ICD Code	Precision	Recall	F1 Score	Support
5990:UTI	0.82	0.75	0.78	1057
2875:Thrombo	0.60	0.39	0.47	493
486:Pneumonia	0.59	0.49	0.53	756
51881:Acute Resp	0.70	0.71	0.71	1173
2859:Anemia	0.49	0.42	0.45	884
4275:Cardiac Arrest	0.63	0.37	0.47	209
49390:Asthma	0.72	0.47	0.57	354
7140:Rh Arthritis	0.78	0.28	0.41	112
4160:Pulm Hyper	0.06	0.02	0.03	56
avg / total	0.65	0.56	0.60	5094

ICD Code	First	Second	Third
5990:UTI	UTI	Urinary Tract Infection	Urinary Tract
2875:Thrombo	Thrombocytopenia	HIT	Antibody
486:Pneumonia	Pneumonia	HCAP	Hospital Acquired Pneumonia
51881:Acute Resp	Respiratory Failure	Intubation	Extubation
2859:Anemia	Anemia	Normocytic	Dilution
4275:Cardiac Arrest	Arrest	PEA	CPR
49390:Asthma	Asthma	Albuterol	Singulair
7140:Rh Arthritis	Rheumatoid	Rheumatoid Arthritis	Arthritis
4160:Pulm Hyper	Moderate	Contrast	Although

Medical Record Understanding

Justin Fu
Daniel Thirman

Background

- Medical records are terse notes recorded by a doctor. Includes information such as:
  - Past medical history. Ex. "Patient is a 55 yo w/ T1D. Prev received R BKA ..."
  - Hospital course. Ex. "Presented hypertension and UTI. He/she was treated with cipro, levofloxacin. Vitals 101, 100, 130/90."
  - Diagnoses, medications prescribed, discharge condition.
- Records are labeled with ICD codes, a comprehensive system which labels symptoms, diseases, medical procedures performed, etc.
  - (ICD10) W55.42XA: Struck by a pig, initial encounter
  - (ICD9) V1582: Personal history of tobacco use
  - (ICD9) 2764: Insertion of palatal implant
- We used the MIMIC III dataset, which has emergency room records labeled with codes by doctors

Challenges

- Recent work has not produced usable results (~15-50 F1, depending on # of codes). Many features are hand-engineered rules for specific codes.
- Large # of labels (About 17,000 total ICD9 codes)
- Label distribution is very lopsided, many codes have little training support.
  - 80% of labels in the top 500 codes (out of ~7000 in our dataset)
  - About 1500 of these codes have 1 training example
- Labels are *extremely* noisy
  - One study showed that for Alzheimer's disease, 65% of notes were not coded (false negative) and 5% coded without evidence in note. Similar numbers for other diseases (Parkinson's 75%/4%, Breast cancer 31%/39%)
- Abbreviations, synonyms, and medical jargon are widespread
  - Very domain-specific language, so we probably can't use generic models trained on abundant sources of data such as Wikipedia.

Baseline Model

- We trained a bag-of-words/phrases model with a multi-label linear classifier (logistic regression + L1 regularization).
  - Chosen for ease of interpretability.
  - Short phrases were often more descriptive than individual words (ex. "urinary tract infection")
- Results: 10 labels: **60%** F1 Score // 30 labels: **45%** F1 score
- Learned features were generally quite relevant. Top features for "asthma":
  - asthma, albuterol, singulair, inhalation, flovent, reactive airway disease, inhaler
- Revealed
  - **Sparsity** - Relevant information is typically contained in 1-2 sentences in the entire document
  - **Multi-word understanding** - Some labels are implied by test results, or vaguely described.
  - We performed notably worse on some codes with little training support

Full Model

- To tackle the first two problems, we have chosen to implement an simple attention-based model.
- **Sparsity**
  - The document is segmented, and each segment is scored using either an RNN or a bag-of-words model. We use either the top scoring segment (non-differentiable, but more computationally efficient) or use the scores as a weight for the classification model
- **Multi-word understanding**
  - The segment is encoded as word vectors, and we train either an RNN to classify the code from the segment.

Conclusions/Future Work

- Problems left to tackle after this project
  - Exploiting hierarchy of codes
    - This has been explored in some recent work, and some ideas can be incorporated into our models
  - Using world knowledge (ex. wikipedia, code descriptions) to predict codes with little training support.
    - However, it is not clear whether there is significant practical benefit to predict rare codes accurately
  - Noisy/incomplete labels. We hope to explore bootstrapping methods and semi-supervised learning to deal with this.
    - This is a serious, and known problem.