Lexical Resources for Named Entity Recognition

State-of-the-art:
- Gazetteers to boost performance (Torii et al., 2009)
- Covers few semantic categories

Problems:
- Lexical variation poses a problem for gazetteers
- No resource perfectly matches an entity class (Wang et al., 2009)

Approximate String Matching

Observations:
- We have large collections of lexical resources for various classes
- Recent advances in approximate string matching makes approximate string matching computationally feasible (Okazaki and Tsujii, 2010)

Idea:
- Approximate instead of strict gazetteers
- Use large collections to disambiguate semantic categories

SimString Features

Feature generation:
1. Query each lexical resource using cosine measure and a sliding similarity threshold \([1.0, 0.7]\) with a step of 0.1
2. If the query returns a match, assign a feature uniquely identifying that data and threshold, and all subsequent lower thresholds

To note:
- Threshold 1.0 is equivalent to strict matching
- Cut-off motivated by the fact that low thresholds will match similarities even at a superficial level

Models

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>Span-internal features used in previous work on NER</td>
</tr>
<tr>
<td>Internal+Gazetteer</td>
<td>Internal features and gazetteer features</td>
</tr>
<tr>
<td>Internal+SimString</td>
<td>Internal features and SimString features</td>
</tr>
</tbody>
</table>

Table: Models used in our experimental setting

Task Setting

Given a textual span (denoted by \([\ldots]\)), assign the semantic category:

\[
\text{Histone H3 protein methylated at lys36 aminoacid was [catalysed] by HMT protein.}
\]

Experimental Results

Evaluation Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>Semantic Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioNLP/NLPBA 2004 Shared Task Corpus</td>
<td>NLPBAC</td>
<td>5</td>
</tr>
<tr>
<td>Gene Regulation Event Corpus</td>
<td>GREC</td>
<td>64 (5 collapsed)</td>
</tr>
<tr>
<td>Collaborative Annotation of a Large Biomedical Corpus</td>
<td>CALBC CII</td>
<td>4</td>
</tr>
<tr>
<td>Epigenetics and Post-Translational Modifications</td>
<td>EPI</td>
<td>17</td>
</tr>
<tr>
<td>Infectious Diseases Corpus</td>
<td>ID</td>
<td>16</td>
</tr>
<tr>
<td>Gene Event Corpus</td>
<td>GENIA</td>
<td>11</td>
</tr>
</tbody>
</table>

Table: Corpora used for evaluation

Conclusions

- Can not establish a clear benefit for all datasets, but works very well for one dataset
- The method appears to have potential but merits further investigations as to when it is applicable

Future Work

- Evaluate non-cosine measures for approximate string matching, which take biological knowledge into account
- Investigate as to why certain datasets although similar on a semantic category level yield different different results
- Contribution of individual resources towards overall performance

Availability

Source code, lexical resources, additional results and future research is/will be available at:

http://github.com/ninjin/simsem/

Feel free to use, derive and/or complain.