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

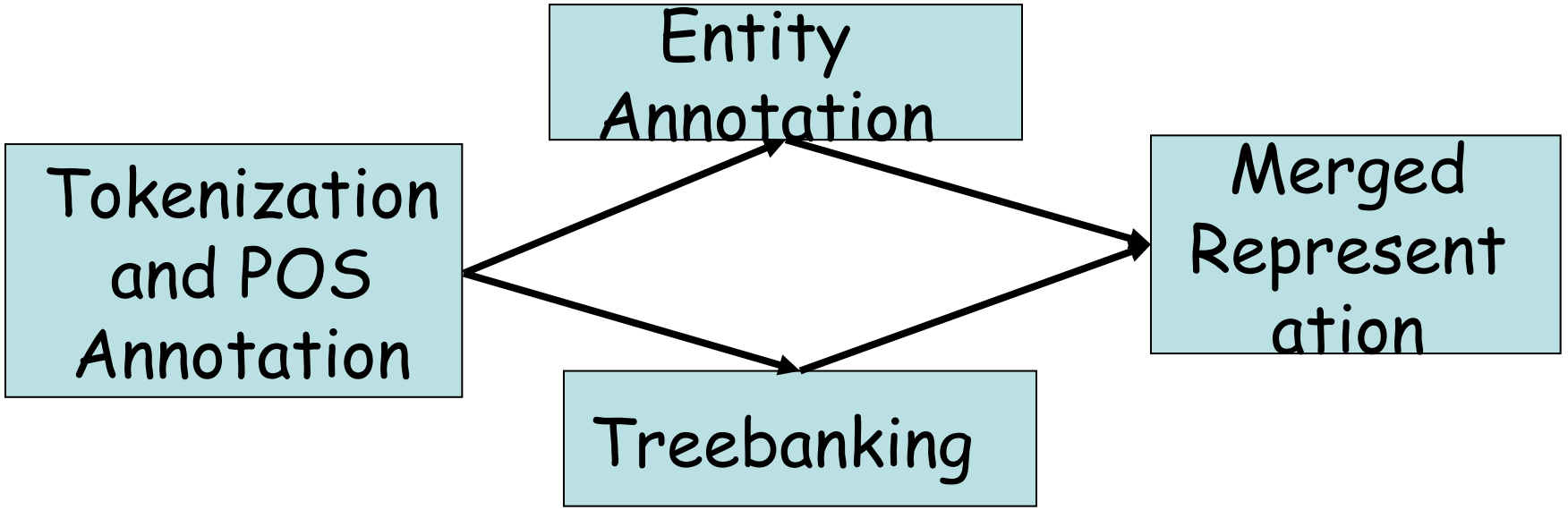
Approach

- 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 *(labelled constituent structure)*
  - Entities and relations
    (relevant to oncology and enzyme inhibition projects)
  - Predicate/argument relations, co-reference
  - Integration:
    textual **entity-mentions** ≈ syntactic **constituents**
- Develop IE tools using the corpus
- Integrate IE with existing bioinformatics databases

Project Workflow



(recently revised to a flat pipeline)

Task	Started	abstracts	words	Software	tagger
Tok + POS	8/22/03	1317	292K	Wordfreak	yes
Entity	9/12/03	1367	308K	Wordfreak	starting
Treebanking	1/8/04	295	70K	TreeEditor	retraining

Integration Issues (1)

- Modifications to Penn Treeebank 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

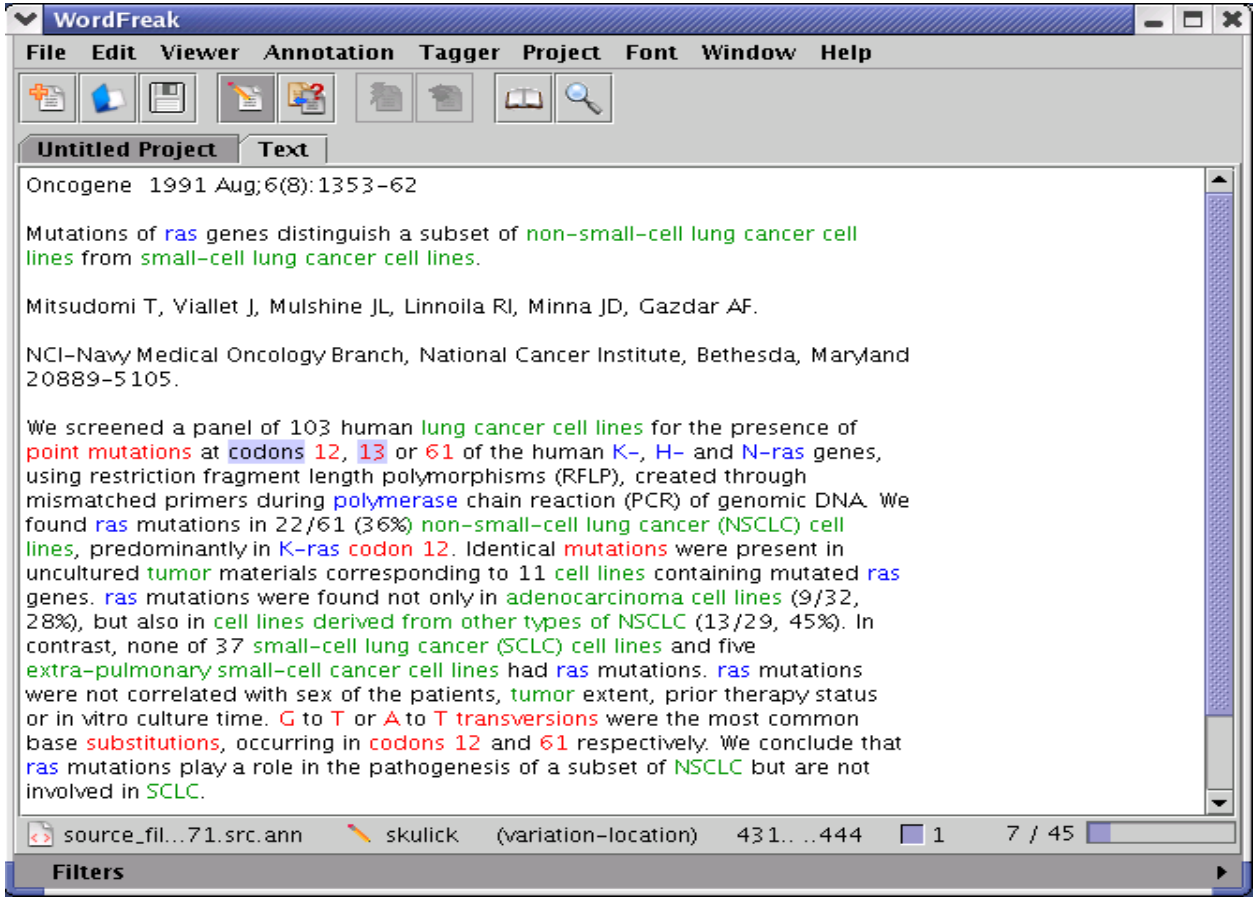
Integration Issues (2)

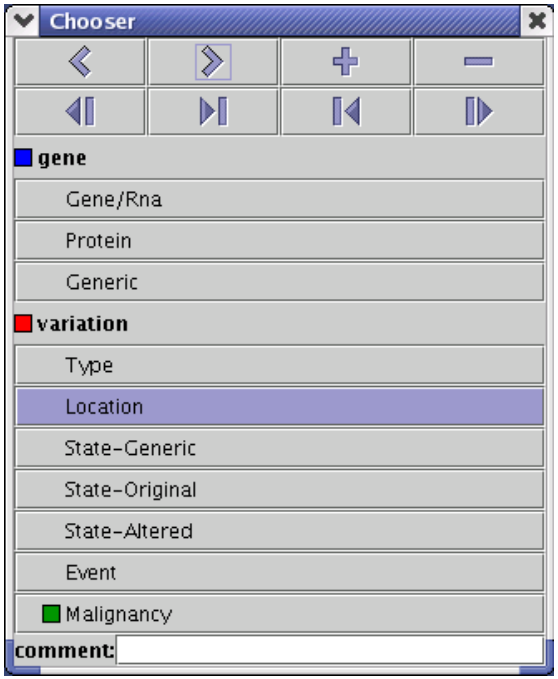
- Coordinated entities
  - "point mutations at codons 12, 13 or 61 of the human K-, H- and N-ras genes"
  - *Wordfreak* allows for discontinous entities
  - Treebank guidelines modified, e.g.:

```
(NP (NOM-1 codons) 12) ,
(NP (NOM-1     *P*   ) 13)  or
(NP (NOM-1     *P*   ) 61)
```

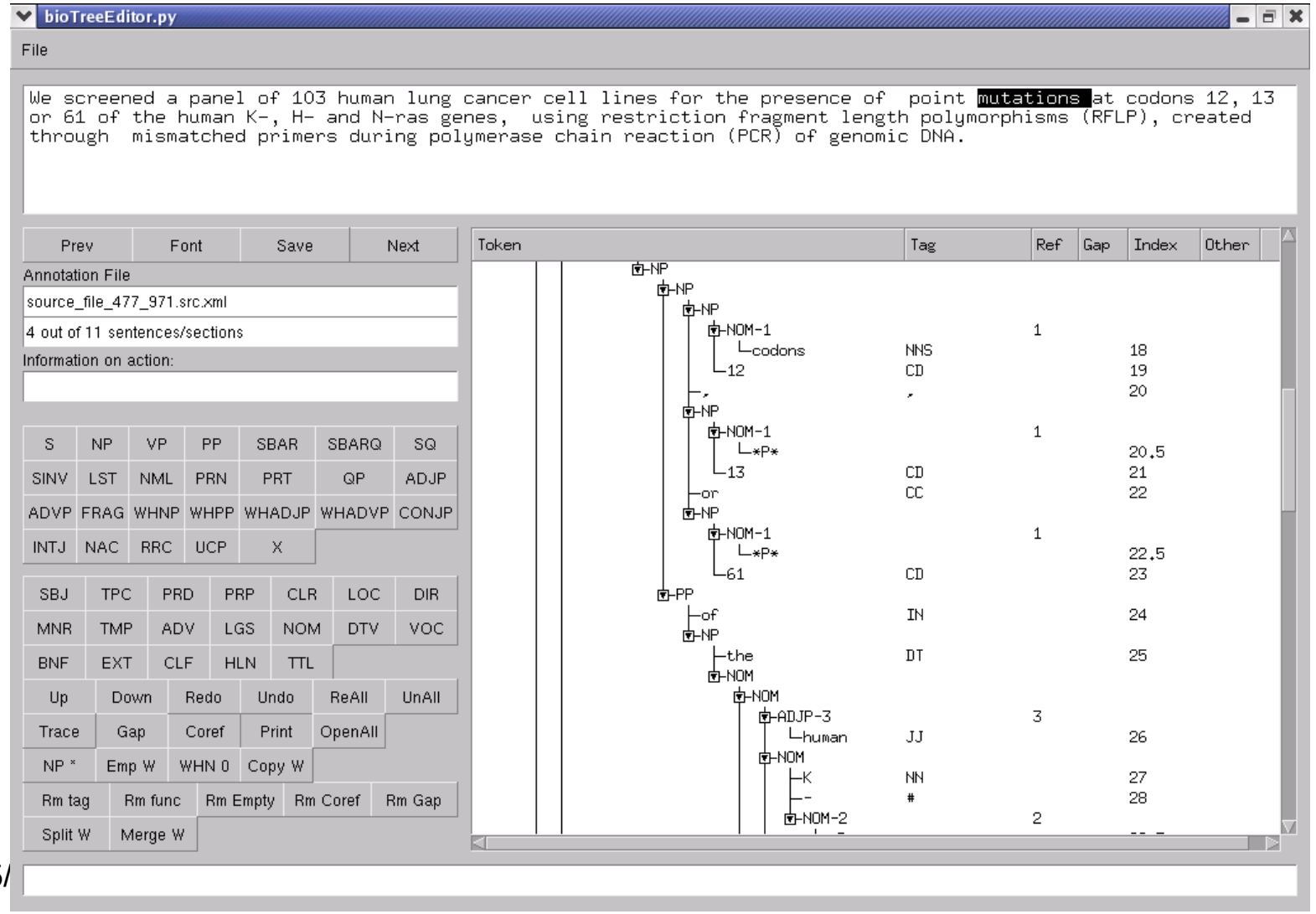
  - Modification works recursively

Entity Annotation





Treebanking



Tagger Development (1)

POS tagger retrained 2/10:

Tagger	Training Material	Tokens
Old	PTB sections 00-15	773832
New	315 abstracts	104159

Tagger	Overall Accuracy	#Unseen Instances	Accuracy Unseen	Accuracy Seen
Old	88.53%	14542	58.80%	95.53%
New	97.33%	4096	85.05%	98.02%

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

Tagger Development (2)

entity	Precision	Recall	F
Variation type	0.8556	0.7990	0.8263
Variation loc	0.8695	0.7722	0.8180
Variation state-init	0.8430	0.8286	0.8357
Variation state-sub	0.8035	0.7809	0.7920
Variation overall	0.8541	0.7870	0.8192
Chemical tagger	0.87	0.73	0.79
Gene tagger	0.93	0.60	0.73

(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