Building a framework for handling clinical abbreviations – a long journey of understanding shortened words

Yonghui Wu¹ PhD, Joshua C. Denny² MD MS, S. Trent Rosenbloom² MD MPH, Randolph A. Miller² MD, Dario A. Giuse² Dr.Ing, Hua Xu¹,² PhD

¹ School of Biomedical Informatics, The University of Texas
² Department of Biomedical Informatics, Vanderbilt University
Abbreviations in Clinical Notes

53 yo woman with HTN, DM2, obesity, depression/anxiety, H. Pylori gastritis, GERD, h/o pancreatitis in 2002 presumed 2/2 passed gallstone (lipase 431/amylase 181).

• **Pervasive** – 17.1% of total word tokens
• **Important** – 33% abbreviations are diseases/symptoms
• **Ambiguous** - 33.1% of abbreviations found in the UMLS 2001 were ambiguous (e.g., pt - patient, pt - physical therapy)
• **Dynamic** – different institutions, different types of notes, over the time
A Corpus-based Framework for Handling Clinical Abbreviations

1. Detect abbreviations from a corpus
   e.g. if “ca” is an abbreviation

2. Build corpus-specific sense inventories
   e.g. “ca” - Carcinoma
       - Calcium

3. Disambiguate ambiguous abbreviations
   e.g. ... Diagnosed with Colon CA after ...
   ✔ Carcinoma
   ✗ Calcium
Detect abbreviations using machine learning classifiers

- Data set: 70 discharge summaries with abbreviations annotated
- Features including word formation, dictionary, corpus frequency
- Different machine learning algorithms including DT, RF, SVM, and ensembles
- Best performance – F-measure 95.7%

Build sense inventory using clustering

Evaluation of clustering-based sense inventory

- 13 abbreviations, 52 senses

<table>
<thead>
<tr>
<th>Annotation Cost</th>
<th>Average Sense completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>EM</td>
</tr>
<tr>
<td>10</td>
<td>66%</td>
</tr>
<tr>
<td>20</td>
<td>69%</td>
</tr>
</tbody>
</table>

Profile-based disambiguation using vector space model

... She was ambulating on room air without shortness of breath...

room air | ...She was ambulating on ra without shortness of breath...

Figure 2: An example of the sense tagging step for the profile-based disambiguation method.

Figure 3: An example of the transformation step for the profile-based disambiguation method.
Combine profile score and estimated sense frequency from clustering analysis

<table>
<thead>
<tr>
<th>Sense frequency based on clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>“right atrium”</td>
</tr>
<tr>
<td>“room air”</td>
</tr>
<tr>
<td>“rheumatoid arthritis”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Profile-based similarity scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>“right atrium”</td>
</tr>
<tr>
<td>“room air”</td>
</tr>
<tr>
<td>“rheumatoid arthritis”</td>
</tr>
</tbody>
</table>

- 0.07 + 0.745 = 0.8154
- 0.185 + 0.0226 = 1.0076

- “rheumatoid arthritis”

Building comprehensive abbreviation sense inventories from Vanderbilt corpus

<table>
<thead>
<tr>
<th>Notes</th>
<th>Abbreviation detected</th>
<th># of clusters for 1000 abbreviations</th>
<th>Final # of abbreviations</th>
<th>Final # of senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discharge Summary (123K)</td>
<td>22,546</td>
<td>15,536</td>
<td>913 (221)</td>
<td>1,299</td>
</tr>
<tr>
<td>Clinic Visit (261K)</td>
<td>107,303</td>
<td>20,101</td>
<td>954 (283)</td>
<td>1,499</td>
</tr>
</tbody>
</table>
Deliverables

- The open source CARD (Clinical Abbreviation Recognition and Disambiguation) Framework
  - Abbreviation Detection
  - Sense Clustering
  - Sense disambiguation
  - List of abbreviations
  - Sense clusters
  - Sense Inventory

- Two comprehensive sense inventories: discharge summaries and clinic visit notes
- Abbreviation wrappers for existing NLP systems: MetaMap and cTAKES