Analyzing Differences between Chinese and English Clinical Text

Yonghui Wu, Jianbo Lei, Wei-Qi Wei, Buzhou Tang, Joshua C. Denny, S. Trent Rosenbloomc, Randolph A. Miller, Dario A. Giuse, Kai Zheng, Hua Xu
Jianbo.lei@uth.tmc.edu
Outline

• Background
• Motivation and Objective
• Method
• Result
• Discussion
Background

• Health insurance now covers 95.6% of the population in China
• The Chinese Ministry of Health (MOH) proposed a series of templates covering EMR basic architectures and data standards
• Chinese government allocated 3.9 billion RMB (approximately $600 million US) in 2011 to implement EMRs in about 200 hospitals
Motivation & Objective

• Motivation
  – A large amount of EHR data generated worldwide
  – Rapid growth of EHR systems in China
  – Collaborations between US and China
  – Differences in culture and practice patterns affect EHR data

• Objective
  – Understand system and cultural differences that exist between Chinese and English clinical documents
Method

• Collect inpatient discharge summaries from one Chinese and three US institutions

• Focus on: Problems, Tests, and Treatments

• Conduct content analysis using Charmaz’s grounded theory approach

• Compare the differences at the document-level and section-level
Organizations

- English
  - University of Pittsburgh Medical Center (UPMC)
  - Partners Healthcare (PARTNERS)
  - Beth Israel Deaconess Medical Center (BETH)
- Chinese
  - Peking Union Medical College Hospital (PUMCH)
Overview of the Study Design

I2B2-English (646)

- UPMC
- Discharge notes

- PARTNERS
- Discharge notes

- BETH
- Discharge notes

Chinese (400)

- PUMCH
- Discharge notes

Problem, Treatment, Test

Extract/Mapping Section headers

Statistical analysis
Result

- English vs Chinese on vocabulary

Stanford Word Segmenter trained on “Penn Chinese Treebank corpus”

Zipf’s distribution
Result cont.

- Normalized distribution of annotated entities

![Graph showing normalized frequency of entities across ranks. The graph indicates that the English corpus used a more diverse vocabulary of entities compared to the Chinese corpus.](image)
Result cont.

- Distribution of different types of entities

<table>
<thead>
<tr>
<th>Corpus</th>
<th># of Doc</th>
<th>Type</th>
<th># of Entity</th>
<th>Average # of entity per note</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPMC (English)</td>
<td>220</td>
<td>Prob</td>
<td>5805</td>
<td>26.39</td>
<td>43.76%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test</td>
<td>2762</td>
<td>12.55</td>
<td>20.82%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treat</td>
<td>4700</td>
<td>21.36</td>
<td>35.43%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>13267</td>
<td><strong>60.30</strong></td>
<td>--</td>
</tr>
<tr>
<td>PARTNERS (English)</td>
<td>235</td>
<td>Prob</td>
<td>8542</td>
<td>36.35</td>
<td>44.69%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test</td>
<td>4884</td>
<td>20.78</td>
<td>25.55%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treat</td>
<td>5686</td>
<td>24.20</td>
<td>29.75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>19112</td>
<td><strong>81.33</strong></td>
<td>--</td>
</tr>
<tr>
<td>BETH (English)</td>
<td>191</td>
<td>Prob</td>
<td>11122</td>
<td>58.23</td>
<td>38.93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test</td>
<td>8947</td>
<td>46.84</td>
<td>31.32%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treat</td>
<td>8499</td>
<td>44.50</td>
<td>29.75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>28568</td>
<td><strong>149.57</strong></td>
<td>--</td>
</tr>
<tr>
<td>PUMCH (Chinese)</td>
<td>400</td>
<td>Prob</td>
<td>20159</td>
<td>50.40</td>
<td>51.25%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test</td>
<td>12114</td>
<td>30.29</td>
<td>30.80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treat</td>
<td>7061</td>
<td>17.65</td>
<td><strong>17.95%</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>39334</td>
<td><strong>98.34</strong></td>
<td>--</td>
</tr>
</tbody>
</table>

Entity number varied greatly even for US institutions

Fewer Treatment entity in Chinese EHRs
Result- cont.

• Relative frequency of entities
**Result cont.**

- **Section-level**

<table>
<thead>
<tr>
<th>Section</th>
<th>UPMCD (English)</th>
<th>PARTNERS (English)</th>
<th>BETH (English)</th>
<th>PUMCH (Chinese)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Doc</td>
<td>Entity</td>
<td>Ave</td>
<td>Doc</td>
</tr>
<tr>
<td>PS</td>
<td>131</td>
<td>4453</td>
<td>33.99</td>
<td>174</td>
</tr>
<tr>
<td>DM</td>
<td>95</td>
<td>1224</td>
<td>12.88</td>
<td>138</td>
</tr>
<tr>
<td>DI</td>
<td>47</td>
<td>314</td>
<td>6.68</td>
<td>54</td>
</tr>
<tr>
<td>CC</td>
<td>33</td>
<td>377</td>
<td>11.42</td>
<td>34</td>
</tr>
<tr>
<td>DD</td>
<td>105</td>
<td>1005</td>
<td>9.57</td>
<td>35</td>
</tr>
<tr>
<td>HOPI</td>
<td>30</td>
<td>486</td>
<td>16.20</td>
<td>151</td>
</tr>
<tr>
<td>PE</td>
<td>25</td>
<td>479</td>
<td>19.16</td>
<td>142</td>
</tr>
<tr>
<td>PMH</td>
<td>59</td>
<td>659</td>
<td>11.17</td>
<td>140</td>
</tr>
<tr>
<td>PL</td>
<td>48</td>
<td>187</td>
<td>3.90</td>
<td>41</td>
</tr>
</tbody>
</table>
Discussion

• The number of clinical entities per document varied widely among different institutions
• The Chinese discharge summaries contained fewer Treatment clinical entities than any US institution’s discharge summaries
• Difference in section
  – Social History and current medication are not found. Medication information could be recorded in a patient’s Past Medical History section, e.g., "the patient was diagnosed with HTN in 1995. She is taking a beta blocker (Metoprolol) and her BP is normal”
  – PS\DI\DM, densities of entities are at least twice as those of Chinese clinical notes
Limitations

• Only one Chinese institution involved
• I2b2 notes lacked information about the clinical settings in which the notes were generated
Acknowledgments

• Grant: NLM R01LM010681.
• The 2010 i2b2/VA challenge
Thanks

Q & A

Hua.xu@uth.tmc.edu
Jianbo.lei@uth.tmc.edu