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

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

Rajkomar et al., 2018, *Nature* [1]

- Data
  - 216,221 patients
  - Timelines in FHIR standard
  - 46 billion data points
- Architectures
  - LSTM
  - Attention-based TANN
  - NN with boosted time-based decision stumps
- Accuracy (in AUROC)
  - In-hospital mortality: 0.93–0.94
  - 30-day unplanned readmission: 0.75–0.76
  - Prolonged length of stay: 0.85–0.86
  - All of a patient's final discharge diagnoses: 0.90

Current Approaches

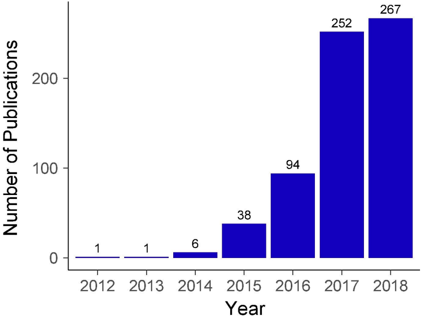


Figure: Occurrence of studies on SemanticScholar with "deep learning" AND "electronic health records" OR "electronic medical records" [2]

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

Architectures

- CNN
- RNN
- LSTM
- BERT attempts
- CRF

Free-Form Medical Text Analysis

Slavic Languages

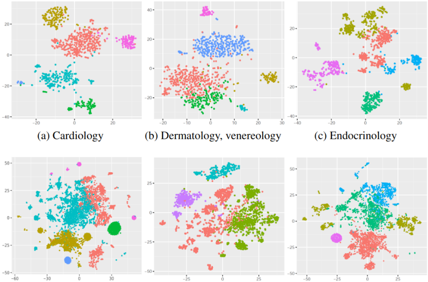
- Slower adoption of interoperable technologies
- Compared to state of the art in English, there is a lack of resources at every level (hospital software, medical ontologies and terminologies for processing, large specialized training corpora, ready-made NLP tools)
- GDPR

Free-Form Medical Text Analysis

Polish

Dobrakowski et al., 2019 [4]

- "Do patients with similar conditions get similar diagnoses?"
- Clustering of patient visits based on word embeddings
- Corpus of 100,000 patient visits



(a) Cardiology (b) Dermatology, venereology (c) Endocrinology

(d) Family medicine (e) Gynecology (f) Internal medicine

pl_ehr_cardio: New Polish Dataset

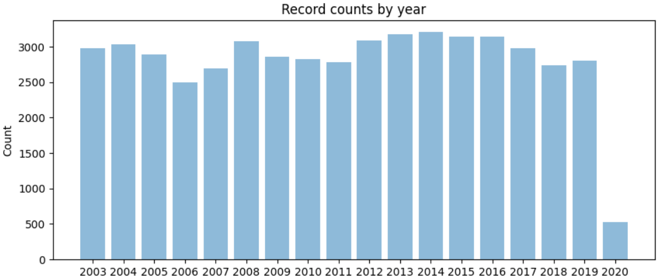
Basic information

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

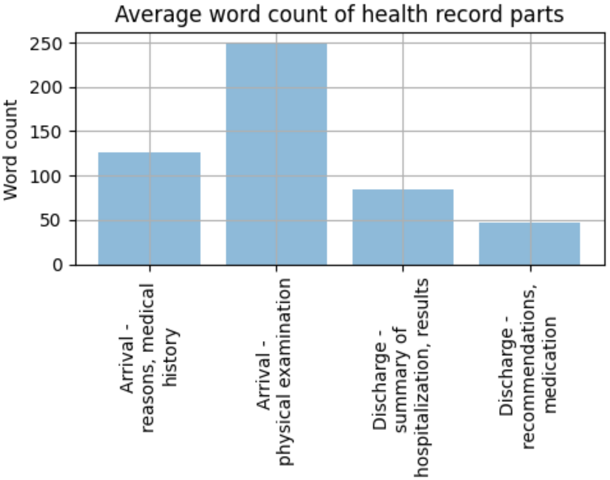
pl_ehr_cardio: Characteristics

Tokens	34,315,153
Words	23,831,785
Sentences	2,583,087
Average sentence length	9.226
Unique word forms	160,042
Unique word forms (lowercase)	141,685
Unique lemmas	124,727
Unique lemmas (lowercase)	114,556

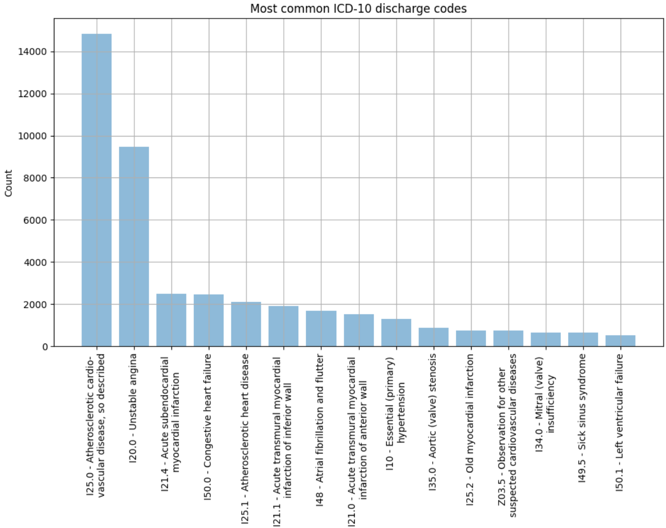
pl_ehr_cardio: Characteristics



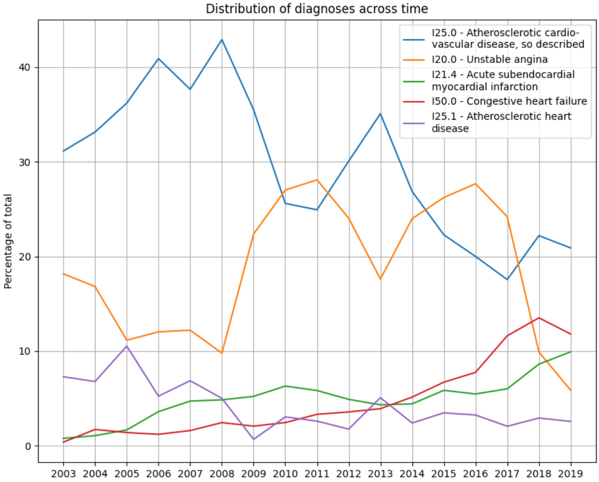
pl_ehr_cardio: Characteristics



pl_ehr_cardio: Characteristics



pl_ehr_cardio: Characteristics



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 I

[1] Rajkomar A., Oren E., and Chen K. et al. “Scalable and accurate deep learning with electronic health records”. In: *npj Digital Med* 1.18 (2018).

[2] Jose Roberto Ayala Solares et al. “Deep learning for electronic health records: A comparative review of multiple deep neural architectures”. In: *Journal of Biomedical Informatics* 101 (2020), p. 103337. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2019.103337. URL: http://www.sciencedirect.com/science/article/pii/S1532046419302564.

[3] Tim Benson. *Principles of Health Interoperability HL7 and SNOMED*. Springer, 2012.

Bibliography II

[4] Adam Gabriel Dobrakowski et al. “Clustering of Medical Free-Text Records Based on Word Embeddings”. In: *CoRR* abs/1907.04152 (2019). arXiv: 1907.04152. URL: http://arxiv.org/abs/1907.04152.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS