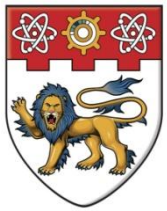


NANYANG TECHNOLOGICAL UNIVERSITY

Discovering Cross-Ontology Subsumption Relationships by Using Ontological Annotations on Biomedical Literature

Watson W.K. Chua and Jung-jae Kim
School of Computer Engineering

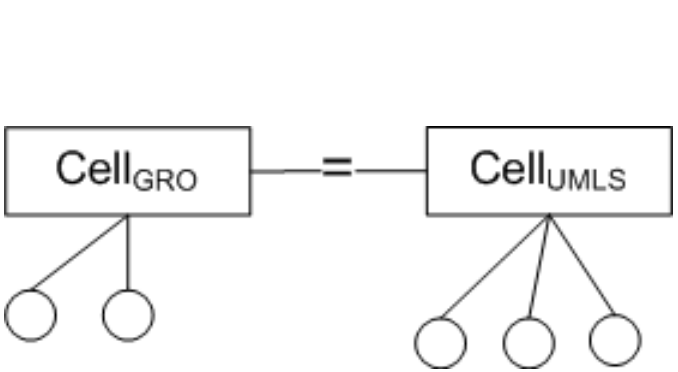
International Conference of Biomedical Ontology
22nd – 25th July 2012

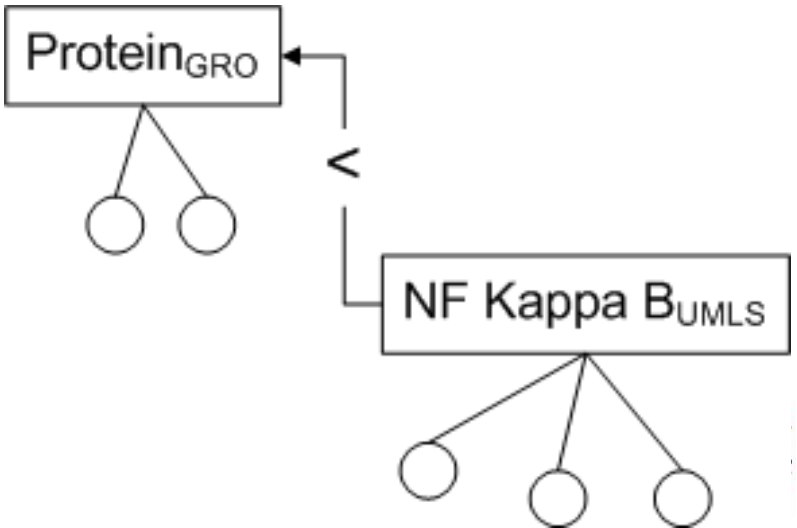
Motivation

- Ontology Alignment:
  - Finding relationships between entities of different ontologies
  - Facilitates interoperability

- Majority of existing research focuses on finding equivalence relationships

- Subsumption relationships complement equivalences

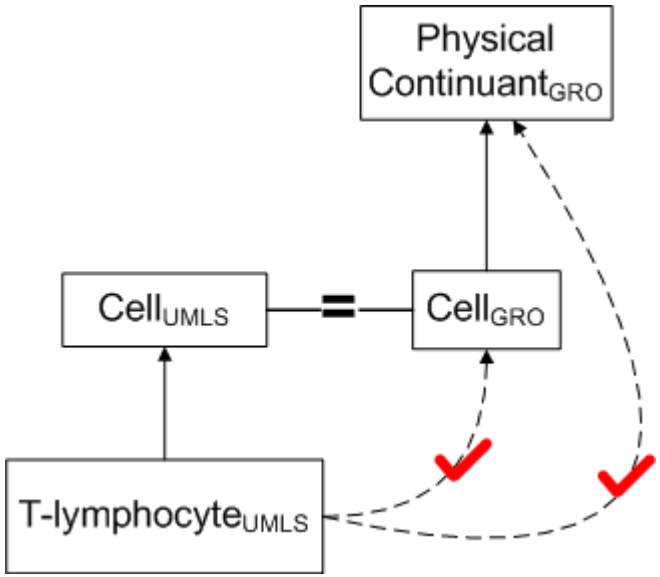


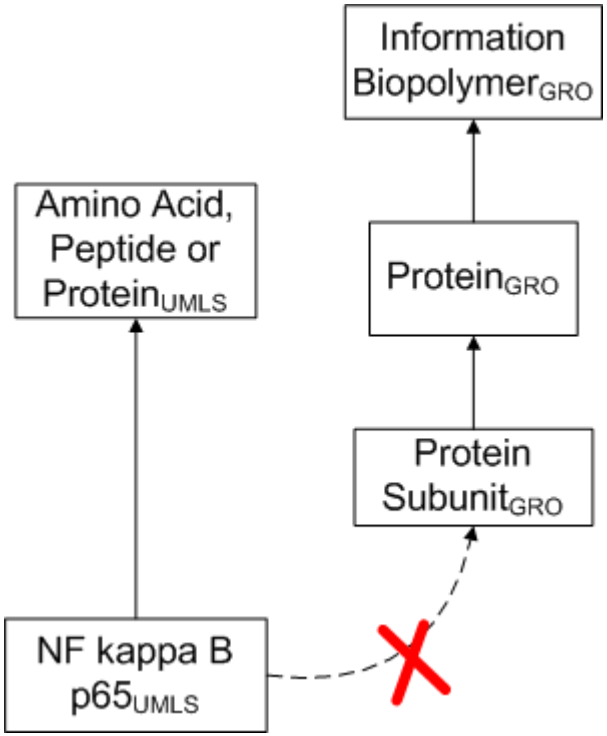


Motivation

- Subsumptions commonly inferred using equivalences but ..
  - equivalences does not always exist as ‘bridges’

- Need to find these subsumption relationships **directly**.





NANYANG TECHNOLOGICAL UNIVERSITY

Related Work

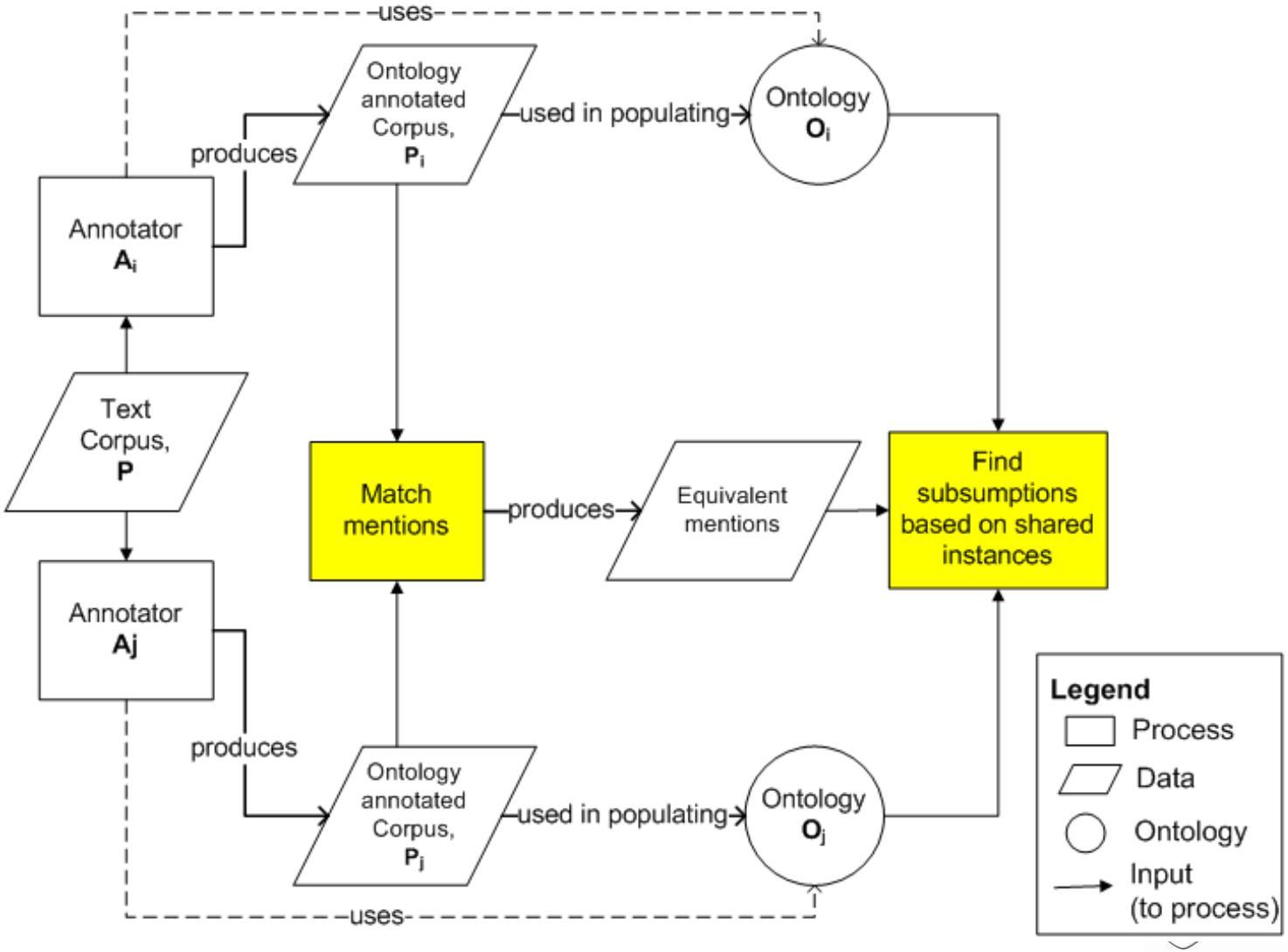
- Hearst Patterns [*van Hage et. al., 2005*]
  - C_i *such as/ especially/ including* C_j => C_j is-a C_i
  - Expressions are not explicitly used when readers have background knowledge

- Machine Learning Methods [*Spiliopoulos et al., 2010*]
  - Training using intra-ontology subsumption relationships
  - Uses trained models to find cross-ontology subsumptions
  - Effective only if both ontologies have similar hierarchical structures

- Instance-based methods [*Doan et al., 2004; Kirsten et al.,2007*]
  - Effective but face lack of shared instances between ontologies

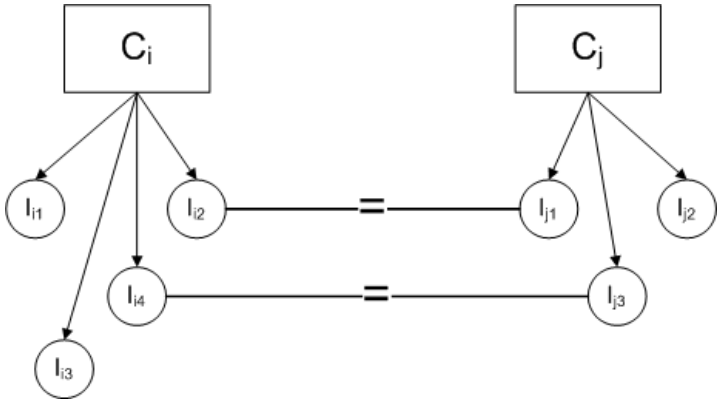
Methodology: SUbsumption Relationships Discovery (SURD) Overview

- Instance-based approach



Matching mentions (instances)

- Used to find shared instances



- Not always straightforward to find
  - different annotators have different guidelines for annotation

"spi-B, like spi-1, was found to be expressed in various murine and human hematopoietic cell lines..."

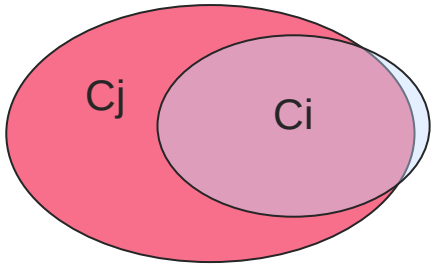
"human hematopoietic cell lines"
[GRO: Cell]

"cell lines"
[UMLS: Cell Lines]

- Two instances are equivalent if they share the same head words.

Discovering subsumptions and equivalences based on shared instances: Co-Annotation Ratio

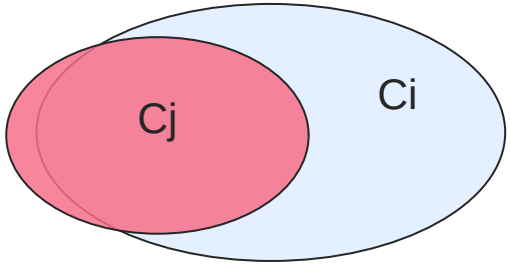
$$CAR_{ij} = \frac{|C_i \cap C_j|}{|C_i|}$$



$CAR_{ij} : High$

$CAR_{ji} : Low$

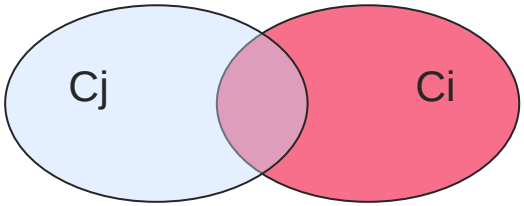
$\rightarrow C_i \subseteq C_j$



$CAR_{ij} : Low$

$CAR_{ji} : High$

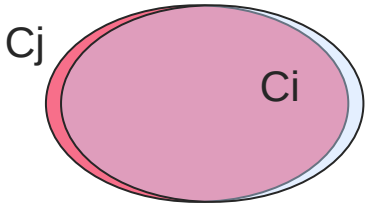
$\rightarrow C_j \subseteq C_i$



$CAR_{ij} : Low$

$CAR_{ji} : Low$

$\rightarrow C_i \neq C_j$



$CAR_{ij} : High$

$CAR_{ji} : High$

$\rightarrow C_j \equiv C_i$

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment

- Dataset 1: GRO Corpus
  - 200 PubMed abstracts
  - manually annotated using Gene Regulation Ontology (GRO)

- Dataset 2: GENIA Corpus
  - 1000 PubMed abstracts
  - manually annotated using GENIA Ontology

- Both corpora were also automatically annotated using MetaMap with the UMLS Metathesaurus

- BOAT was used to find equivalences to infer ‘trivial’ subsumptions for comparison

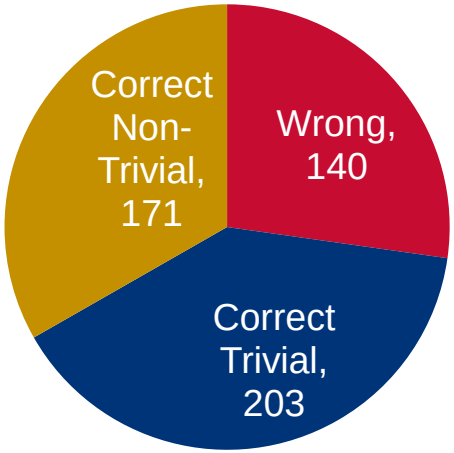
Ontologies

- UMLS Metathesaurus
  - Large, multi-purpose thesaurus
  - Wide coverage, fine-granularity
  - Very specific leaf concepts (e.g. p56 and Glucose)
- GRO
  - Describes gene regulation events
  - Relatively coarse-grained compared to UMLS
  - Has very specific concepts regarding the domain of gene regulation.
- GENIA
  - Formal model of cell signaling reactions in human
  - Coarse-granularity (leaf concepts like Protein Molecule and Carbohydrate)
- Can expect to find many subsumptions between Metathesaurus and each of the two ontologies.

Results

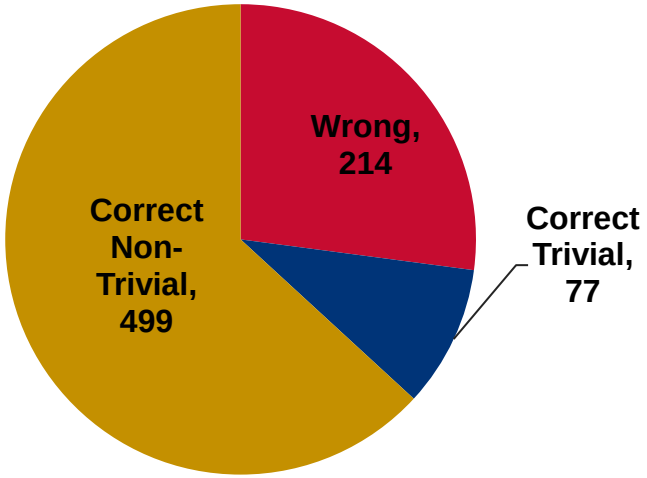
GRO-UMLS

- 1952 subsumptions found
- 514 randomly validated
- Precision: 0.79
- 42.3% of correct subsumptions are non-trivial



GENIA-UMLS

- 5200 subsumptions found
- 790 randomly validated
- Precision: 0.73
- 86.6% of correct subsumptions are non-trivial



NANYANG TECHNOLOGICAL UNIVERSITY

Application: Automatic Cross-Ontology Corpus Annotation

- Use automatic annotators customised to perform annotation with O_2, to perform annotation with O_1
  - E.g.: Using MetaMap to annotate a corpora with GRO

The upstream region of the human homeobox gene HOX3D is a target for regulation by retinoic acid and HOX homeoproteins.

1) Annotate sentence with MetaMap

The upstream region of the human homeobox gene
[UMLS: *C1415679 (HOX3D)*] **HOX3D** is a target for regulation by
[UMLS: *C0040845 (Retinoic Acid)*] **retinoic acid** and
[UMLS: *C0242617 (Homeoproteins)*] **HOX homeoproteins**.

2) Look up subsumptions between UMLS and GRO concepts

UMLS: *C1415679 (HOX3D)*	<	*GRO: Homeobox*
UMLS: *C0040845 (Retinoic Acid)*	<	*GRO: Retinoic Acid*
UMLS: *C0242617 (Homeoproteins)*	<	GRO: Protein

3) Convert MetaMap annotations to GRO annotations using subsumptions

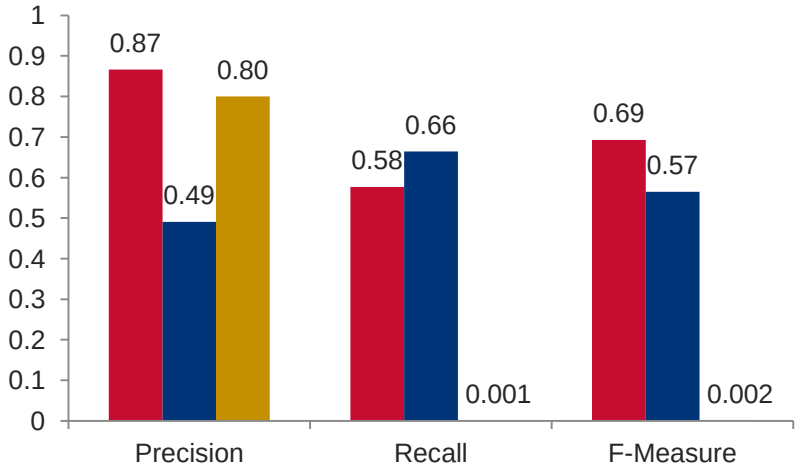
The upstream region of the human homeobox gene
[GRO: *Homeobox*] **HOX3D** is a target for regulation by
[*GRO: Organic Chemical*] **retinoic acid** and
[*GRO: Protein*] **HOX homeoproteins**.

Evaluation of Automatic Corpus Annotation

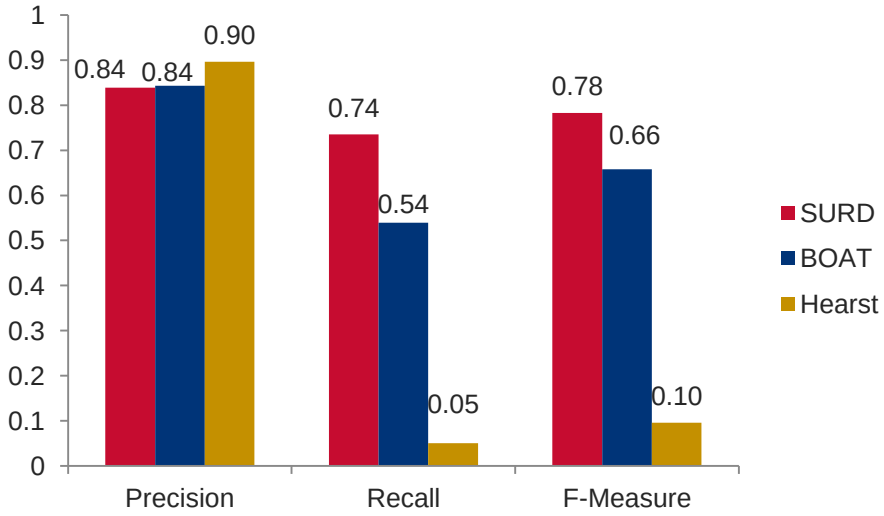
- 10-fold Cross-Validation
- Compare translations with original manual annotation to find
  - Precision
  - Recall
  - F-Measure
- Comparison is done between translations utilising subsumptions:
  1. found using SURD
  2. found using Hearst patterns
  3. inferred from equivalences found by BOAT

Automatic Cross-Ontology Corpus Annotation Results

MetaThesaurus-> GRO



MetaThesaurus->GENIA



NANYANG TECHNOLOGICAL UNIVERSITY

Conclusion

- Instance-based Subsumption Relation Identification
- Uses text annotations as concept instances
- Can be used for automatic cross-ontology corpus annotation
- Future work:
  - Extend to other ontologies used in annotation corpora (E.g. CRAFT)
  - Use subsumptions to complement BOAT's equivalences for integrating ontologies

Thank you for your attention.

Questions?