High-precision biological event extraction with a concept recognizer

K. Bretonnel Cohen*, Karin Verspoor*, Helen L. Johnson, Chris Roeder, Philip V. Ogren, William A. Baumgartner Jr., Elizabeth White, Hannah Tipney, and Lawrence Hunter

U. Colorado Denver School of Medicine

*These two authors contributed equally to the work reported here

Introduction

- All three tasks approached as concept recognition and analysis using the OpenDMAP semantic parser
- Manually-written rules
- Achieved highest precision for two of the three tasks; recall was low

Methods

- Named entity recognition
  - Proteins: provided by organizers
  - Other semantic classes:
    - LingPipe with GENIA model
    - ConceptMapper (Gene Ontology cellular components, Cell Type Ontology, BRENDA Tissue Ontology, and Sequence Ontology)
      - Only modifications: remove *cell* from CTO and add synonym *nuclear* to GOCC

Methods

- Coördination
  - Retrained OpenNLP constituent parser with…
    - 500 abstracts from beta version of GENIA treebank
    - 10 full-text articles from the CRAFT corpus
  - Distributed meaning assumed
  - Used for proteins only

Methods

- OpenDMAP (Hunter et al. 2008)
  - Uses ontology as central organizing structure
    - Commitment to using community-consensus ontologies (mostly OBO)
  - Semantic grammars allow mixing of terminals and semantically typed non-terminals
  - Context-free power with variable ordering
  - Slot fillers constrained by ontology

Methods

Protein_transport :=

[TRANSPORTED-ENTITY] translocation

@(from {DET}? [TRANSPORT-ORIGIN])

@(to {DET}? [TRANSPORT-DESTINATION])

Bax translocation to mitochondria from the cytosol

Bax translocation from the cytosol to the mitochondria

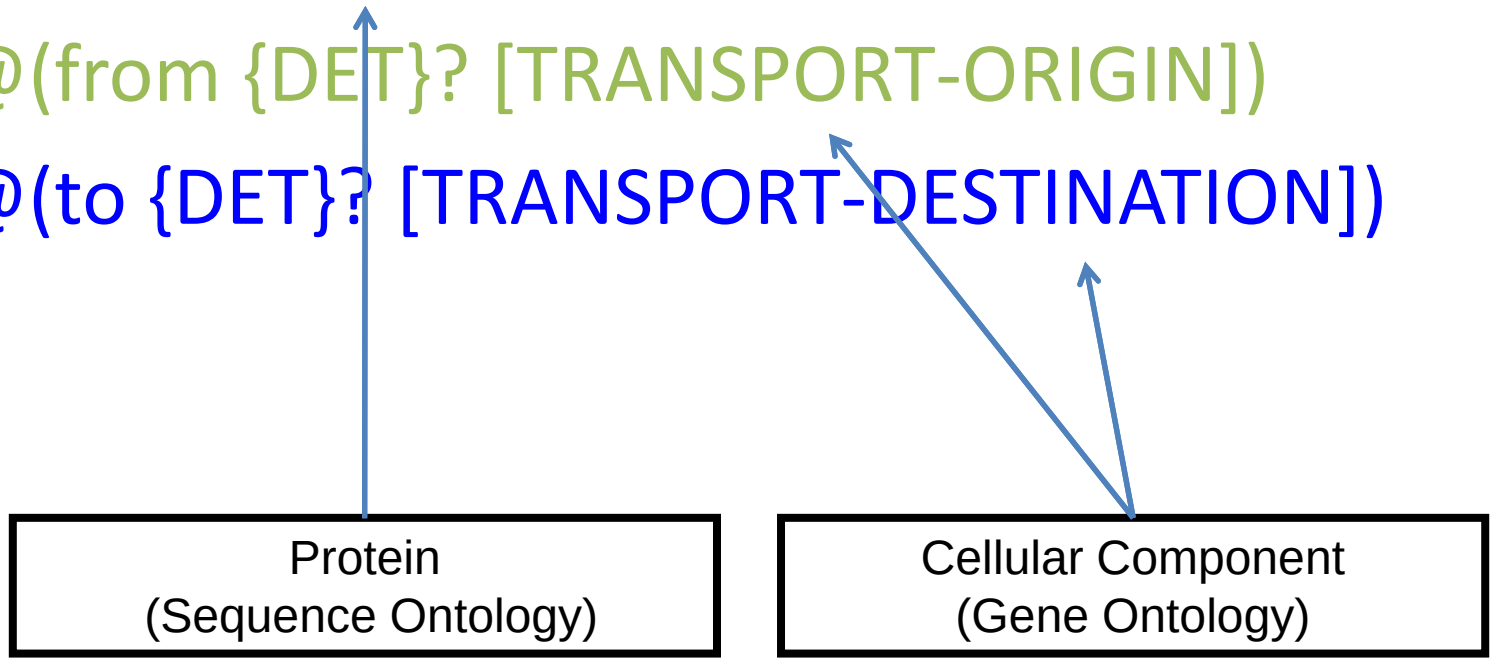
Methods

Protein_transport :=

[TRANSPORTED-ENTITY] translocation

@(from {DET}? [TRANSPORT-ORIGIN])

@(to {DET}? [TRANSPORT-DESTINATION])

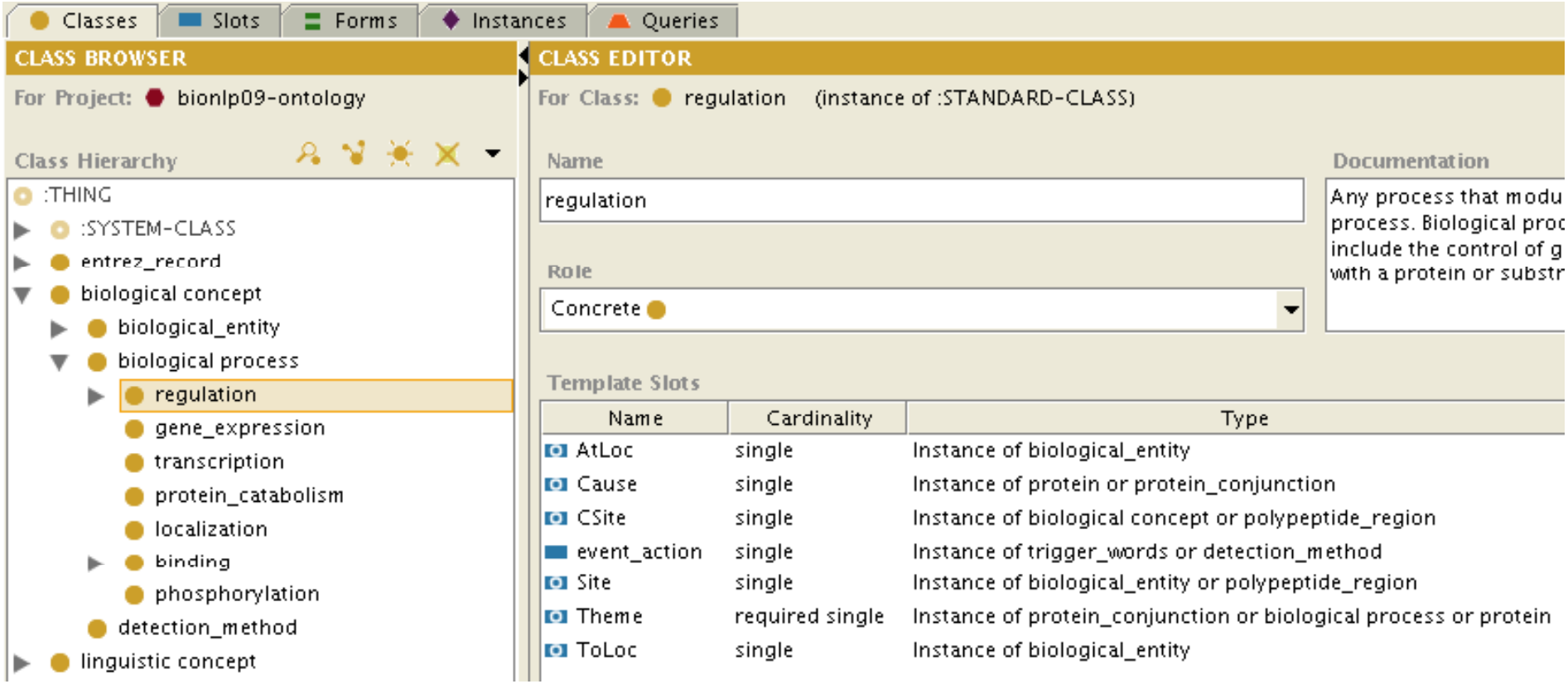


Methods

- All event types represented as frames
  - Elements from ontology constrain every slot

EVENT TYPE: REGULATION
AtLoc: instance of biological_entity
Cause: instance of protein
CSite: instance of biological_concept or polypeptide_region
Event_action: instance of trigger_word or detection_method
Site: instance of biological concept or polypeptide region
Theme: instance of protein or biological_process
ToLoc: instance of biological entity

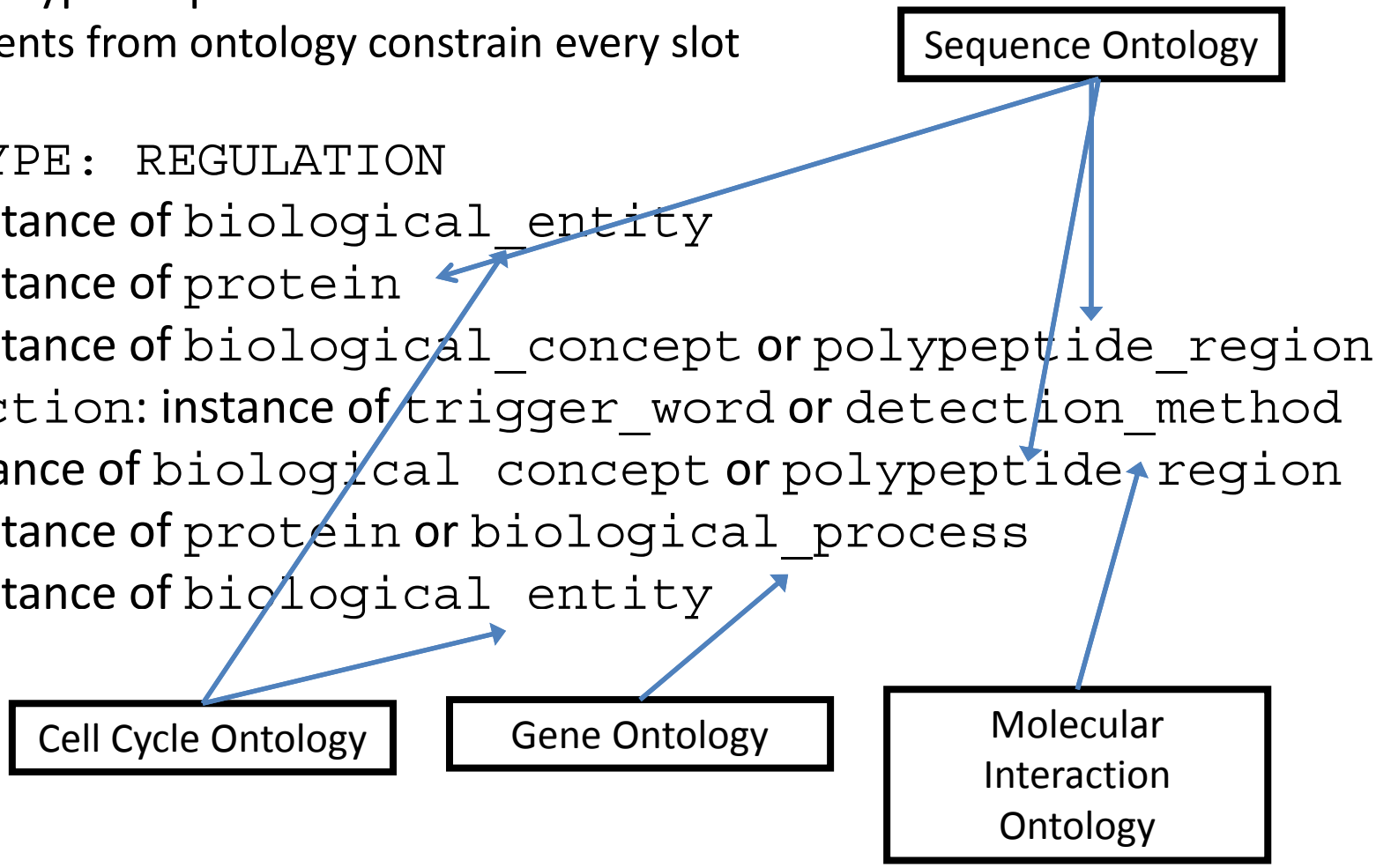
Methods


Classes
Slots
Forms
Instances
Queries
CLASS BROWSER
For Project: bionlp09-ontology
Class Hierarchy
:THING
:SYSTEM-CLASS
entrez_record
biological concept
biological_entity
biological process
regulation
gene_expression
transcription
protein_catabolism
localization
binding
phosphorylation
detection_method
linguistic concept
CLASS EDITOR
For Class: regulation (instance of :STANDARD-CLASS)
Name
regulation
Documentation
Any process that modu
process. Biological proc
include the control of g
with a protein or substr
Role
Concrete
Template Slots
Name
Cardinality
Type
AtLoc
single
Instance of biological_entity
Cause
single
Instance of protein or protein_conjunction
CSite
single
Instance of biological concept or polypeptide_region
event_action
single
Instance of trigger_words or detection_method
Site
single
Instance of biological_entity or polypeptide_region
Theme
required single
Instance of protein_conjunction or biological process or protein
ToLoc
single
Instance of biological_entity

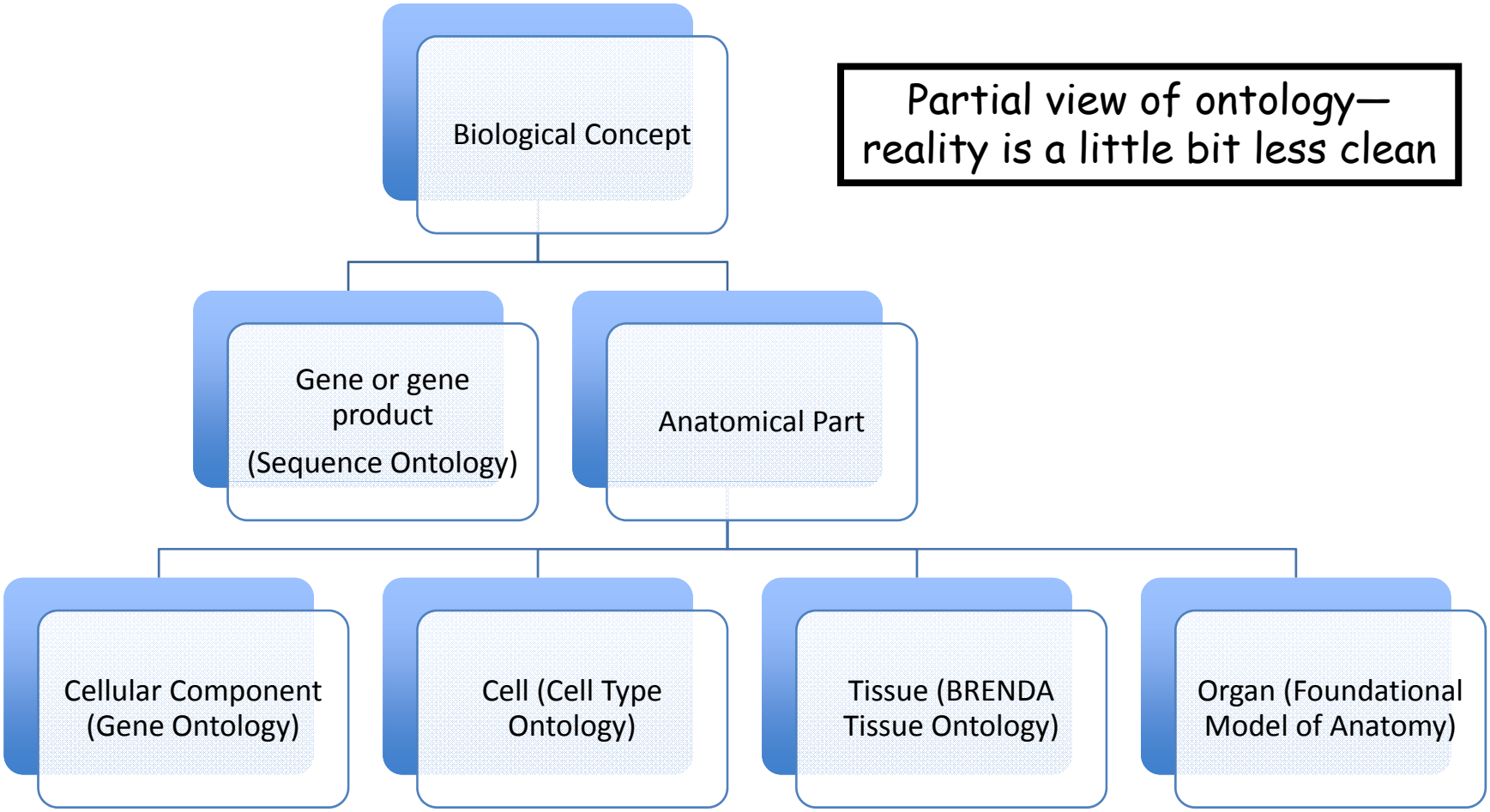
Methods

- All event types represented as frames
  - Elements from ontology constrain every slot

EVENT TYPE: REGULATION
AtLoc: instance of biological_entity
Cause: instance of protein
CSite: instance of biological_concept or polypeptide_region
Event_action: instance of trigger_word or detection_method
Site: instance of biological concept or polypeptide region
Theme: instance of protein or biological_process
ToLoc: instance of biological entity



Methods



Partial view of ontology—
reality is a little bit less clean

Methods

Event type	Site	AtLoc	ToLoc
Binding	protein domain (SO), binding site (SO), DNA (SO), chromosome (SO)		
Gene expression	gene (SO), biological entity (CCO)	tissue (BTO), cell type (CTO), cellular component (GO)	
Localization		cellular component (GO)	cellular component (GO)
Phosphorylation	amino acid (FMA), polypeptide region (SO)		
Protein catabolism	cellular component (GO)		
Transcription	gene (SO), biological entity (CCO)		

BTO: BRENDA Tissue Ontology
CCO: Cell Cycle Ontology
CTO: Cell Type Ontology
GO: Gene Ontology
SO: Sequence Ontology

Methods

- Manual pattern-writing
  - Before availability of training data: based on native speaker intuitions, examples from PubMed, and variations on same, as in Cohen et al. (2004)
  - After release of training data: based on examination of corpus data, targeting high-frequency predicates only
  - Nominalizations predominated; used insights from Cohen et al. (2008) regarding Theme placement
  - Protein binding rules re-used from BioCreative II protein-protein interaction task
  - Eschewed use of wildcards

Results

	Our system			Best team			Best P/R/F		
	P	R	F	P	R	F	P	R	F
Task 1	**71.81**	13.45	22.66	58.48	**46.73**	**51.95**	**71.81**	**46.73**	**51.95**
Task 2	**70.97**	13.25	43.12	54.08	**35.86**	**43.12**	**70.97**	**35.86**	**43.12**
Task 3	57.40	12.33	20.30	**60.83**	**32.68**	**42.52**	**60.83**	**32.68**	**42.52**

Task 1: P 10 points higher than second-highest
Task 2: P 14 points higher than second-highest
Task 3: P 3.4 points lower than highest (3/6)

Results

Unofficial results: contribution of bug repairs

	P	R	F
Official results	71.81	13.45	22.66
With bug fixes	67.19	17.38	27.10

Still the highest precision
(#2 was 62.21)

Results

- Contribution of coördination-handling
  - Bug-fixed results: F 27.62 (Task 1)
  - Without coördination-handling: F 24.72
  - Decrease in F of 2.9 without coördination-handling

Results

- Error analysis: false negatives
  - Intervening material between trigger words and arguments
  - Coördination of things other than protein names
  - Low coverage of trigger words
  - Anaphora and coreference
  - Appositive gene names and symbols

Results

- Intervening material between trigger words and arguments

to efficiently [express] in developing thymocytes a mutant form of the [NF-kappa B inhibitor] (PMID 10092801)

- Solutions: use dependency syntax, multiword expressions

Results

- Coördination of things other than protein names

[transcription] and subsequent synthesis and secretion of [galectin-3] (PMID 8623933)

Results

- Low coverage of trigger words
  - Sharing of some trigger words across multiple event types
  - Time constraints of the shared task
- Similarly, limited attempts to deal with intervening material

Results

- Anaphora and coreference

Although 2 early lytic transcripts, [BZLF1] and [BHRF1], were also detected in 13 and 10 cases, respectively, the lack of ZEBRA staining in any case indicates that these lytic transcripts are most likely [expressed] by rare cells in the biopsies entering lytic cycle (PMID 8903467)

Results

- Appositive gene names and symbols

Rule: {gene_expression} :=
expression of {det}? [Theme];

[expression] of Fas ligand ([FasL]) (PMID 10092076)

Intervening material

False Positive Analysis

-Training data
-Biologist examined all events scored as FP in 10 documents per event type
- 42% judged as actually TP

	Analyzed	TP	FP
Gene Expression	11	7	4
Binding	21	15	6
Transcription	4	1	3
Protein Catabolism	1	1	0
Localization	4	1	3
Phosphorylation	3	2	1
Regulation	13	1	12
Positive Regulation	18	7	11
Negative Regulation	13	2	11

False Positive Analysis

Key:
trigger word
Theme

- Incorrect Theme chosen:
  - "IL-10 production by **gp41**"[10089566]
  - "induction of **I kappa B alpha** phosphorylation"[7499266]
- Missed a required entity:
  - "induction of **IL-10 production** by gp41"[10089566]
- Misrecognized trigger word:
  - "upstream of the **GM-CSF** transcription initiation site"[7478534]
  - "effects of **IL-11** were associated with reduced **NF-kappaB** activation"[10411003]
  - "up-regulation of **CD80 Ag**"[8690900]
- Sentence parse error:
  - "was suppressed by **alpha B2**. Coexpression of alpha B1"[7605990]

Discussion and Conclusions

- Recall is a function of the rule set, not of the approach
- Can this high-precision, apparently low-recall approach scale to practical performance levels?
  - Redundancy in individual papers
  - These results make no use of syntactic analysis
  - Exploring rapid adaptation to new tasks
    - Rule/template inheritance
    - Extend coördination handling beyond proteins
    - Using dependency parses
    - Mappings from frames to WordNet for better trigger word coverage

Discussion and Conclusions

- Not just building system for shared tasks—input to other tasks, such as high-throughput data analysis, which it has proven useful for

Acknowledgements

- Michael Bada for help with loading the Sequence Ontology into Protégé
- Alias-I for LingPipe
- Anonymous ConceptMapper creator
- Shared task organizers for organizing the task and answering many questions
- Work supported by:
  - NIH grants R01LM009254, R01GM083649, and R01LM008111 to Larry Hunter
  - NIH grant T15LM009451 to Philip Ogren

Availability

- OpenDMAP semantic parser and rule set:

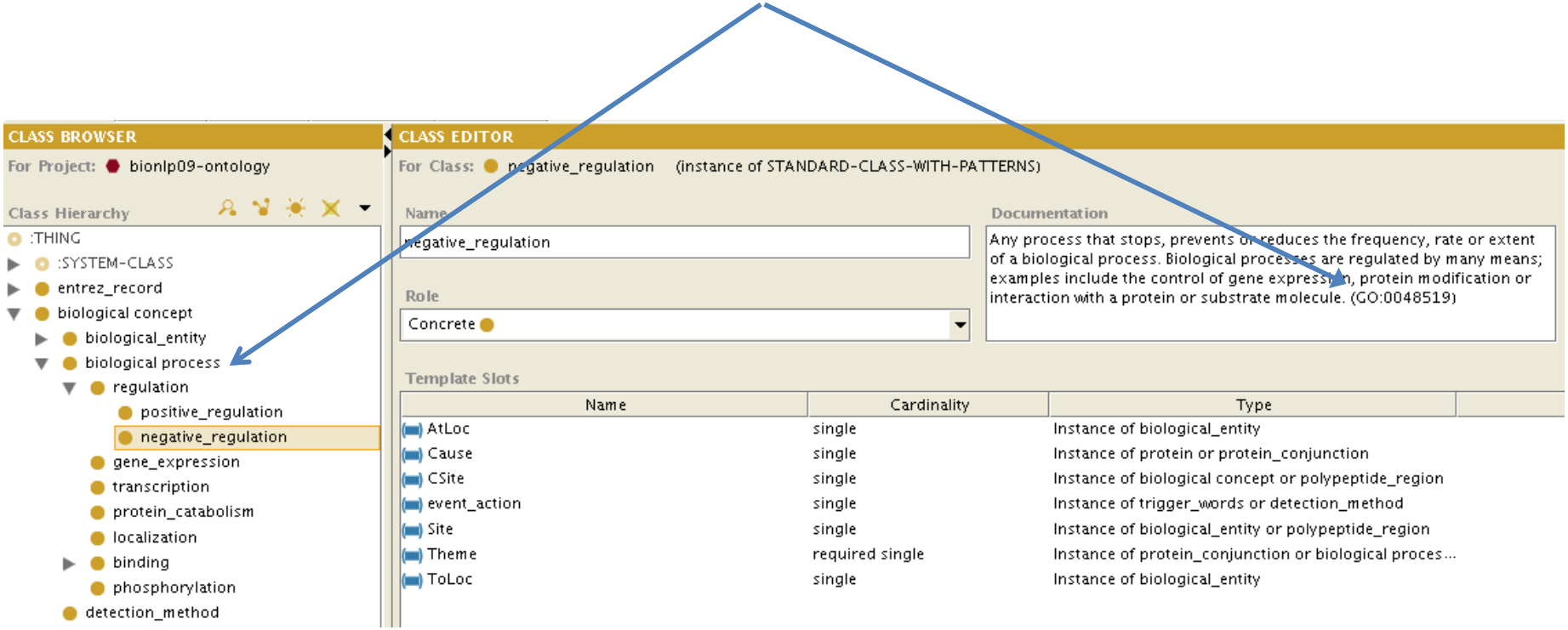
`bionlp.sourceforge.net`

Questions to be prepared for

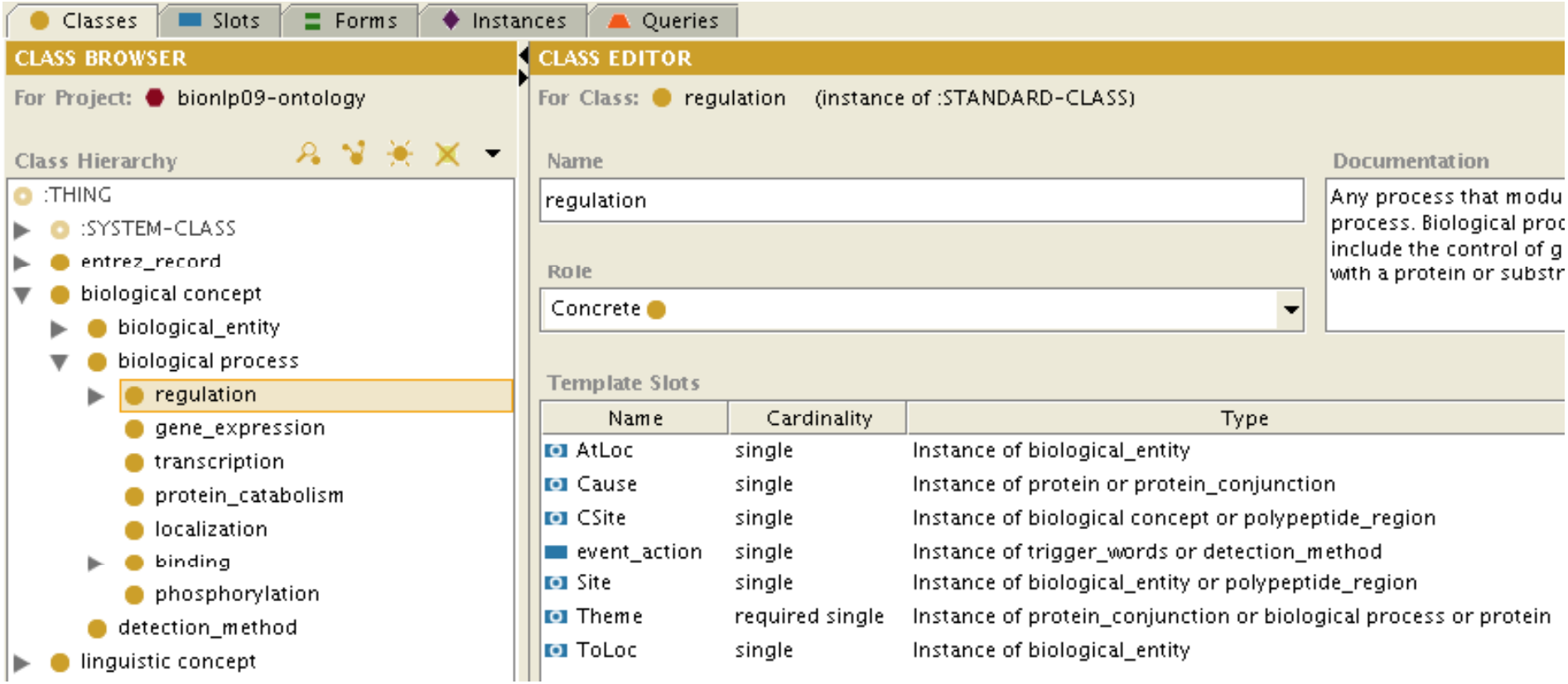
- What does it mean to have an ontology be the central organizing structure of a semantic parser?
- Why didn’t we get the top precision on Task 3?
- You didn’t say anything about how you tackled negation and speculation

Ontology as the central organizing principle

Ontology specifies the targets of information extraction

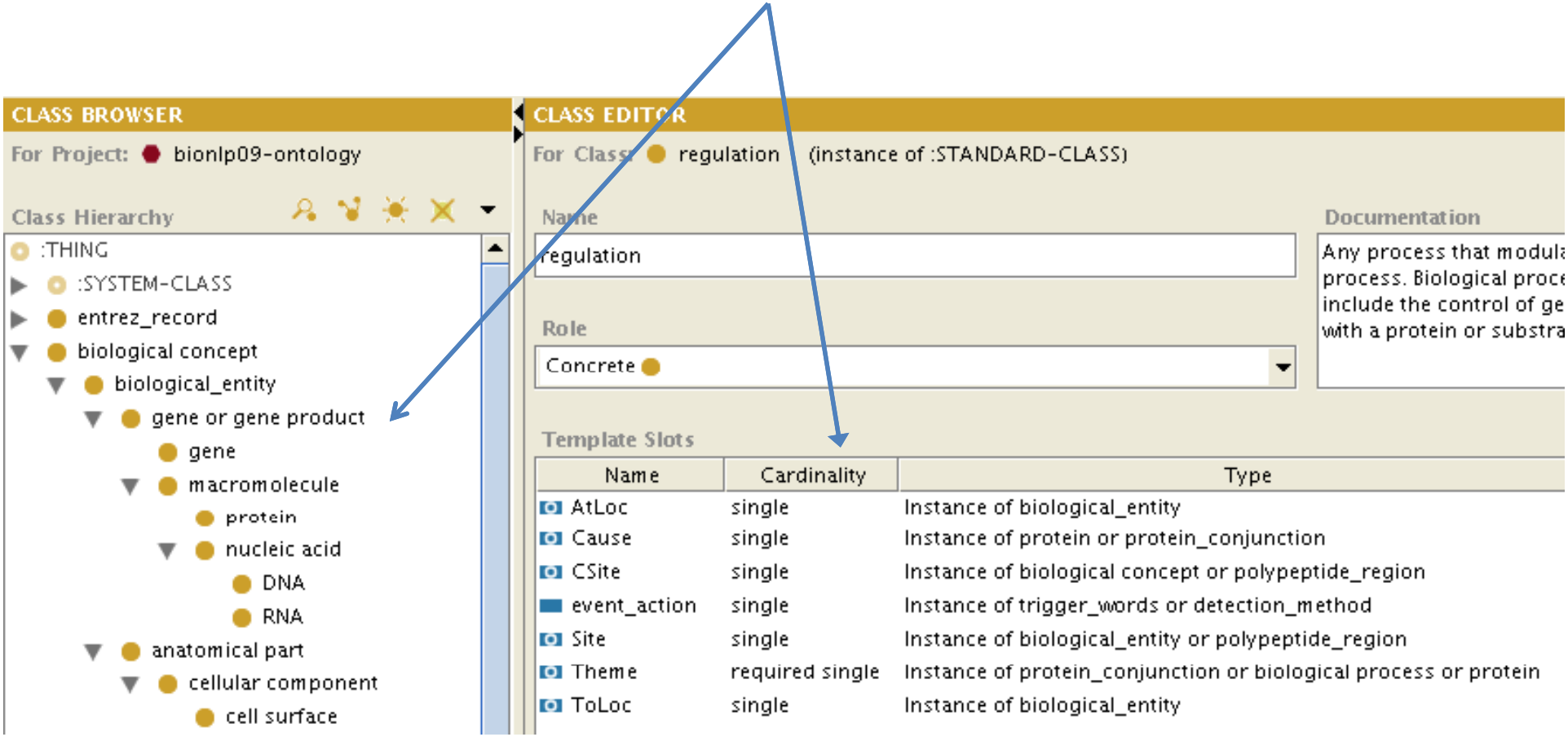


Ontology as the central organizing principle



Ontology as the central organizing principle

Ontology constrains slot fillers in rules



Methods

- Negation and speculation handled by string literals
- Negation: *absence of, failure to, not KEYWORD*
- Speculation: *research, study, examine, investigate, suggests, unknown, seems*
- Heavy use of wildcards