NLP Project: Multi-Label Classification Of ICD Codes Using Patient Discharge Summaries

Tom Amsterdam (308228998) and Hadas Shaham (203149596)

Abstract— **The processing of medical records for medical classification and payment handling requires the assignment of diagnosis and procedure codes. The assignment of codes is done based on free-text document such as clinical notes summaries. Wide implementation of Electronic Health Records (EHRs) opens the possibility for automated reading and analyzing the clinical notes and code assignment. We present an Hierarchical Neural model for automated coding of clinical notes. The model is comprised of an embedding layer, a recurrent layer and convolutional layers with additional two attention layers at the word and sentence levels. Our Model achieves 0.4702 Micro F1 score and 0.9569 AUC. These results are better the the baseline comparison.**

I. INTRODUCTION

As part of the medical and payment process, there is a need to assign multiple labels (ICD, International Statistical Classification of Diseases, dodes) that represent procedures performed and diagnoses of a specific case. In order to assign the correct subset of ICD codes to a hospital admission, human medical coders read the clinical notes summaries and then decide on the appropriate coding labels based on pre-defined guidelines. In our project we refer to the guidelines of the United State government for ICD-9 codes. ICD-9 standard is a list of codes mapping diagnoses and procedures recorded in hospital care in the US. These codes are written into the patients EHRs and are used in the future for diagnostics, billing and reporting purposes. Each code is composed of 3, 4 or 5 characters with latent hierarchical structure. The first three code characters represent the category while the other digits present a more fine detailing of the diagnosis/procedure. Medical coders assign manually a set of appropriate ICD-9 codes after reviewing the information of a patient record. This labelling requires professional knowledge in medicine and medical procedures hence the process is very expensive and is prone to human error. Diagnosis assignment is a particularly hard task due to two main reasons. The first reason is the need to base the assignment on human-made long clinical notes that are not structured and prune to mistakes. The second reason is the massive size of the label set, the ICD-9 codes amount to approximately 14,000 possible codes. It is estimated that the cost of ICD coding of clinical records and correcting related errors is around 25B USD a year in the US. [3] In our project we experiment with leveraging multi-labeling assignment of unstructured text, the note summaries portion of the EHR, along with different applications of neural architecture. Successful solution for the automation of the process with high accuracy can reduce massive costs and simplify that process management in the medical system.

II. RELATED WORK

The examination of previous works shows two different fields of work that project on our specific problem:

- Multi Label Classification - Approaches to NLP Multi-Label Classification tasks that are not necessarily specific or related to the medical field in general or patients record classification in particular.
- Patient Record Labelling - Research regarding the NLP task of processing and classifying labels based on medical datasets of patients record labelling.

There has been significant work on the task of automated coding of clinical notes since 1990. The first approaches tackled the task using pattern matching and rule based systems that leveraged specific structures, words and idioms such as the work of [1]Crammer et al. Earlier Rule-Based approaches were based on manually crafted rules that capture lexical features from the results of text processing algorithms such as n-grams, s-grams and lexical algorithms. More advanced approaches leveraged non-neural machine learning based classification techniques. [7] Perotte et al. (2013) implemented and hierarchical SVM models corresponding to the hierarchical structure of the ICD-9 codes. Each model classified a document with a specific ICD-9 code. They achieved Micro-F1 score of 0.293. [6] Kavuluru et al. (2015) built SVM based classifiers for ICD-9 diagnosis based on UKY EHR medical recorders. They performed strong feature selection and achieved Micro-F1 Score of 0.48. Although receiving impressive results metric wise on a specific dataset, most of the ML methods were not able to generalize well because the over fitting performed by the manual crafted features that failed to capture the complexity of a medical summary note. The most recent and advanced work in the field is tackling the task with deep learning neural methods. Deep learning has the potential to overcome the problems that arise when trying to create manual features or rules. Baumel Baumel et al. 2017 work is based on the MIMIC-III dataset and is dataset we chose to base our work upon as well. In their work they proposed an hierarchical model that incorporates label attention mechanism. They used similar a model as we used, two-level bidirectional GRU encoders, but with different architecture and tasks. The first layer operates over word tokens and embeds them into a sentence. The second layer uses the encoded sentences to create a representation of the entire document. Using the model, they achieved Micro-F1 score of 0.41. The state of the art results were achieved by [4] Mullenbach et al. (2018). Their model included an embedding and a CNN layer with two individual attention layers. Similarly to our model, Mullenbach et al. used a pre-trained embedding matrix. Their best mode obtained a micro F1-score of 0.52.

III. Methodology

Our project formulates the ICD codes classification task as a multi-task binary classification problem. For each hospital admission, every ICD code can be present or absent (labeled 1 and 0 accordingly). For each code the models outputs the probability for the association of the code with the input document. Our approach is comprised from the following process:

- **Preprocessing** - Processing the dataset in order to clean and tokenize the dataset for future stages. In the processing stage we put an emphasize on maintaining the textual construct of the note summary.
- **Word Embedding** - Mapping of the words to their continuous embedding space while maintaining context and semantic relationship of words. We used pre-trained embedding and tuned the embeddings over our dataset.
- **Neural Model** - We used deep learning methods leveraging an hierarchical model with word and sentence encoders to receive a single vector representation of each medical note.
- **Labels Embedding** - Instead of using the ICD-9 code as labels, we use their word description. Through this embedding we received a label representation vector that represents contextual meaning and used this to compute the attention.

IV. The Dataset

We used the public MIMIC-III data-set which contains anonymized EHR records of 58,976 patients visits at the Beth Israel Deaconess Medical Center in the years 2001-2012. For our project we used the discharge summaries only, they contain the most informative diagnostical information.

A. General Properties

The dataset includes 6,918 unique ICD diagnosis codes, however 5 codes we invalid and we omitted them. Due to inherent popularity properties of diseases, the codes in the data-set are distributed such that a small number of the codes represents the majority of the distribution of label occurrences. A single note summary has on average 10-15 associated set of codes.

B. Unique Challenges

- **Label Distribution** - The high number of labels and their unbalanced distribution will cause the model's predictive accuracy of rare codes to be low.

- **Long Unstructured Text** - The discharge summaries can be very long. The average number of words per note is 1,728.
- **Contextual and Semantical Structure** - The notes are usually divided into section such as 'medical history', 'hospital course' and 'history of present illness'.
- **Medical Language** - The notes are hard to process as they are written in medical language using medical abbreviations and contain many spelling mistakes.

C. Specific Model Dataset

Due to the size of the dataset the task raises processing time and memory space challenges. Therefore, we reduced the dataset size by splitting it to 4 parts and using a quarter of the samples. Our dataset is comprised of 52,726 different admission ID discharge summaries. Over the 52,726 samples we performed Train, Test and Development split and made sure that the splits are patient independent according to the following ratio 70:20:10. The statistics regarding our dataset and each specific split can be found below.

Dataset	Number Admissions	Number Patients	Avg Labels Per Admission	Avg number words per note
Total	52,726	41,127	16.1	1,728
Train	36,908	28,139	15.2	1,821
Test	10,546	8,632	18.3	1,576
Dev	5,272	4,357	18	1,382

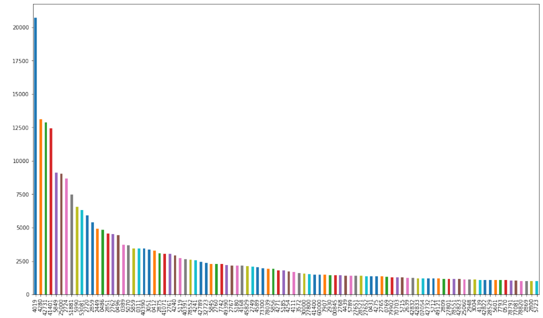


Fig. 1. ICD-9 labels distributions over the dataset

V. Baseline Model

In order to be able to compare our model and results we used past results as a baseline for our work. The baseline model results we used were of [2] Fu and Thirman from the Stanford NLP class who used RNN based document embedding and logistic regression layers. They achieved 0.42 micro F1 score on an easier problem formulation. We use their results as the lower threshold for our model. In addition to that, we examine our result in comparison with the state-of-the-art result achieved by [4] Mullenbach et al. 2018. that is a 0.56 micro F1 score.

VI. The Model

A. Preprocessing and Tokenization

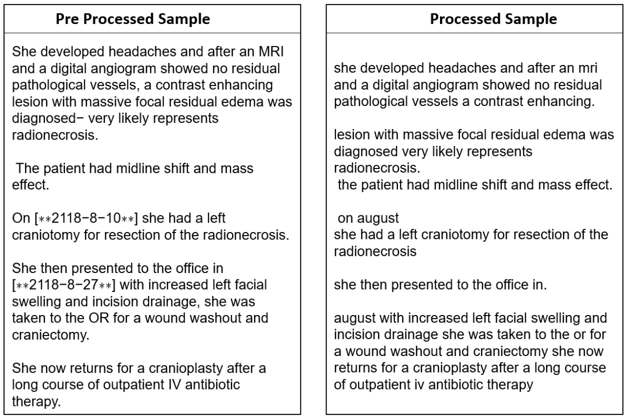
The clinical notes contained in the MIMIC-III dataset require preprocessing before they can be effectively tokenized due to several textual reasons.

- **Medical English** - The language used in the notes differs from common English due to the presence of medical expressions, abbreviations, vital signs transcriptions and drug dispensing report.
- **Spelling Mistakes** - The notes were human written and contain high amounts of misspellings.
- **Anonymized Expressions** - Personal information with anonymization tokens that need to be dealt with. Anonymized information is contained between special symbol [*...**].

We performed preprocessing using the spaCy library to tokenize the text while incorporating custom rules and rules from external projects to handle correctly medical abbreviations. We mapped the anonymized expressions to special tokens representing the type of expression (e.g. "firstname-token"). We built a vocabulary of words using all the tokens that appeared at least 3 times in the training set, any out of vocabulary word was replaced by a special token representing unknown words. We recognise issues with our preprocessing methods that need to be addressed in the future and we believe would lead to improved results but didn't have the time to experiment with:

- **Out Of Vocabulary Words -** We believe that instead of replacing the words with unknown token, a solution to replace with the closest word according to a pre-defined distance function (Nearest Neighbour) will lead to improved results.

- **Sentence Boundaries** - Many sentences were incorrectly split into two or more smaller sentences and thus hurting the semantic meaning and future representation of the sentences. We believe that the fixing of sentence boundaries issue will also lead to improved result.



Pre Processed Sample

She developed headaches and after an MRI and a digital angiogram showed no residual pathological vessels, a contrast enhancing lesion with massive focal residual edema was diagnosed− very likely represents radionecrosis.

The patient had midline shift and mass effect.

On [**2118−8−10**] she had a left craniotomy for resection of the radionecrosis.

She then presented to the office in [**2118−8−27**] with increased left facial swelling and incision drainage, she was taken to the OR for a wound washout and craniectomy.

She now returns for a cranioplasty after a long course of outpatient IV antibiotic therapy.

Processed Sample

she developed headaches and after an mri and a digital angiogram showed no residual pathological vessels a contrast enhancing.

lesion with massive focal residual edema was diagnosed very likely represents radionecrosis.
the patient had midline shift and mass effect.

on august
she had a left craniotomy for resection of the radionecrosis

she then presented to the office in.

august with increased left facial swelling and incision drainage she was taken to the or for a wound washout and craniectomy she now returns for a cranioplasty after a long course of outpatient iv antibiotic therapy

Fig. 2. Processing sample of text - dealing with medical terms while maintaining structure

B. Word Embedding

In contrary to other projects that used Bag Of Words embedding that were trained on the MIMIC dataset itself, we used the GloVe model, word embeddings that were pre-trained on an external English corpus. The rational is that we initialize the embedding layer with the pre-trained vectors and then fine-tune them on medical English during the training. We used d=300 with the pre-trained embedding that were trained on CommonCrawl.

C. Model Architecture

We experimented with an hierarchical model that is based on the one used by [5]Baumel et al.(2017) from BGU with few modifications. Due to the fact that each document is comprised of many words (average of 1728) we split the encoding to be hierarchical: word encoding and sentence encoding.

The model is comprised of 7 layers:

1) **Word Embedding Layers** - Using GloVe pre-trained embedding matrix. Each word is mapped into a vector. We note w_{is} to be the embedded word in index i and sentence s.
2) **Code Label Embedding Layer** - We use a Bidirectional GRU to embed the textual code description for each ICD code. For example for ICD label code '4010' we use the description 'Malignant essential hypertension'. We use the embedding matrix based on the pre-trained GloVe to embed the words and the GRU to represent the entire description sentence as a vector u_l to be used later as a contextual attention vector for the Sentence Attention layer.
3) **Word Encoder** - The tokenized embedded sentences are the input sequence to a bidirectional GRU which encodes information from both directions at the word level while incorporating contextual information into the representation. Using the sequence of hidden states we obtain embedding representation of a sentence in the note summary h_i. We note h_{is} the new word embedding representation for word i in sentence s.
4) **Word Attention** - We used attention for every word h_i to obtain sentence embedding s_i by computing the weighted sum of the word representation h_{is} scaled by its attention weights. The attention similarity vector is initialized randomly and learned during the training.
5) **Sentence Encoder** - A layer that combines the adjacent sentences that a note summary contains. Using the layers to obtain a representation of the note summary D where D_i is the sentence in index i in the document. This is done using a convolutional layer due to processing speed.
6) **Sentence Attention** (For Each ICD Code) - We compute attention similarly to the attention computed at the word level. The difference is that we use an attention vector for each ICD label l using u_l the attention vector relevant for label l and a softmax function. this layer outputs a a document representation vector for each ICD label code noted as T_l
7) **Classification Layer** - From the sentence attention layer we receive a document vector

T_l for each label l. This vector is an high level representation of the note summary in respect to label l. We use this vector as a feature vector for the classification task by feeding it to a fully connected layer and a Sigmoid activation function.

The model training is performed using the binary cross-entropy loss for each label that is computed on the output of the last layer.

D. Detailed Architecture

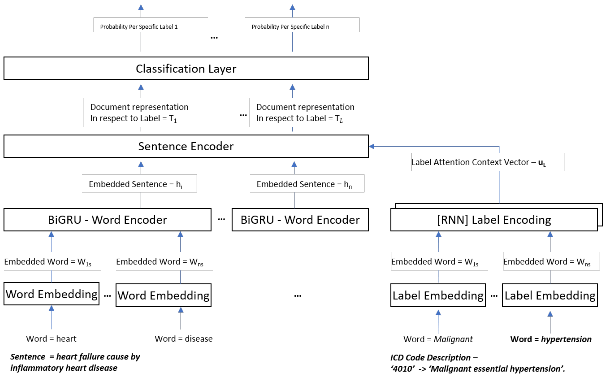

Probability Per Specific Label 1
Probability Per Specific Label n
Classification Layer
Document representation In respect to Label = T_1
Document representation In respect to Label = T_L
Sentence Encoder
Label Attention Context Vector – u_L
Embedded Sentence = h_i
Embedded Sentence = h_n
BiGRU - Word Encoder
BiGRU - Word Encoder
[RNN] Label Encoding
Embedded Word = W_{1s}
Embedded Word = W_{ns}
Embedded Word = W_{1s}
Embedded Word = W_{ns}
Word Embedding
Word Embedding
Label Embedding
Label Embedding
Word = heart
Word = disease
Word = *Malignant*
Word = *hypertension*
Sentence = heart failure cause by inflammatory heart disease
ICD Code Description – '4010' -> 'Malignant essential hypertension'.

Fig. 3. High level architecture of the entire model.

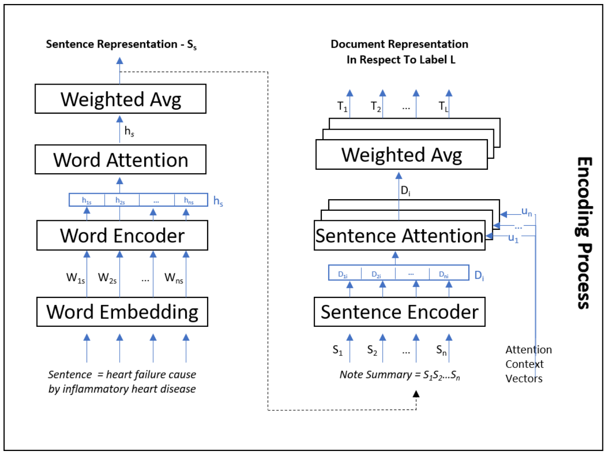

Sentence Representation - S_s
Weighted Avg
h_s
Word Attention
h_{1s} h_{2s} ... h_{ns}
h_s
Word Encoder
W_{1s} W_{2s} ... W_{ns}
Word Embedding
Sentence = heart failure cause by inflammatory heart disease
Document Representation In Respect To Label L
T_1 T_2 ... T_L
Weighted Avg
D_i
Sentence Attention
u_n
u_1
D_{1i} D_{2i} ... D_{ni}
D_i
Sentence Encoder
S_1 S_2 ... S_n
Note Summary = $S_1S_2...S_n$
Attention Context Vectors
Encoding Process

Fig. 4. Word and sentence encoding process to obtain a document vector representation.

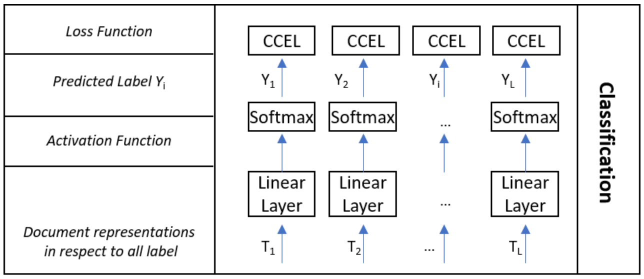


Fig. 5. The classification process in respect to each label

VII. Evaluation Metrics

The ICD code label set is inherently sparse, for a specific note summary most codes labels are labelled as false while only few are true. Therefore when considering the evaluation metric we give more weights to micro evaluation metrics as they are more informative for the specific task. In order to be able to compare the results to prior work of others, we focus on Micro-Average F1 Score, ROC and AOC.

1) F1 Score - The harmonic mean of precision and recall. It is popularly used to evaluate the performance of binary classifiers.
2) AUC Score - The score measures the probability that the model assigns higher score for a positive instance than a negative one. In our specific task the AUC tends to be very high because it measures true-negatives that are very frequent in this task.

VIII. Results

The results table shows the evaluation of the model over the chosen metrics compared to the baseline model and the state-of-the-art models. We achieved significant improvement in respect to the baseline model in the comparable metrics. This improvement is even more impressive when considering that our problem formulation is harder than that of their problem formulation as we didn't shrink the target label space. A major limitation to our work was the compute and memory resources needed to tackle the challenge, due to the amount of training data. Each training Epoch takes a lot of time even with parallel processing. Comparing to the state-of-the-art results achieved by Mullenbach et al. shows that our model is still inferior in comparison to their CAML model (Convolutional Attention for Multi Label classification). This can also be explained by the differences in architecture complexity and processing resources needed to train their model.

Mode	Embedding	Micro Precision	Micro Recall	Micro F1	Micro AUC
Baseline	Word2Vec	—	—	0.42	—
HA-GRU	Word2Vec	—	—-	0.464	—
CAML	MIMIC-III	0.6322	0.4428	0.5208	0.9842
Our Model	**GloVe**	**0.5381**	**0.4223**	**0.4702**	**0.9569**

IX. CONCLUSIONS

We presented a Neural model for the task of medical free text classification of ICD-9 codes. We achieved comparable but slightly inferior results of the state-of-the-art model. To the best of our knowledge we use a new novel architecture for this specific task by leveraging a GRU based word encoder with a CNN based sentence encoder therefore enjoying the benefits from both words - better long-term memory at the sentence level than Mullenbach et al. and faster processing time than Baumel et al. We find this result important as they benchmark well in respect to real-world human coding error mistakes rate. It is estimated in the medical coding industry that 0.35-0.45 of human coded note summaries contain ICD code assignment mistakes. Therefore, this field of study is getting closer to becoming applicable as Decision Support Systems of human medical coders. Regardless of the model decisions described below, we believe that running the training with a stronger infrastructure with more compute and memory resources will allow to achieve better results. We see a number of directions going forward that we believe will help increase the quality of the model:

- Improved Preprocessing - better handling of out-of-vocabulary words, our suggested direction is to define a distance function to find the words contextual Nearest Neighbour. Fixing the sentence boundary issues we encountered in the tokenizaiton process. Building a better medical words dictionary with spelling mistakes capabilities for unique medical words.
- Better Utilization of Code Structure - We believe that it is possible to use hierarchical structure of the ICD codes and the fact that they are built with internal properties of conjunction and aggregation to create an hierarchical classifier layer that can yield better results.

REFERENCES

[1] Koby Crammer, Mark Dredze and Kuzman Ganchev and Partha Pratim Talukdar Automatic Code Assignment to Medical Text

[2] CS224N Final Project - Medical Record Understanding

[3] Richard Farkas, Gyorgy Szarvas. ”Automatic construction of rule-based ICD-9-CM coding systems”. BMC Bioinformatics 2008.

[4] Mullenbach, James, Wiegreffe, Sarah, Duke, Jon, Sun, Jimeng, Eisenstein, Jacob. 2018. Explainable Prediction of Medical Codes from Clinical Text.

[5] Baumel, Tal, Nassour-Kassis, Jumana, Elhadad, Michael,Elhadad, Noemie. 2017. Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment.

[6] Kavuluru, Ramakanth, Rios, Anthony, Lu, Yuan. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records.

[7] Perotte, Adler, Weiskopf, Nicole, Elhadad, Noemie, Pivovarov, Rimma, Natarajan, Karthik, and Wood, Frank. 2013. Diagnosis code assignment: models and evaluation metrics.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[9] Yitao Zhang A Hierarchical Approach to Encoding Medical Concepts for Clinical Notes Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 67–72

[10] Alan R. Aronson1, Olivier Bodenreider1, Dina Demner-Fushman1, Kin Wah Fung1, Vivian K. Lee1,2, James G. Mork1, Aurelie Neveol1, Lee Peters1, Willie J. Rogers From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. BioNLP 2007: Biological, translational, and clinical language processing, pages 105–112

[11] [11] Zhang, M. and Zhi-Hua Z. (2014) A review on multi-label learning algorithms. Knowledge and Data Engineering, IEEE Transactions 26.8: 1819-1837.

[12] Priyanka Nigam Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records cs224d Class paper presentation. 2015