NLP-Based Mapping of Textbook Pathology to the Ontology for General Medical Science (OGMS)

Lindsay Cowell and Richard Scheuermann
University of Texas Southwestern Medical Center

Sanda Harabagiu, Bryan Rink, and Kirk Roberts
University of Texas at Dallas

Mathias Brochausen and Bill Hogan
University of Arkansas for Medical Sciences

Project Goal

- Information about physiology and pathology exists primarily as natural language

- Some computable representations (e.g. FMA, SNoMed), but …
  - Not connected
  - Not interoperable
  - Critical gaps in coverage

Project Goal

- Develop the Human Pathology Network (HPN)
  - a computable representation of human pathology

  - Accommodates different disease types
  - Spans biological scales – from molecules to clinical phenotypes
  - Connects pathological entities to their normal counterparts

Approach

- Ontology-driven NLP applied to “Robbins & Cotran Pathologic Basis of Disease”
  - Manual annotation
  - Active learning NLP
  - Ontology based representation of extracted information

Approach

- Ontology-driven NLP applied to “Robbins & Cotran Pathologic Basis of Disease”
  - Manual annotation
  - Active learning NLP
  - Ontology based representation of extracted information

Approach

- Ontology-driven NLP applied to “Robbins & Cotran Pathologic Basis of Disease”
  - Manual annotation
  - Active learning NLP
  - Ontology based representation of extracted information

Approach

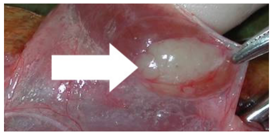
- Ontology-driven NLP applied to “Robbins & Cotran Pathologic Basis of Disease”
  - Manual annotation
  - Active learning NLP
  - Ontology based representation of extracted information

Approach

- Ontology-driven NLP applied to "Robbins & Cotran Pathologic Basis of Disease"
  - Manual annotation
  - Active learning NLP
  - Ontology based representation of extracted information

Example

Relevant Entities

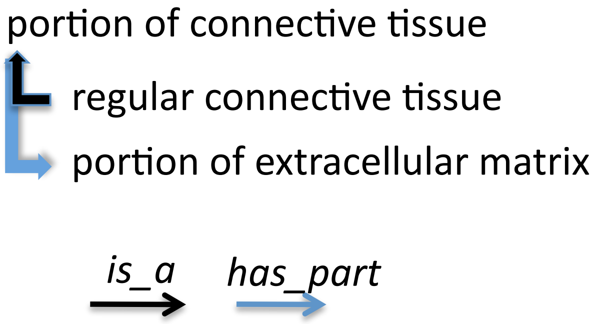


Textual Description of Such Entities

All tumors, benign and malignant, have two basic components: (1) clonal neoplastic cells that constitute their parenchyma and (2) reactive stroma made up of connective tissue, blood vessels, and variable numbers of macrophages and lymphocytes.

Ontologies

e.g. Foundational Model of Anatomy



portion of connective tissue

regular connective tissue

portion of extracellular matrix

is_a *has_part*

Linguistic Resources

e.g. FrameNet

•**Inclusion frame**: A total has a part, either as a member of an aggregate or as a constituent part of a simple entity.

- •Frame elements: part, total
- •Lexical units: contain (verb), exclude (verb), excluding (preposition), have (verb), include (verb), including (preposition), inclusive (adjective), incorporate (verb), integrated (adjective)

Manual Annotation

All tumors, benign and malignant, have two basic components:(1) clonal neoplastic cells that constitute their parenchyma and (2) reactive stroma made up of connective tissue, blood vessels, and variable numbers of macrophages and lymphocytes.

	Term from Text	**Ontology Term**
FMA	parenchyma	parenchyma
	stroma	stroma
	connective tissue	portion of connective tissue
	blood vessel	vein
CL	macrophage	macrophage
	lymphocyte	lymphocyte
NCIT	tumor	neoplasm
	benign	benign neoplasm
	malignant	malignant neoplasm
	neoplastic cell	neoplastic cell

Manual Annotation

All tumors, benign and malignant, have two basic components:(1) clonal neoplastic cells that constitute their parenchyma and (2) reactive stroma made up of connective tissue, blood vessels, and variable numbers of macrophages and lymphocytes.

has_part	*part_of*
→	→

Manual Annotation

[[**All**]QUANTIFIER **tumors**]**PART-OF_FRAME(1):FE=WHOLE**, benign and malignant, **have** [[**two**]CARDINAL-1 **basic components**]**PART-OF_FRAME (1):LU** : [(1)ORDINAL@CARDINAL-1 clonal neoplastic cells that constitute their [**parenchyma**]**PART-OF_FRAME(1):FE=PART**] and [(2) ORDINAL@CARDINAL-1 [**reactive stroma**]**PART-OF_FRAME(1):FE=PART**] made up of connective tissue, blood vessels, and variable numbers of macrophages and lymphocytes.

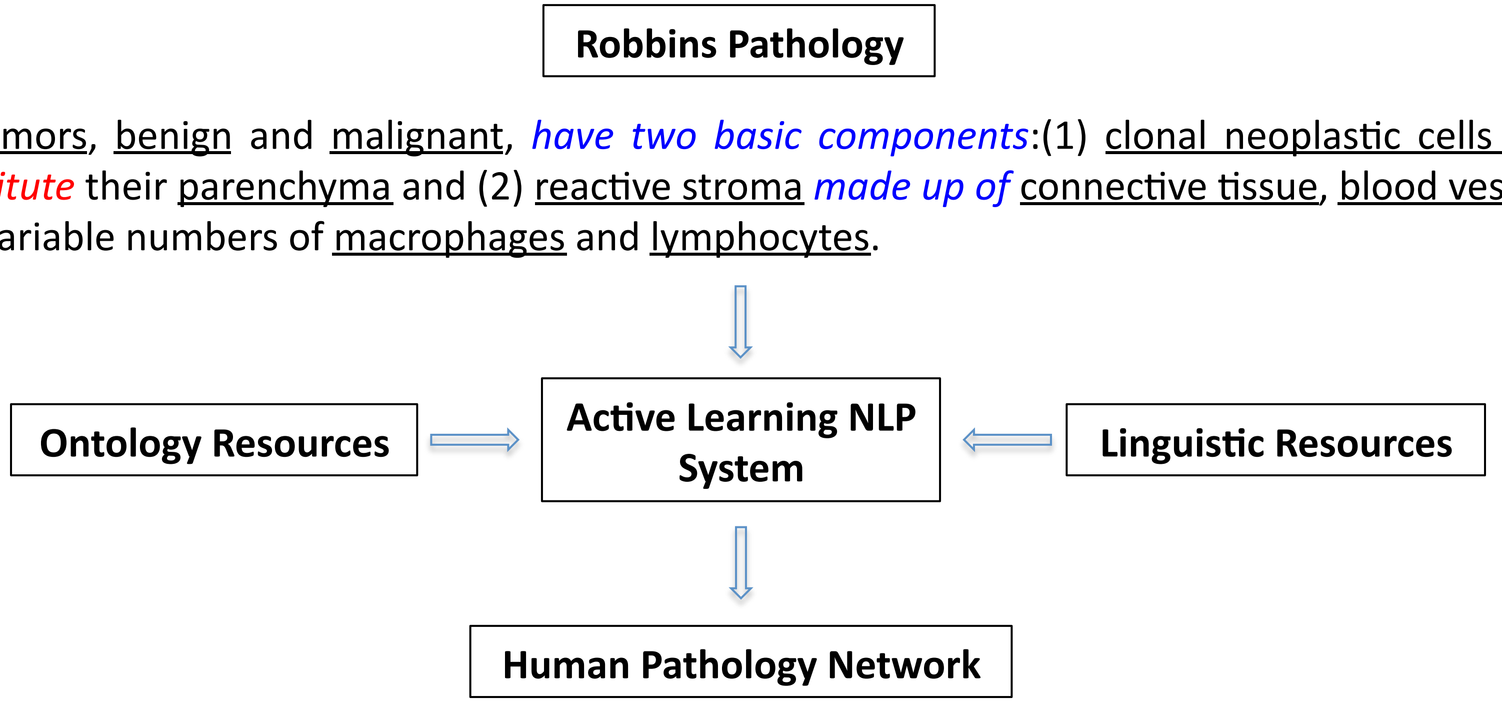
•**Inclusion frame**: A total has a part, either as a member of an aggregate or as a constituent part of a simple entity.

- •Frame elements: part, total
- •Lexical units: contain (verb), exclude (verb), excluding (preposition), have (verb), include (verb), including (preposition), inclusive (adjective), incorporate (verb), integrated (adjective)

Active Learning NLP

- Annotations (ours, Genia Corpus, …)
- AL guides the annotation process
  - Selection
  - Presentation
  - Validation or correction
  - Incremental training
- Selection based on informativeness
- Machine learning methods
  - Developed by UTD team for the i2b2 2010 and 2011

Representation of Extracted Text



Robbins Pathology

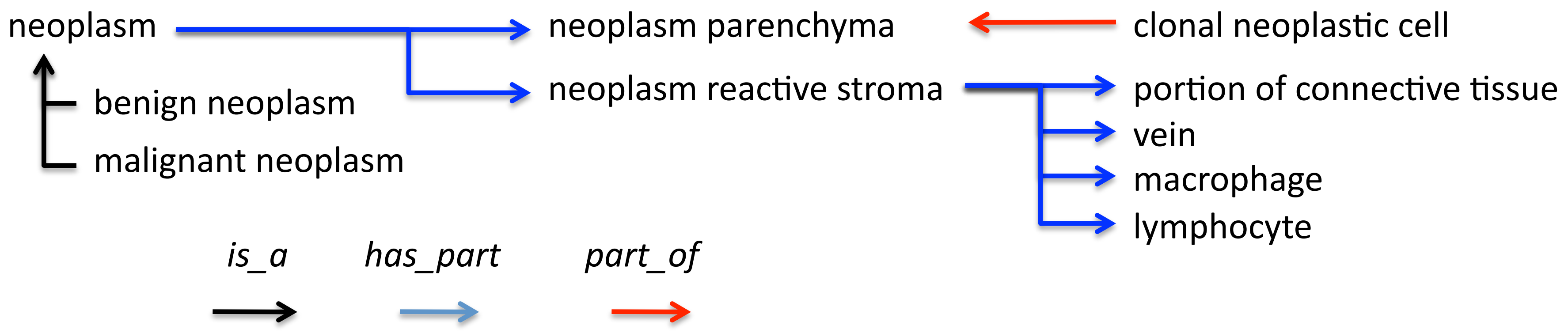
All tumors, benign and malignant, *have two basic components*:(1) clonal neoplastic cells that *constitute* their parenchyma and (2) reactive stroma *made up of* connective tissue, blood vessels, and variable numbers of macrophages and lymphocytes.

Ontology Resources

Active Learning NLP System

Linguistic Resources

Human Pathology Network



neoplasm

benign neoplasm

malignant neoplasm

neoplasm parenchyma

neoplasm reactive stroma

clonal neoplastic cell

portion of connective tissue

vein

macrophage

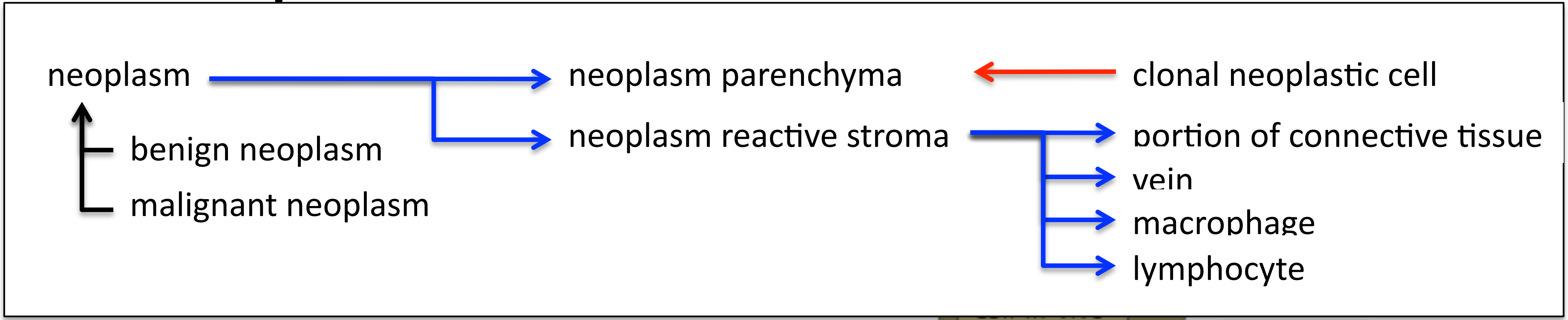
lymphocyte

is_a

has_part

part_of

Representation of Extracted Text



neoplasm

benign neoplasm

malignant neoplasm

neoplasm parenchyma

neoplasm reactive stroma

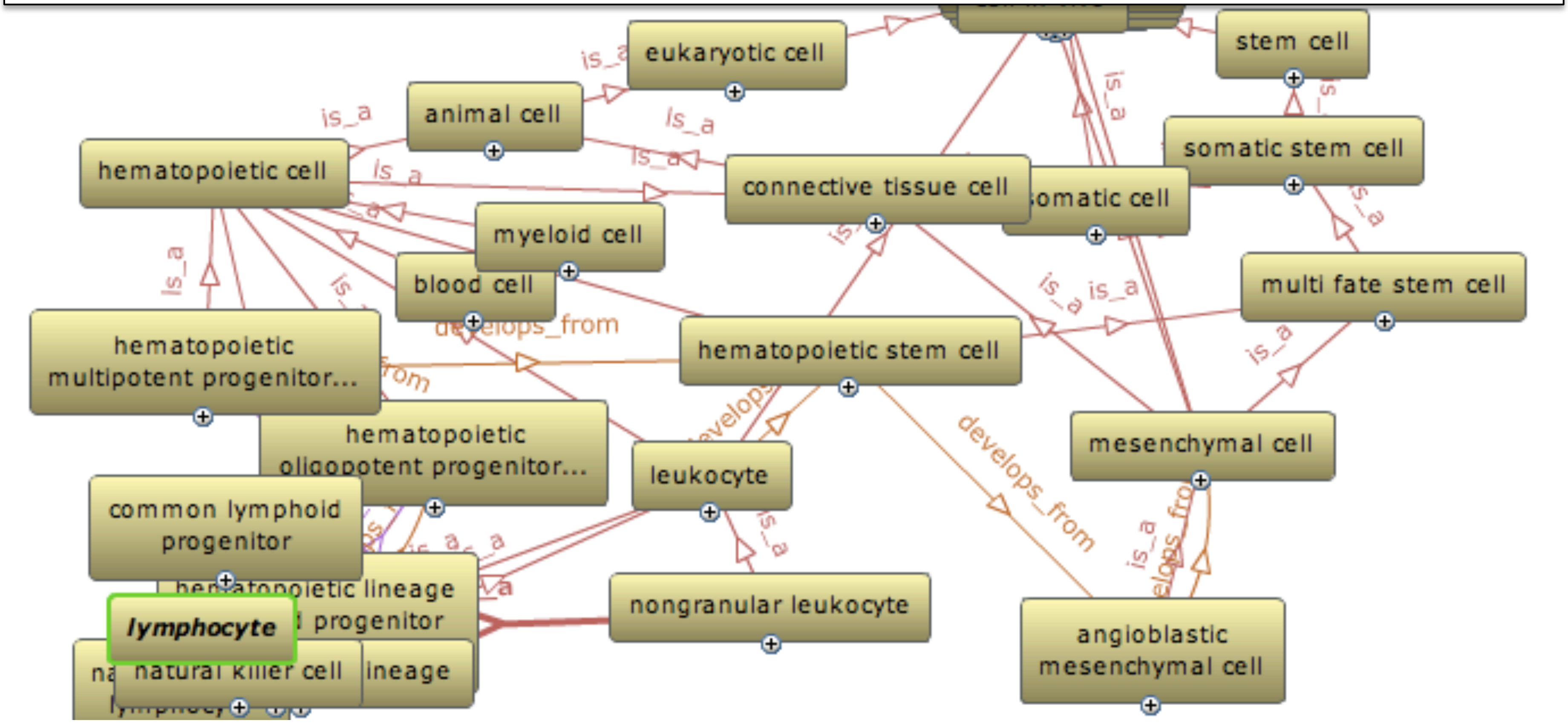
clonal neoplastic cell

portion of connective tissue

vein

macrophage

lymphocyte



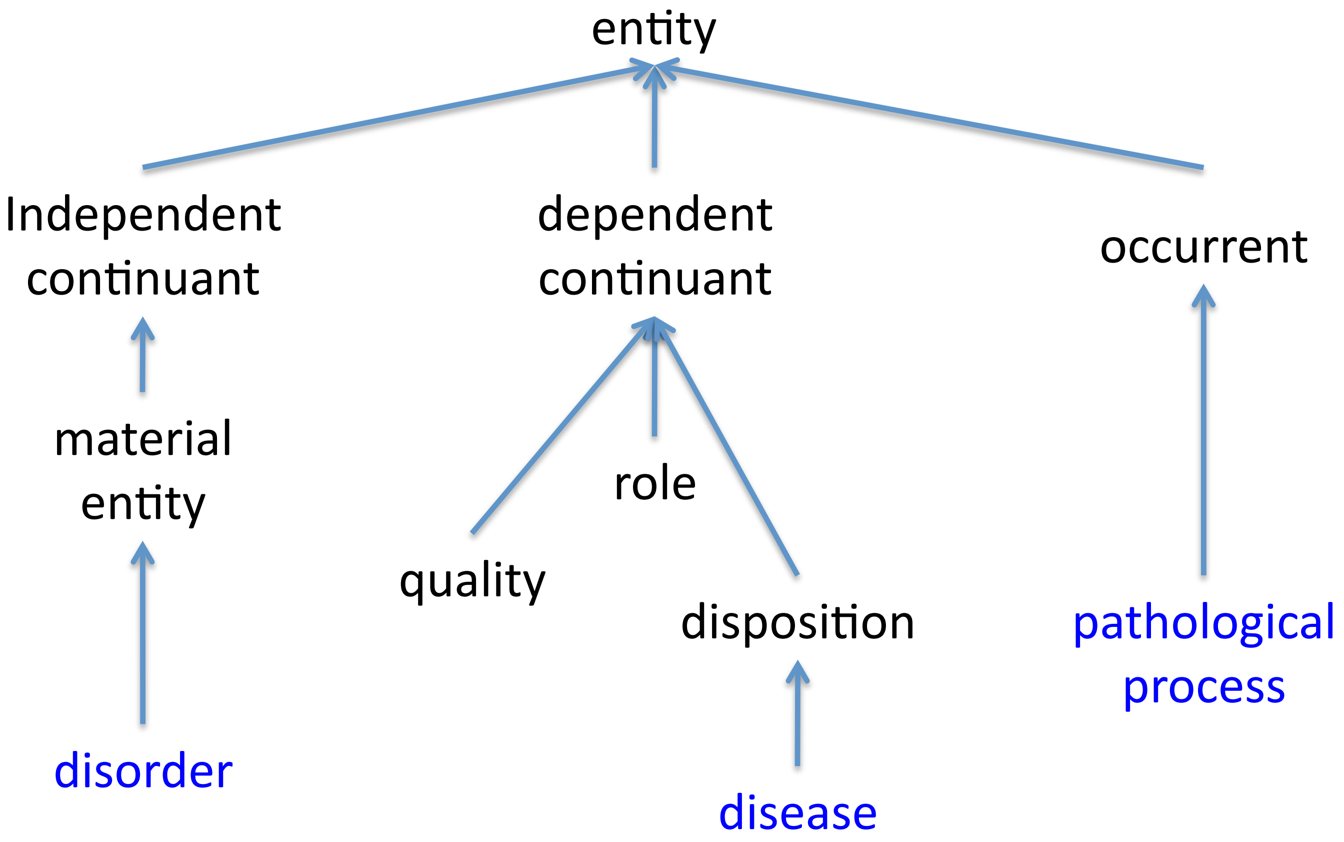
Creation of Ontology Resources

- Pathology Foundational Ontology (PFO)
- Biomedical Relation Ontology (BRO)

PFO

- Developed within the OBO Foundry framework
  - Utilize the Basic Formal Ontology
  - Use existing ontologies where possible
    - Emphasize OBOF ontologies
    - Import terms via MIREOT (Courtot et al 2010)
- Expand iteratively as encounter terms for annotation
- Challenge: integrating terms from clinical terminologies and ontologies

BFO and OGMS



entity

Independent continuant

dependent continuant

occurrent

material entity

role

quality

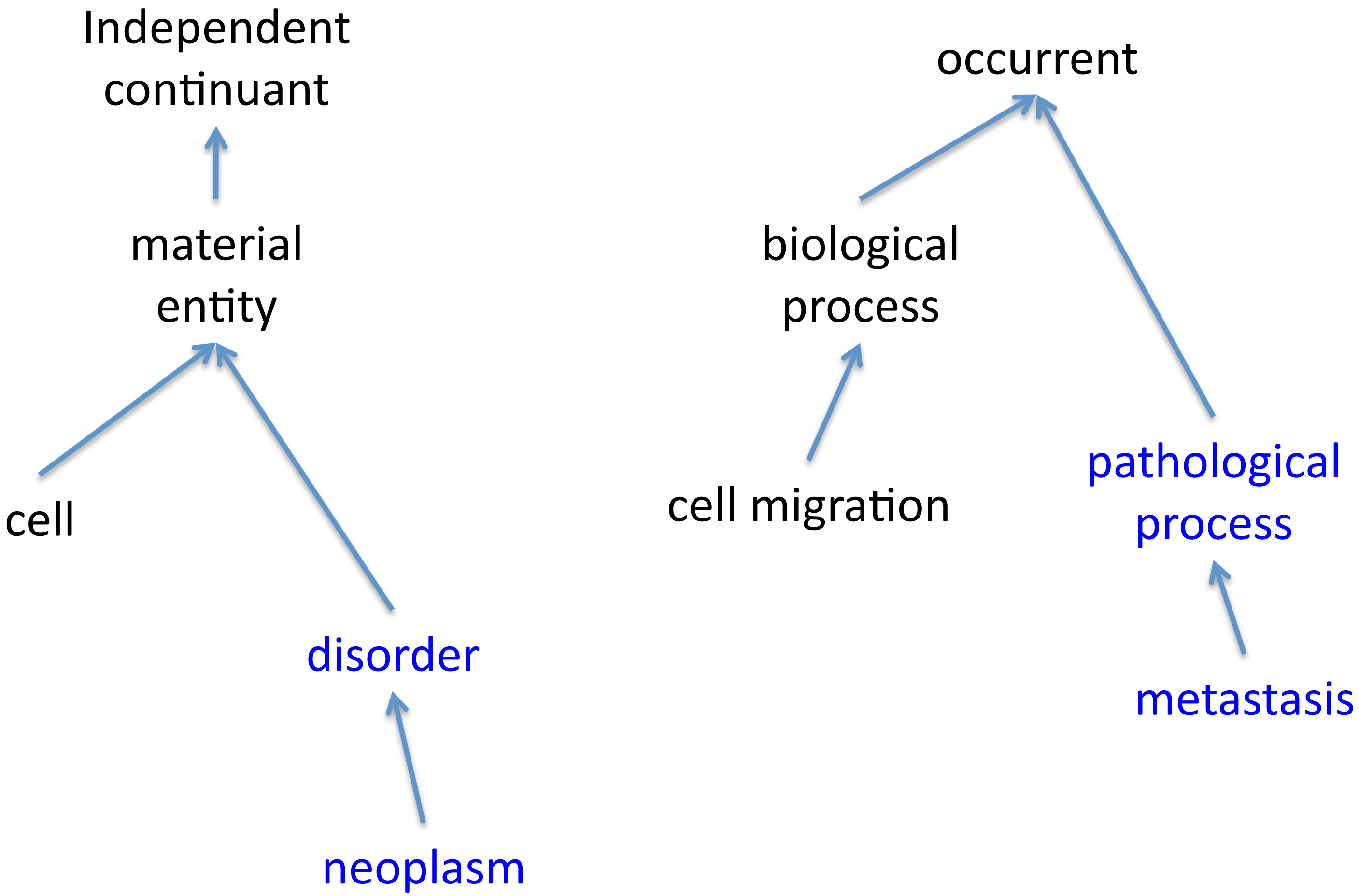
disposition

disorder

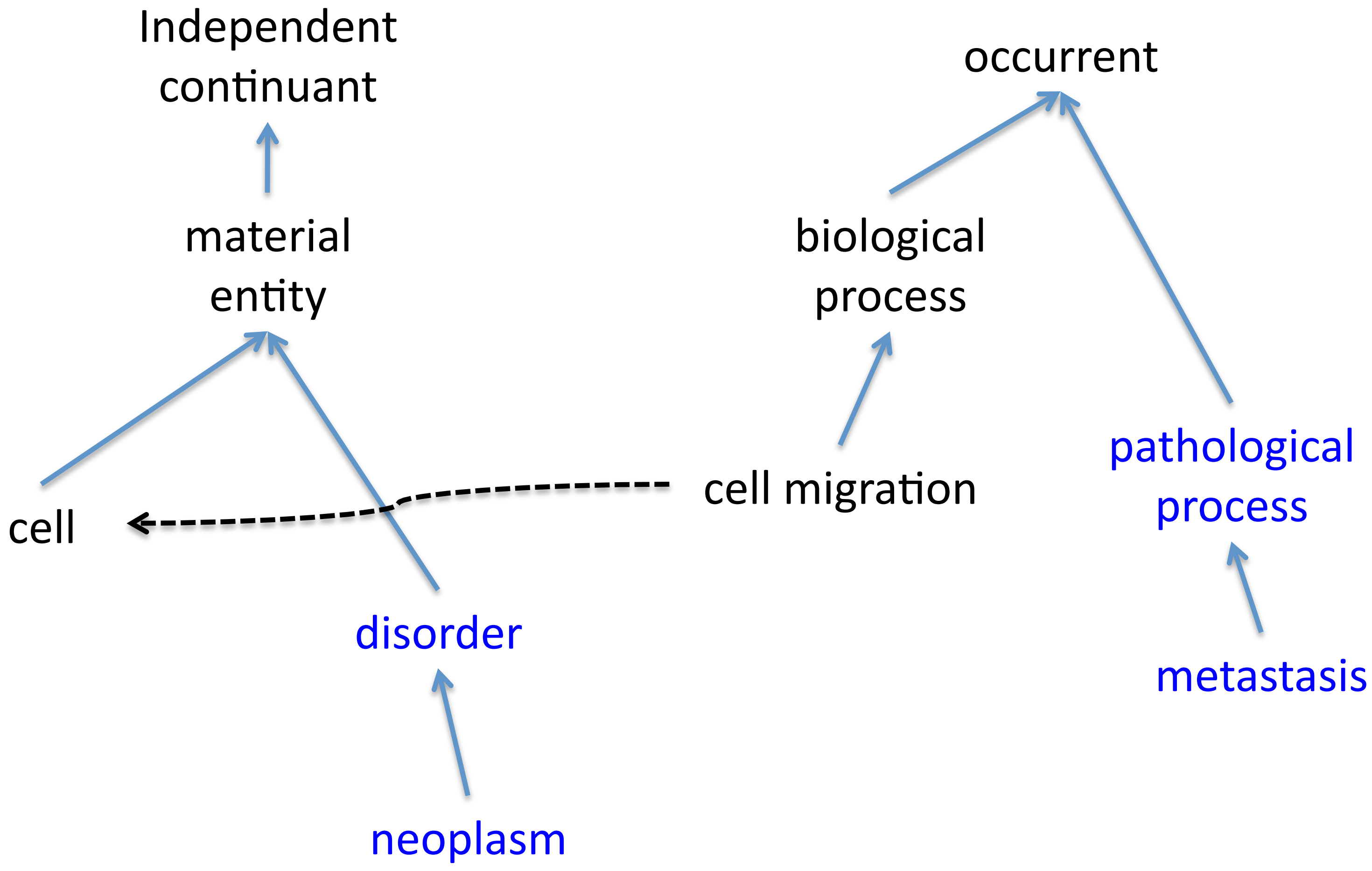
disease

pathological process

PFO


Independent continuant
material entity
cell
disorder
neoplasm
occurrent
biological process
cell migration
pathological process
metastasis

PFO



Independent continuant

material entity

cell

disorder

neoplasm

occurrent

biological process

cell migration

pathological process

metastasis

Development of BRO

- Text analysis to identify core set of relations
- Map to RO where possible.
- Of the remaining relations, determine which are foundational and which are domain-specific.
- Develop formal, first order logic definitions for new relations
  - Define in terms of existing RO relations
  - Define domain-specific relations in terms of foundational relations
- Provide a representation for each relation as an OWL object property.

Current Work

- Cellular Responses to Stress and Toxic Insults: Adaptation, Injury, and Death
- Infectious Diseases
- Neoplasia
- The Heart
- The Lung
- Diseases of the Immune System

Initial survey

- Initial survey of the text showed reference to a relatively small number of relations
  - (We assume that is due to the fact that the text stems from a textbook)
- A large number of relations are in accordance with OBO RO.

Relations annotated in the first 2 subsections :

has_integral_part*	causes	is_integral_part_of*
adjacent_to*	contained_in*	disorder_leads_to_disease
promotes	realizes	transmitted_by
is_a	depends_on	unfolds_in
bearer_of	infects	aggregates_to
larger_than		

* Relations represented by RO release version 1.01

Challenges

- Scale
  - Protégé, OWL reasoners, …
- Inter-annotator consistency

Interesting Questions

- At what level of specificity should we annotate?
  - Do we annotate ‘lymphocyte’ with the CL term ‘lymphocyte’ or ‘cell’?
- In what ways can the ontology relations be used to improve NLP?

NLP-Based Mapping of Textbook Pathology to the Ontology for General Medical Science (OGMS)

Lindsay Cowell and Richard Scheuermann
University of Texas Southwestern Medical Center

Sanda Harabagiu, Bryan Rink, and Kirk Roberts
University of Texas at Dallas

Mathias Brochausen and Bill Hogan
University of Arkansas for Medical Sciences