Integrated Annotation for Biomedical IE

Mining the Bibliome:
Information Extraction from the Biomedical Literature

NSF ITR grant EIA-0205448

- 5-year grant, now 1.5 years from start
- University of Pennsylvania
  Institute for Research in Cognitive Science (IRCS)
- subcontract to Children’s Hospital of Philadelphia (CHOP)
- cooperation with GlaxoSmithKline (GSK)
Two Areas of Exploration

1. Genetic variation in malignancy (CHOP)  
   Genomic entity X is varied by process Y in malignancy Z  
   Ki-ras mutations were detected in 17.2% of the adenomas.  
   Entities: Gene, Variation*, Malignancy* 
      (*relations among sub-components)

2. Cytochrome P450 inhibition (GSK)  
   Compound X inhibits CYP450 protein Y to degree Z  
   Amiodarone weakly inhibited CYP3A4-mediated activities with Ki = 45.1 μM  
   Entities: Cyp450, Substance, quant-name, quant-value, quant-units
Approach

• Build hand-annotated corpora in order to train automated analyzers
• Mutual constraint of form and content:
  – parsing helps overcome diversity and complexity of relational expressions
  – entity types and relations help constrain parsing
• Shallow semantics integrated with syntax
  – entity types, standardized reference, co-reference
  – predicate-argument relations
• Requires significant changes in both syntactic and semantic annotation
• Benefits:
  – automated analysis works better
  – patterns for “fact extraction” are simpler
Project Goals

• Create and publish corpora integrating different kinds of annotation:
  – Part of Speech tags
  – Treebanking \textit{(labelled constituent structure)}
  – Entities and relations \textit{(relevant to oncology and enzyme inhibition projects)}
  – Predicate/argument relations, co-reference
  – Integration:
    \begin{itemize}
    \item textual \textbf{entity-mentions} $\approx$ syntactic \textbf{constituents}
    \end{itemize}

• Develop IE tools using the corpus

• Integrate IE with existing bioinformatics databases
Project Workflow

Tokenization and POS Annotation

Entity Annotation

Merged Representation

Treebanking

(recently revised to a flat pipeline)

<table>
<thead>
<tr>
<th>Task</th>
<th>Started</th>
<th>abstracts</th>
<th>words</th>
<th>Software</th>
<th>tagger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tok + POS</td>
<td>8/22/03</td>
<td>1317</td>
<td>292K</td>
<td>Wordfreak</td>
<td>yes</td>
</tr>
<tr>
<td>Entity</td>
<td>9/12/03</td>
<td>1367</td>
<td>308K</td>
<td>Wordfreak</td>
<td>starting</td>
</tr>
<tr>
<td>Treebanking</td>
<td>1/8/04</td>
<td>295</td>
<td>70K</td>
<td>TreeEditor</td>
<td>retraining</td>
</tr>
</tbody>
</table>
Integration Issues (1)

• Modifications to Penn Treebank guidelines
  (for tokenization, POS tagging, treebanking)
    – to deal with biomedical text
    – to allow for syntactic/semantic integration
    – to be correct!

• Example: Prenominal Modifiers
  old way:

  the breast cancer-associated autoimmune antigen
  DT   NN      JJ               JJ       NN
  (NP.................................................................................)

  new way:

  the breast cancer - associated autoimmune antigen
  DT   NN      NN      -   VBN      JJ       NN
  (NML...............)
  (ADJP........................................)  (NML..........................)*
  (NP..........................................................)

  *implicit
Integration Issues (2)

• Coordinated entities
  – "point mutations at codons 12, 13 or 61 of the human K-, H- and N-ras genes"
  – *Wordfreak* allows for discontinuous entities
  – Treebank guidelines modified, e.g.:
    (NP (NOM-1 codons) 12) ,
    (NP (NOM-1 *P* ) 13) or
    (NP (NOM-1 *P* ) 61)
  – Modification works recursively
Entity Annotation

Mutations of ras genes distinguish a subset of non-small-cell lung cancer cell lines from small-cell lung cancer cell lines.

Mitsudomi T, Viallet J, Mulshine JL, Linnola RI, Minna JD, Gazdar AF.

NCI-Navy Medical Oncology Branch, National Cancer Institute, Bethesda, Maryland 20893-5105.

We screened a panel of 103 human lung cancer cell lines for the presence of point mutations at codons 12, 13 or 61 of the human K-, H- and N-ras genes, using restriction fragment length polymorphisms (RFLP). Using mismatched primers during polymerase chain reaction (PCR) of genomic DNA. We found ras mutations in 22/61 (36%) non-small-cell lung cancer (NSCLC) cell lines, predominantly in K-ras codon 12. Identical mutations were present in uncultured tumor materials corresponding to 11 cell lines containing mutated ras genes. ras mutations were found not only in adenocarcinoma cell lines (9/32, 28%), but also in cell lines derived from other types of NSCLC (13/29, 45%). In contrast, none of 37 small-cell lung cancer (SCLC) cell lines and five extra-pulmonary small-cell cancer cell lines had ras mutations. ras mutations were not correlated with sex of the patients, tumor extent, prior therapy status or in vitro culture time. G to T or A to T transversions were the most common base substitutions, occurring in codons 12 and 61 respectively. We conclude that ras mutations play a role in the pathogenesis of a subset of NSCLC but are not involved in SCLC.
We screened a panel of 103 human lung cancer cell lines for the presence of point mutations at codons 12, 13 or 61 of the human K-, H- and N-ras genes, using restriction fragment length polymorphisms (RFLP), created through mismatched primers during polymerase chain reaction (PCR) of genomic DNA.
Tagger Development (1)

POS tagger retrained 2/10:

<table>
<thead>
<tr>
<th>Tagger</th>
<th>Training Material</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>PTB sections 00-15</td>
<td>773832</td>
</tr>
<tr>
<td>New</td>
<td>315 abstracts</td>
<td>104159</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tagger</th>
<th>Overall Accuracy</th>
<th>#Unseen Instances</th>
<th>Accuracy Unseen</th>
<th>Accuracy Seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>88.53%</td>
<td>14542</td>
<td>58.80%</td>
<td>95.53%</td>
</tr>
<tr>
<td>New</td>
<td>97.33%</td>
<td>4096</td>
<td>85.05%</td>
<td>98.02%</td>
</tr>
</tbody>
</table>

Tokenizer also retrained -- new tokenizer used in both cases
Tagger Development (2)

<table>
<thead>
<tr>
<th>entity</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation type</td>
<td>0.8556</td>
<td>0.7990</td>
<td>0.8263</td>
</tr>
<tr>
<td>Variation loc</td>
<td>0.8695</td>
<td>0.7722</td>
<td>0.8180</td>
</tr>
<tr>
<td>Variation state-init</td>
<td>0.8430</td>
<td>0.8286</td>
<td>0.8357</td>
</tr>
<tr>
<td>Variation state-sub</td>
<td>0.8035</td>
<td>0.7809</td>
<td>0.7920</td>
</tr>
<tr>
<td>Variation overall</td>
<td>0.8541</td>
<td>0.7870</td>
<td>0.8192</td>
</tr>
<tr>
<td>Chemical tagger</td>
<td>0.87</td>
<td>0.73</td>
<td>0.79</td>
</tr>
<tr>
<td>Gene tagger</td>
<td>0.93</td>
<td>0.60</td>
<td>0.73</td>
</tr>
</tbody>
</table>

(Precision & recall from 10-fold cross-validation, **exact string match**)

Taggers are being integrated into the annotation process.
References

• Project homepage: http://ldc.upenn.edu/myl/ITR
• Annotation info: http://www.cis.upenn.edu/~mamandel/annotators/
• Wordfreak: http://www.sf.net/projects/wordfreak
• Taggers: http://www.cis.upenn.edu/datamining/software_dist/biosfier/
• Integration analysis (entities and treebanking): http://www.cis.upenn.edu/~skulick/biomerge.html
• LAW http://www.sf.net/projects/law