Analyzing Structured and Unstructured Data in Electronic Health Records

Jing Zhao & Aron Henriksson
Department of Computer and Systems Sciences (DSV)
Stockholm University
Electronic Health Records

• Administrative and clinical patient data collected systematically and routinely
• Capture and integrate data on all aspects of care over time
• Comprise various data types
• Sensitive data
Electronic Health Records

- Administrative and clinical patient data collected systematically and routinely
- Capture and integrate data on all aspects of care over time
- Comprise various data types
- Sensitive data

- Health records from Karolinska University Hospital
  - 5 years: 2006-2010
  - A large variety of clinics in the Stockholm region
  - 1M patients
  - 10K diagnosis codes, 3M diagnoses
  - 1.3K drugs, 2M prescriptions
  - 700 clinical measurement types, 15M measurements
  - 4M token types, 1.5B tokens
Benefits & Challenges of Using EHRs

- Large amounts of longitudinal observations
- Holistic perspective of patients’ clinical conditions
- Systematically collected and archived data
- Clinical notes complementing structured information

- Under-reporting (especially under-coding) still exists
- High dimensionality and sparsity
- Integrating various types of data
- Clinical notes noisy
Clinical Text - A Peculiar Genre

- A flexible communication tool, documenting healthcare across
  - space — between clinicians
  - time — “memos”

- Written under serious time constraints
Clinical Text - A Peculiar Genre

- A flexible communication tool, documenting healthcare across
  - space — between clinicians
  - time — “memos”
- Written under serious time constraints

**Clinical Text - A Peculiar Genre**

- **telegraphic, ungrammatical sentences**
  - Status:
    - Heart: regular rythm 80/min. BP: 120/70.
    - Lungs: normal. Joint destructions in PIP and DIP in all digits.
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.

- **non-standard shorthand**
  - Status:
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.

- **misspellings**
  - Status:
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.

- **synonyms**
  - Status:
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.

- **multiword terms**
  - Status:
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.

- **negation**
  - Status:
  - Assessment:
    - 75-year-old woman with sivere rheumatoid arthritis. Needs further rehabilitation with physiotherapy. Will see our occ therapist tom.
Clinical Text - A Peculiar Genre

<table>
<thead>
<tr>
<th></th>
<th>Types</th>
<th>Tokens</th>
<th>Type/Token Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Corpus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUC 3.0</td>
<td>0.1M</td>
<td>1M</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Medical Corpora</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vårdguiden</td>
<td>0.05M</td>
<td>2.8M</td>
<td>0.0160</td>
</tr>
<tr>
<td>Läkartidningen</td>
<td>0.4M</td>
<td>20M</td>
<td>0.0230</td>
</tr>
<tr>
<td>Stockholm EPR Corpus</td>
<td>3.9M</td>
<td>1582M</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Stockholm EPR Corpus: **72% verbless** and **10% subjectless** sentences!

- **11%** in SUC 3.0
- **15%** in Läkartidningen
- **0.4%** in SUC 3.0
- **0.05%** in Läkartidningen
Clinical Language Processing: Challenges

NLP Systems

- part-of-speech taggers
- syntactic parsers
- named entity recognizers
- co-reference resolvers

SUC 3.0

Stockholm EPR Corpus

genereal domain

clinical domain
Clinical Language Processing: Challenges

- Text data, modeled as a *bag of words*, is very high-dimensional and sparse
  - 4M types → 4M features?
  - Exacerbated by lexical variability of concepts

- Alternatives to the bag-of-words approach?

---

**synonymy**

**homonymy**

---

<table>
<thead>
<tr>
<th></th>
<th>patient</th>
<th>patient</th>
<th>pat</th>
<th>pathological</th>
<th>fever</th>
<th>pain</th>
<th>cold</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>Ex2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Ex3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
Distributed Word Representations

- Model lexical semantics based on co-occurrence information
  - words that appear in similar contexts tend to have similar meanings (Zellig Harris, John R. Firth, Ludwig Wittgenstein, …)

<table>
<thead>
<tr>
<th></th>
<th>itchy</th>
<th>asthma</th>
<th>chemotherapy</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>allergy</td>
<td>87</td>
<td>44</td>
<td>2</td>
<td>…</td>
</tr>
<tr>
<td>hypersensitivity</td>
<td>46</td>
<td>55</td>
<td>4</td>
<td>…</td>
</tr>
<tr>
<td>cancer</td>
<td>0</td>
<td>8</td>
<td>84</td>
<td>…</td>
</tr>
</tbody>
</table>

- Vector representations in reduced-dimensional semantic space (d << vocabulary size)
Random Indexing

- Efficient and scalable algorithm for creating semantic spaces
  - No dimensionality reduction applied to term-term matrix
  - Incrementally obtains co-occurrence information in *pre-reduced* semantic vectors

- Two types of vectors (of dimensionality $d \ll$ vocabulary size):
  - index vectors — used only in construction phase
  - semantic vectors — word meaning representations; make up semantic space

- Each unique word $w_j$ assigned an index vector $\mathbf{IV}_j$ and a semantic vector $\mathbf{SV}_j$
  - $\mathbf{IV}$: sparse, randomly assigned a small number of 1s and -1s
  - $\mathbf{SV}$: sum of IVs of the words that $w_j$ co-occurs with within a given window size
**Data Science Workshop at Stockholm University: Jing Zhao, Aron Henriksson - December 4, 2014**

Analyzing Structured and Unstructured Data in Electronic Health Records

- **Patient experiences a tremor in right hand from time to time**
- **2+2 context window**

\[
\langle 0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Applications of Distributional Semantics

1. Medical terminology development
   - handling lexical variability by identifying synonyms
   - distributionally similar words are candidate synonyms

2. Assignment of diagnosis codes
   - semantic space comprising diagnosis codes and words
   - discover distributional similarities between words used in notes and diagnosis codes assigned to them
   - recommend codes to assign based on words used in clinical notes

3. Creating features for named entity recognition
   - exploit large amounts of unlabeled text data to support learning with small (labeled) datasets


Data Science Workshop at Stockholm University: Jing Zhao, Aron Henriksson - December 4, 2014
Analyzing Structured and Unstructured Data in Electronic Health Records
Prototype Vectors for Named Entity Recognition

- Task: identify protected health information (PHI) in clinical text
- Proposed method presupposes the existence of two resources:
  - an annotated (named entity) corpus
  - a (large) unannotated corpus


Data Science Workshop at Stockholm University: Jing Zhao, Aron Henriksson - December 4, 2014
Analyzing Structured and Unstructured Data in Electronic Health Records
Prototype Vectors for Named Entity Recognition


Data Science Workshop at Stockholm University: Jing Zhao, Aron Henriksson - December 4, 2014
Analyzing Structured and Unstructured Data in Electronic Health Records
Prototype Vectors for Named Entity Recognition

- **Prototype vector**: abstract representation of a target (named entity) class
- Centroid of annotated instances’ semantic vectors
  - words annotated as belonging to a particular class
- Centroid defined as column-wise *median* values
  - reduce impact of noise

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>d7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>-14.5</td>
<td>12.0</td>
<td>0</td>
<td>110.1</td>
<td>-25.4</td>
<td>120.0</td>
<td>-2.2</td>
<td>...</td>
</tr>
<tr>
<td>Anna</td>
<td>-10.7</td>
<td>0</td>
<td>4.0</td>
<td>75.4</td>
<td>-5.5</td>
<td>12.0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Peter</td>
<td>40.4</td>
<td>5.0</td>
<td>10.2</td>
<td>50.4</td>
<td>-5.0</td>
<td>42.5</td>
<td>-12.0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototype</td>
<td>-10.7</td>
<td>5.0</td>
<td>4.0</td>
<td>75.4</td>
<td>-5.5</td>
<td>42.0</td>
<td>-2.2</td>
<td></td>
</tr>
</tbody>
</table>

Prototype Vectors for Named Entity Recognition

- Use prototype vectors to generate features for all instances in dataset

- Binary features: True if cosine similarity between prototype vector and semantic vector > threshold

- Threshold based on distances between labeled instances’ semantic vectors and their corresponding prototype vector

- Set to maximize $F_\beta$-score on training data
  - Positive examples: instances belonging to one named entity
  - Negative examples: instances belonging to all other named entities

\[
\arg\max_{t \in V} \left( (1 + \beta^2) \frac{P(t) \cdot R(t)}{(\beta^2 \cdot P(t)) + R(t)} \right),
\]

\[V = (0, 0.0001, 0.0002, ..., 1)\]
Prototype Vectors for Named Entity Recognition

- Comparison of two feature sets provided to the learning algorithm
  - Baseline: set of standard orthographic and syntactic features
  - Baseline features + distributional semantic features
- Significant improvements obtained with generated features
- Further improvements obtained when combining multiple semantic spaces created with different models

Analyzing Structured Data for Pharmacovigilance

• Traditional data sources
  ‣ Pre-marketing surveillance
    ✦ Clinical trials: Limited sample size and observation period
  ‣ Post-marketing surveillance
    ✦ Spontaneous reports: Largely under-reported, Reliability and compliance, No exposure information

• Emerging alternatives for post-marketing surveillance
  ‣ Electronic health records: High recording rate, Patient medical history, Exposure information recorded
  ‣ Social media
Using EHRs is not Unproblematic!

|     | ADE | Diag1 | Drug1 | CM1 | Diag2 | Drug2 | CM2 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>yes</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>5.6</td>
<td>...</td>
</tr>
<tr>
<td>P2</td>
<td>yes</td>
<td>NA</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>...</td>
</tr>
<tr>
<td>P3</td>
<td>no</td>
<td>0</td>
<td>?</td>
<td>30</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>...</td>
</tr>
</tbody>
</table>
|       | ADE | Diag1 | Drug1 | CM1 | Diag2 | Drug2 | CM2 |...
|-------|-----|-------|-------|-----|-------|-------|-----|-----|
| P1    | yes | 1     | 0     | NA  | 0     | 1     | 5.6 |...
| P2    | yes | NA    | 0     | 12  | 1     | 0     | ?   |...
<p>| P3    | no  | 0     | ?     | 30  | 1     | NA    | NA  |...|</p>
<table>
<thead>
<tr>
<th></th>
<th>ADE</th>
<th>Diag1</th>
<th>Drug1</th>
<th>CM1</th>
<th>Diag2</th>
<th>Drug2</th>
<th>CM2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>yes</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>5.6</td>
<td>...</td>
</tr>
<tr>
<td>P2</td>
<td>yes</td>
<td>NA</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>...</td>
</tr>
<tr>
<td>P3</td>
<td>no</td>
<td>0</td>
<td>?</td>
<td>12</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADE</td>
<td>Diag1</td>
<td>Drug1</td>
<td>CM1</td>
<td>Diag2</td>
<td>Drug2</td>
<td>CM2</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>P1</td>
<td>yes</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>yes</td>
<td>NA</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>no</td>
<td>0</td>
<td>?</td>
<td>30</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

Drug1 was prescribed to P3 more than once

CM2 was taken P2 more than once

Drug1 was prescribed to P3 more than once

CM2 was taken P2 more than once
ADE Detection Using Structured EHRs

- Adapting disproportionality methods for detecting ADE signals
- Using machine learning algorithms to detect missing ADE codes in patients’ medical history
Adapting Disproportionality Methods to EHR Data

Disproportionality methods:
PRR, ROR, BCPNN, GPS

\[ \text{PRR} = \frac{a}{a+b} / \frac{c}{c+d} \]

Event level: count drug-event pairs

Patient level: count distinct patients who experienced the same drug-event pair

Zhao, J., Karlsson, I., Asker, L. and Boström, H. Applying Methods for Signal Detection in Spontaneous Reports to Electronic Patient Records. In 19th Knowledge Discovery and Data Mining (KDD) Conference’s Workshop on Data Mining for Healthcare (DMH), August 11-14, 2013, Chicago, USA.
Adapting Disproportionality Methods to EHR Data

Zhao, J., Karlsson, I., Asker, L. and Boström, H. Applying Methods for Signal Detection in Spontaneous Reports to Electronic Patient Records. In 19th Knowledge Discovery and Data Mining (KDD) Conference’s Workshop on Data Mining for Healthcare (DMH), August 11-14, 2013, Chicago, USA.
Using Machine Learning Algorithms to Detect Missing ADE Codes in Patients’ Medical History
Two Studies: Experimental Setup

27 Datasets -> Stratified 10-fold cross validation -> Random Forest -> Accuracy AUC
Two Studies: Experimental Setup

27 Datasets

Stratified 10-fold cross validation

Random Forest

Accuracy AUC

Binary classification

- Patients diagnosed with an ADE-related code (e.g., D64.2, drug induced anemia)

- Patients diagnosed with a similar code (e.g., D64.3, other anemia)
Exploiting Concept Hierarchies of Clinical Codes

Exploiting Concept Hierarchies of Clinical Codes

## Multiple Representations of Clinical Measurements

<table>
<thead>
<tr>
<th>Multiple Representations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>Baseline</td>
</tr>
<tr>
<td>Mean</td>
<td>Average of repeated measurements</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>SD of repeated measurements</td>
</tr>
<tr>
<td>Slope</td>
<td>Difference between first and last measurement over time span</td>
</tr>
<tr>
<td>Existence</td>
<td>Whether a measurement exists</td>
</tr>
<tr>
<td>Count</td>
<td>How many times a measurement was repeated</td>
</tr>
<tr>
<td>Combined</td>
<td>Combination of above</td>
</tr>
</tbody>
</table>

## Multiple Representations of Clinical Measurements

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>79.33</td>
<td>0.5</td>
</tr>
<tr>
<td>Mean</td>
<td>81.97</td>
<td>0.67</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>81.41</td>
<td>0.66</td>
</tr>
<tr>
<td>Slope</td>
<td>80.72</td>
<td>0.63</td>
</tr>
<tr>
<td>Existence</td>
<td>82.30</td>
<td>0.68</td>
</tr>
<tr>
<td>Count</td>
<td>82.67</td>
<td>0.70</td>
</tr>
<tr>
<td>Combined</td>
<td>82.92</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Multiple Representations of Clinical Measurements

Combining Structured and Unstructured EHRs

Random Forests

Random Forest

Random Forest

Random Forest

Codes Space

Features

Features

Features

Measurements Space

Notes Space

Feature set

Random Forest

Data Science Workshop at Stockholm University: Jing Zhao, Aron Henriksson - December 4, 2014
Analyzing Structured and Unstructured Data in Electronic Health Records
Thanks for your attention!

A short demo

aDEX & aDET