERT) and two data augmentation techniques, word-level transformation and text generation methods, for synthesizing labeled-data. Models based on data augmentation pipelines achieved the best performances in comparison to the other proposed models for both tasks on the majority of evaluation metrics.

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 Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2019)
2. Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., Zwerdling, N.: Do not have enough data? deep learning to the rescue! In: AAAI Conference on Artificial Intelligence. pp. 7383–7390 (2020)
3. Campbell, S., Giadresco, K.: Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Information Management Journal **49**(1), 5–18 (2020)
4. Catling, F., Spithourakis, G.P., Riedel, S.: Towards automated clinical coding. International journal of medical informatics **120**, 50–61 (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR **arXiv:1810.04805** (2018)
6. Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the clef ehealth 2019 multilingual information extraction (2019)
7. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Gonzales, G.S., 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) . LNCS Volume number: 12260 (2020)
8. Kavuluru, R., Rios, A., Lu, Y.: An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artificial intelligence in medicine **65**(2), 155–166 (2015)
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. Bioinformatics **36**(4), 1234–1240 (2020)
10. Li, F., Yu, H.: Icd coding from clinical text using multi-filter residual convolutional neural network. In: AAAI Conference on Artificial Intelligence. pp. 8180–8187 (2020)
11. 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 ehealth clef 2020. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)
12. O'Dowd, A.: Coding errors in nhs cause up to£ 1bn worth of inaccurate payments. BMJ: British Medical Journal (Online) **341** (2010)
13. 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 (2014)
14. Ruder, S.: Neural Transfer Learning for Natural Language Processing. Ph.D. thesis, NATIONAL UNIVERSITY OF IRELAND, GALWAY (2019)
15. 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 Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2019)

16. Shi, H., Xie, P., Hu, Z., Zhang, M., Xing, E.P.: Towards automated icd coding using deep learning. CoRR **arXiv:1711.04075** (2017)
17. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. CoRR **arXiv:1901.11196** (2019)