Beyond Dictionary Match for Disease Named Entity Recognition

Robert Leaman
Department of Computer Science

Chris Miller & Graciela Gonzalez
Department of Biomedical Informatics

Arizona State University, Tempe AZ
Motivation

- Named entity recognition (NER)
  - Find all entities of a specified type (diseases, genes, drugs, etc) in a given text
  - Not trivial: many terms ambiguous
    - Neurofibromatosis 2 (Nf 2); AD
    - Dictionary insufficient

- Building block task
  - e.g. gene/disease relationship extraction
  - High performance required

- State of the art technique: sequence-based machine learning e.g. conditional random fields
Related Work & Corpus

• Disease corpus by Jimeno et. al. (EBI)
  − Evaluation centered on dictionary matching
  − Did not contain mention locations - required for applying sequence-based machine learning


• Our effort: update corpus
  − Re-annotate for mention location
  − Annotate corpus sentences not prev. annotated
  − Currently 2427 sentences, 3367 disease annotations
**Methods**

- **BANNER - biomedical NER system**
  - Portable, trainable (CRF), configurable
  - Open source: [http://banner.sourceforge.net](http://banner.sourceforge.net)

- **Disease dictionary**
  - Extracted from UMLS MetaThesaurus
  - 167,944 entries
  - Added to BANNER as an input feature

- Evaluation by 5x2 cross-validation
## Evaluation Results

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BANNER with UMLS-based dictionary</td>
<td>80.78</td>
<td>70.66</td>
<td>75.36</td>
</tr>
<tr>
<td>- with CRF order = 1</td>
<td>80.17</td>
<td>70.95</td>
<td>75.27</td>
</tr>
<tr>
<td>- without UMLS-based dictionary</td>
<td>78.55</td>
<td>66.77</td>
<td>72.16</td>
</tr>
<tr>
<td>UMLS-based disease dictionary</td>
<td>62.72</td>
<td>61.78</td>
<td>62.24</td>
</tr>
<tr>
<td>MetaMap (Jimeno et.al.)</td>
<td>31.27</td>
<td>30.07</td>
<td>30.66</td>
</tr>
<tr>
<td>Statistical (Jimeno et.al.)</td>
<td>26.07</td>
<td>30.18</td>
<td>27.97</td>
</tr>
</tbody>
</table>
Ablation Study
How many sentences do we need?

BANNER, with dictionary

BANNER, without dictionary

Percentage of training data used