Project Proposal - W266 Summer 2017

Zenobia Liendo (Group 1), Guillaume De Roo (Group 2), Amit Karmakar (Group 1)

Introduction

Electronic Health Records (EHRs) bring an unprecedented amount of patient health information in both structured (20%) and unstructured formats (80%), with key health information on clinical notes. Applying NLP on the unstructured data, we can extract information and insights to improve clinical diagnosis and the understanding of health and diseases in general.

Objective

The Objective of this project is to apply Natural Language Processing (NLP) techniques on patient EHRs to assign the corresponding ICD-9 label (International Classification of Diseases).

Dataset

MIMIC-III is an openly available dataset developed by the MIT Lab for Computational Physiology, it contains de-identified health records from about 53,000 patients, who stayed in critical care units between 2001 and 2012. MIMIC-III includes several types of clinical notes, including discharge summaries (n = 52,746) and nursing notes (n=812,128), and labelled by ICD-9 codes.

Algorithms and Metrics

From our literature review, the approach followed for ICD-9 classification is generally:

- Preprocessing is a combination of
    - Tokenization, using Ctakes
    - POS tagging, shallow parse tree for grammatical structure
    - Named-entity recognition(NER) to get concept unique identifiers (CUIs) in UMLS (Unified Medical Language System)
    - TF-IDF rather than binary scores
- Baseline is often a Logistic Regression or Random Forest modeling
- More advanced algorithms include a convolutional neural network (CNN) [1], or a RNN/LSTM with a single layer [2]

We are planning to review these algorithms and try to improve them by testing variants on different parts of the process, e.g. embedding vectors based on NER, logistic regression with cross validation and regularization, multilayer LSTM.

We will use precision, recall and F1 as metrics for performance improvement.

References

[1] Comparing Rule-Based and Deep Learning Models for Patient Phenotyping
(MIT, Harvard, Tufts, Washington University School of Medicine, etc)
Patrick D Tyler, Leo Anthony Celi. March 25, 2017

[2] Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records
Priyanka Nigam, Department of Computer Science, Stanford University
June 2016

[3] Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries
Jonathan Brauer. Institut fur Informatik. January 25, 2017

[4] Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES)
Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, Christopher G Chute, 2010