Multilingual ICD-10 Code Assignment with Transformer Architectures using MIMIC-III Discharge Summaries

FHDO Biomedical Computer Science Group (BCSG)

Henning Schäfer[1][0000−0002−4123−0406], and
Christoph M. Friedrich[1,2][0000−0001−7906−0038]

[1] Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Emil-Figge Str. 42, 44227 Dortmund, Germany
hesch024@stud.fh-dortmund.de
christoph.friedrich@fh-dortmund.de
[2] Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Essen, Germany

Abstract. In this work, we present the participation of FHDO Biomedical Computer Science Group (BCSG) to the CLEF eHealth challenge 2020 Task 1 on automatic assignment of ICD-10 codes (CIE-10 in the Spanish translation) to clinical case studies. Training data has been augmented with documents from the Medical Information Mart for Intensive Care (MIMIC-III), a critical care database. ICD-10 CM General Equivalence Mappings (GEMs) were subsequently used to convert the codification from ICD-9 to ICD-10. Recent state-of-the-art Transformer-based models, such as BioBERT and ClinicalBERT are compared to the Generalized Autoregressive Pretraining for Language Understanding (XLNet) model. Finally, the apriori algorithm has been applied to build association rules by finding frequent item sets. An ensemble of BioBERT and XLNet achieved a mean Average Precision (MAP) score of 0.259 (0.306 for the subset of codes only present in the training and validation sets).

Keywords: BioBERT· MIMIC-III · Apriori · XLNet · ICD-10 Code Conversion

1 Introduction

This paper describes the participation of FHDO Biomedical Computer Science Group (BCSG) to the Conference and Labs of the Evaluation Forum (CLEF) eHealth 2020 Task 1 Subtask 1 on Multilingual Information Extraction (IE), which focuses on International Statistical Classification of Diseases (ICD) coding for clinical textual data in Spanish [20,12]. Diagnostic codes are used as a billing mechanism in the Electronic Health Record (EHR) and can be used for automatic semantic indexing of clinical documents, but also to facilitate decision support systems that aim to help clinical coders by suggesting a relevant subset of potential codes for selection. The problem can be described as a mapping from natural language free-texts to medical concepts such that, given a new document, the system can assign multiple codes to it.

In terms of application in the biomedical field, Bidirectional Encoder Representations from Transformers (BERT) has only recently been used for ICD code assignment tasks, such as classifying German animal experiments in CLEF eHealth 2019 [3,27,25]. While it has proven to work well on assigning a smaller subset of ICD codes, it is uncertain how Transformer architecture models can perform on arbitrary long clinical text and in solving extreme multi-label classification problems with a high average amount of assigned codes per document.

CLEF eHealth tracks feature the classification of multilingual clinical documents using ICD codes since 2016 [22,23,24,25]. This work enriches training data with the Medical Information Mart for Intensive Care (MIMIC-III) database and compares BERT based models with XLNet [32].

2 Related Work

A hierarchy-based approach with Support Vector Machines (SVM) [8], using the 'is-a' relationship between ICD-9 codes to model label dependencies has been an early approach to ICD coding [26]. The hierarchy-based classifier surpassed the flat SVM, which did not consider code dependencies. Other approaches identified label density and label noise as useful features [29], while others empirically evaluated the simultaneous occurrence of labels [16].

ML-NET [10] followed the hierarchy-based approach and extended the coding of documents. Its deep neural network consists of an additional network for estimating the number of labels. Instead of separating relevant vs. irrelevant labels by a threshold value, a network for predicting the number of labels was built by using the document vector as input.

Baumel et al. [4] evaluated 4 different models for ICD code assignment using data from MIMIC-II and MIMIC-III data sets. They presented a continuous bag-of-word model [19] (CBOW), a convolutional neural network, an SVM one-versus-all model and a bidirectional gated-recurrent unit model with hierarchical attention (HA-GRU).

Another proposed model is a code-wise attention network [21], where attention mechanisms are used to extract n-grams from the text that are influential in predicting each code.

Unified Medical Language System (UMLS) [5] mapping and word embeddings have shown to be effective within text classification in the biomedical domain and improved results in automatic ICD coding [28]. The embeddings were selected by sequentially mapping discharge summaries to UMLS biomedical concepts in an approach to enrich word representations and to eliminate variations caused by tense, abbreviations and/or spelling mistakes.

3 Dataset

For training data, two different sources were used: The offical CodiEsp dataset[3] with manually generated ICD-10 codifications, and the MIMIC-III database with the older ICD-9 classification system in use, which are mappable to discharge summaries [15]. When exploring other additional resources, such as the abstracts collected from Lilacs and Ibecs[4], the MIMIC-III database was selected as the main data source for augmentation, because it seems to be the most similar database compared to the CodiEsp corpus. Among the free text narrative structured documents describing hospital courses, it has a high average amount of manually assigned codes per document coming from real-world EHRs. With the decision to use the MIMIC-III dataset for augmentation it was also decided to focus on the English translated documents of CodiEsp corpus. A key difference between the two data sources is that the codification for CodiEsp is a semantic mapping of concepts, where the assigned codes do not have to be based on medical outcome. For example, a negative serum test (as seen in Listing 1.1) for CodiEsp still results in appropriately assigned ICD-10 codes, whereas it would not appear on MIMIC-III.

Listing 1.1. Excerpt of CodiESP Document with id S0211-69952009000500014-1, showing results of a blood serum test and codification (Assigned Codes List: r80.9, r20.2, b19.20, b19.10, r23.8, r60.0, r10.9, r19.7, m25.50, l98.9, b20).

```
[...]
On physical examination: blood pressure 104/76 mmHg,
BMI 27, minimal edema in lower limbs and papules in
elbows and arms. Blood count and coagulation were
normal, creatinine 0.9 mg/dl, total cholesterol
238 mg/dl, triglycerides 104 mg/dl, total protein
6.5 g/dl and albumin 3.6 g/dl. Anticardiolipin
antibodies anticardiolipin: Serology against HBV,
HCV and HIV was negative.
[...]
```

3.1 CodiEsp Corpus

The CodiEsp corpus consists of 1,000 clinical case studies manually selected by a practicing physician and a clinical documentarian [20]. The training and development dataset comprises 750 documents with an average of 11.09 codes assigned per document. The test set contains 250 documents and was provided with an additional collection of more than 2,000 documents (background set) to prevent manual corrections. Within the CodiEsp training and development dataset there are 26,696 unique tokens with an average of 301 tokens and 19 sentences per document. It contains 2,557 distinct codes in total of which 363

[3] `https://doi.org/10.5281/zenodo.3625746`, last accessed 2020-07-17

[4] `https://doi.org/10.5281/zenodo.3606625`, last accessed 2020-07-17

unseen codes appear in the test set as seen in Figure 1 (a). 68.24 % of the codes are explainable with the CodiEsp training and development dataset.

3.2 MIMIC-III Corpus

The MIMIC-III database comprises de-identified records from Beth Israel Deaconess Medical Center intensive care unit (ICU) stays, collected between 2001 and 2012. It contains 59,652 discharge summaries with an average of 11.48 codes assigned per document. It has 119,171 unique tokens with 1,947 tokens and 112 sentences on average. The dataset is in principle very well suited but has some characteristics that need to be adapted. The coding system is ICD-9, which has to be converted to ICD-10 accordingly to match the CodiEsp codification. In addition, the dataset only contains summaries of intensive care unit stays, which on average exceed the maximum length of tokens available for Transformer architectures. After conversion, the dataset contains 5,447 distinct codes as seen in Figure 1 (b).

Segmentation For BERT [9] models, the maximum length of a sequence after tokenizing is 512, resulting in an effective limit of 510 tokens for the input layer after subtracting the [CLS] and [SEP] tokens. Because MIMIC-III discharge summaries have an average length of 1,947 tokens (see Table 1) with only 11.67 % of all documents not exceeding 510 tokens, the data has to be truncated in order to fit into the Transformer model.

A simple approach as supposed by Sun et. al. [30] would be to only use the first 510 tokens (head-only) or to use the last 510 tokens (tail-only) of a document, but none of them seem to be appropriate for truncating clinical text without losing relevant information.

When inspecting the summaries, even though they are free text narratives, a fixed structure has been identified in most of the documents: They usually start with a *Chief Complaint* followed by a *historical Background* section, which may include *History of Present Illness*, *Past Medical History*, *Social History* and *Family History*. During *Diagnostics* and *Pertinent Results*, the structure is no longer as consistent and different sections appear, which are more dependent on the individual case. From the middle towards the end of the documents there is a section called *Brief Hospital Course*, which summarizes the ICU stay followed by discharge condition instructions and/or followup instructions.

In early experiments, the effect of using different segments was evaluated. Here, it was found that using the first 510 tokens (head-only) of discharge summaries decreased the performance in comparison to using the last 510 tokens (tail-only). It can be assumed that this is because the background history, which comes at the top of the documents, is not as relevant to the clinical coding as the narrative over the actual present hospital course. It was decided to remove content up to the *Brief Hospital Course* section and sequentially use the remaining document up to whatever fits into 510 tokens. $7,822$ documents were omitted where this section was not present, resulting in 13 % loss of data. Descriptive statistics of the segmented corpus can be seen in Table 1.

Table 1. MIMIC-III, CodiESP Training and Development dataset descriptive statistics. (*) denotes the segmented corpus.

	MIMIC-III	MIMIC-III*	CodiESP Train_Dev
Number of records with ICD code	59,652	51,830	750
Number of unique tokens	1,091,025	276,500	26,696
Number of bigrams	10,609,279	2,846,377	114,846
Number of trigrams	27,814,651	7,873,155	180,650
Avg. number of tokens / record	1,947	427	301
Avg. number of sentences / record	112	39	19
Avg. number of labels / record	11.48	11.45	11.09

ICD-9 code Conversion with General Equivalence Mappings ICD-9 codes of the MIMIC-III database have been converted to ICD-10 using the publicly available ICD-10 CM General Equivalence Mappings (GEMs) [6]. Turer et al. assessed the reliability of conversion between ICD-9 and ICD-10 and found that manual coding from the forward GEMs and backward GEMs were reproducible by 85.2 % and 90.4 % respectively [31].

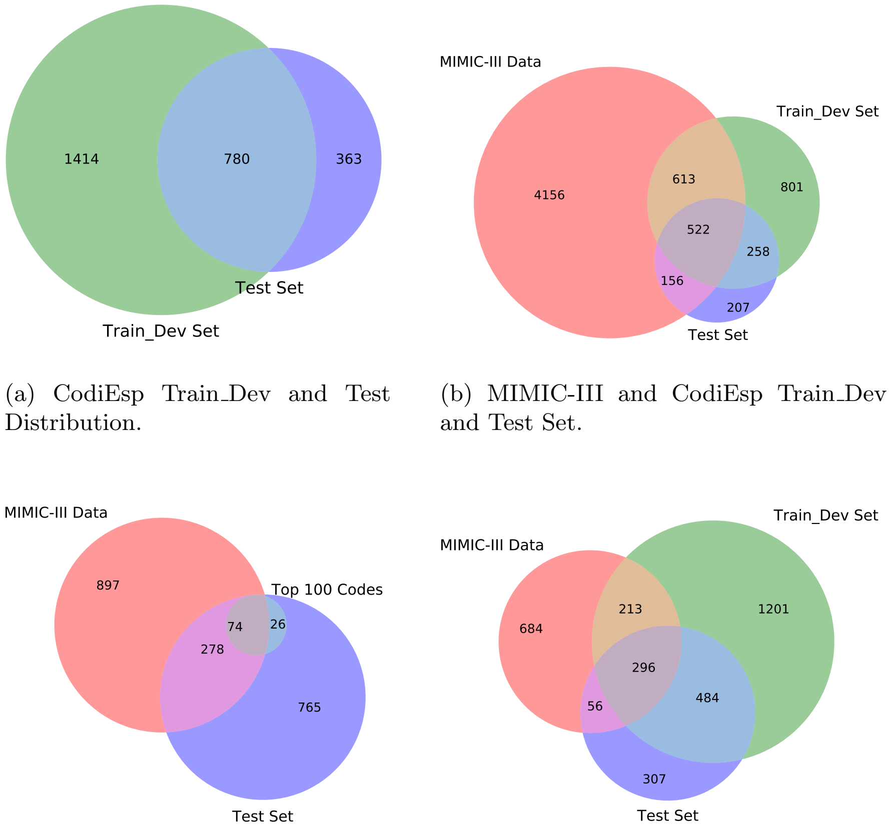
Data Selection Because of the different data sources and MIMIC-III being limited to ICU cases, both datasets have been compared in terms of their distinct code subsets. As seen in Figure 1 (b), the MIMIC-III data contains 4,156 unique ICD-10 codes that are not present in the CodiEsp train, development, and testset. These codes are less generic, apply to the ICU cases and are not covered by the smaller CodiEsp corpus. To make the data augmentation more practical, only documents where 50 % or more of the assigned codes are present in the Top 100, Top 250 or Top 500 frequent codes of the CodiEsp training and development set were used (impact on training size can be seen in Table 3). Only discharge summaries containing the *Brief Hospital Course* section were selected by using a regular expression match, resulting in 51,830 out of 59,652 available documents.

Available data augmentation increases when changing the Top frequent code amount, because the criteria/matching rule, if a document has 50 % or more codes is less strict, resulting in more MIMIC-III documents ending up in training data. Increasing the augmentation data in that way increases recall, but reduces precision (see Table 4). A good compromise was to create a model that is able to predict the Top 100 frequent codes in CodiEsp.

4 Methods

4.1 Transformer architecture and BERT

BERT and Transformer [9] have proven to be extremely effective in many downstream natural language processing (NLP) tasks. While it works well on assigning a smaller subset of ICD codes [3,27], it is uncertain how BERT models can work



(a) CodiEsp Train_Dev and Test Distribution.

(b) MIMIC-III and CodiEsp Train_Dev and Test Set.

(c) MIMIC-III with 50 % in Top 100 CodiEsp and Test Set.

(d) MIMIC-III with 50 % in Top 100 CodiEsp and Train_Dev and Test Set.

Fig. 1. Venn diagrams showing the distribution of the number of distinct ICD-10 codes for different datasets and subsets.

with clinical texts of any length and in solving extreme multi-label classification problems with a high average number of assigned codes per document.

Though the MIMIC-III augmentation does not fit into the token limitation without clipping documents, the Transformer architecture offers good innovations that can be practical in the classification of clinical text. The word tokenizer allows words that are outside the vocabulary to be represented by word pieces instead of simply assigning them to an unknown token, which is why it was selected for the first tests. This feature is particularly useful for discharge summaries, as spelling mistakes and non-standard abbreviations are common.

Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) [18] and ClinicalBERT [2] have the same architecture but are pre-trained on large-scale biomedical corpora. BioBERT has been pre-trained on PubMed abstracts[5] and PMC[6] full-text articles. Bio_ClinicalBERT[7] is an extended model that was also pre-trained on all notes from MIMIC-III (880M words). The Bio_ClinicalBERT model was selected because of the larger pre-training.

4.2 XLNet

The recently proposed Generalized Autoregressive Pretraining for Language Understanding (XLNet) model [32] is an autoregressive language model (LM). It is important to note that although BERT and XLNet have many similarities, there are some differences that need to be explained. Here, autoregressive means that XLNet makes use of the TransformerXL [14] to capture information from previous sequences in order to process the current sequence, and achieving the *regressive* effect at the sequence level. XLNet uses relative position coding and a permutation LM, by factorizing the output with all possible permutations.

The permutation effect is limited to words which are "attended" to. This is done by changing the attention mask prior to the attention softmax while keeping track of the positional information in a sequence. For example, during pre-training, to predict a token t, the attention mask is set to minimum numbers for tokens that appear after position $i > t$. Only the tokens before and including t on the current factorization are used to compute the attention. The advantage is that the tokens that come before t change with each permutation, but their positions within the sequence are kept constant, allowing XLNet to capture bidirectional context.

XLNet implements the Multi-head attention, which is slightly different from the one in BERT, where it is known that it generates a query Q, a key K, and a value V projection of each word in the input sentence. For each query Q, the Multi-head attention Layer uses K to compute an attention score for each value vector V and then sums the value vectors into a single representation using the attention weights [7].

For XLNet, linear layers are used to map the input to the Multi-head attention layer directly. This results in mapping the input into smaller space with the same number of dimensions that add up to the original dimension as known for BERT. This allows each word to attend more to other words and not only to itself, which results in a final richer representation of each word.

4.3 Preprocessing

Because the discharge summaries were de-identified free text narratives, additional pre-processing steps were taken to convert them into a sequence of sentences, removing all numbers, and name placeholders. Leading and trailing

[5] `https://pubmed.ncbi.nlm.nih.gov/`, last accessed 2020-07-17

[6] `https://www.ncbi.nlm.nih.gov/pmc/`, last accessed 2020-07-17

[7] `https://huggingface.co/emilyalsentzer/Bio_Discharge_Summary_BERT`, last accessed 2020-07-17

spaces, quotations and semicolons have also been removed. For the CodiESP corpus, no pre-processing was applied.

4.4 Training the models

The experiments were done with the PyTorch-transformers implementations of BERT and XLNet[8]. The overall end-to-end training process can be seen in Figure 2. The models were fine-tuned on all layers without freezing. As proposed by the original papers [9,32], Adam [17] was used in early experiments as the optimizier, but was then replaced by the Layerwise Adaptive Large Batch (LAMB) optimizer [33] because it resulted in a slightly reduced training time. The hyperparameters have been selected and optimized based on the development set performance. Using a *learning rate* of 7*e*-4 or 6*e*-4 resulted in the best scores, though the Transformer model seems to react very sensitively to the use of different learning rates, because selecting different settings often led to poor results.

Different *warmup schedules* were tried, but had no impact on the results. Among the two versions of BERT *cased* and *uncased*, it was found that overall the *uncased* version works slightly better. However, the difference is still very small. For XLNet, the only available version is *cased*. The base version of XLNet was preferred over the large version due to computational expense. The training *batch size* was 8 for XLNet and 16 for BERT models. To produce the ranking of the codes, Binary cross-entropy with logits was used to obtain confidence for each ICD-10 code during inference. They were then ordered by confidence and cut off with a threshold of $t = 0.4$. The prediction pipeline of the BERT model including the association rules is shown in figure 3.

4.5 Apriori Association Rules

The apriori algorithm [1] has been used to find frequent itemsets in a list of transactions but recently has also been in use to find association rules and label co-occurrences in clinical text, such as in autopsy reports [11]. Association rules can be obtained with the *support* and *confidence* parameter, where the support of a set of items is the probability that this set of items occurs in a transaction. Confidence refers to the likelihood that an item B will also be purchased when item A is purchased. It can be calculated by dividing the number of transactions where A and B are bought together by the total number of transactions where A is bought. To identify and explore co-occurrences, a low min support (0.02) value has been used on the CodiEsp train and development set. The resulting apriori association rules as seen in Figure 6 have been plotted with the *arulesviz* [13] R package. The graph shows 59 rules.

One example for a relation is Hepatitis B and C as shown by the rule that connects b19.10 and b19.20. When exploring the data, it was found that this rule refers to serology tests, that often include test results for different viruses, such as hepatitis B and C. An example can be seen in Listing 1.1. Another confident

[8] `https://github.com/huggingface/transformers`, last accessed 2020-07-17

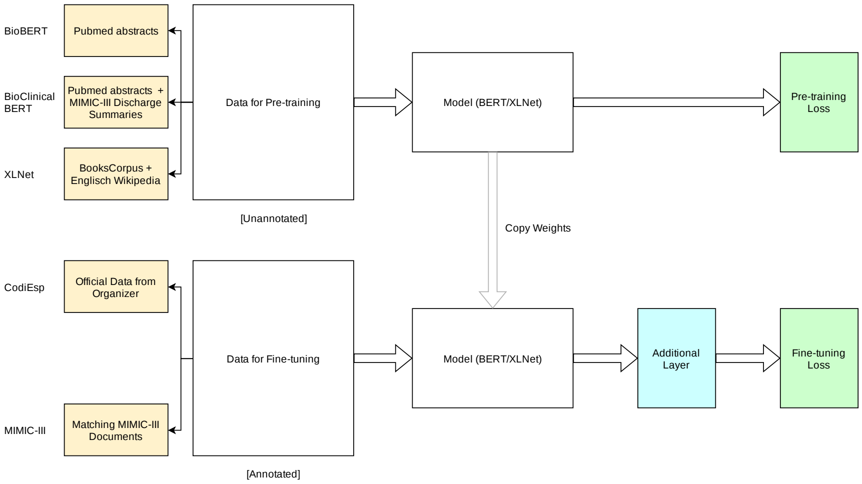


Fig. 2. Workflow of end-to-end training process. Unannotated pre-training (on large corpora) and annotated fine-tuning with combined resources (CodiEsp and MIMIC-III). The weights from pre-training phase were transferred.

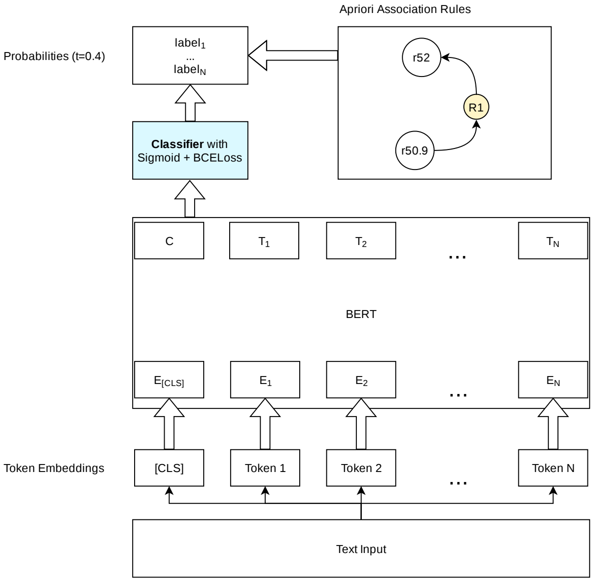


Fig. 3. Inference for BERT models with apriori association rules. The text-input is classified with BCELoss function to get probabilities for all available ICD codes. The confidence must be at least $t = 0.4$ (threshold) to count as a positive ICD code.

rule is that localized enlarged lymph nodes (r59.0 and r59.9) links to unspecified fever (r50.9), which then links to unspecified pain (r52). As such rules should be covered by the trained model, not that many different rulesets have been tested and added during inference.

However, the 11-ruleset as seen in Figure 5 improved the mean Average Precision (MAP) results on the development set between 0 % to 1.2 % depending on the model and was therefore added to the final submission if missed out. The submission guideline requires that the prediction is ordered by confidence. Because the predicted confidence cannot be compared with apriori support or confidence values and because the confidence of the primary model was not high enough, the association rule codes were added at the end. They were ranked by highest level of support.

5 Results and Discussion

Figure 4 shows experimental runs on the development set for the tested models with different pre-trained embeddings and different frequent Top code subsets. This results in different enriched training data and also in a different amount of labels a model is able to predict. A comparison of how many documents end up in the training data can be seen in Table 3. The final best results on the development set for each model can be seen in Table 2.

While the F1-Score is superior on models which are only able to predict the Top 50 frequent codes, the MAP score penalises this behaviour on the full set, because not only the classification but also the positional ranking is taken into consideration. When matching the Top 50 most frequent codes with MIMIC-III there is not enough data available for augmentation (363 additional documents). Starting with the Top 100 most frequent codes, improvements coming from the additional data can be seen. The augmentation improves the reported MAP score by 0.097 (0.128 F1) for the XLNet model. Increasing the training data further increases recall, but decreases precision.

The final test set results for evaluation were reported by the task organisers and can be seen in Table 4. On the test set, the Bio_ClinicalBERT model achieved the overall best performance for a single model with a MAP score of 0.259. XLNet on Top 100 frequent codes achieved the best performance in precision.

When the goldstandard for the test set was released, it was evaluated how many of the unseen codes would have been explainable by keeping the remaining annotated codes of each MIMIC-III document within the training data (Knowledge Discovery). Figure 1 (d) shows that for the Top 100 most frequent codes training set, 56 distinct unseen codes would have been explainable. Here, a small performance improvement can be expected, but it is noteworthy that only a few of the codes were seen more than once in the test data (76 appearances in total). Because they were unseen before, it can be assumed that these are codes with rare appearances. It can be concluded that more resources are needed to be able to explain the full code set.

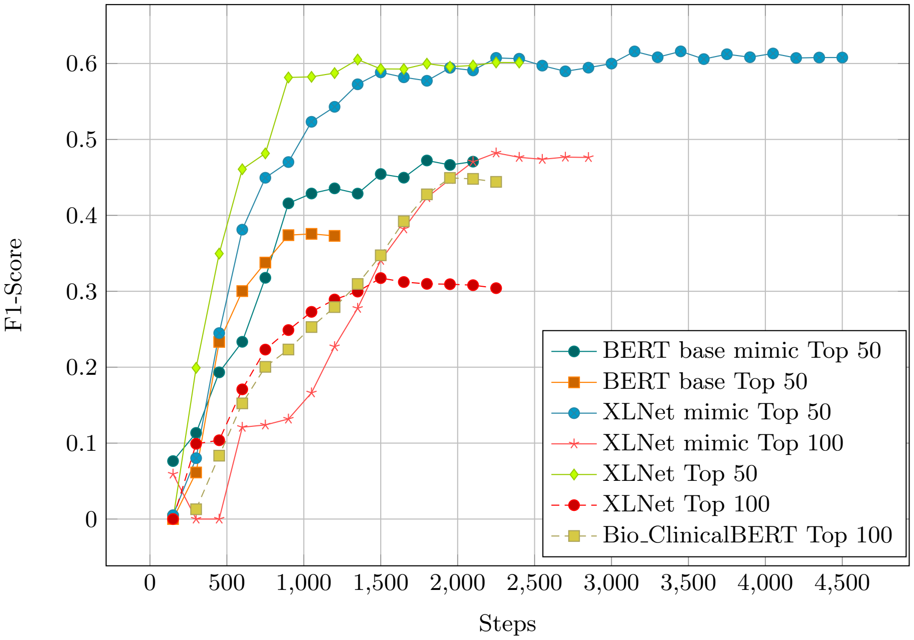


Fig. 4. Experimental runs on the development set for tested models with different pre-trained embeddings and different frequent Top code subsets. F1-Score is being reported at intervals of 500 steps during training.

Table 2. Results of the evaluation performed on the development set. MAP and F1-Score are being reported, where bold indicates best results for a category.

Model	MAP	F1
XLNet_base_cased + MIMIC-III - Top 50	0.232	**0.608**
XLNet_base_cased - Top 50	0.216	0.602
BERT_base_uncased + MIMIC-III - Top 50	0.143	0.47
BERT_base_uncased - Top 50	0.165	0.372
XLNet_base_cased + MIMIC-III - Top 100	**0.247**	0.432
XLNet_base - Top 100	0.15	0.304
Bio_Clinical_BERT + MIMIC-III Top 100	0.244	0.361

Table 3. Model size comparison on final submission.

Model	Training Data Size	Model Size
XLNet_mimic_500	19,484 documents	459.78M
XLNet_mimic_250	10,754 documents	459.03M
XLNet_mimc_100	3,286 documents	458.58M
Bio_Clinical_BERT_100	3,286 documents	423.43M

Table 4. Results of the final evaluation performed by the task organisers. They report MAP, Precision, Recall and F1 scores. (*_Cat) denotes that the score has been computed for categories determined as the first three digits of a code. (*_Codes) denotes that the score has been computed for the subset of codes only present in the train and validation sets. (*) denotes ensemble of Bio_ClinicalBERT_mimic_100 and XLNet_mimic_100. Bold indicates best results for the category. (BERT†) denotes that the Bio_ClinicalBERT version was used.

Model	MAP	MAP_Codes	P	R	F1	F1_Codes	F1_Cat
BERT†_mimic_100_apriori	0.242	0.288	0.375	0.285	0.324	0.352	0.373
XLNet_BERT†_ensemble*	**0.259**	**0.306**	0.407	**0.287**	**0.337**	**0.367**	**0.387**
XLNet_mimic_100_apriori	0.231	0.275	**0.457**	0.244	0.318	0.351	0.366
XLNet_mimic_250_apriori	0.21	0.244	0.342	0.28	0.308	0.334	0.366
XLNet_mimic_500_apriori	0.128	0.149	0.235	0.215	0.225	0.243	0.276

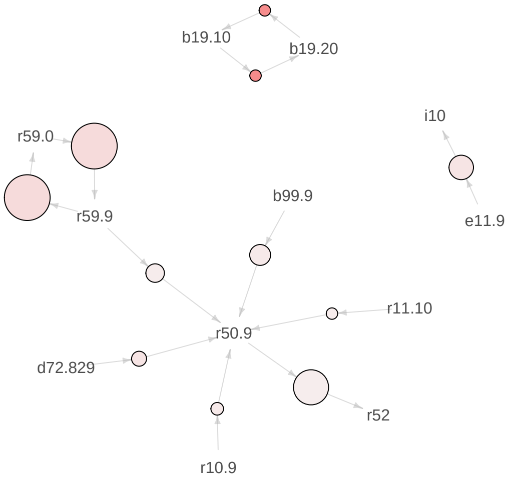


Fig. 5. Graph for 11 ICD-10 apriori association rules. Size: min_support(0.03) min_confidence(0.3), Color: lift(1.393-20.294).

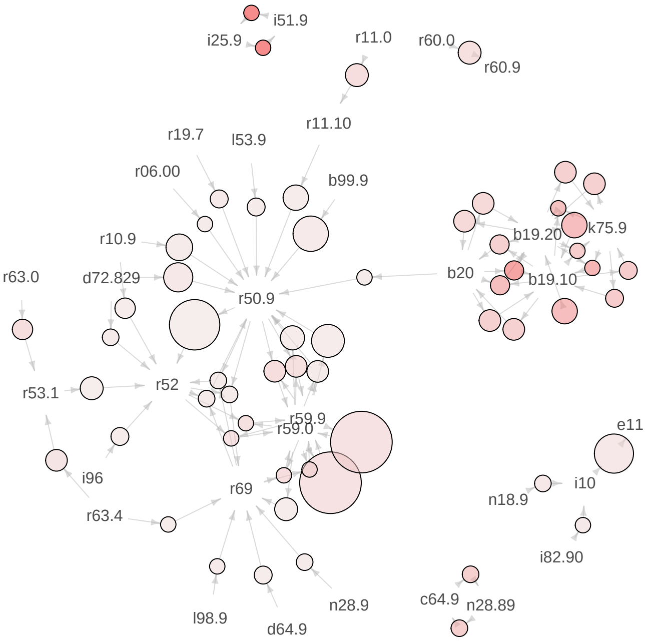


Fig. 6. Graph for 59 ICD-10 apriori association rules. Size: min_support(0.02) min_confidence(0.3), Color: lift(1.393-20.294).

6 Conclusions

This work compared BERT based models with XLNet. The effect of enriching training data with documents from MIMIC-III was evaluated. Here, it was found that the MIMIC-III augmentation with code conversion was able to improve the results compared to using only the stock data set. The apriori algorithm has been applied to build and explore association rules by finding frequent item sets. The 11-ruleset was able to improve the mean Average Precision (MAP) results on the development set between 0 % and 1.2 %.

Among the submitted models, the ensemble of BioBERT and XLNet achieved the highest mean Average Precision (MAP) score of 0.259 (0.306 for the subset of codes only present in the train and validation sets). In terms of single model performance, the Bio_ClinicalBERT model achieved overall best performance. The XLNet, even though pre-trained on generic text has the highest precision value on the test set and overall best performance on the development set.

Though the models are still far from achieving good results on the full label set, the task has been very challenging with many possible labels, given only a relatively small dataset. It was found that the large MIMIC-III database is not able to cover all unseen codes, so it can be concluded that more resources are needed to be able to explain the full code set.

In future work, XLNets attention should be further evaluated because the sequence dependency on the hidden states of previous sequences can be adjusted by a memory length hyper-parameter. It will be interesting to tune and see the impact of this parameter, but also to test and see how a domain-specific XLNet model performs when pre-trained on large biomedical data.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB). vol. 1215, pp. 487–499 (1994)
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, 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 (Jun 2019). https://doi.org/10.18653/v1/W19-1909
3. 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 (2019)
4. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., Elhadad, N.: Multi-label classification of patient notes: case study on ICD code assignment. In: Workshops at the thirty-second AAAI conference on artificial intelligence (2018)
5. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
6. Butler, R.R.: ICD-10 general equivalence mappings: Bridging the translation gap from ICD-9. Journal of AHIMA **78**(9), 84–86 (2007)
7. Clark, K., Khandelwal, U., Levy, O., Manning, C.: What Does BERT Look at? An Analysis of BERT's Attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 276–286 (01 2019). https://doi.org/10.18653/v1/W19-4828
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
10. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., Lu, Z.: ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association **26**(11), 1279–1285 (2019). https://doi.org/10.1093/jamia/ocz085
11. Duarte, F., Martins, B., Pinto, C.S., Silva, M.J.: Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. Journal of Biomedical Informatics **80**, 64 – 77 (2018). https://doi.org/10.1016/j.jbi.2018.02.011
12. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez Gonzales, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth Evaluation Lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . LNCS Volume number: 12260 (2020)

13. Hahsler, M., Chelluboina, S.: arulesviz: Visualizing association rules and frequent itemsets. R package version 0.1-5 (2012)
14. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339 (01 2018). https://doi.org/10.18653/v1/P18-1031
15. Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
16. 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** (05 2015). https://doi.org/10.1016/j.artmed.2015.04.007
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
18. 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 (2019). https://doi.org/10.1093/bioinformatics/btz682
19. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: 1st International Conference on Learning Representations (ICLR). vol. abs/1301.3781. Scottsdale, Arizona, USA (2013)
20. 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)
21. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable Prediction of Medical Codes from Clinical Text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1101–1111 (2018). https://doi.org/10.18653/v1/N18-1100
22. Névéol, A., Cohen, K.B., Grouin, C., 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. CEUR Workshop Proceedings **1609**, 28–42 (2016)
23. Névéol, A., Robert, A., Anderson, R., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
24. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., Ramadier, L., Rey, G., Zweigenbaum, P.: CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2018)
25. Neves, M.L., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the CLEF eHealth 2019 multilingual information extraction. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2019)

26. 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)
27. 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 (2019)
28. Schäfer, H., Friedrich, C.M.: UMLS mapping and Word embeddings for ICD code assignment using the MIMIC-III intensive care database. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 6089–6092. IEEE (2019)
29. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science **292**, 135–151 (2013)
30. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: China National Conference on Chinese Computational Linguistics. pp. 194–206. Springer (2019)
31. Turer, R.W., Zuckowsky, T.D., Causey, H.J., Rosenbloom, S.T.: ICD-10-CM Crosswalks in the primary care setting: assessing reliability of the GEMs and reimbursement mappings. Journal of the American Medical Informatics Association **22**(2), 417–425 (2015)
32. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: NeurIPS 2019. pp. 5754–5764. Vancouver, BC, Canada (2019)
33. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In: International Conference on Learning Representations (ICLR) (2020)