CS779A

Group 12-Biomedical/Clinical Text Processing

Saksham Gupta (170613)
Sathvik Bhagavan (170638)
Harshit Gupta (170252)

Contents

What is Biomedical/Clinical Text Processing?

- Biomedical/Clinical text mining (**BioNLP**) is the study of how NLP techniques are applied to texts and literature of the biomedical and molecular biology domains.

- Clinical reports (**EHRs**) have a lot of information about patients: family background, disease and treatment results, interpretation of test images, behaviour and much more.

Tasks and Challenges in BioNLP

Negation Detection: It is an important problem in BioNLP as many clinical statements are written as the absence of certain disease. Formally, it is a classification task.

Word Sense Disambiguation: Abbreviations are common in biomedical documents, and many are ambiguous in the sense that they have several potential expansions.

Information Retrieval: This is the task of extracting and encoding information from clinical narratives, journals, EHRs, discharge summaries, or medical reports

Named Entity Recognition: Identification of entities like diseases, genes, chemicals, drugs and symptoms in the given text. It is particularly a complex problem as entities in biomedical domain are often described using long phrases consisting of punctuation and characters.

Clinical Coding: The task of translating clinical statements into a set of codes, as defined in international standards. We will see more about Clinical Coding in upcoming slides.

References: WSD, NER

Datasets in BioNLP

Dataset	Description
MIMIC-III Dataset	Notes, Procedures, ICD Codes and more
I2b2 datasets	Clinical notes for tasks like Relationship Extraction, Negation detection, Temporal relations etc.
MeDal Corpus	Dataset for Abbreviation Disambiguation
PubMed PICO Element Detection	Dataset for Participant(P), Intervention(I), Comparison(C), Outcome(O)
UFAL Medical Corpus	Parallel corpus for translation of medical texts
.... and many more.	

NOTE: We use **CodiEsp 2020 Corpus** for our mini research project, a freely available dataset.

References: MIMIC, i2b2, MeDal, PubMed, UFAL, CodiEsp

What is Clinical Coding?

- It involves translating clinical texts into a set of codes as defined in International standards.

- The major amongst them are: **ICD10** and **CPT**.

- Formally, we can define clinical coding in NLP to be a **Multi-Label Classification** task

- Use of NLP in Clinical Coding will help in automatic extraction of codes. It will save time and can even outperform human coders.

Motivation and Related Research

Reasons we need Clinical Coding:

1. Standardization (language independent)
2. Statistical analysis and hence good decision making
3. Intricately related to billing and insurance.

Reasons we need automated NLP systems for Clinical Coding

1. It reduces human effort
2. Accurately built systems can outperform humans
3. Speed of processing medical records

Several NLP systems have been proposed to solve the task of Clinical Coding. Next, we will look briefly at the three major types of models used in Clinical Coding.

Example of a Clinical Coding Dataset (from)

Input: we describe the case of a 37-year-old man with a previous active life who complained of osteoarticular pain of variable location in the last month and fever in the last week with peaks (morning and evening) of 40 c……

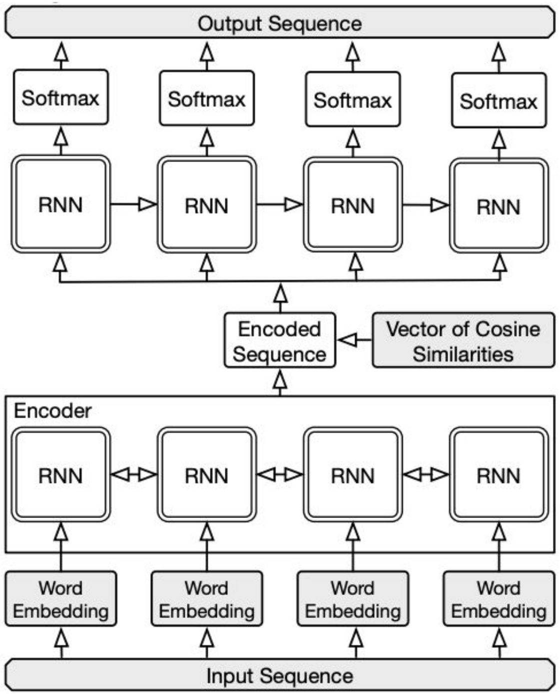
Target: n44.8, z20.818, r60.9, r52, a23.9, i83.90, i87.8, r50.9, n45.3, m25.50 [A list of codes]

One-hot encoded vector of size N,
where N is number of uniques codes possible in the data

Seq2seq Architecture for ICD-10 coding

- Model details-
  - ➢ Encoder: A layer of bidirectional LSTMs
  - ➢ Decoder: A layer of left to right LSTMs
  - ➢ A Cosine Similarity vector concatenated to encoded state
- Dataset- **CepiDC** (Cause of Death Corpus): It contains causes of death, as reported by Physicians in free text description format.
- Results Obtained: Compared the model (Precision, Recall, F1 measure) with the average results of the competition.

	Precision	Recall	F1 measure
Proposed Model	0.891	0.812	0.850
Avg of other models	0.670	0.582	0.622

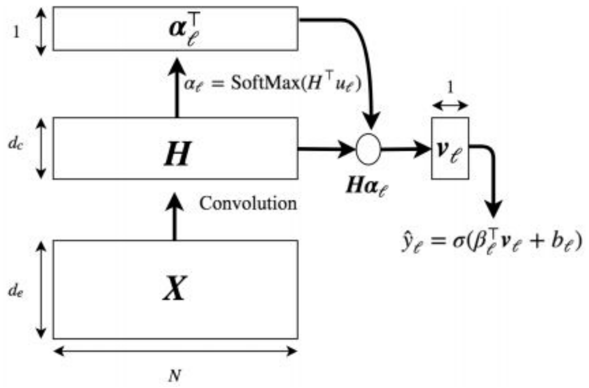


References: Paper

CNNs for ICD-10 Coding

- Model description-
  - ➢ Built on top of the **CAML** model, which used CNNs to build document representations, used for prediction of ICD codes
  - ➢ Utilized multi filter and multi residual CNN layers to capture text patterns of varying lengths.
- Dataset used- **MIMIC-II**, **MIMIC-III** datasets
- Results-
  - ➢ The paper compared the model with the earlier SoTA (**CAML** and **DR-CAML**)
  - ➢ A part of the results is summarized below. It shows comparisons on the micro and macro averaged F1 scores, and the precision @8, 15

CAML architecture [2]



Results

	Macro F1	Micro F1	P@5
CAML	0.532	0.614	0.609
MultiRes CNN	0.606	0.670	0.641

Note: P@K: Proportion of the correctly predicted labels, in the top-k predicted labels

References: [1] MultiResCNN paper, [2] CAML paper

Transformers

BertXML:

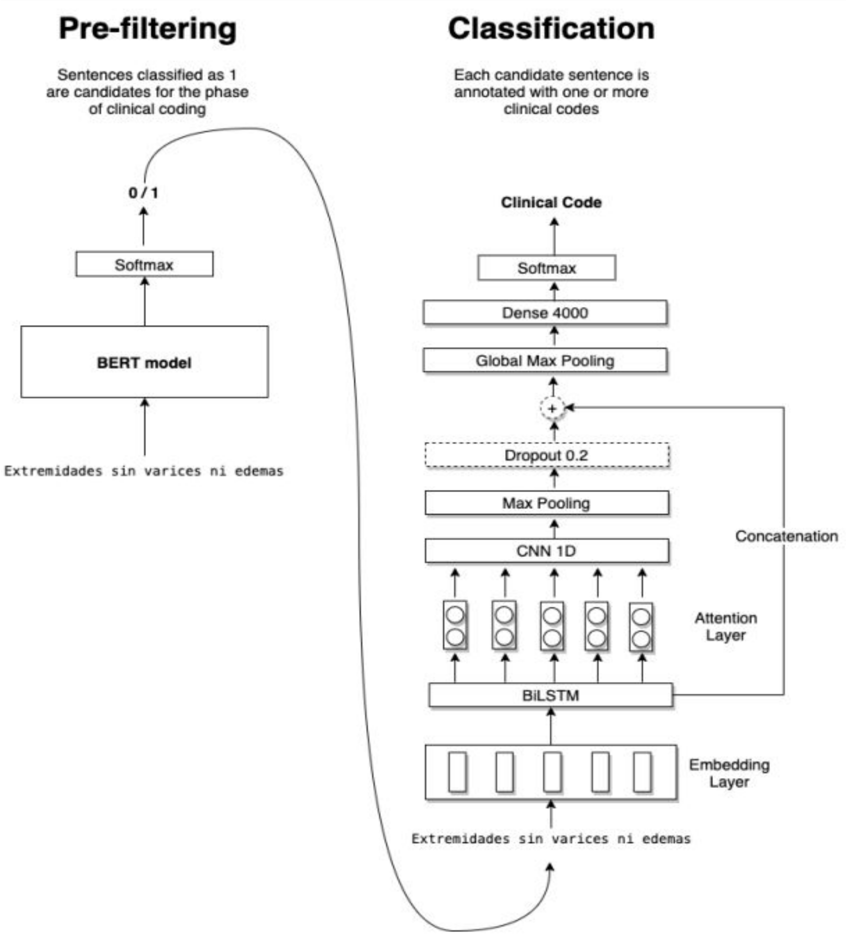
- Trained a Bert model on EHR notes from scratch
- Doubled the input size of as compared BioBert to allow a complete documents as input.
- Learnt vocabulary better suited for EHR tasks, thus outperforming other methods.
- **SOTA**: The model outperformed BioBert and ClinicalBert in ICD code classification.

Bert with LSTMs and CNNs:

- Used Bert as a prefiltering step
- Output indicated probability of sentence containing references to clinical code
- High probability sentences were passed into the Classifier
- Different classifiers composed of LSTMs, CNNs with(out) attention were used
- Architecture shown on the next slide

References: BertXML, Bert with LSTMs and CNNs

BERT with LSTMs/ CNNs



References: Paper

Our Research and Experimentation

Dataset - CodiEsp Corpus

Corpus: Information and Processing

Text Data Analysis:

- The CodiEsp data consists of annotated clinical documents.
- The documents are originally in Spanish; we also have machine translated English texts.
- The annotations are of two types: Procedural (P) and Diagnostic (D) [Two Subtasks].
- The distribution of number of documents is as follows:
  - Training: 500 | Development: 250 | Test: 250
- The documents are long with following statistics for tokenized sequence length:
  - Mean: 342.63 | Median: 318.5 | Standard deviation: 161.12

Data Pre-Processing:

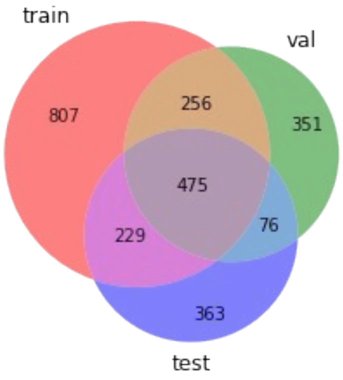
- Text is converted to lowercase.
- All stopwords are removed using nltk library.

References: Dataset

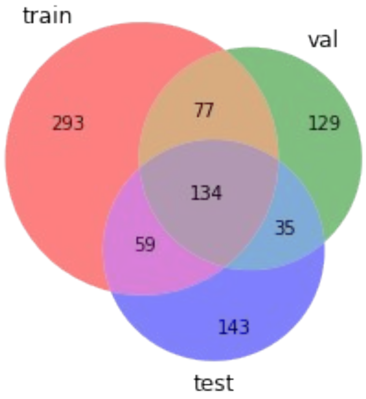
ICD Codes Analysis:

- The corpus has 18,483 annotated codes, of which, 3427 are unique.
- **Interestingly, not all ICD codes are present in the training dataset.** The label class distribution is shown in the diagrams below:

Number of ICD10 Codes in train, val, test sets and their overlap for D subtask



Number of ICD10 Codes in train, val, test sets and their overlap for P subtask

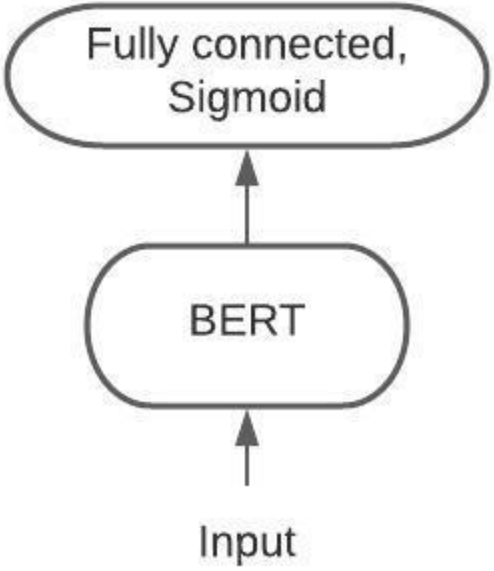


Baseline Models

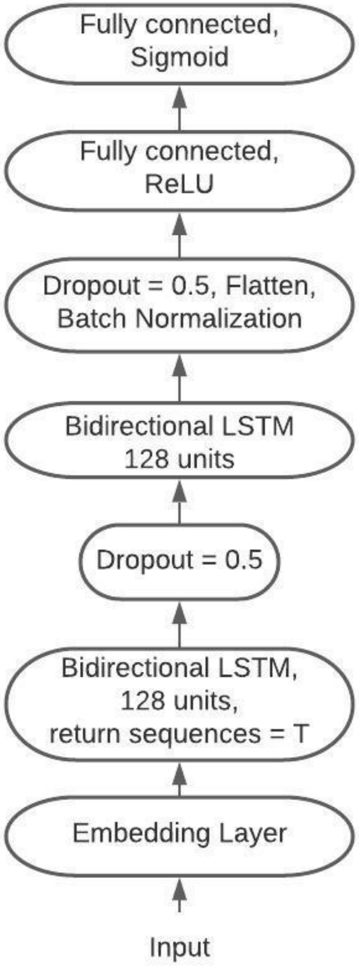
Model Name [Text fed using Bag of Words method (performed better than TF-IDF)]	Subtask - D Test Set		Subtask - P Test Set	
	Hamming Score	**F1 Score**	**Hamming Score**	**F1 Score**
Multinomial Naive Bayesian	0.0227	0.0015	0.0311	0.0016
XGBoost Classifier	0.2091	0.0245	0.1936	0.0206
Support Vector Classifier	0.0056	0.0002	0.0031	0.0003
Random Forest Classifier	0.0081	0.0005	0.0061	0.0003
Logistic Regression	0.0651	0.0078	0.0771	0.0068
AdaBoost Classifier	**0.2848**	**0.0722**	**0.2012**	**0.0344**

References: sklearn

Model Architectures



BERT Transformer



2 layer Bi-LSTM model

References: HuggingFace

DNN and Transformer Model Results

Model Name [Text is pre-processed as explained on Slide-12]	Subtask - D Test Set		Subtask - P Test Set	
	Hamming Score	F1 Score	Hamming Score	F1 Score
Feed Forward NN	0.0227	0.0015	0.0311	0.0016
2 layer Bi-LSTM Model	0.0590	0.0008	-	-
BERT Transformer	0.0054	0.0082	0.0070	0.0085
Clinical-BERT Transformer	0.0061	0.0077	-	-
Bio-BERT Transformer	0.0058	0.0079	-	-

NOTE: In the competition, the winner achieved F1 score equal to **0.687** in task D and **0.522** in task P.

References: HuggingFace

Future Work and Ideas

- Due to very small dataset, the traditional NLP techniques outperform Deep Neural methods (LSTMs and Transformers).

- A logical direction is to explore F**ew Shot Learning** techniques.

- After getting access to **MIMIC-III** dataset, we can test our various architectures on it.

- We believe the results will be significantly better as MIMIC dataset is quite large.

Contribution of Team Members

	Work	Saksham Gupta	Sathvik Bhagavan	Harshit Gupta
Code Implementation	Data Preprocessing	✅	✅	
	Baseline Models	✅		✅
	Deep Learning based models		✅	
	Transformer based models	✅	✅	
Report & Presentation	Introduction and Conclusion	✅	✅	✅
	Related Research To Clinical Coding			✅
	Experimental models and Future Direction	✅	✅	✅

Thank You!

Suggestions and Feedback?

References

- Disambiguation of Biomedical Abbreviations
- Biomedical named entity recognition using deep neural networks with contextual information
- KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks
- ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network
- Explainable Prediction of Medical Codes from Clinical Text
- BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining
- A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020
- Biomedical Named Entity Recognition at Scale
- Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review

Continued...

- Entity recognition from clinical texts via recurrent neural network
- Medical Entity Recognition: A Comparison of Semantic and Statistical Methods
- Clinical Named Entity Recognition Using Deep Learning Models
- Embedding Strategies for Specialized Domains: Application to Clinical Entity Recognition
- AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification
- Named Entity Recognition in Biomedical Texts using an HMM Model