Text-mining the Estonian National Electronic Health Record

Raul Sirel
rsirel@ut.ee
13.11.2015
• Electronic Health Records & Text Mining
• De-identifying the Texts
• Resolving the Abbreviations
• Terminology EXtraction and Text Analytics (TEXTA) Toolkit
Electronic Health Record (EHR)

Estonian National Health Information System (ENHIS)

- A nation-wide electronic health record
- All healthcare providers are obligated by law to forward their medical data to the ENHIS
- The main unit of data is the **epicrisis**, which contains information about:
  - the reason the patient arrived (anamnesis)
  - conducted procedures
  - medications
  - etc.
## The Data

<table>
<thead>
<tr>
<th>Epicrisis type</th>
<th>2012</th>
<th>2013</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outpatient consultation summaries</td>
<td>1,216,400</td>
<td>1,975,016</td>
<td>3,191,416</td>
</tr>
<tr>
<td>Discharge summaries</td>
<td>214,874</td>
<td>208,171</td>
<td>423,045</td>
</tr>
<tr>
<td>Total</td>
<td>1,431,274</td>
<td>2,183,187</td>
<td>3,614,461</td>
</tr>
</tbody>
</table>

- 2 years
- ~1 million patients
Why Text Mining?

- Significant portion (~50%) of the digital health data is *unstructured* (Hicks 2003)!
Patient complaints

Pulse

Blood Pressure Measurements
Why Text Mining?

- Significant portion (~50%) of the digital health data is unstructured (Hicks 2003)!

- In order to do something useful with the data, we need to analyse the unstructured data!
Outline

- Electronic Health Records & Text Mining
- De-identifying the Texts
- Resolving the Abbreviations
- Terminology EXtraction and Text Analytics (TEXTA) Toolkit
De-identifying the Texts

- Medical records contain sensitive information

- Identity-related information often found among the unstructured data

- Prior to releasing the data to researchers, the identify-related information needs to be removed:
  - names
  - national identity numbers
  - phone numbers
  - etc.
De-identifying the Texts

**Input**


**De-identified text**

Under the Bonnet

- Motivation from Named Entity Recognition
- CRF learning algorithm
- Surrounding words and grammatical attributes (case, number, etc.) as features

<table>
<thead>
<tr>
<th></th>
<th>CRF-based System</th>
<th>Dictionary-based system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>97%</td>
<td>40%</td>
</tr>
<tr>
<td>Recall</td>
<td>95%</td>
<td>70%</td>
</tr>
</tbody>
</table>
Outline

• Electronic Health Records & Text Mining

• De-identifying the Texts

• Resolving the Abbreviations

• Terminology EXtraction and Text Analytics (TEXTA) Toolkit
Resolving the Abbreviations

- Up to 13%* of all tokens are abbreviations:

* Short functional words removed prior to analysis
For each abbreviation in text:
- produce all possible full forms (rule-based model)
- select the most probable variant (statistical language model)

<table>
<thead>
<tr>
<th>Context</th>
<th>Full form</th>
</tr>
</thead>
<tbody>
<tr>
<td>... p silm ei näe ...</td>
<td>parem</td>
</tr>
<tr>
<td>... kolmas p palavik ...</td>
<td>päev</td>
</tr>
<tr>
<td>... vähene p pleurareaktsiooni riba ...</td>
<td>parietaalne</td>
</tr>
<tr>
<td>...</td>
<td>pupill</td>
</tr>
</tbody>
</table>

Full form: parietaalne | Score: 92%
parem | 4%
päev | 3%
pupill | 0,3%
• Electronic Health Records & Text Mining
• De-identifying the Texts
• Resolving the Abbreviations
• Terminology EXtraction and Text Analytics (TEXTA) Toolkit
Understanding the Text

For patients above five years with mild asthma inhaled steroids are the most effective preventer drug.

http://www.cvast.tuwien.ac.at/projects/iUMLS
Problems

• **Language specificity:**
  
  *most of the existing methodologies in NLP are usually language-specific and therefore not applicable for processing other languages*

• **Domain specificity:**
  
  *most of the research in NLP is currently focused on general language (e.g. newspaper articles)*

• **Lack of semantic resources:**
  
  *when working with sublanguages, it is often the problem that lexical resources (e.g. dictionaries or thesauri) built for general language correspond poorly to the actual language usage*

• **Scalability:**
  
  *existing NLP methods usually require large scale resources in order to be used in big data analysis*
The Objective

- The aim was to build a system for exploratory text analytics which:
  - is robust (and scalable)
  - is domain independent
  - doesn’t require language-specific resources
  - doesn’t require external semantic resources
Terminology Extraction and Text Analytics (TEXTA) Toolkit

- A system for:
  - describing domain terminologies
  - exploring and analysing the data using the defined terminologies

- For each subtask, the toolkit provides a corresponding tool
Base lexicon – a list of words describing some topic or semantic property, e.g.:

- **symptoms**: pain, nausea, queasiness, cut, etc.
- **anatomical**: head, hand, arm, leg, lung, etc.
- **locations**: left, right, central, lower, upper, medial, etc.
- **etc.**
1. User enters some words

2. User is supported with „similar“ words
Under the Bonnet

- **Distributional semantics:**
  - „You shall know a word by the company it keeps“ (Firth 1957)
  - Distributional hypothesis: words with similar distributional properties are semantically similar
  - Language modelling, word-vector modelling

![Bar chart](chart.png)
Under the Bonnet

- „Semantic“ similarity in word-vector models using cosine similarity
**TEXTA: Semantic Grouping of Words**

- The aim is to group together words with similar meaning:
  - *headache* - *migraine*
  - *pain* – *ache*
  - *etc.*

- The user is supported with an interactive 2-D projection of the base lexicons:
  - PCA
  - MDS
  - t-SNE
TEXTA: Semantic Grouping of Words

- PCA plot of a base lexicon containing patient complaints:
TEXTA: Semantic Grouping of Words

- PCA plot of a base lexicon containing patient complaints:

The user can now group similar words into concepts (groups of words with similar meanings)

constipation-related words

nausea-related words

pain-related words
More complex concepts are represented as multi-word expressions:

Text Corpus

- Patient complaints pain in left arm.
- Motorcycle accident – deep cut in right leg.

Base lexicons
- Complaints: pain, cut, ...
- Anatomical: head, arm, ...
- Locations: left, right, ...

Multi-word expressions
A **k-partite graph** is a graph whose vertices are partitioned into *k* different independent sets.

- **k = number of base lexicons**
- **A multi-word expression** is a path with a length of *n* (*n*≤*k*), whose vertices are located in different sets (the path is acyclic)
<table>
<thead>
<tr>
<th>Field: lemmas</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bool: and</td>
<td></td>
</tr>
<tr>
<td>Match type:</td>
<td></td>
</tr>
<tr>
<td>Slope:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Search took <strong>0.39</strong> seconds. Showing <strong>50</strong> of total <strong>631</strong> matches.</td>
</tr>
<tr>
<td>Field: lemmas</td>
<td><strong>anonym</strong></td>
</tr>
<tr>
<td></td>
<td><strong>08.09.2012 - Tuttavate koer hammustas Koer hammustas paremast ölast</strong></td>
</tr>
<tr>
<td></td>
<td><strong>02.02.2013 - oma koer hammustas, koer on vaksineeritud</strong></td>
</tr>
<tr>
<td></td>
<td><strong>04.11.201316 : 47 - , - D07925 - E600 - üldarstabi Hammustas koer Vasakul reie ülemises osas koer hammustuses marrastus dT ei mäleta</strong></td>
</tr>
<tr>
<td></td>
<td><strong>30.07.2013 - oma koer hammustas koer ja laps vaksineeritud</strong></td>
</tr>
<tr>
<td></td>
<td><strong>29.06.2013 - Ninal ja parema silmaaluselt pindmised koera hammustushaavad</strong></td>
</tr>
<tr>
<td></td>
<td><strong>04.11.201316 : 47 - , - D07925 - E600 - üldarstabi Hammustas koer Vasakul reie ülemises osas koera hammustuses marrastus dT ei mäleta</strong></td>
</tr>
<tr>
<td></td>
<td><strong>30.07.2012 - oma koer hammustas koer on vaksineerimata sellel aastal, laps on vaksineeritud</strong></td>
</tr>
<tr>
<td></td>
<td><strong>04.11.2013 : PÖ - tänana homnikul koolimaja trepil sai tundmatult koeralt hammustada ( koer nägi valja hookudud, korraik , ketjupp kaelas )</strong></td>
</tr>
<tr>
<td></td>
<td><strong>06.05.2012 - Vasemal põsel 3 pindmist koera hammustushaava</strong></td>
</tr>
<tr>
<td></td>
<td><strong>06.05.2012 - Vasemal põsel 3 pindmist koera hammustushaava</strong></td>
</tr>
<tr>
<td></td>
<td><strong>05.07.2012 - Õmmeldud vasaku sääre koera hammustushaava Põletik Õmlblused eemaldatud Haavade vahel mõõdukas punetus</strong></td>
</tr>
<tr>
<td></td>
<td><strong>10.06.2013 - Laupäeval õmmeldud koera hammustushaava paremal õlavareel</strong></td>
</tr>
</tbody>
</table>
**TEXTA: Aggregating the Matches**

- Matching documents can be aggregated over any field in the dataset.
- Bite-related documents aggregated over time:

![Graph showing hit count over time for 'putukas hammastas', 'kass hammastas', and 'koer hammastas'.]
TEXTA: Aggregating the Matches

- Bite-related documents aggregated over diagnoses:

  - Open wound of unspecified body region
  - Venom of other arthropods
  - Need for immunization against rabies
  - Multiple open wounds of wrist and hand
  - Cellulitis of other parts of limb
  - Lyme disease (Borreliosis)
  - Localized oedema
  - Urticaria, unspecified
TEXTA: Aggregating the Matches

- Bite-related documents aggregated over:
  - significant words:
    - to bite (verb)
    - bite wound
    - dog
    - tick
    - neighbour
    - anti-rabic
**TEXTA: Aggregating the Matches**

**Bite-related documents aggregated over:**

- **significant words:**
  - to bite (verb)
  - bite wound
  - dog
  - tick
  - neighbour
  - anti-rabic

- **gender:**
  - Female
  - Male
**TEXTA: Conclusion**

- **TEXTA** – A toolkit for performing text mining

- **Toolkit’s workflow is based on:**
  - describing domain terminologies
  - exploring and analysing the data using the defined terminologies

- **The sales pitch:**
  - it’s robust (and scalable)
  - it’s domain independent
  - it doesn’t require language-specific resources
  - it doesn’t require external semantic resources
TEXTA: Demo

- [https://ehr.stacc.ee/public/texta](https://ehr.stacc.ee/public/texta)
Overall Conclusion

• The general aim is to provide resources for increasing the meaningful usage of unstructured data:
  • clinical research
  • quality of care assessments
  • clinical decision support
  • personalised medicine
  • etc.
Thank You for listening!
References

