Electronic Health Records Processing

Whys and Hows

Krištof Anetta
xanetta@fi.muni.cz

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

December 2, 2020
What Are Electronic Health Records?

All data collected about a patient in machine-processable format (no OCR, except as a separate problem)

- medical history, diagnoses, medications, treatment plans, immunization dates, allergies
- radiology images
- laboratory and test results
- vital signs during hospitalizations
Introduction

Why Process EHRs?

Core motivation

The global accumulation of billions of EHRs contains latent knowledge about medical science and global health

- Harnessing this data and statistically processing it may bring about a paradigm shift in how medical scientific studies are done
- Discovering patterns in the data using deep learning has the potential to transform expert systems and predictive medicine
Who Benefits?

- Populations
  - Statistical information for population-based studies
  - Comparison of populations
- Individual patients
  - Automatic identification of risk groups
  - Prediction in general
  - Outlier detection, error notification
Healthcare Standardization

- **FHIR (Fast Healthcare Interoperability Resources)**
  - Standard describing data formats
  - API

- **SNOMED CT (Clinical Terms)**
  - The most comprehensive clinical healthcare terminology in the world [3]
  - Multilingual

- **UMLS (Unified Medical Language System)**
  - Compendium of many controlled vocabularies in the biomedical sciences
  - Metathesaurus, semantic network
Structured Data

- Temporal order
- Numeric values from measurements
- Categorical variables (nominal/ordinal)

Use of structured data

Pros:
- Limited number of observed variables
- Ready for deep learning

Cons:
- Only a small subset of medical procedures generates it
- Crucial context and demographics are hard to structure
**Unstructured Data**

- Free-form text entered by doctors (estimated to be 85% of all patient data)
- Images

---

**Use of unstructured data**

**Pros:**

- Large amounts available
- Contains key context with the most detail: interview, admission examination, symptoms, and recommendations

**Cons:**

- Large number of dimensions, sparsely filled
- Difficult annotation process
## Current Approaches


<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>216,221 patients</td>
</tr>
<tr>
<td>Timelines in FHIR standard</td>
</tr>
<tr>
<td>46 billion data points</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
</tr>
<tr>
<td>Attention-based TANN</td>
</tr>
<tr>
<td>NN with boosted time-based decision stumps</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy (in AUROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-hospital mortality: 0.93–0.94</td>
</tr>
<tr>
<td>30-day unplanned readmission: 0.75–0.76</td>
</tr>
<tr>
<td>Prolonged length of stay: 0.85–0.86</td>
</tr>
<tr>
<td>All of a patient’s final discharge diagnoses: 0.90</td>
</tr>
</tbody>
</table>
Figure: Occurrence of studies on SemanticScholar with "deep learning" AND "electronic health records" OR "electronic medical records" [2]
Free-Form Medical Text Analysis

Central Tasks

What we need to do with EHR text

- Entity recognition
  - Symptoms, examination findings, diagnoses, medications, measurements
- Relation extraction
Free-Form Medical Text Analysis

Challenges for NLP

Problematic text characteristics

- Latin elements in bi-/multilingual text
- Incomplete syntax, often relying on specific conventions (omission, telegraphic style)
- Errors (fast typing) and sloppy punctuation and capitalization
- Abbreviations
- Shifted meaning of ordinary words (categorical variables)
- Numbers
- Language- and doctor-specific conventions
Free-Form Medical Text Analysis

Existing Frameworks

Apache cTAKES: clinical Text Analysis and Knowledge Extraction System

- NLP system that extracts clinical information from the unstructured text of EHRs
- built with OpenNLP and UIMA (Unstructured Information Management Architecture framework)

MetaMap (for biomedical text in general)
- Tool that links concepts in a text to the UMLS Metathesaurus
Free-Form Medical Text Analysis

State of the Art

Deep learning from unstructured notes
- Youth depression, prostate cancer, smoking, adverse drug events

<table>
<thead>
<tr>
<th>Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
</tr>
<tr>
<td>RNN</td>
</tr>
<tr>
<td>LSTM</td>
</tr>
<tr>
<td>BERT attempts</td>
</tr>
<tr>
<td>CRF</td>
</tr>
</tbody>
</table>
Free-Form Medical Text Analysis
Slavic Languages

- Slower adoption of interoperable technologies
- Compared to state of the art in English, there is a lack of resources at every level (hospital software, medical ontologies and terminologies for processing, large specialized training corpora, ready-made NLP tools)
- GDPR
"Do patients with similar conditions get similar diagnoses?"

Clustering of patient visits based on word embeddings

Corpus of 100,000 patient visits
pl_ehr_cardio: New Polish Dataset

Basic information

- Cardiology (specialization advantage)
- 50,465 patient hospitalizations
- 2003 to 2020
- Includes ICD-10 diagnosis codes
- Separate hospitalizations only - no longitudinal data on individual patients
New Polish Dataset

**pl_ehr_cardio: Characteristics**

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>34,315,153</td>
</tr>
<tr>
<td>Words</td>
<td>23,831,785</td>
</tr>
<tr>
<td>Sentences</td>
<td>2,583,087</td>
</tr>
<tr>
<td>Average sentence length</td>
<td>9.226</td>
</tr>
<tr>
<td>Unique word forms</td>
<td>160,042</td>
</tr>
<tr>
<td>Unique word forms (lowercase)</td>
<td>141,685</td>
</tr>
<tr>
<td>Unique lemmas</td>
<td>124,727</td>
</tr>
<tr>
<td>Unique lemmas (lowercase)</td>
<td>114,556</td>
</tr>
</tbody>
</table>
pl_ehr_cardio: Characteristics
pl_ehr_cardio: Characteristics

![Average word count of health record parts](image)

- Arrival - reasons, medical history
- Arrival - physical examination
- Discharge - summary of hospitalization, results
- Discharge - recommendations, medication
New Polish Dataset

pl_ehr_cardio: Characteristics

![Bar chart showing the most common ICD-10 discharge codes related to cardiovascular diseases.]
New Polish Dataset

pl_ehr_cardio: Characteristics

Distribution of diagnoses across time

- I25.0 - Atherosclerotic cardiovascular disease, so described
- I20.0 - Unstable angina
- I21.4 - Acute subendocardial myocardial infarction
- I50.0 - Congestive heart failure
- I25.1 - Atherosclerotic heart disease

Percentage of total

Years: 2003 - 2019
New Polish Dataset

pl_ehr_cardio: Demo

Initial processing

- spaCy NER
- pl_core_news_lg, biggest statistical model for Polish
  - 500k keys, 500k unique vectors (300 dimensions)
  - NER F-score: 85.67
- Results clearly demonstrate the specificity of EHR language
Bibliography


Bibliography II

Thank You for Your Attention!
MUNI
FACULTY
OF INFORMATICS