Syntactic-semantic analysis for information extraction in biomedicine

Sérgio Matos\textsuperscript{1}, Anabela Barreiro\textsuperscript{2}
\textsuperscript{1}IEETA, Universidade de Aveiro
\textsuperscript{2}Centro de Linguística, Universidade do Porto

aleixomatos@ua.pt; barreiro_anabela@hotmail.com

June 2009
Outline

• Background
• Text Mining and Information Extraction in Biomedicine
• Objectives
• Implementation
• Results
• Conclusions
Background

- Genomics and Proteomics are fast-growing fields
- Literature grows exponentially
  - MEDLINE/PubMed ~ 18m citations

- Researchers need to contextualize their theories and findings
  - Interactions between genes/proteins
  - Involvement in biological processes and in disease
  - And many other factors...

- How to keep up-to-date with new knowledge in the field?
Background

• Manually curated biomedical databases are a good source of information
  – Publications are reviewed and important information added to DBs (e.g. protein interactions)
  – Impossible to keep DBs up-to-date due to increased volume of publications

• Text Mining can be useful for
  – Information retrieval (IR)
  – Information extraction (IE)
  – DB curators and end-users (researchers)
Text Mining and Information Extraction in Biomedicine

• Text mining deals with the automated processing of texts to derive high quality information
• Information Extraction can be seen as one application of TM

• Different processing levels
  • Entity Recognition (ER)  genes, proteins, etc.
  • Normalization  ATF2 - GeneID 1386
                  ATF-2 – Uniprot P15336
  • Relation extraction  PPI, gene/disease
  • Event extraction  gene expression, regulation

+ semantics  + domain knowledge
Text Mining and Information Extraction in Biomedicine

• Good results for NER, but limited to a few entity types
  – 80%-90% for recognition of genes/proteins
  – Need to include more entities, like chemical compounds, diseases, experimental conditions

• Relation extraction has focused mostly on PPI

• Inter-concept relations not too explored
  – e.g. gene/disease, drug/target
  – mostly based on co-occurrence statistics
Text Mining and Information Extraction in Biomedicine

- Recent interest towards extraction of events
  - BioNLP shared task and BioCreaTive II.5

- ... and other entities / facts
  - e.g. Experimental conditions, lab techniques, measurements

- ... Discourse analysis
  - “indicating/suggesting that…”, “in contrast…”

- Full-text vs. Abstracts
  - Complexity in grammar
Linguistic Resources for Biomedical TM

• UMLS Metathesaurus
  – various terms, all linked to same concept (e.g. ‘Hypertension’)
  – semantic information provided by the UMLS Semantic Network

• BioLexicon
  – Includes domain relevant verbs (localize, bind, express, ...)

• Lexical resources can be created from available online DBs
  – NCBI Entrez Gene for gene names
  – UniProt for proteins
  – OMIM for diseases
  – Various ontologies
Objectives

• Extract phrases indicating a biomolecular event from scientific text

• Biomolecular events include various types
  – Examples
    • “phosphorylation of TRAF2”
    • “localization of beta-catenin”
    • “TRADD interacts with TES2”

• BioNLP'09 Shared Task on Event Extraction
  – http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/
Objectives

• Six event types considered
  – Localization, Binding, Gene expression, Transcription, Protein Catabolism, Phosphorylation

• Training data
  – Annotation of genes/proteins occurring in each input text, including the text section (start and end characters)
  – Annotation of the events, including the event type, the participating entities and the corresponding trigger word (with start and end times)

• Test data
  – Annotation of participating genes/proteins is given
  – Create annotation of events for the given entities
Implementation

• General approach
  – Create syntactic grammars to detect phrases that indicate events
  – Grammars are based only on NEs and domain verbs (and derived names)

• Requisites
  – Grammars outputs should indicate the event type

• Solution
  – Event types can be associated with the trigger word using the semantic properties in NooJ dictionaries
  – Event types associated with each trigger word are derived from training data
Implementation

• Resources
  – Entity dictionary
    • Create dictionary with list of entities occurring in the texts
# Implementation

<table>
<thead>
<tr>
<th>Lemma</th>
<th>PoS</th>
<th>FLX</th>
<th>Semantic properties</th>
<th>ID</th>
<th>TAXID</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>N</td>
<td>TABLE</td>
<td>ORGANISM</td>
<td>9606</td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>N</td>
<td></td>
<td>ORGANISM</td>
<td>9606</td>
<td></td>
</tr>
<tr>
<td>Mus musculus</td>
<td>N</td>
<td></td>
<td>ORGANISM</td>
<td>10090</td>
<td></td>
</tr>
<tr>
<td>Breast cancer type 1 susceptibility protein</td>
<td>N</td>
<td></td>
<td>PROTEIN</td>
<td>P3839 8</td>
<td>9606</td>
</tr>
<tr>
<td>BRCA1</td>
<td>N</td>
<td></td>
<td>PROTEIN</td>
<td>P3839 8</td>
<td>9606</td>
</tr>
<tr>
<td>BRCA1</td>
<td>N</td>
<td></td>
<td>PROTEIN</td>
<td>P4875 4</td>
<td>10090</td>
</tr>
<tr>
<td>BRCA1</td>
<td>N</td>
<td></td>
<td>GENE</td>
<td>672</td>
<td>9606</td>
</tr>
<tr>
<td>RNFL53</td>
<td>N</td>
<td></td>
<td>GENE</td>
<td>672</td>
<td>9606</td>
</tr>
</tbody>
</table>
Implementation

• Resources
  – Entity dictionary
    • Create dictionary with list of entities occurring in the texts
  – BioLexicon verb dictionary
    • Adapted to include event type
      – From the training data, extract the verbs associated with events
      – Add a semantic property to the dictionary entry indicating the event type
      – Example: “express,V+EventType=Gene_Expression”
  • Added inflectional and derivation rules
    – The inflected and derivated forms inherit the verb’s semantic properties
Implementation

• Verb dictionary

<table>
<thead>
<tr>
<th>Lemma</th>
<th>PoS</th>
<th>DRV</th>
<th>FLX</th>
<th>EventType</th>
</tr>
</thead>
<tbody>
<tr>
<td>express</td>
<td>V</td>
<td>ION:TABLE</td>
<td>ABOLISH</td>
<td>Gene_expression</td>
</tr>
<tr>
<td>ligate</td>
<td>V</td>
<td>TION:TABLE</td>
<td>SMILE</td>
<td>Binding</td>
</tr>
<tr>
<td>stimulate</td>
<td>V</td>
<td>TION:TABLE</td>
<td>SMILE</td>
<td>Positive_regulation</td>
</tr>
</tbody>
</table>
Implementation

• Syntactic grammars
  – Sentences from training set used to generate surface patterns
  – Manual procedure
  – Seven grammars created
  – Example:

  “stimulation of human CD4”
Implementation

Stimulation of human CD4

\(<\text{EVENT}+\text{PROTEIN}=$CD4$+\text{EXP}=\text{Stimulation}+\text{TYPE}=\text{Positive\_regulation}>\)
Results

- Example patterns extracted from texts

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Concordance in text</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;entity&gt; [&lt;entity_type&gt;] &lt;nominalization&gt;</code></td>
<td><em>HSP gene expression</em></td>
</tr>
<tr>
<td><code>&lt;nominalization&gt; “of” [&lt;entity_type&gt;] &lt;entity&gt;</code></td>
<td><em>upregulation of Fas</em></td>
</tr>
<tr>
<td><code>&lt;entity&gt; [&lt;entity_type&gt;] &lt;be&gt; [“not”] [&lt;adverb&gt;] &lt;verb&gt;</code></td>
<td><em>IL-2R stimulation was totally inhibited</em></td>
</tr>
<tr>
<td><code>&lt;verb&gt; &lt;preposition&gt; &lt;entity&gt;</code></td>
<td><em>binding of TRAF2</em></td>
</tr>
<tr>
<td><code>&lt;verb&gt; &lt;nominalization&gt; “of” &lt;entity&gt;</code></td>
<td><em>suppressing activation of STAT6</em></td>
</tr>
</tbody>
</table>
Results

- Average results

<table>
<thead>
<tr>
<th>Event type</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localization</td>
<td>35.63</td>
<td>70.45</td>
<td>47.33</td>
</tr>
<tr>
<td>Binding</td>
<td>13.54</td>
<td>34.06</td>
<td>19.38</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>46.40</td>
<td>78.45</td>
<td>58.31</td>
</tr>
<tr>
<td>Transcription</td>
<td>33.58</td>
<td>41.07</td>
<td>36.95</td>
</tr>
<tr>
<td>Protein Catabolism</td>
<td>35.71</td>
<td>62.50</td>
<td>45.45</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>49.63</td>
<td>79.76</td>
<td>61.19</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>36.76</strong></td>
<td><strong>65.58</strong></td>
<td><strong>47.11</strong></td>
</tr>
</tbody>
</table>
Conclusions

• NooJ syntactic grammars for IE
  – Simple and flexible approach
  – Takes advantage of semantic properties and inflectional and derivational morphology in NooJ dictionaries

• Pattern identification
  – Manual method is limited
  – How to generate new patterns automatically?

• Gene regulatory events
  – Described by complex constructions
  – Can syntactic grammars be used for this type of events?
References and Acknowledgments

• BioLexicon was developed within the BOOTStrep project
  – http://www.nactem.ac.uk/biolexicon/
  – http://www.bootstrep.eu/bin/view/Extern/WebHome

• Data set from the BioNLP’09 Shared Task on Event Extraction
  – http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

Sérgio Matos is funded by Fundação para a Ciência e Tecnologia (FCT)
under the Ciência2007 programme.