CITY UNIVERSITY
LONDON

The University for business and the professions

CITY UNIVERSITY
LONDON

Annotating biomedical terms in GATE

Phil Gooch

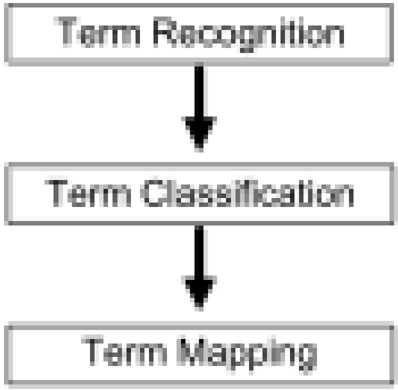
Philip.Gooch.1@city.ac.uk

10 May 2010

Background

- Term recognition: *identification* of lexical units (generally noun phrases) that are related to domain concepts. (Krauthammer et al 2004)
  - Differentiate between terms and non-terms.

- Concept recognition: *mapping* of text strings to an ontology or thesaurus of inter-related, classified concepts (Shah et al 2009)
  - Need to provide an *unambiguous* semantic representation of what the string denotes – not enough to say that a text string is a gene or a disease (Baumgartner et al 2008)



Term Recognition

Term Classification

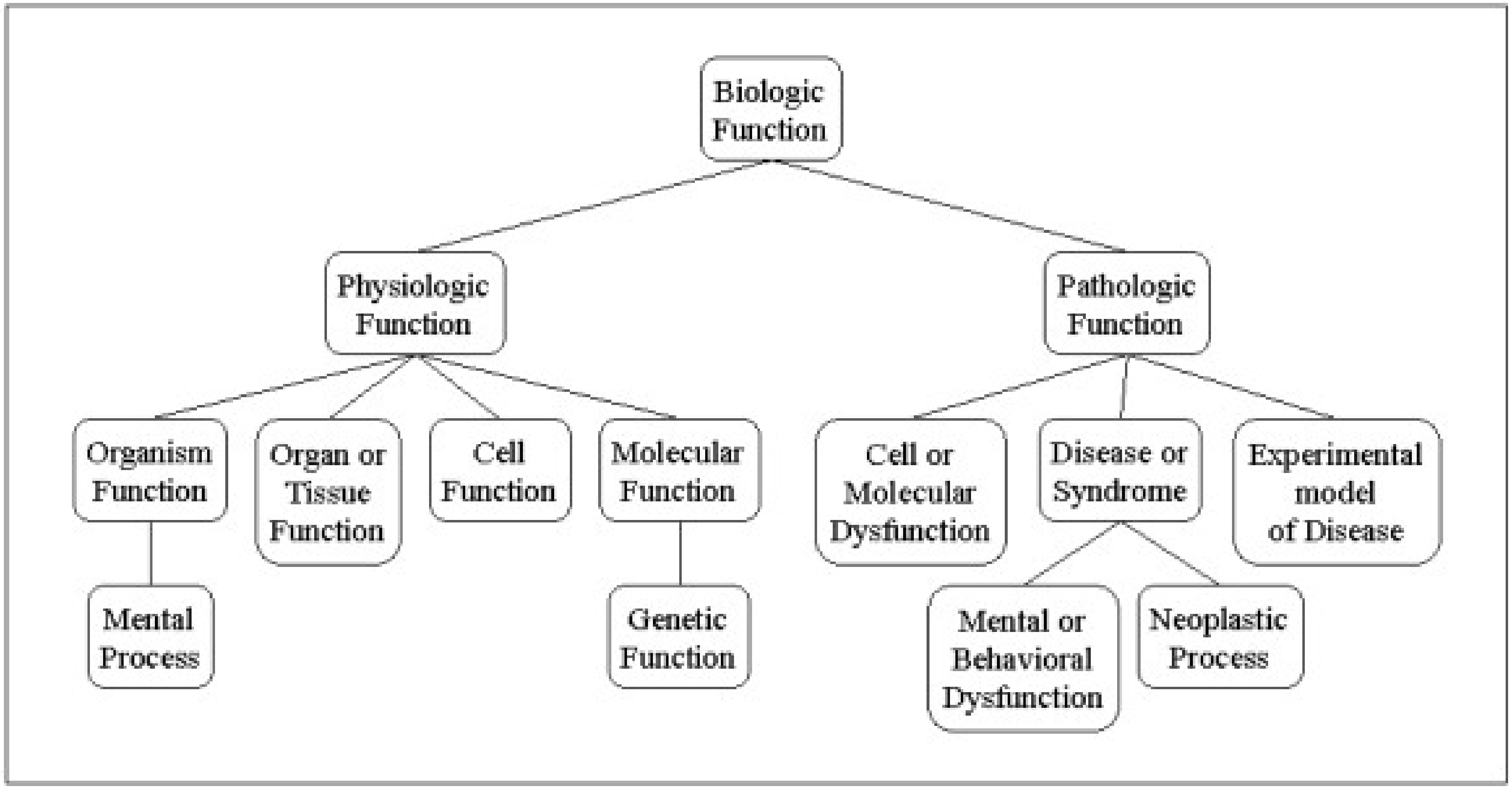
Term Mapping

Three steps to term identification: Krauthammer et al 2004, p. 513

1. Recognise the text string as being a (possible) term
2. Classify the text string (e.g. disease, drug)
3. Map the term to a concept(s) within an agreed data source

(ontology, thesaurus) by assigning concept identifier from the data source

- United Medical Language System (UMLS) Metathesaurus is one of the most comprehensive and widely used thesauri in the biomedical domain.
- Integrates and cross-references a number of **source vocabularies**, such as SNOMED-CT (clinical terms), LOINC (lab observations), FMA (anatomy), RxNorm (drug names) and many others
- Concepts are categorised according to a set of broad **semantic types**
- Each concept has a 'concept unique identifier' CUID (although the concept may exist in many source vocabularies with their own interal IDs)



Biologic Function

Physiologic Function

Pathologic Function

Organism Function

Organ or Tissue Function

Cell Function

Molecular Function

Cell or Molecular Dysfunction

Disease or Syndrome

Experimental model of Disease

Mental Process

Genetic Function

Mental or Behavioral Dysfunction

Neoplastic Process

http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?

book=nlmumls&part=ch05&rendertype=figure&id=ch05.F1

- National Library of Medicine **MetaMap** software is considered to be the 'gold standard' for recognising UMLS concepts

  + full text parsing (sentence, phrase, token, POS)

  + NegEx algorithm (Chapman et al 2002)

  + can limit annotation to specific vocabularies and specific semantic types

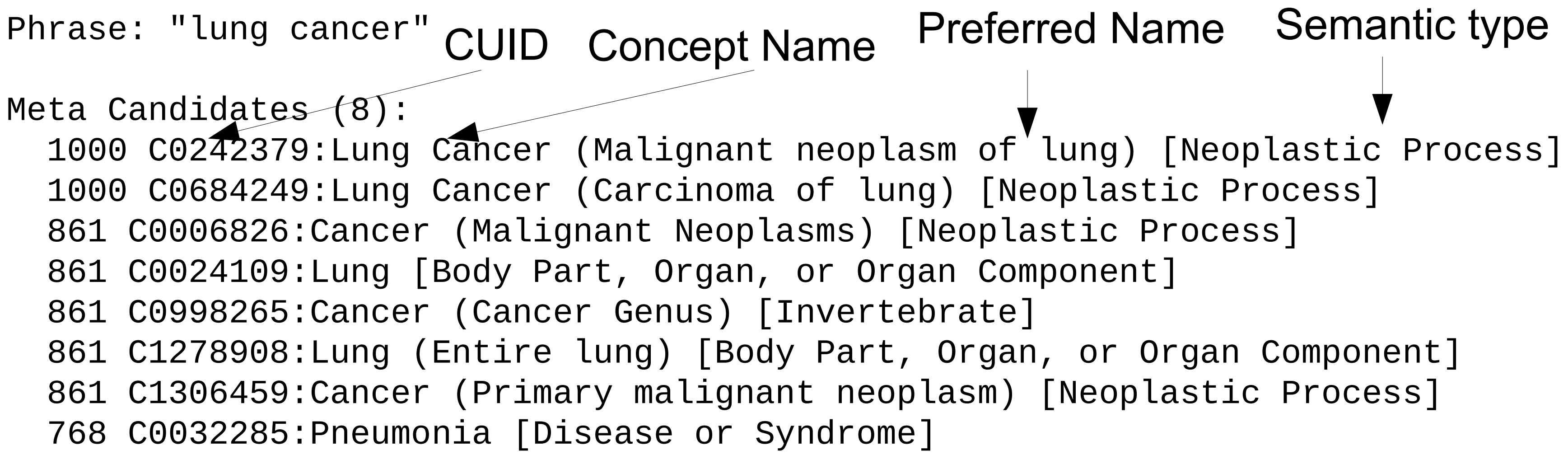
  + Java API

  - ASCII input text only

  - *Nix binaries only

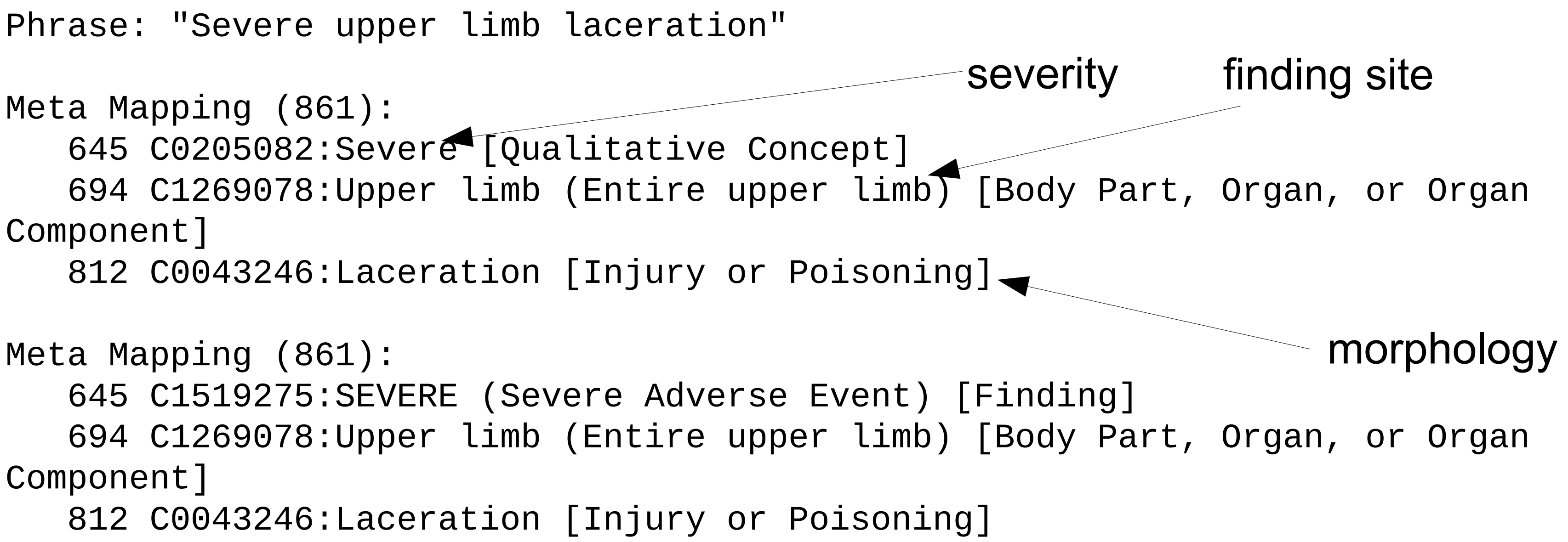
  - performance, memory requirements, input (MEDLINE abstracts)

Example MetaMap mapping: noun phrase to noun phrase



```
Phrase: "lung cancer"

Meta Candidates (8):
  1000 C0242379:Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
  1000 C0684249:Lung Cancer (Carcinoma of lung) [Neoplastic Process]
  861 C0006826:Cancer (Malignant Neoplasms) [Neoplastic Process]
  861 C0024109:Lung [Body Part, Organ, or Organ Component]
  861 C0998265:Cancer (Cancer Genus) [Invertebrate]
  861 C1278908:Lung (Entire lung) [Body Part, Organ, or Organ Component]
  861 C1306459:Cancer (Primary malignant neoplasm) [Neoplastic Process]
  768 C0032285:Pneumonia [Disease or Syndrome]

Meta Mapping (1000):
  1000 C0684249:Lung Cancer (Carcinoma of lung) [Neoplastic Process]
Meta Mapping (1000):
  1000 C0242379:Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
```

Example MetaMap mapping: noun phrase to individual terms

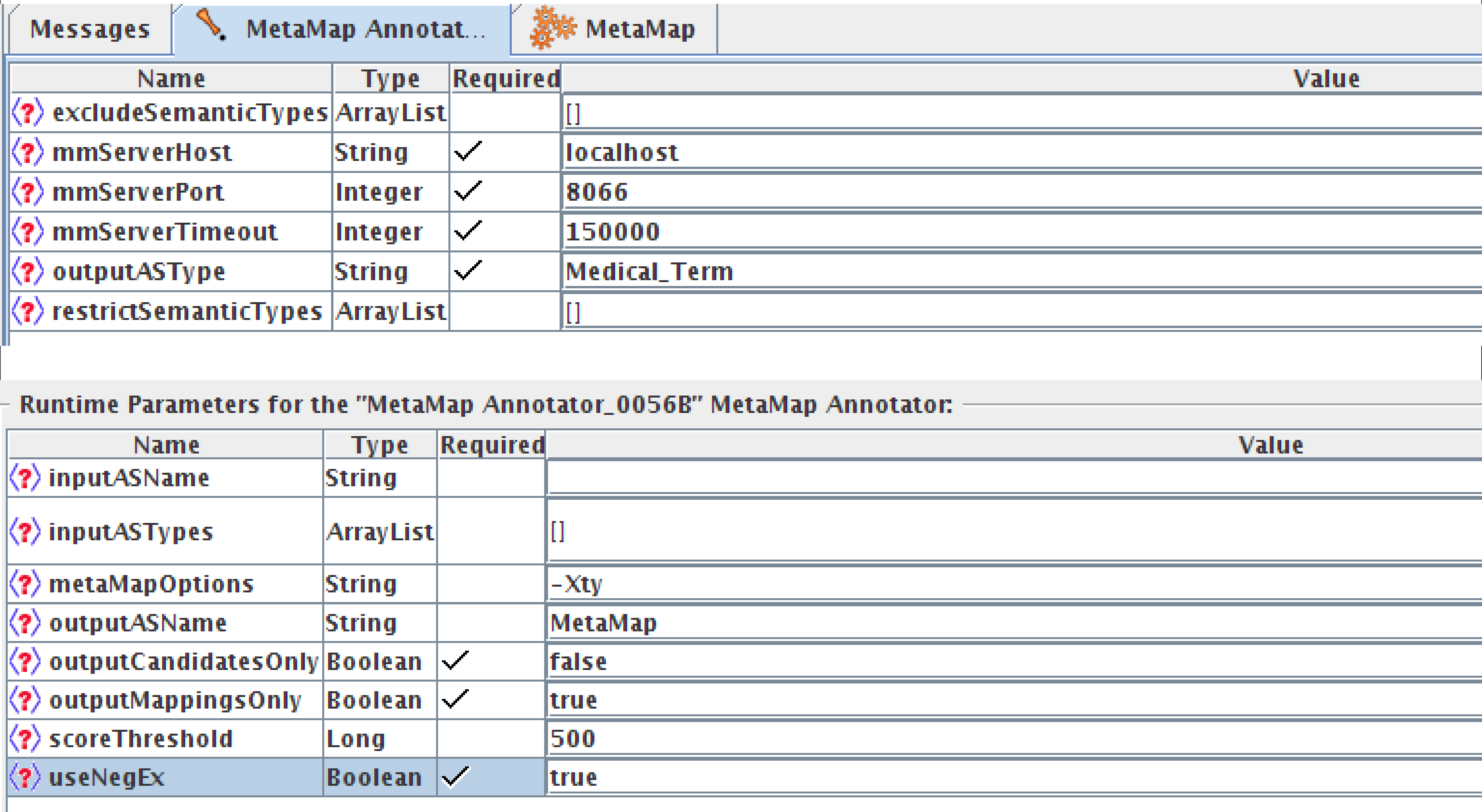


```
Phrase: "Severe upper limb laceration"

Meta Mapping (861):
   645 C0205082:Severe [Qualitative Concept]
   694 C1269078:Upper limb (Entire upper limb) [Body Part, Organ, or Organ
Component]
   812 C0043246:Laceration [Injury or Poisoning]

Meta Mapping (861):
   645 C1519275:SEVERE (Severe Adverse Event) [Finding]
   694 C1269078:Upper limb (Entire upper limb) [Body Part, Organ, or Organ
Component]
   812 C0043246:Laceration [Injury or Poisoning]
```

This can be useful for *post-coordination* of terms already identified, but problematic for generating annotations from free text as a single noun phrase can generate > 1 annotation (one for each MetaMap mapping)

CITY UNIVERSITY
LONDON

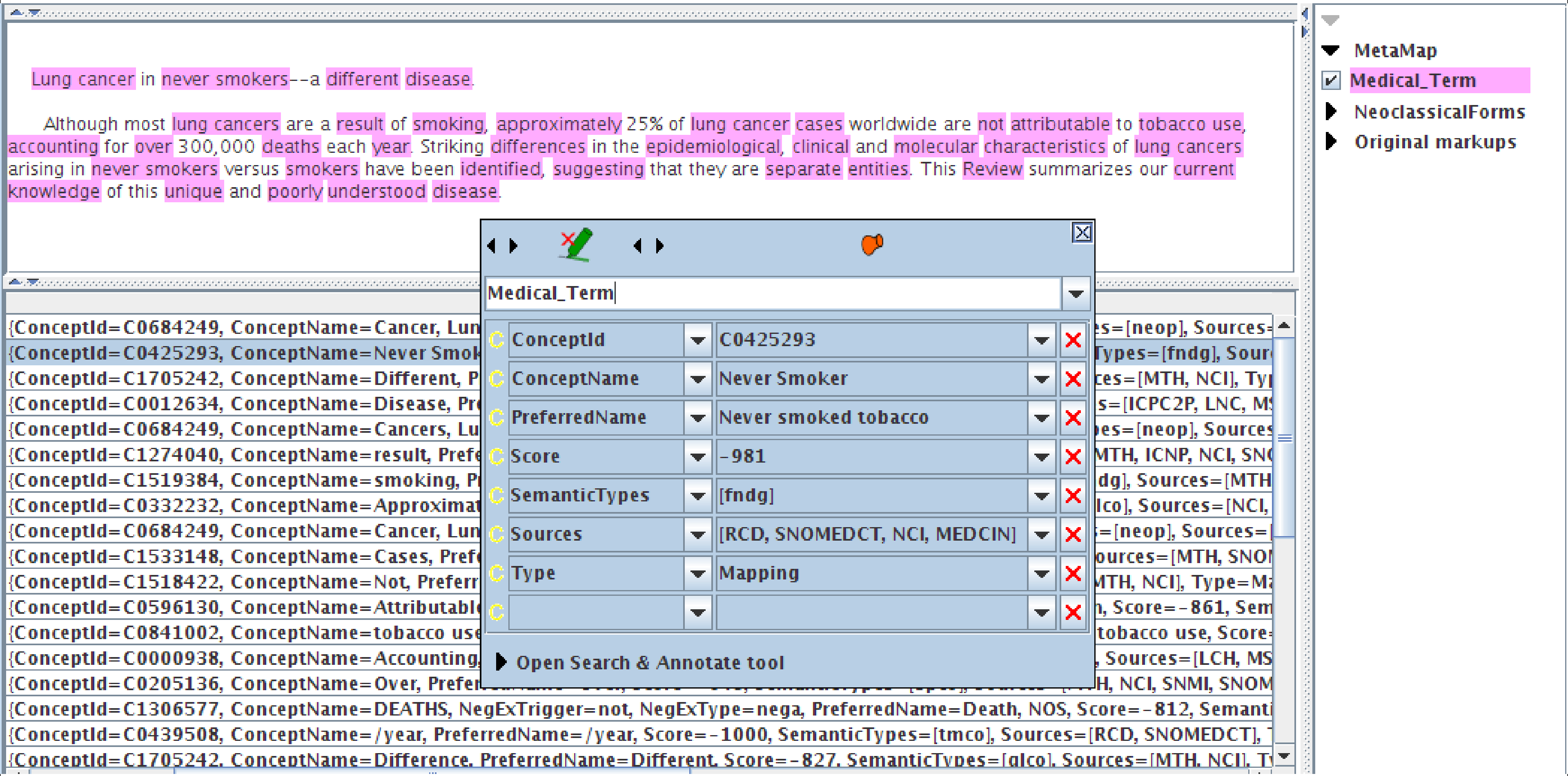
Demonstration of MetaMap GATE plugin



Name	Type	Required	Value
excludeSemanticTypes	ArrayList		[]
mmServerHost	String	✓	localhost
mmServerPort	Integer	✓	8066
mmServerTimeout	Integer	✓	150000
outputASType	String	✓	Medical_Term
restrictSemanticTypes	ArrayList		[]

Runtime Parameters for the "MetaMap Annotator_0056B" MetaMap Annotator:

Name	Type	Required	Value
inputASName	String		
inputASTypes	ArrayList		[]
metaMapOptions	String		-Xty
outputASName	String		MetaMap
outputCandidatesOnly	Boolean	✓	false
outputMappingsOnly	Boolean	✓	true
scoreThreshold	Long		500
useNegEx	Boolean	✓	true



Lung cancer in never smokers--a different disease.

Although most lung cancers are a result of smoking, approximately 25% of lung cancer cases worldwide are not attributable to tobacco use, accounting for over 300,000 deaths each year. Striking differences in the epidemiological, clinical and molecular characteristics of lung cancers arising in never smokers versus smokers have been identified, suggesting that they are separate entities. This Review summarizes our current knowledge of this unique and poorly understood disease.

- MetaMap
- ☑ Medical_Term
- NeoclassicalForms
- Original markups

Medical_Term

ConceptId	C0425293
ConceptName	Never Smoker
PreferredName	Never smoked tobacco
Score	-981
SemanticTypes	[fndg]
Sources	[RCD, SNOMEDCT, NCI, MEDCIN]
Type	Mapping

Open Search & Annotate tool

```
{ConceptId=C0684249, ConceptName=Cancer, Lun                                            es=[neop], Sources=
{ConceptId=C0425293, ConceptName=Never Smok                                             Types=[fndg], Sour
{ConceptId=C1705242, ConceptName=Different, P                                           ces=[MTH, NCI], Ty
{ConceptId=C0012634, ConceptName=Disease, Pr                                            s=[ICPC2P, LNC, MS
{ConceptId=C0684249, ConceptName=Cancers, Lu                                            pes=[neop], Sources
{ConceptId=C1274040, ConceptName=result, Prefe                                          MTH, ICNP, NCI, SN
{ConceptId=C1519384, ConceptName=smoking, Pr                                            dg], Sources=[MTH
{ConceptId=C0332232, ConceptName=Approximat                                             lco], Sources=[NCI,
{ConceptId=C0684249, ConceptName=Cancer, Lun                                            =[neop], Sources=
{ConceptId=C1533148, ConceptName=Cases, Prefe                                           ources=[MTH, SNO
{ConceptId=C1518422, ConceptName=Not, Preferr                                           MTH, NCI], Type=M
{ConceptId=C0596130, ConceptName=Attributabl                                            n, Score=-861, Sem
{ConceptId=C0841002, ConceptName=tobacco use                                            tobacco use, Score=
{ConceptId=C0000938, ConceptName=Accounting,                                            , Sources=[LCH, MS
{ConceptId=C0205136, ConceptName=Over, Prefer                                           H, NCI, SNMI, SNOM
{ConceptId=C1306577, ConceptName=DEATHS, NegExTrigger=not, NegExType=nega, PreferredName=Death, NOS, Score=-812, Semanti
{ConceptId=C0439508, ConceptName=/year, PreferredName=/year, Score=-1000, SemanticTypes=[tmco], Sources=[RCD, SNOMEDCT], T
{ConceptId=C1705242, ConceptName=Difference, PreferredName=Different, Score=-827, SemanticTypes=[glco], Sources=[MTH, NCI], Ty
```

Idea ...

- If we could pre-identify 'candidate' biomedical noun-phrase terms, we could pass each of these to MetaMap for validation, addition of metadata (UMLS CUID, preferred name, semantic type etc), and post-coordination
- Useful for large texts where biomedical terms are quite sparse
- Useful for disambiguation (text might contain non-biomedical terms, e.g. organisations, people, places)

Term identification – differentiate the drug names

zaclovir

oxymoron

bactiflox

propanolol

xylophone

xylocaine

triumvir

oxymycin

orthodox

protocol

How did you decide?

Term identification – prefixes and suffixes?

- **oxy-mycin**
- **oxy**moron
- zaclo-**vir**
- trium**vir**

- **bacti**-fl-**ox**
- orthod**ox**
- **propan**-ol-**ol**
- protoc**ol**

- **xyl-**o-**caine**
- **xyl**ophone

Neoclassical combining forms (NCF)

- Since 16th C, naming of scientific terms (chemical, biological) has involved use of Latin and Greek *morphemes* (linguistic unit that has semantic meaning)
- **Leukocytosis**: leuk- (white) -cyte (cell) -osis (disease)
- Computational analysis of NCF has been around since early 1980s
  - can narrow range of possible meanings
  - provide semantic classification (e.g. -itis, inflammation => disease)
  - can help identify unknown words
  - provides a 'best guess' (i.e. provide candidate terms) for human review

  (McCray 1988)

Source of Latin/Greek morphemes for NCF

- Wikipedia:
http://en.wikipedia.org/wiki/List_of_medical_roots,_suffixes_and_prefixes

- NLM Specialist Lexicon Database (NC.DB)
http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html

  - separates NCF morpheme types into *prefix*, *root*, *terminal*
  - prefix (normally) precedes root and cannot attach directly to a terminal

abdomin(o)|abdomen|root
ab|away from|prefix
desis|binding|terminal

Strategy for NCF term identification in GATE

- Create gazetteer of NCF morphemes
- Write JAPE rules to identify Tokens that contain them in the correct order (prefix*, root+, terminal?)
- Annotate the surrounding NP that contains the candidate NCF Token
- Pass the candidate NPs to MetaMap
- Convert the MetaMap output to GATE Annotations and features
- Use Corpus QA tools to measure recall and precision by comparing candidate NCF annotations with MetaMap annotations

Gazetteer with features; wholeWordsOnly=false, longestMatch=true

```
dipl;rel=two;type=root
dipso;rel=thirst;type=root
dips;rel=thirst;type=root
disco;rel=disk shaped;type=root
disc;rel=disk shaped;type=root
disko;rel=disk shaped;type=root
disk;rel=disk shaped;type=root
dis;rel=negate;type=prefix
diverticulo;rel=diverticulum;type=root
diverticul;rel=diverticulum;type=root
di;rel=two;type=prefix
```

JAPE rules

Simple approach:

```
{Token.category == NN, Lookup.majorType == neoclassical_forms,
Lookup.type == prefix,
Token contains {Lookup.type == root},
Token contains {Lookup.type == terminal}
} |
{Token.category == NN, Lookup.majorType == neoclassical_forms,
Lookup.type == root,
Token contains {Lookup.type == terminal}
} |
...
```

JAPE rules

Better approach: create new JAPE operators startsWith and endsWith

{Token **startsWith** {Lookup.type == prefix}, Token contains {Lookup.type == root}, Token **endsWith** {Lookup.type == terminal} }

Or use Java on the RHS to check for the correct positioning of prefix, root and terminal Lookups within the Token

Implementation issues

- Overlapping lookups:

  leuko-cyte

  nano-particle

- Morphemes have multiple roles (sometimes a root, sometimes a suffix)

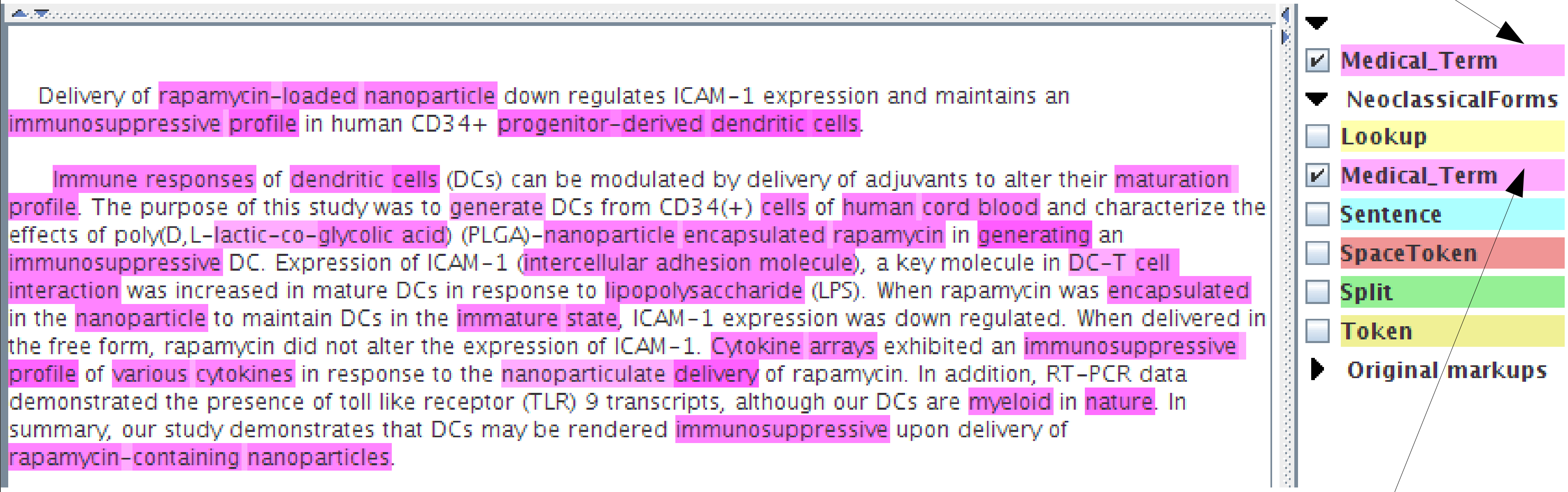
  ad (prefix: **ad**duction)   ad (suffix: dors**ad**, mon**ad**)

- Some roots are common non-NCF morphemes:

em (blood)	oo (egg)	or (mouth)
embolism	oocyte	orifice
emphasis	oozing	orchestra

Results

MetaMap

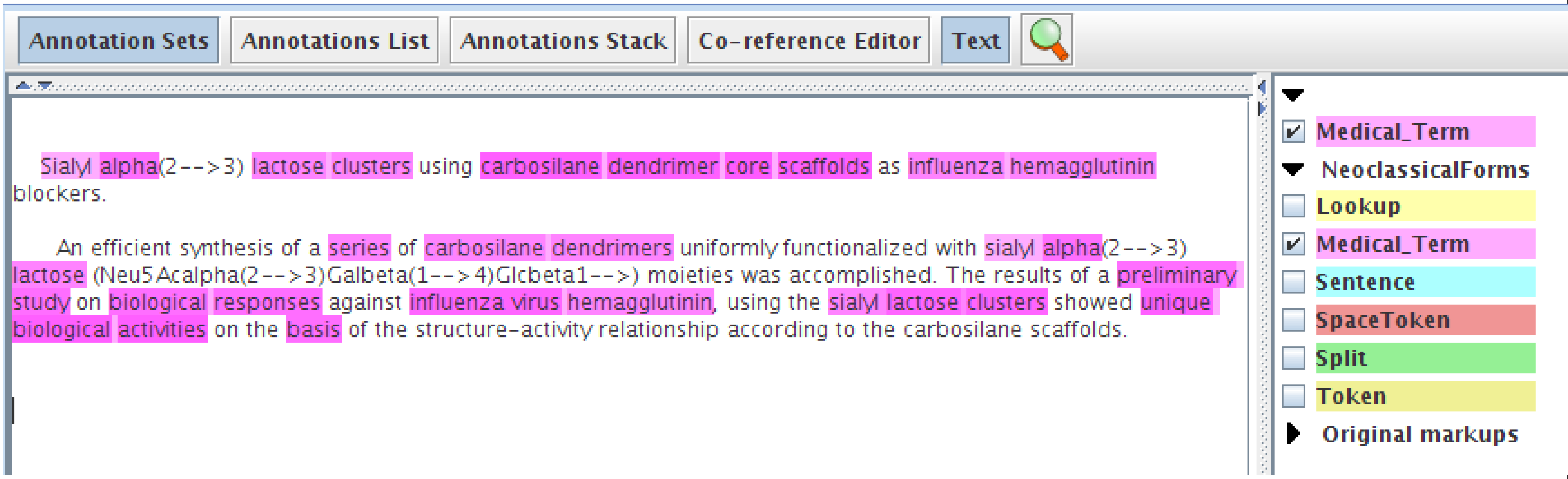


Delivery of rapamycin-loaded nanoparticle down regulates ICAM-1 expression and maintains an immunosuppressive profile in human CD34+ progenitor-derived dendritic cells.

Immune responses of dendritic cells (DCs) can be modulated by delivery of adjuvants to alter their maturation profile. The purpose of this study was to generate DCs from CD34(+) cells of human cord blood and characterize the effects of poly(D,L-lactic-co-glycolic acid) (PLGA)-nanoparticle encapsulated rapamycin in generating an immunosuppressive DC. Expression of ICAM-1 (intercellular adhesion molecule), a key molecule in DC-T cell interaction was increased in mature DCs in response to lipopolysaccharide (LPS). When rapamycin was encapsulated in the nanoparticle to maintain DCs in the immature state, ICAM-1 expression was down regulated. When delivered in the free form, rapamycin did not alter the expression of ICAM-1. Cytokine arrays exhibited an immunosuppressive profile of various cytokines in response to the nanoparticulate delivery of rapamycin. In addition, RT-PCR data demonstrated the presence of toll like receptor (TLR) 9 transcripts, although our DCs are myeloid in nature. In summary, our study demonstrates that DCs may be rendered immunosuppressive upon delivery of rapamycin-containing nanoparticles.

Neoclassic combining forms

Results



Sialyl alpha(2-->3) lactose clusters using carbosilane dendrimer core scaffolds as influenza hemagglutinin blockers.

An efficient synthesis of a series of carbosilane dendrimers uniformly functionalized with sialyl alpha(2-->3) lactose (Neu5Acalpha(2-->3)Galbeta(1-->4)Glcbeta1-->) moieties was accomplished. The results of a preliminary study on biological responses against influenza virus hemagglutinin, using the sialyl lactose clusters showed unique biological activities on the basis of the structure-activity relationship according to the carbosilane scaffolds.

▼
☑ Medical_Term
▼ NeoclassicalForms
☐ Lookup
☑ Medical_Term
☐ Sentence
☐ SpaceToken
☐ Split
☐ Token
▶ Original markups

Results: NCF NP terms vs MetaMap mappings

Annotation	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1.0-lenient
Lookup	0	0	142807	0	1.00	0.00	0.00
Medical_Term	5694	82726	890	6922	**0.13**	**0.93**	0.23

Results: NCF NP terms validated against MetaMap

Annotation	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1.0-lenient
Lookup	0	0	142807	0	1.00	0.00	0.00
Medical_Term	6317	19498	812	6377	**0.39**	**0.94**	0.56

Further work

- Classify NCF morphemes into 'strong' and 'weak' roots and terminals

  - *Strong* roots: presence on their own strongly indicates a term, e.g. 'cyte', 'cyto' → cytoplasm, leukocyte

  - *Weak* roots: requires a strong terminal or a co-occurring strong root to indicate a term, e.g. 'oo' + 'cyte', 'cyto' → oocyte, oocytosis

- Classify NCF morphemes into semantic types: parts of body, symptoms, procedures. E.g. -ectomy (excision → procedure), hepat- (liver → organ), -phyma (swelling → symptom)

Further work

- Combine NCF patterns with Hearst patterns (NP, such as NP*)
- Combine NCF patterns with abbreviation-matching heuristics, e.g. ALICE (Ao & Takagi 2005)

gastro-oesophageal reflux disease (GORD)

Acknowledgements

Paul Appleby (introducing GATE)

Angus Roberts (NLM Specialist Lexicon and MetaMap)

Diana Maynard (JAPE, papers on annotating without gazetteers)

Ian Roberts (GATE Java API)

Mark Greenwood (writing GATE plugins)

Niraj Aswani (OntoRoot Gazetteer)

And everyone in the GATE team and GATE user list

References

Ao H & Takagi T (2005) ALICE: An Algorithm to Extract Abbreviations from MEDLINE. JAMIA 12(5) 576-586

Baumgartner WA, Lu Z et al (2008) Concept recognition for extracting protein interaction relations from biomedical text. Genome Biology 9(Suppl): 59

Chapman W, Bridewell W et al (2002) A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inf 34: 301-310

Krauthammer M, Nenadic G (2004) Term identification in the biomedical literature. J Biomed Inf 37: 512-526

McCray AT, Browne AC, Moore DL (1988) The Semantic Structure of Neo-Classical Compounds. Proc of Annual Symp on Comp Appl in Med Care; 1998. p. 165–8

Shah NH, Bhatia N, et al (2009) Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics 10(Suppl): S14