Translational BioNLP

Kevin Bretonnel Cohen

Biomedical Text Mining Group Lead,
Computational Bioscience Program,
U. of Colorado School of Medicine

Adjunct assistant professor,
Department of Linguistics,
U. of Colorado at Boulder

kevin.cohen@gmail.com
http://compbio.ucdenver.edu/Hunter_lab/Cohen

- Opportunity for translational work in biomedical natural language processing—getting results from biological domain into the clinical domain
- Background knowledge and ontologies are now rich enough to leverage in natural language processing

What can 14 years of research into journal articles tell us about clinical documents?

Early observation on clinical versus published text

- Clinical text is noun-dominated, biological journal articles are verb-dominated. (Friedman, Kra, and Rzhetsky 2002)

Definitions

- Nominalization: noun derived from a verb
  - Verbal nominalization: *activation, inhibition, induction*
  - Argument nominalization: *activator, inhibitor, inducer, mutant*

Nominalizations are dominant in biomedical journal articles

Predicate	Nominalization	All verb forms
Express	**2,909**	1,233
Develop	**1,408**	597
Analyze	**1,565**	364
Observe	185	**809**
Differentiate	**737**	166
Describe	10	**621**
Compare	185	**668**
Lose	**556**	74
Perform	86	**599**
Form	**533**	511

Cohen et al. (2008)

Nominalizations are key to translational research

- Translational research: bench-to-bedside
- Nominalizations are frequent and sometimes prevalent in journal articles
- Nominalizations are argument-bearing and an important pivot point in information extraction (Kilicoglu et al. 2010)
- Linguistic study reveals highly complex and interesting patterns of nominalization argument behavior

Definitions

- Argument: participant in or qualifier of the action of the predicate

Argument		Associated prepositions
Arg0	Causer of increase	*after, by, during, in, of*
Arg1	Thing increasing	*in, for, of, with*
Arg2	Amount increased by	*by, in, of, up, with*
Arg3	Start point	*From*
Arg4	End point	*to, with*

Our representation of this predicate is the same as PropBank's.
doi:10.1371/journal.pone.0003158.t001

Cohen et al. (2008)

Argument		Associated prepositions
Arg0	Causer of increase	*after, by, during, in, of*
Arg1	Thing increasing	*in, for, of, with*
Arg2	Amount increased by	*by, in, of, up, with*
Arg3	Start point	*From*
Arg4	End point	*to, with*

Our representation of this predicate is the same as PropBank's.
doi:10.1371/journal.pone.0003158.t001

D0870, an azole antifungal agent[Arg0], *produced dose-related* ***increases*** *in total cytochrome P450 and aldrin epoxidase*[Arg1]

Cohen et al. (2008)

Argument		Associated prepositions
Arg0	Causer of increase	*after, by, during, in, of*
Arg1	Thing increasing	*in, for, of, with*
Arg2	Amount increased by	*by, in, of, up, with*
Arg3	Start point	*From*
Arg4	End point	*to, with*

Our representation of this predicate is the same as PropBank's.
doi:10.1371/journal.pone.0003158.t001

Increase *in phosphorylation of APP* [Arg1] *by overexpression of the nerve growth factor receptor Trk A* [Arg0]

Cohen et al. (2008)

Definitions

- Alternation: variations in the surface syntactic form of predicates and their arguments
  - Active/passive
    - X phosphorylates Y
    - Y is phosphorylated by X
  - Transitive/intransitive
    - X decreases Y
    - Y decreases

Alternations of nominalizations: positions of arguments

- Any combination of the set of positions for each argument of a nominalization
  - Pre-nominal: *phenobarbital* ***induction***, *trkA* ***expression***
  - Post-nominal: ***increases*** *of oxygen*
  - No argument present: ***Induction*** *followed a slower kinetic…*
  - Noun-phrase-external: *this enzyme can undergo* ***activation***

Interesting case: pre-nominal arguments

- Agent (Arg0)
  - cytochrome(s) P-450 **mediation**
  - interferon-gamma **inhibition** of VSV replication
  - phenobarbital **treatment**
- Patient (Arg1, ≅ logical object)
  - trkA **expression**
  - agonist **association**
  - cancer **treatment**

Interesting case: pre-nominal arguments

- Agent (Arg0)
  - cytochrome(s) P-450 **mediation**
  - interferon-gamma **inhibition** of VSV replication
  - phenobarbital **treatment**
- Patient (Arg1, ≅ logical object)
  - trkA **expression**
  - agonist **association**
  - cancer **treatment**

Hypotheses investigated

- Within sublanguage, we should expect a limited number of alternations
- Argument semantic types should be predictable from the restricted semantics of the domain

- *In the [biomedical] sublanguage many of the science-specific verbs have only one or two object possibilities, fewer than their use in English as a whole.* (Sager 1982)
- Subject/Object possibilities are describable as “word classes”:
  - BODY-PART
  - SIGN-SYMPTOM
  - LAB-RESULT (Hirschman and Sager 1982)

Corpus study

- Materials
  - Release 0.9 of the PennBioIE corpus (collection of abstracts of journal articles, annotated with parts of speech, syntactic structure, and entities)
- Methods
  - Marked arguments for 746 tokens of nominalizations of the 10 most common verbs
  - Second annotator marked 15% of these to calculate interannotator agreement

Result1: attested alternations are extraordinarily diverse

- *Inhibition,* a 3-argument predicate—only Arguments 0 and 1 shown

		Arg0			
		Pre	**Post**	**Ext**	**Abs**
Arg1	**Pre**	–	2	8	4
	Post	1	15	16	26
	Ext	1	3	5	1
	Abs	3	2	2	6

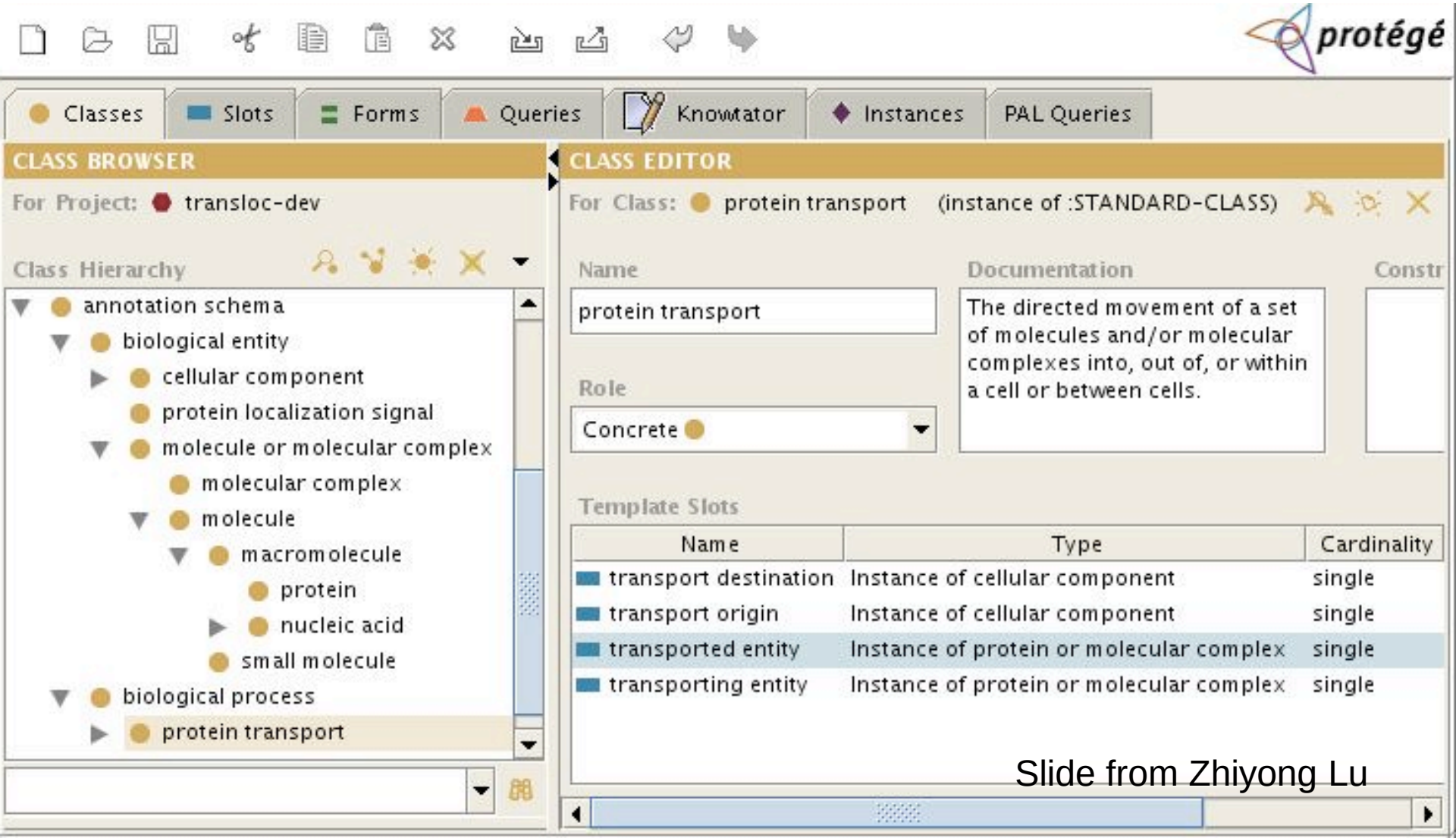
Data is combined from both parts of the BioIE corpus. 24/64 possible patterns are attested in 95 tokens (5 can't-tell).
doi:10.1371/journal.pone.0003158.t032

Cohen et al. (2008)

OpenDMAP

- Rule-based semantic parser
- Open Source
  - Implemented in Java
  - Available at www.sourceforge.net
- All aspects structured by Open Access, community-consensus ontologies—both frames and slot-fillers
- Rules are sophisticated
  - Incorporate syntactic and semantic information
  - Highly flexible with respect to ordering of text
- Strong performance in BioCreative shared task twice, highest precision in BioNLP

Transport Frame: from Gene Ontology



Slide from Zhiyong Lu

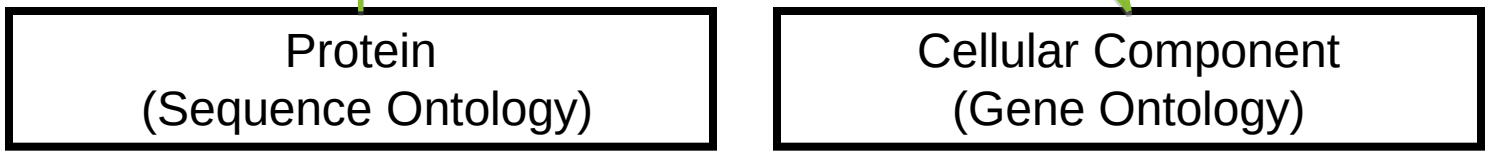
Slot-fillers defined by ontologies

Protein_transport :=

[TRANSPORTED-ENTITY] translocation

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

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



Flexible ordering of text

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

Even a little bit of knowledge helps

- Goal: information extraction about gene activation events
- Initial system: restrict enzyme and substrate to be of semantic type *protein*
- Using external knowledge: restrict slot-fillers as follows:
  - Enzyme: must be *protein* with GOA annotation of *catalytic activity*
  - Substrate: must be protein with GOA annotation of *receptor activity*

Even a little bit of knowledge helps

		Original	Knowledge	Difference
Enzyme	P	0.24	0.37	0.13
	R	0.27	0.20	-0.07
	F	0.26	0.26	0.00
Substrate	P	0.08	0.34	0.26
	R	0.17	0.12	-0.05
	F	0.11	0.18	0.07
Total	P	0.16	0.36	0.20
	R	0.24	0.18	-0.06
	F	0.19	0.24	0.05

Livingston et al. (2010)

Classic question about knowledge-based approaches: Can they scale?

- Agent (Arg0)
  - cytochrome(s) P-450 **mediation**
  - interferon-gamma **inhibition** of VSV replication
  - phenobarbital **treatment**
- Patient (Arg1, ≅ logical object)
  - trkA **expression**
  - agonist **association**
  - cancer **treatment**

UMLS Metathesaurus

⊞ **Concept: [C0747568] PHENOBARBITAL TREATMENT**

⊟ **Semantic Types**

Therapeutic or Preventive Procedure [T061]

⊟ **Atoms (2)** string [AUI / RSAB / TTY / Code]

⊞ PHENOBARBITAL TREATMENT [A1622420/CCPSS/PT/1011824]

⊞ phenobarbital treatment [A18568214/CHV/PT/0000047988]

UMLS Metathesaurus

Therapeutic or Preventive Procedure

Definition

A procedure, method, or technique designed to prevent a disease or a disorder, or to improve

Properties

Unique Identifier: T061

Tree Number: B1.3.1.3

Parents

Health Care Activity

Relations

Therapeutic or Preventive Procedure treats Sign or Symptom (DNI)

Therapeutic or Preventive Procedure prevents Pathologic Function (DNI)

Therapeutic or Preventive Procedure affects Patient or Disabled Group (DNI)

Therapeutic or Preventive Procedure uses Manufactured Object (DNI)

Therapeutic or Preventive Procedure complicates Biologic Function (DNI)

Therapeutic or Preventive Procedure uses Food (DNI)

Therapeutic or Preventive Procedure uses Pharmacologic Substance (DNI)

Therapeutic or Preventive Procedure isa Health Care Activity (DNI)

Therapeutic or Preventive Procedure treats Injury or Poisoning (DNI)

Therapeutic or Preventive Procedure treats Anatomical Abnormality (DNI)

Therapeutic or Preventive Procedure treats Pathologic Function (DNI)

Therapeutic or Preventive Procedure method_of Therapeutic or Preventive Procedure (DNI)

UMLS Metathesaurus

⊞ **Concept: [C0920425] cancer treatment**

⊟ **Semantic Types**

Therapeutic or Preventive Procedure [T061]

⊟ **Definitions**

NCI/PT | Any intervention for management of a malignant neoplasm.

NCI/NCI-GLOSSPT | Treatment to stop or prevent cancer.

⊟ **Atoms (13)** string [AUI / RSAB / TTY / Code]

- ⊞ cancer therapies [A18606744/CHV/SY/0000052576]
- ⊞ cancer therapy [A18569495/CHV/PT/0000052576]
- ⊞ cancer therapy [A1962166/CSP/ET/2030-1247]
- ⊞ Cancer treatment [A19509121/LNC/CN/MTHU038130]
- ⊞ Cancer treatment [A19510443/LNC/LPN/LP114959-2]
- ⊞ cancer treatment [A15933670/MEDCIN/PT/304181]

cancer treatment (treatment) [A15967700/MEDCIN/FN/304181]

UMLS Metathesaurus

Therapeutic or Preventive Procedure

Definition

A procedure, method, or technique designed to prevent a disease or a disorder, or to improve

Properties

Unique Identifier: T061

Tree Number: B1.3.1.3

Parents

Health Care Activity

Relations

Therapeutic or Preventive Procedure treats Sign or Symptom (DNI)

Therapeutic or Preventive Procedure prevents Pathologic Function (DNI)

Therapeutic or Preventive Procedure affects Patient or Disabled Group (DNI)

Therapeutic or Preventive Procedure uses Manufactured Object (DNI)

Therapeutic or Preventive Procedure complicates Biologic Function (DNI)

Therapeutic or Preventive Procedure uses Food (DNI)

Therapeutic or Preventive Procedure uses Pharmacologic Substance (DNI)

Therapeutic or Preventive Procedure isa Health Care Activity (DNI)

Therapeutic or Preventive Procedure treats Injury or Poisoning (DNI)

Therapeutic or Preventive Procedure treats Anatomical Abnormality (DNI)

Therapeutic or Preventive Procedure treats Pathologic Function (DNI)

Therapeutic or Preventive Procedure method_of Therapeutic or Preventive Procedure (DNI)

Corresponding rules

- *Treat(ment)*
  - Arg0: Provider
  - Arg1: Condition
  - Arg2: Treatment

  - *phenobarbital treatment*
  - [pharmacologic substance] treatment
  - Arg2: phenobarbital
-
  - *cancer treatment*
  - [disease] treatment
  - Arg1: cancer

Nota bene

- The point isn't that *phenobarbital treatment* and *drug treatment* and *cancer treatment* are in the UMLS Metathesaurus—the point is that *Therapeutic or Preventive Procedure* has the *treats* and *uses* relations and maps them to different semantic categories.

Nominalizations and relational nouns turn out to be important in clinical text, too

- Domain knowledge allows recovery of missing arguments—clinical language ellipsis is different from General English ellipsis
  - *Had previous hospitalization.*
- Relational nouns allow inference of metonomycally implied arguments necessary for full representation of meaning
  - *Dorsal spine shows* implies *X-ray, MRI,* etc.

(Hirschman and Sager

- MedLEE and GENIES: same processing engine, two sublanguages.

Conclusions

- Linguistic approaches are revealing and provide means for translating insights from history of research on journal articles into future work on clinical documents
- Knowledge-based approaches are now feasible in the biomedical domain
- Both leverage considerable previous and current NIH investment in NLP, ontology, and knowledge base development

Acknowledgments

- NIH 5R01 LM009254-06 (Hunter, PI)
- NIH 5R01 LM008111-07 (Hunter, PI)
- NIH 5R01 GM083649-04 (Hunter, PI)
- Stephen Johnson, Columbia U.
- Michael Kahn, The Children’s Hospital