Ontology Alignment

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Ontologies in biomedical research

- many biomedical ontologies

  e.g. GO, OBO, SNOMED-CT

- practical use of biomedical ontologies

  e.g. databases annotated with GO

GENE ONTOLOGY (GO)

```
immune response
  i- acute-phase response
  i- anaphylaxis
  i- antigen presentation
  i- antigen processing
  i- cellular defense response
  i- cytokine metabolism
    i- cytokine biosynthesis
      synonym cytokine production
      …
    p- regulation of cytokine
      biosynthesis
    …
  …
  i- B-cell activation
    i- B-cell differentiation
    i- B-cell proliferation
  i- cellular defense response
  …
  i-  T-cell activation
    i- activation of natural killer
      cell activity
    …
```

Ontologies with overlapping information

GENE ONTOLOGY (GO)

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
    …
- …
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response
- …
- i- T-cell activation
  - i- activation of natural killer cell activity
    …

SIGNAL-ONTOLOGY (SigO)

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

Ontologies with overlapping information

- Use of multiple ontologies
  - custom-specific ontology + standard ontology
  - different views over same domain
  - overlapping domains

- Bottom-up creation of ontologies

  experts can focus on their domain of expertise

→ important to know the inter-ontology relationships

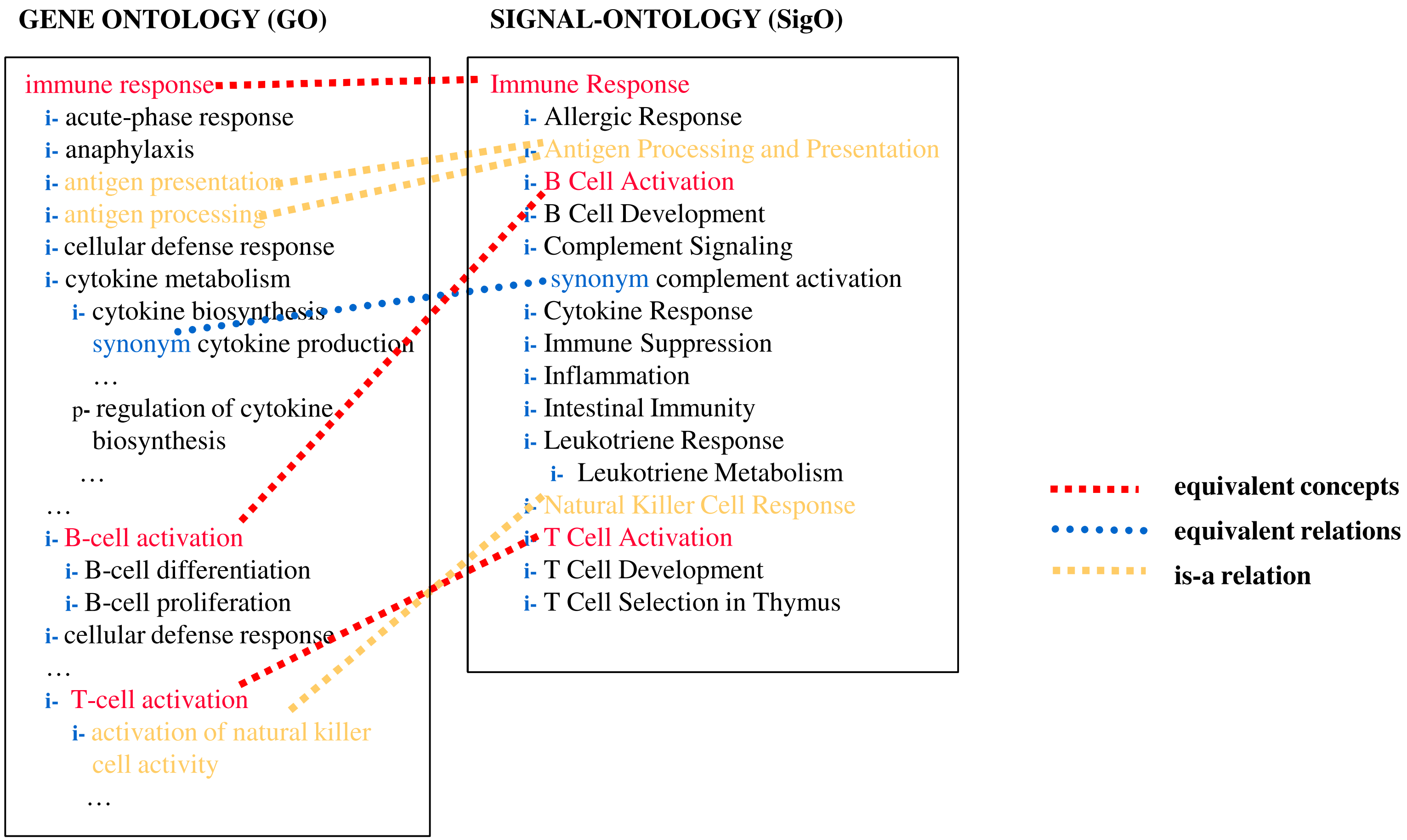
GENE ONTOLOGY (GO)

immune response
- **i-** acute-phase response
- **i-** anaphylaxis
- **i-** antigen presentation
- **i-** antigen processing
- **i-** cellular defense response
- **i-** cytokine metabolism
  - **i-** cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  …
- …
- **i-** B-cell activation
  - **i-** B-cell differentiation
  - **i-** B-cell proliferation
- **i-** cellular defense response
- …
- **i-** T-cell activation
  - **i-** activation of natural killer cell activity
  …

SIGNAL-ONTOLOGY (SigO)

Immune Response
- **i-** Allergic Response
- **i-** Antigen Processing and Presentation
- **i-** B Cell Activation
- **i-** B Cell Development
- **i-** Complement Signaling
  synonym complement activation
- **i-** Cytokine Response
- **i-** Immune Suppression
- **i-** Inflammation
- **i-** Intestinal Immunity
- **i-** Leukotriene Response
  - **i-** Leukotriene Metabolism
- **i-** Natural Killer Cell Response
- **i-** T Cell Activation
- **i-** T Cell Development
- **i-** T Cell Selection in Thymus

Ontology Alignment

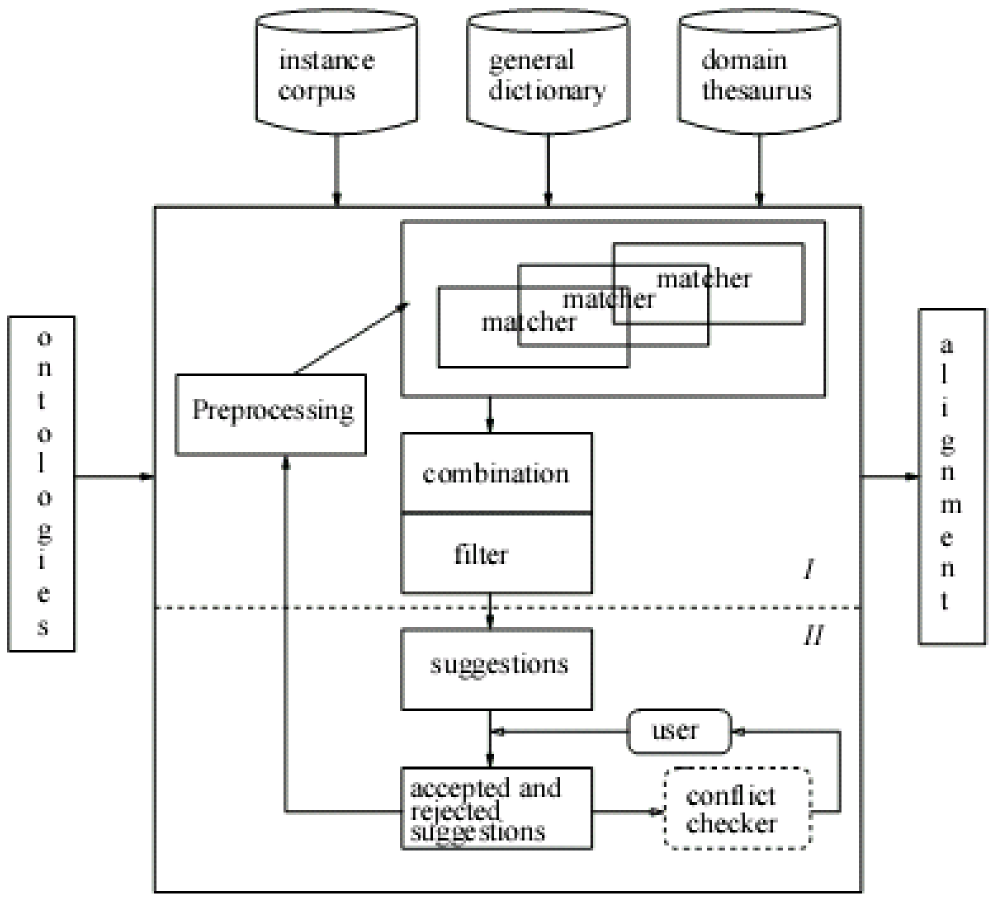


Defining the relations between the terms in different ontologies

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

An Alignment Framework



Classification

- According to input
  - KR: OWL, UML, EER, XML, RDF, …
  - components: concepts, relations, instance, axioms
- According to process
  - What information is used and how?
- According to output
  - 1-1, m-n
  - Similarity vs explicit relations (equivalence, is-a)
  - confidence

Preprocessing

Preprocessing

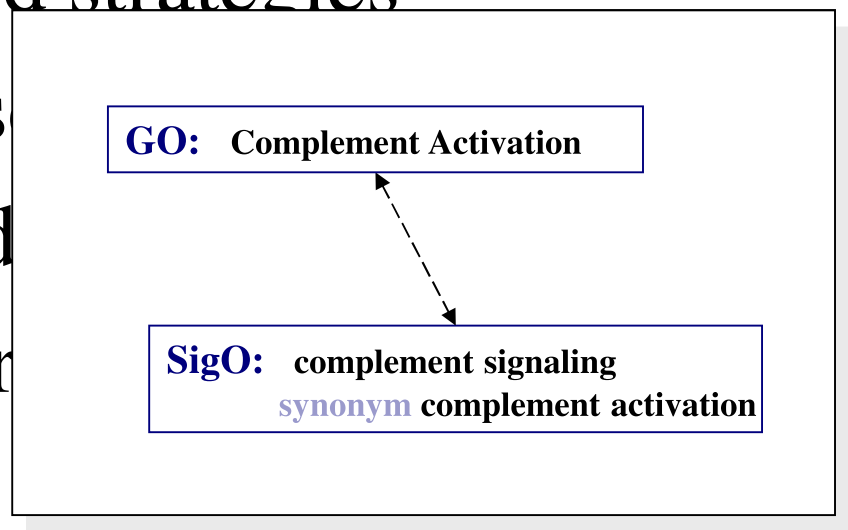
For example,

- Selection of features
- Selection of search space

Matchers

Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-bas
- Instance-based
- Use of auxiliar

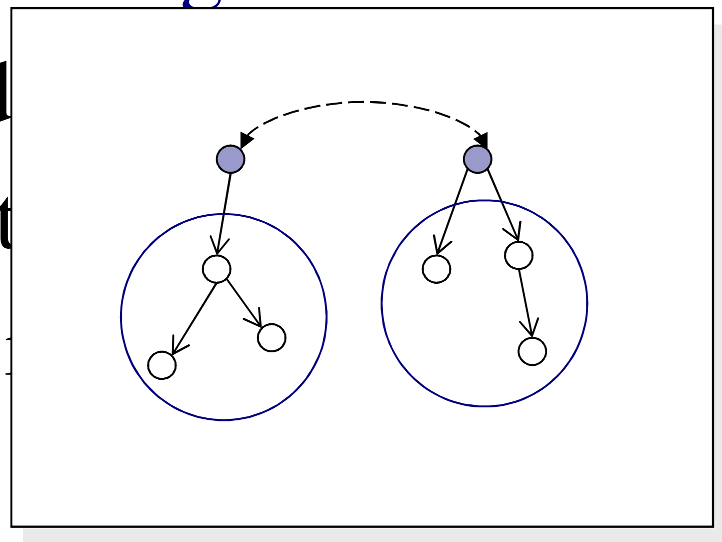


Example matchers

- Edit distance
  - Number of deletions, insertions, substitutions required to transform one string into another
  - aaaa → baab: edit distance 2
- N-gram
  - N-gram : N consecutive characters in a string
  - Similarity based on set comparison of n-grams
  - aaaa : {aa, aa, aa};   baab : {ba, aa, ab}

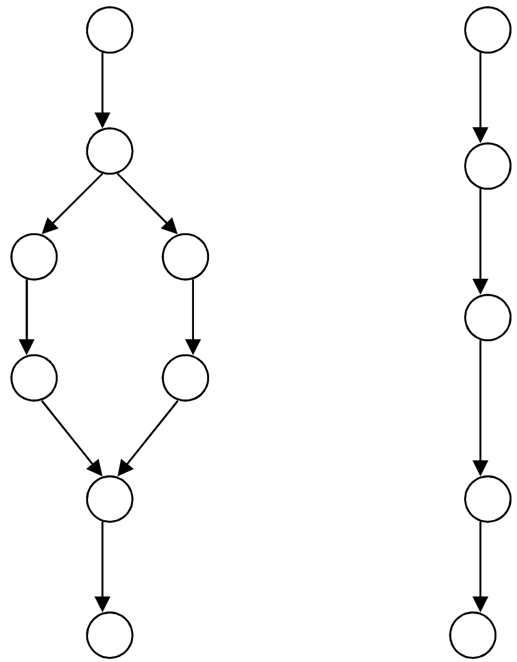
Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based
- Instance-based st
- Use of auxiliary



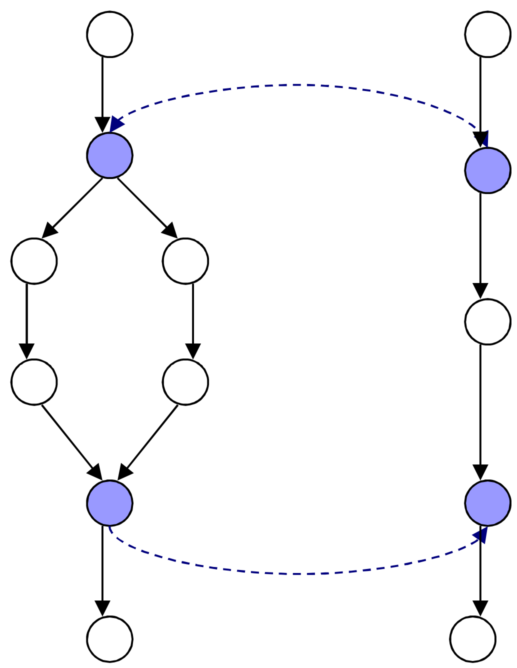
Example matchers

- Propagation of similarity values
- Anchored matching



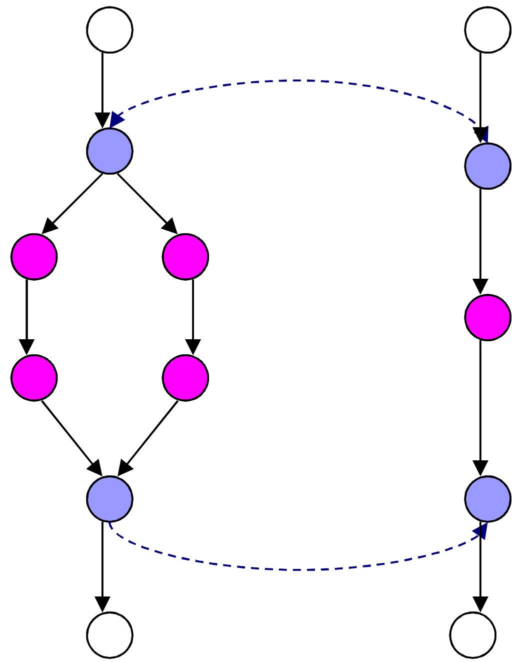
Example matchers

- Propagation of similarity values
- Anchored matching



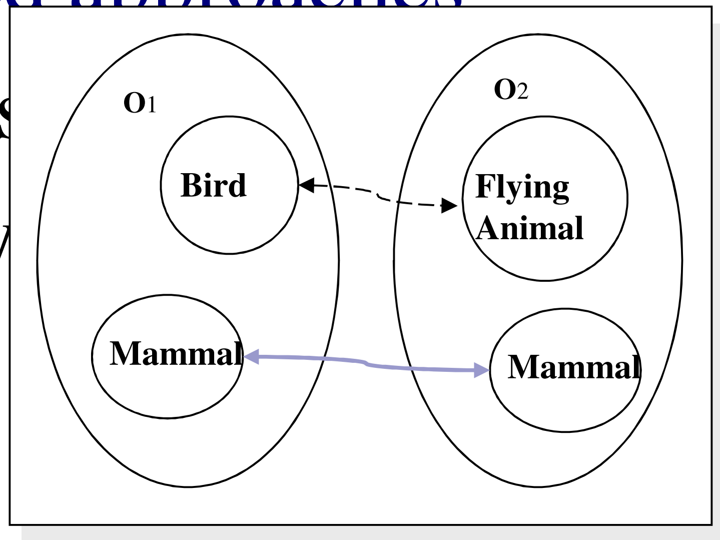
Example matchers

- Propagation of similarity values
- Anchored matching



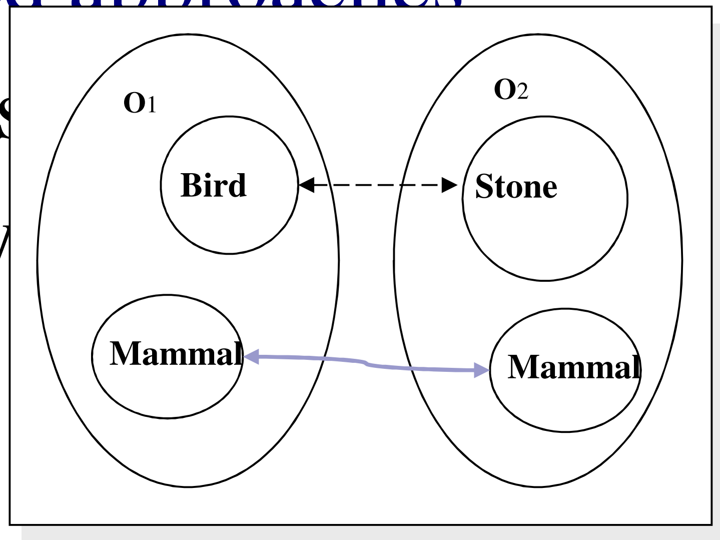
Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary



Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary

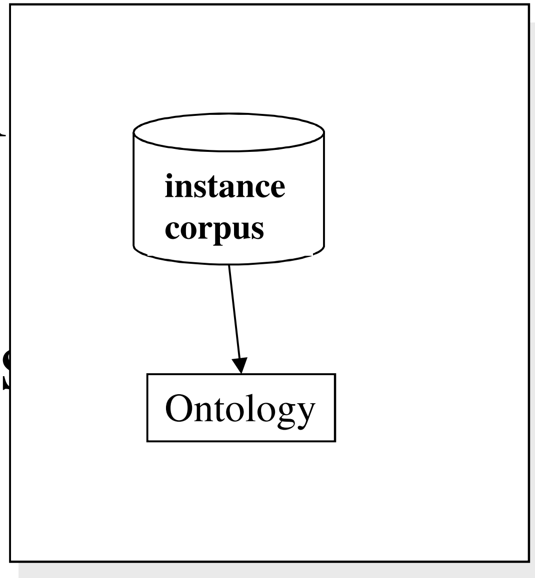


Example matchers

- Similarities between data types
- Similarities based on cardinalities

Matcher Strategies

- Strategies based on linguisti
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information



Example matchers

- Instance-based
- Use life science literature as instances

- Structure-based extensions

Learning matchers – instance-based strategies

- Basic intuition

  A similarity measure between concepts can be computed based on the probability that documents about one concept are also about the other concept and vice versa.

- Intuition for structure-based extensions

  Documents about a concept are also about their super-concepts.

  (No requirement for previous alignment results.)

Learning matchers - steps

- Generate corpora
  - Use concept as query term in PubMed
  - Retrieve most recent PubMed abstracts
- Generate text classifiers
  - One classifier per ontology / One classifier per concept
- Classification
  - Abstracts related to one ontology are classified by the other ontology's classifier(s) and vice versa
- Calculate similarities

Basic Naïve Bayes matcher

- Generate corpora
- Generate classifiers
  - Naive Bayes classifiers, one per ontology
- Classification
  - Abstracts related to one ontology are classified to the concept in the other ontology with highest posterior probability P(C|d)
- Calculate similarities

$$sim(C_1, C_2) = \frac{n_{NBC2}(C_1, C_2) + n_{NBC1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

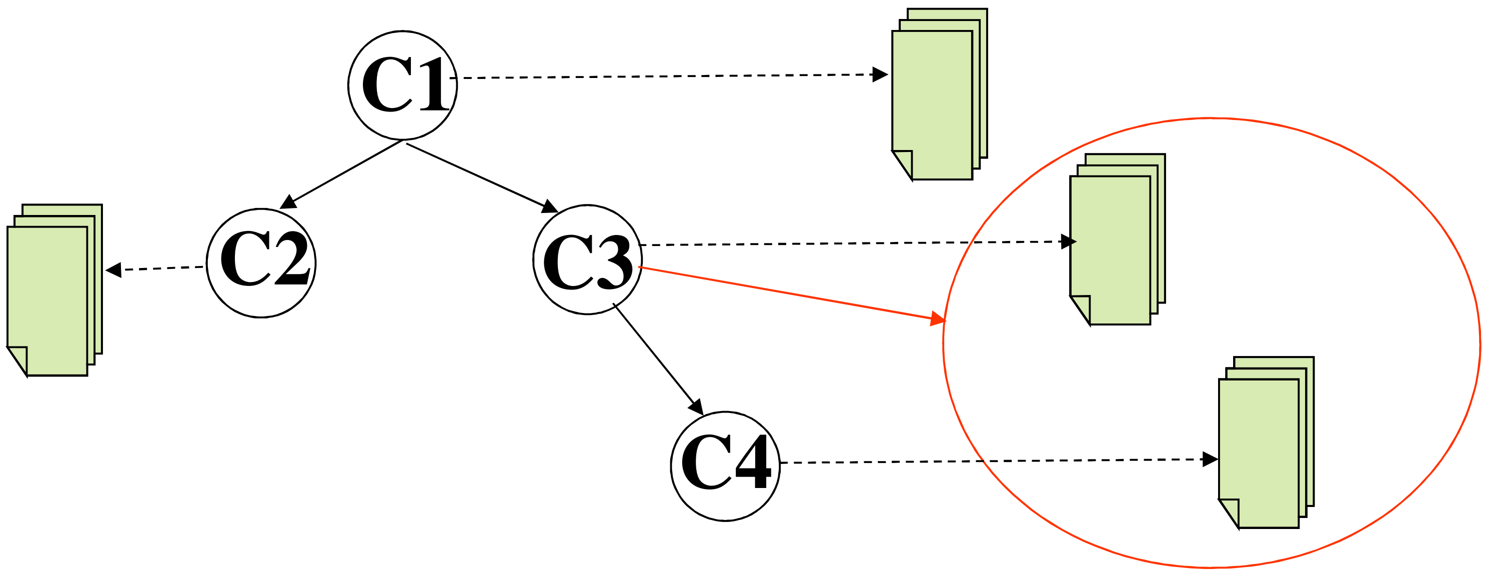
Basic Support Vector Machines matcher

- Generate corpora
- Generate classifiers
  - SVM-based classifiers, one per concept
- Classification
  - Single classification variant: Abstracts related to concepts in one ontology are classified to the concept in the other ontology for which its classifier gives the abstract the highest positive value.
  - Multiple classification variant: Abstracts related to concepts in one ontology are classified all the concepts in the other ontology whose classifiers give the abstract a positive value.
- Calculate similarities

$$\frac{n_{SVMC-C_2}(C_1, C_2) + n_{SVMC-C_1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

Structural extension ‘Cl’

- Generate classifiers
  - Take (is-a) structure of the ontologies into account when building the classifiers
  - Extend the set of abstracts associated to a concept by adding the abstracts related to the sub-concepts



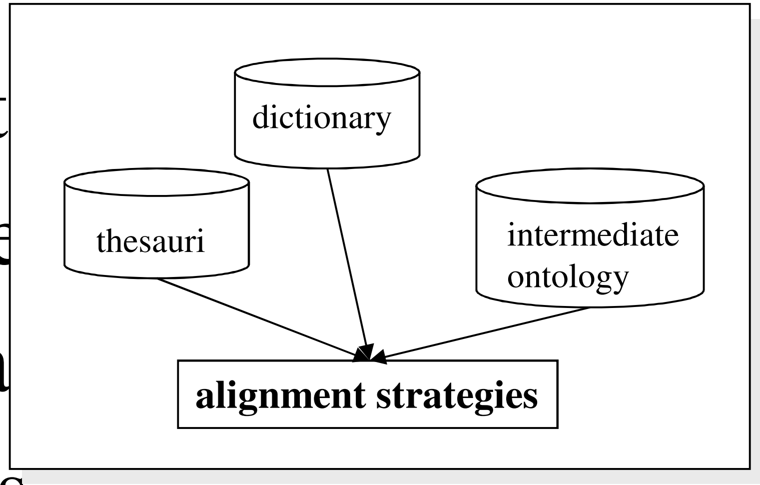
Structural extension ‘Sim’

- Calculate similarities
  - Take structure of the ontologies into account when calculating similarities
  - Similarity is computed based on the classifiers applied to the concepts and their sub-concepts

$$sim_{struct}(C_1, C_2) = \frac{\sum_{C_i \sqsubseteq C_1, C_j \sqsubseteq C_2} n_{NBC2}(C_i, C_j) + \sum_{C_i \sqsubseteq C_1, C_j \sqsubseteq C_2} n_{NBC1}(C_j, C_i)}{\sum_{C_i \sqsubseteq C_1} n_D(C_i) + \sum_{C_j \sqsubseteq C_2} n_D(C_j)}$$

Matcher Strategies

- Strategies based linguist
- Structure-based strategie
- Constraint-based approa
- Instance-based strategies
- Use of auxiliary information



Example matchers

- Use of WordNet
  - Use WordNet to find synonyms
  - Use WordNet to find ancestors and descendants in the is-a hierarchy
- Use of Unified Medical Language System (UMLS)
  - Includes many ontologies
  - Includes many alignments (not complete)
  - Use UMLS alignments in the computation of the similarity values

Ontology Alignment and Mergning Systems

	linguistic	structure	constraints	instances	auxiliary
ArtGen	name	parents, children		domain specific documents	WordNet
ASCO	name, label description	parents, children, siblings, path from root			WordNet
Chimaera	name	parents, children			
FCA-Merge	name			domain specific documents	
FOAM	name, label	parents, children	equivalence		
GLUE	name	neighborhood		instances	
HCONE	name	parents, children			WordNet
IF-Map				instances	a reference ontology
iMapper		leaf, non-leaf, children, related node	domain, range	instances	WordNet
OntoMapper		parents, children		documents	
(Anchor-) PROMPT	name	direct graphs			
SAMBO	name, synonym	is-a and part-of, descendants and ancestors		domain specific documents	WordNet, UMLS
S-Match	label	path from root	semantic relations codified in labels		WordNet

Combinations

Combination Strategies

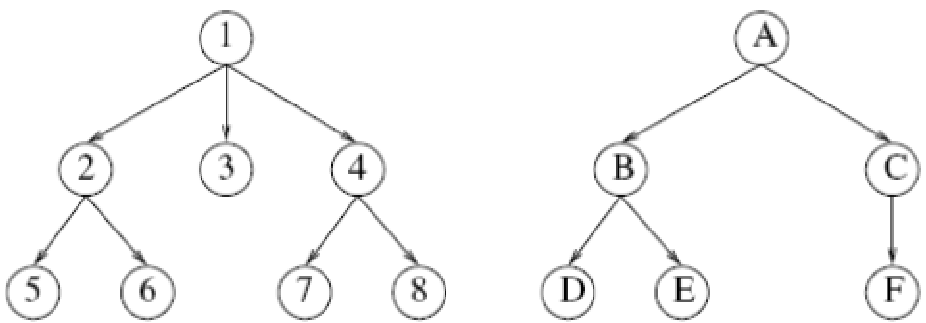
- Usually weighted sum of similarity values of different matchers
- Maximum of similarity values of different matchers

Filtering

Filtering techniques

- Threshold filtering

  Pairs of concepts with similarity higher or equal than threshold are alignment suggestions



sim

(2, *B*)
(3, *F*) suggest
(6, *D*)
(4, *C*)

th

(5, *C*) discard
(5, *E*)

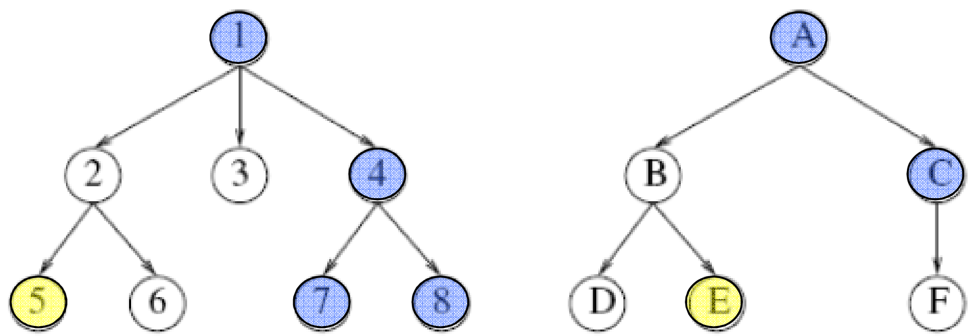
……

Filtering techniques

- Double threshold filtering

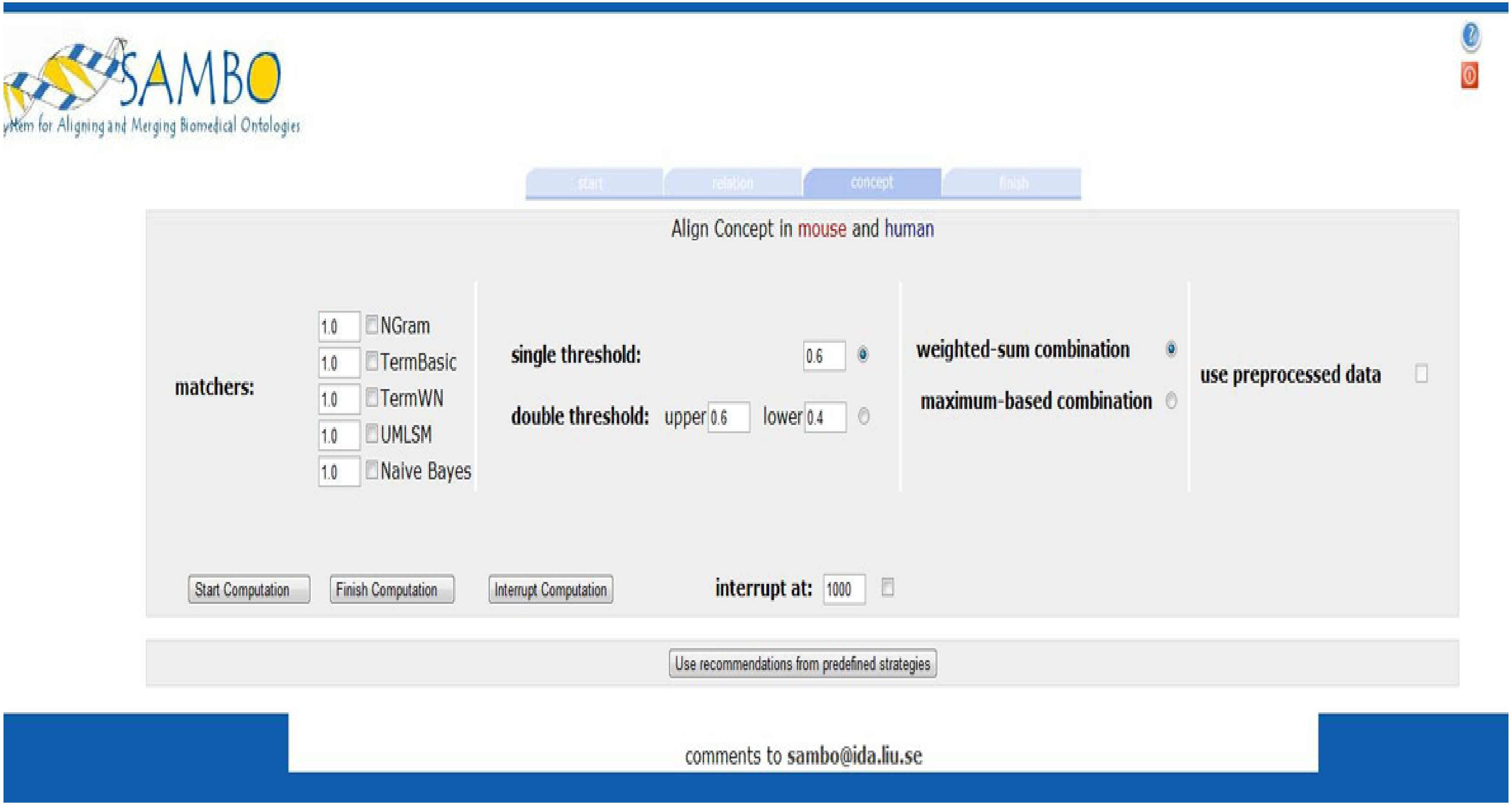
  (1) Pairs of concepts with similarity higher than or equal to **upper** threshold are alignment suggestions

  (2) Pairs of concepts with similarity between **lower** and **upper** thresholds are alignment suggestions if they make sense with respect to the structure of the ontologies and the suggestions according to (1)

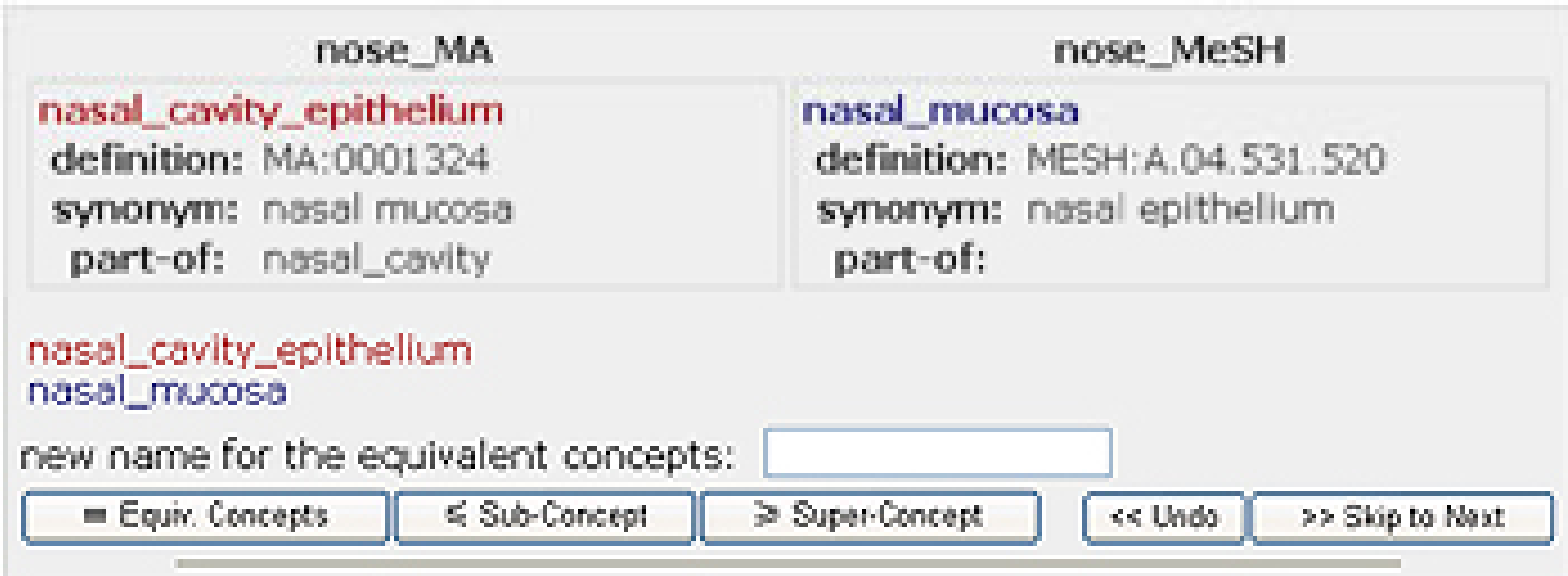


(2, B)
(3, F)
(6, D)
(4, C)
upper-th
(5, C)
(5, E)
lower-th
……

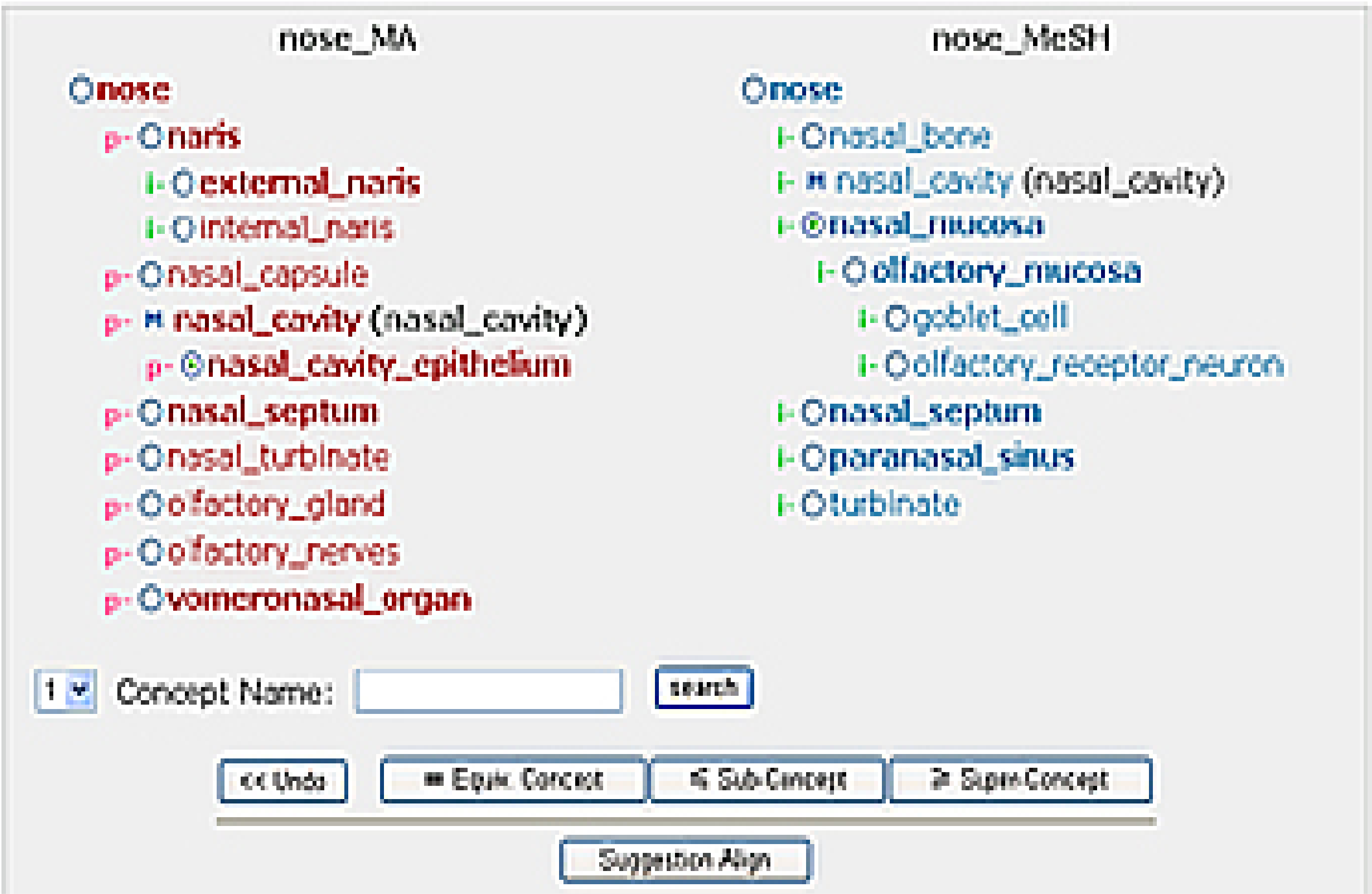
Example alignment system SAMBO – matchers, combination, filter



Example alignment system SAMBO – suggestion mode



Example alignment system SAMBO – manual mode



Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Evaluation measures

- Precision:

$$\frac{\#\ \text{correct mapping suggestions}}{\#\ \text{mapping suggestions}}$$

- Recall:

$$\frac{\#\ \text{correct mapping suggestions}}{\#\ \text{correct mappings}}$$

- F-measure: combination of precision and recall

Ontology Alignment
Evaluation Initiative

OAEI

- Since 2004
- Evaluation of systems
- Different tracks (2014)
  - benchmark
  - expressive: anatomy, conference, large biomedical ontologies
  - multilingual: multifarm (8 languages)
  - directories and thesauri: library
  - interactive
  - instances: identity, similarity

OAEI

- Evaluation measures
  - Precision/recall/f-measure
  - recall of non-trivial mappings

  - full / partial golden standard

OAEI 2007

- 17 systems participated
  - benchmark (13)
    - ASMOV: p = 0.95, r = 0.90
  - anatomy (11)
    - AOAS: f = 0.86, r+ = 0.50
    - SAMBO: f =0.81, r+ = 0.58
  - library (3)
    - Thesaurus merging: FALCON: p = 0.97, r = 0.87
    - Annotation scenario:
      - FALCON: pb =0.65, rb = 0.49, pa = 0.52, ra = 0.36, Ja = 0.30
      - Silas: pb = 0.66, rb= 0.47, pa = 0.53, ra = 0.35, Ja = 0.29
  - directory (9), food (6), environment (2), conference (6)

OAEI 2008 – anatomy track

- Align
  - Mouse anatomy: 2744 terms
  - NCI-anatomy: 3304 terms
  - Mappings: 1544 (of which 934 ‘trivial’)
- Tasks
  - 1. Align and optimize f
  - 2-3. Align and optimize p / r
  - 4. Align when partial reference alignment is given and optimize f

OAEI 2008 – anatomy track#1

- 9 systems participated
- SAMBO
  - p=0.869, r=0.836, r+=0.586, f=0.852
- SAMBOdtf
  - p=0.831, r=0.833, r+=0.579, f=0.832
- Use of TermWN and UMLS

OAEI 2008 – anatomy track#1

Is background knowledge (BK) needed?

Of the non-trivial mappings:

- Ca 50% found by systems using BK and systems not using BK
- Ca 13% found only by systems using BK
- Ca 13% found only by systems not using BK
- Ca 25% not found

Processing time:

hours with BK, minutes without BK

OAEI 2008 – anatomy track#4

Can we use given mappings when computing suggestions?

→ partial reference alignment given with all trivial and 50 non-trivial mappings

- SAMBO
  - p=0.636→0.660, r=0.626→0.624, f=0.631→0.642
- SAMBOdtf
  - p=0.563→0.603, r=0.622→0.630, f=0.591→0.616

(measures computed on non-given part of the reference alignment)

OAEI 2007-2008

- Systems can use only one combination of strategies per task

  → systems use similar strategies

  - text: string matching, tf-idf
  - structure: propagation of similarity to ancestors and/or descendants
  - thesaurus (WordNet)
  - domain knowledge important for anatomy task?

OAEI 2014

- 14 systems
- Anatomy:
  - best system f=0.944, p=0.956, r=0.932, r+=0.822, 28 seconds
  - many systems produce coherent mappings

Evaluation of algorithms

Cases

- GO vs. SigO

GO: 70 terms **GO-immune defense**

SigO: 15 terms **SigO-immune defense**

GO: 60 terms **GO-behavior**

SigO: 10 terms **SigO-behavior**

- MA vs. MeSH

MA: 15 terms **MA-nose**

MeSH: 18 terms **MeSH-nose**

MA: 77 terms **MA-ear**

MeSH: 39 terms **MeSH-ear**

MA: 112terms **MA-eye**

MeSH: 45 terms **MeSH-eye**

Evaluation of matchers

- Matchers

  Term, TermWN, Dom, Learn (Learn+structure), Struc

- Parameters

  Quality of suggestions: precision/recall

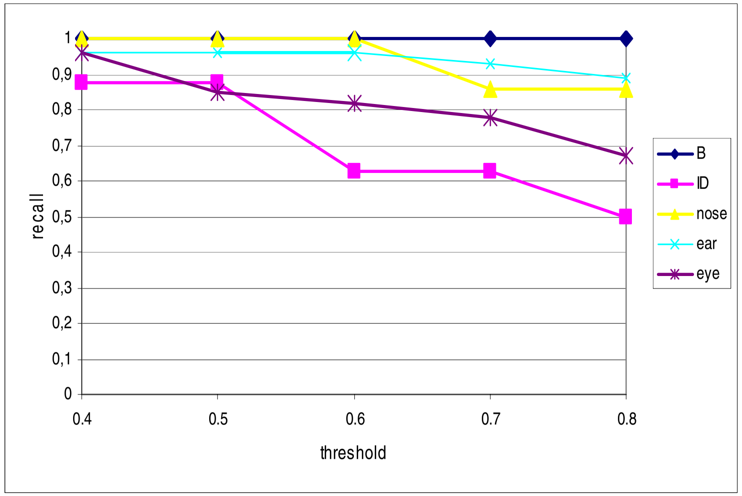
  Threshold filtering : 0.4, 0.5, 0.6, 0.7, 0.8

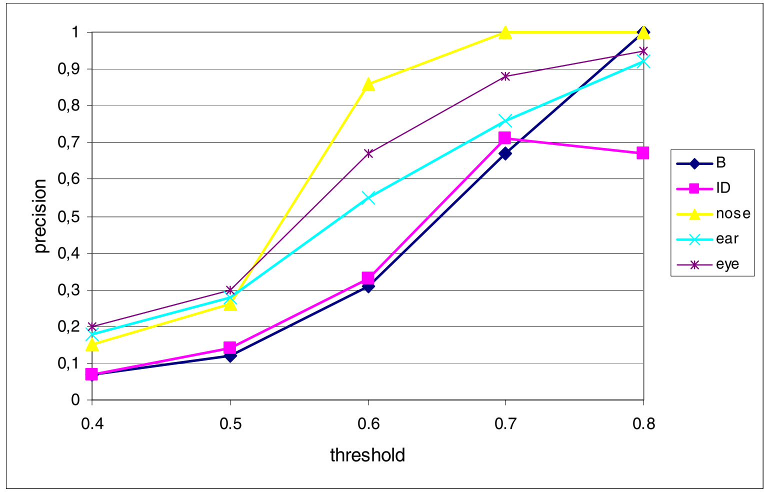
  Weights for combination: 1.0/1.2

KitAMO
(http://www.ida.liu.se/labs/iislab/projects/KitAMO)

Results

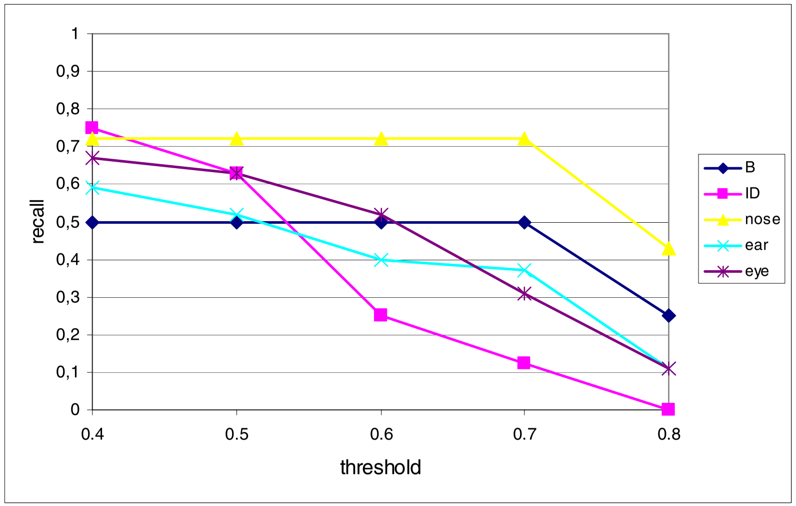
- Terminological matchers

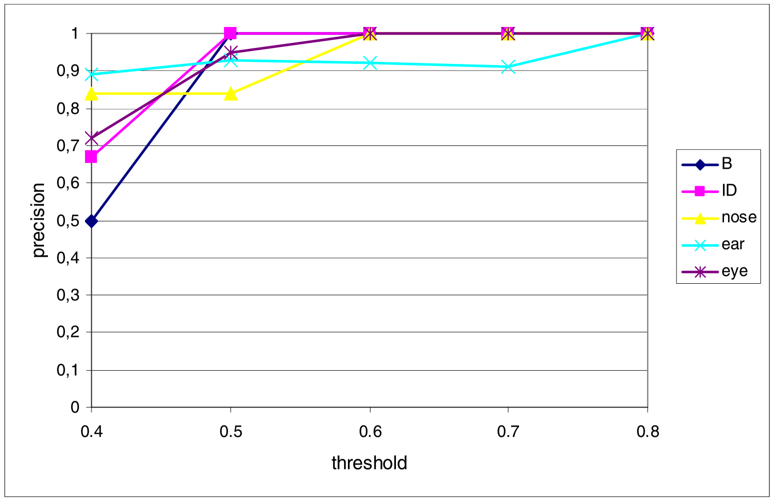




Results

- Basic learning matcher (Naïve Bayes)



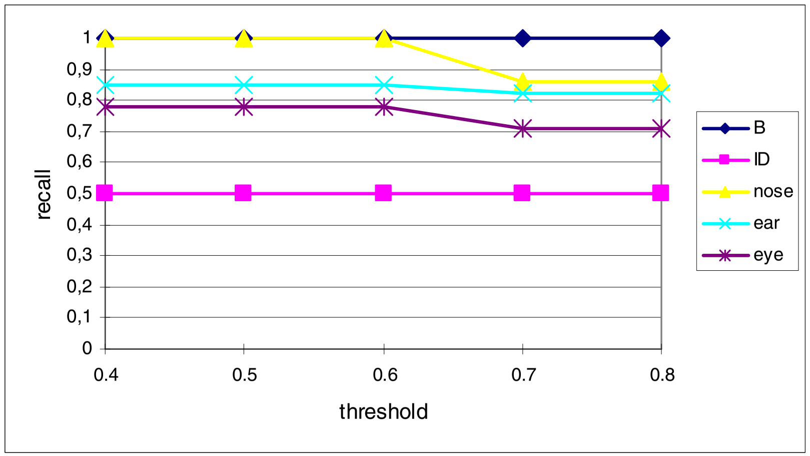


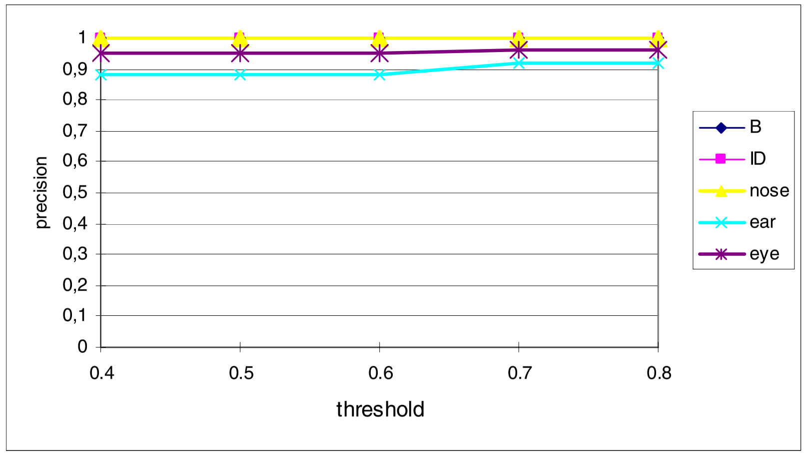
Naive Bayes slightly better recall, but slightly worse precision than SVM-single

SVM-multiple (much) better recall, but worse precision than SVM-single

Results

- Domain matcher (using UMLS)





Results

- Comparison of the matchers

  $CS\_TermWN \supseteq CS\_Dom \supseteq CS\_Learn$

- Combinations of the different matchers
  - combinations give often better results
  - no significant difference on the quality of suggestions for different weight assignments in the combinations

  (but: did not check for large variations for the weights)

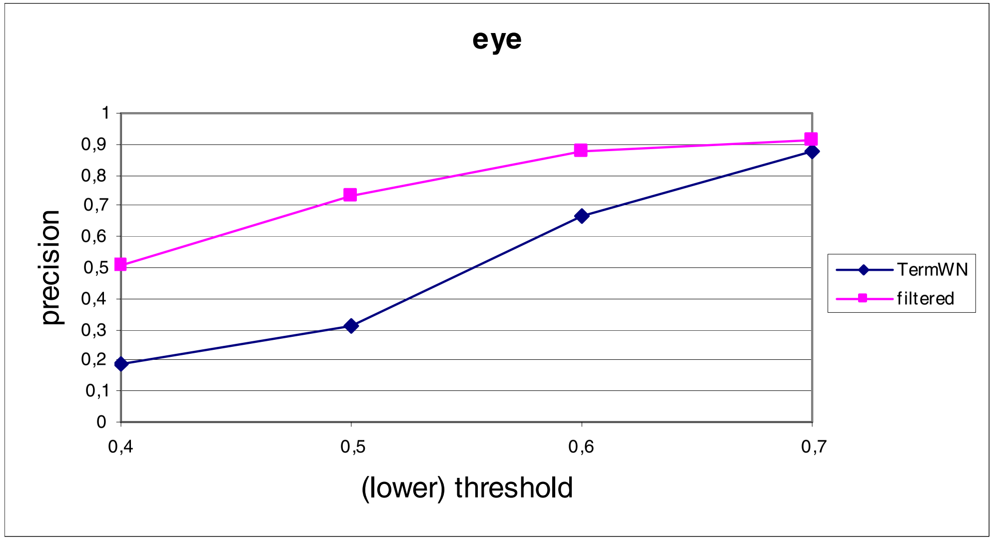
- Structural matcher did not find (many) new correct alignments

  (but: good results for systems biology schemas SBML – PSI MI)

Evaluation of filtering

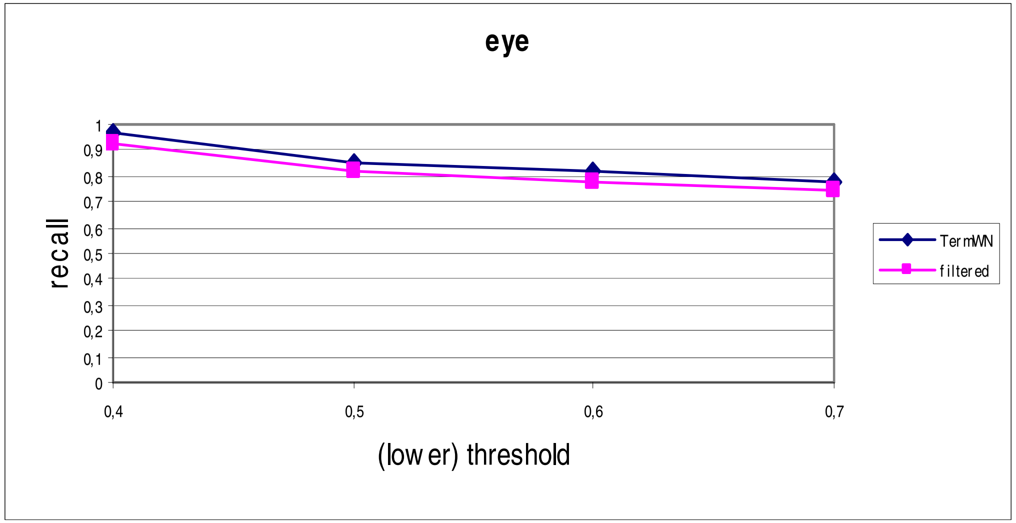
- Matcher
  - TermWN
- Parameters
  - Quality of suggestions: precision/recall
  - Double threshold filtering using structure:
    - Upper threshold: 0.8
    - Lower threshold: 0.4, 0.5, 0.6, 0.7, 0.8

Results



- The precision for double threshold filtering with upper threshold 0.8 and lower threshold T is higher than for threshold filtering with threshold T

Results



- The recall for double threshold filtering with upper threshold 0.8 and lower threshold T is about the same as for threshold filtering with threshold T

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Challenges

- Large-scale matching evaluation
- Efficiency of matching techniques
  - parallellization
  - distribution of computation
  - approximation of matching results (not complete)
  - modularization of ontologies
  - optimization of matching methods

Challenges

- Matching with background knowledge
  - partial alignments
  - reuse of previous matches
  - use of domain-specific corpora
  - use of domain-specific ontologies

- Matcher selection, combination and tuning
  - recommendation of algorithms and settings

Challenges

- User involvement
  - visualization
  - user feedback

- Explanation of matching results
- Social and collaborative matching
- Alignment management: infrastructure and support

Further reading

Starting points for further studies

Further reading ontology alignment

- http://www.ontologymatching.org
(plenty of references to articles and systems)

- Ontology alignment evaluation initiative: http://oaei.ontologymatching.org
(home page of the initiative)

- Euzenat, Shvaiko, *Ontology Matching*, Springer, 2007.

- Shvaiko, Euzenat, Ontology Matching: state of the art and future challenges, *IEEE Transactions on Knowledge and Data Engineering* 25(1):158-176, 2013.

- Lambrix P, Kaliyaperumal R, Contributions of LiU/ADIT to Ontology Alignment, in Lambrix, (ed), *Advances in Secure and Networked Information Systems - The ADIT Perspective*, 97-108, LiU Tryck / LiU Electronic Press, 2012. http://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A573657&dswid=-155

Further reading ontology alignment

Systems at LiU / IDA / ADIT

- Lambrix, Tan, SAMBO – a system for aligning and merging biomedical ontologies, *Journal of Web Semantics*, 4(3):196-206, 2006.

  (description of the SAMBO tool and overview of evaluations of different matchers)

- Lambrix, Tan, A tool for evaluating ontology alignment strategies, *Journal on Data Semantics*, VIII:182-202, 2007.

  (description of the KitAMO tool for evaluating matchers)

- Lambrix P, Kaliyaperumal R, A Session-based Approach for Aligning Large Ontologies, *Tenth Extended Semantic Web Conference - ESWC 2013*, LNCS 7882, 46-60, 2013.

Further reading ontology alignment

- Chen, Tan, Lambrix, Structure-based filtering for ontology alignment,*IEEE WETICE workshop on semantic technologies in collaborative applications,* 364-369, 2006.

(double threshold filtering technique)

- Tan, Lambrix, A method for recommending ontology alignment strategies, *International Semantic Web Conference*, 494-507, 2007.

  Ehrig, Staab, Sure, Bootstrapping ontology alignment methods with APFEL, *International Semantic Web Conference*, 186-200, 2005.

  Mochol, Jentzsch, Euzenat, Applying an analytic method for matching approach selection, *International Workshop on Ontology Matching*, 2006.

(recommendation of alignment strategies)

- Lambrix, Liu, Using partial reference alignments to align ontologies, *European Semantic Web Conference*, 188-202, 2009.

(use of partial alignments in ontology alignment)

Further reading ontology alignment

- Lambrix, Strömbäck, Tan, Information integration in bioinformatics with ontologies and standards, chapter 8 in Bry, Maluszynski (eds), *Semantic Techniques for the Web*, Springer, 2009. ISBN: 978-3-642-04580-6.

(largest overview of systems)

Ontology Debugging

Defects in ontologies

- Syntactic defects
  - E.g. wrong tags or incorrect format
- Semantic defects
  - E.g. unsatisfiable concepts, incoherent and inconsistent ontologies
- Modeling defects
  - E.g. wrong or missing relations

Example - incoherent ontology

- Example: DICE ontology
  - Brain ⊑ CentralNervousSystem ⊓ BodyPart ⊓ ∃systempart.NervousSystem ⊓ ∃ region.HeadAndNeck ⊓ ∀region.HeadAndNeck

    A brain is a central nervous system and a body part which has a system part that is a nervous system and that is in the head and neck region.

  - CentralNervousSystem ⊑ NervousSystem

    A central nervous system is a nervous system.

  - BodyPart ⊑¬NervousSystem

    Nothing can be at the same time a body part and a nervous system.

Slide from G. Qi

Example - inconsistent ontology

- **Example from Foaf:**
  - **Person(timbl)**
  - **Homepage(timbl, http://w3.org/)**
  - **Homepage(w3c, http://w3.org/)**
  - **Organization(w3c)**
  - **InverseFunctionalProperty(Homepage)**
  - **DisjointWith(Organization, Person)**
- **Example from OpenCyc:**
  - **ArtifactualFeatureType(PopulatedPlace)**
  - **ExistingStuffType(PopulatedPlace)**
  - **DisjointWith(ExistingObjectType,ExistingStuffType)**
  - **ArtifactualFeatureType $\sqsubseteq$ ExistingObjectType**

Slide from G. Qi

Example - missing is-a relations

- In 2008 Ontology Alignment Evaluation Initiative (OAEI) Anatomy track, task 4
  - Ontology MA : Adult Mouse Anatomy Dictionary (2744 concepts)
  - Ontology NCI-A : NCI Thesaurus - anatomy (3304 concepts)
  - 988 mappings between MA and NCI-A
    - 121 missing is-a relations in MA
    - 83 missing is-a relations in NCI-A

Influence of missing structure

- Ontology-based querying.

Medical Subject Headings (MeSH)

All MeSH Categories
- **Diseases Category**
  - **Eye Diseases**
    - **Scleral Diseases**
      - **Scleritis**

...

return 1617 articles

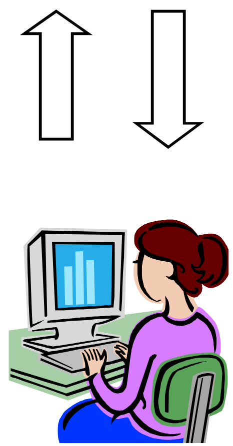
Influence of missing structure

- Incomplete results from ontology-based queries

Medical Subject Headings (MeSH)

All MeSH Categories
- **Diseases Category**
  - **Eye Diseases**
    - **Scleral Diseases**
      - ~~Scleritis~~

...



return 1617 articles

return 695 articles

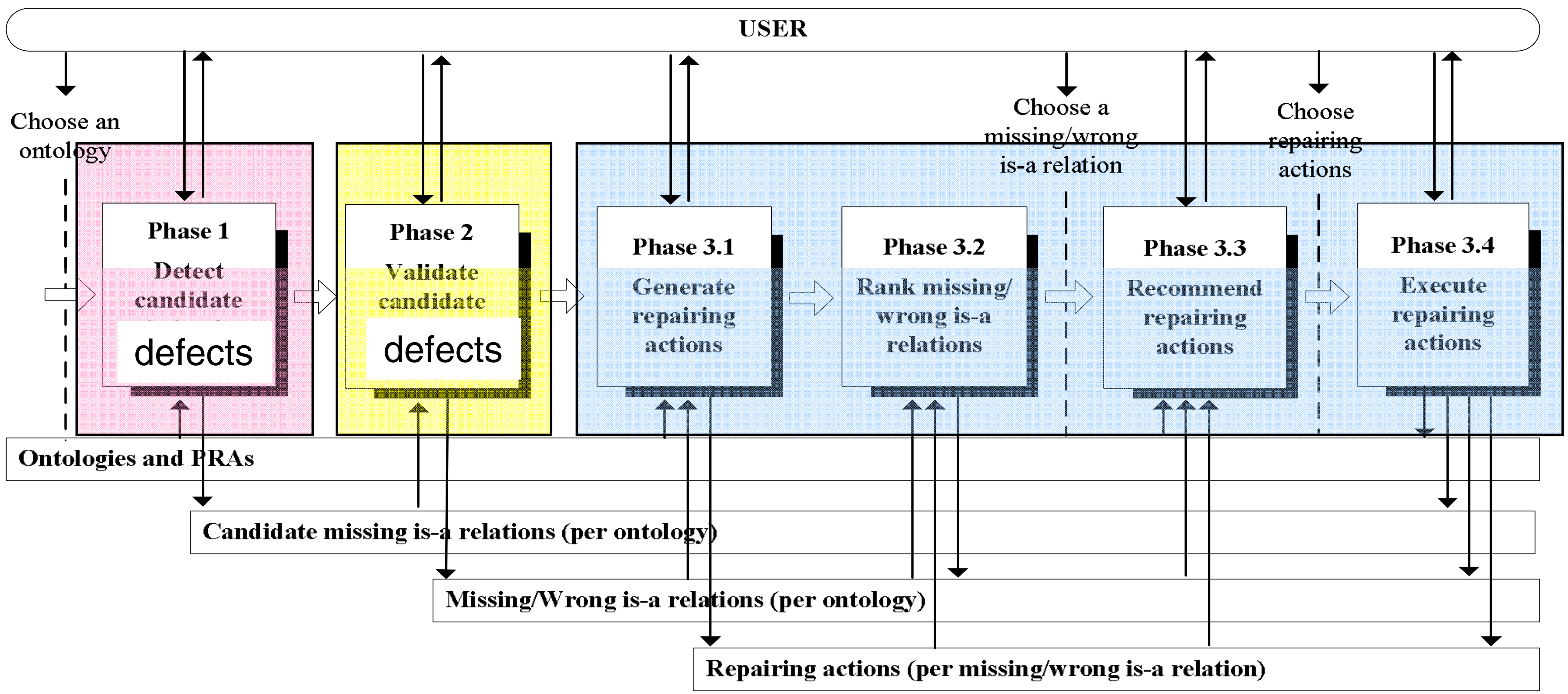
57% results are missed !

Defects in ontologies and ontology networks

- Ontologies and ontology networks with defects, although often useful, also lead to problems when used in semantically-enabled applications.

→ Wrong conclusions may be derived or valid conclusions may be missed.

Overview of debugging approach



Debugging semantic defects

Example : an Incoherent Ontology

Consider the following TBox $\mathcal{T}^*$, where A, B and C are primitive and $A_1, \ldots, A_7$ defined concept names:

$ax_1: A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$	$ax_2: A_2 \dot{\sqsubseteq} A \sqcap A_4$
$ax_3: A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$	$ax_4: A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$
$ax_5: A_5 \dot{\sqsubseteq} \exists s.\neg B$	$ax_6: A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$
$ax_7: A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$	

$\Downarrow$

DL Reasoner $\Rightarrow$

The ontology is incoherent!

The set of unsatisfiable concepts are : $\{A_1, A_3, A_6, A_7\}$.

What are the root causes of these defects?

Explain the Semantic Defects

- We need to identify the sets of axioms which are necessary for causing the logic contradictions.

$ax_1 : A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$	$ax_2 : A_2 \dot{\sqsubseteq} A \sqcap A_4$
$ax_3 : A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$	$ax_4 : A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$
$ax_5 : A_5 \dot{\sqsubseteq} \exists s.\neg B$	$ax_6 : A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$
$ax_7 : A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$	

- For example, for the unsatisfiable concept "A_1", there are two sets of axioms.

$$\boxed{\begin{array}{l} ax_1 : A_1 \dot{\sqsubseteq} \underline{\neg A} \sqcap A_2 \sqcap A_3 \\ ax_2 : A_2 \dot{\sqsubseteq} \underline{A} \sqcap A_4 \end{array}}$$

$$\boxed{\begin{array}{l} ax_1 : A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap \underline{A_3} \\ ax_3 : A_3 \dot{\sqsubseteq} \underline{A_4 \sqcap A_5} \\ ax_4 : A_4 \dot{\sqsubseteq} \underline{\forall s.B} \sqcap C \\ ax_5 : A_5 \dot{\sqsubseteq} \underline{\exists s.\neg B} \end{array}}$$

Minimal Unsatisfiability Preserving Sub-TBoxes (MUPS)

Definition 1 Let A be a concept which is unsatisfiable in a TBox $\mathcal{T}$. A set $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal unsatisfiability-preserving sub-TBox (MUPS)* of $\mathcal{T}$ if

- A is unsatisfiable in $\mathcal{T}'$, and
- A is satisfiable in every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$.

We will abbreviate the set of MUPS of $\mathcal{T}$ and A by $mups(\mathcal{T}, A)$.

$$mups(\mathcal{T}^*, A_1) = \{\{ax_1, ax_2\}, \{ax_1, ax_3, ax_4, ax_5\}\}$$

- The MUPS of an unsatisfiable concept imply the solutions for repairing.

  → Remove at least one concept from each axiom set in the MUPS

Example

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{\cancel{ax_1}, ax_2\}, \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{\cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{\cancel{ax_1}, ax_2, \cancel{ax_4}, ax_6\}, \\
&\quad \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{\cancel{ax_4}, ax_7\}\}
\end{aligned}
$$

- Possible ways of repairing all the unsatisfiable concepts in the ontology:

$$\{ax_1, ax_3, ax_4\}$$

How to represent all these possibilities?

Minimal Incoherence Preserving Sub-TBox (MIPS)

Definition 2 Let $\mathcal{T}$ be an incoherent TBox. A TBox $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal incoherence-preserving sub-TBox (MIPS) of* $\mathcal{T}$ if

- $\mathcal{T}'$ is incoherent, and
- every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$ is coherent.

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{ax_1, \cancel{ax_2}\}, \{ax_1, ax_3, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{ax_3, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{ax_1, \cancel{ax_2}, \cancel{ax_4}, ax_6\}, \\
&\quad\ \{ax_1, ax_3, \cancel{ax_4}, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{\cancel{ax_4}, \cancel{ax_7}\}\}
\end{aligned}
$$

We will abbreviate the set of MIPS of $\mathcal{T}$ by $mips(\mathcal{T})$. For $\mathcal{T}^*$ we get three MIPS:

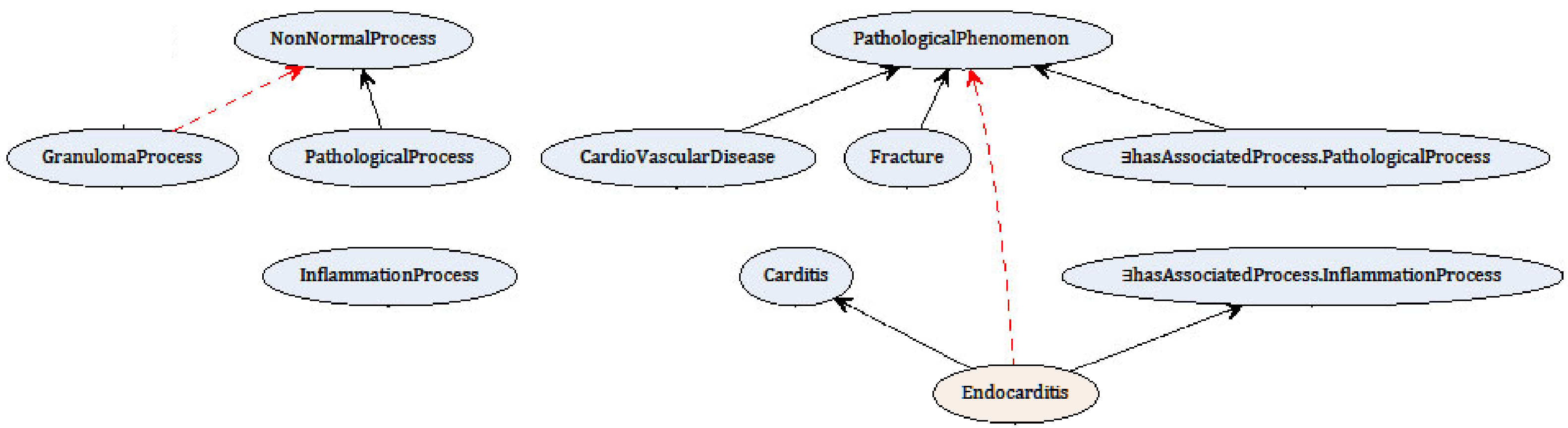
$$mips(\mathcal{T}^*) = \{\{ax_1, ax_2\}, \{ax_3, ax_4, ax_5\}, \{ax_4, ax_7\}\}$$

A possible repairing is $\{ax_i\} \cup \{ax_j\} \cup \{ax_k\}$, where

- $ax_i \in \{ax_1, \cancel{ax_2}\}$
- $ax_j \in \{ax_3, \cancel{ax_4}, ax_5\}$
- $ax_k \in \{ax_4, \cancel{ax_7}\}$

Completing the is-a structure of ontologies

Example



Repairing actions:

$\{\text{Endocarditis} \dot{\sqsubseteq} \text{PathologicalPhenomenon}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{NonNormalProcess}\}$

$\{\text{Carditis} \dot{\sqsubseteq} \text{CardioVascularDisease}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{PathologicalProcess}\}$

$\{\text{Carditis} \dot{\sqsubseteq} \text{Fracture}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{NonNormalProcess}\}$

Description logic EL

- Concepts

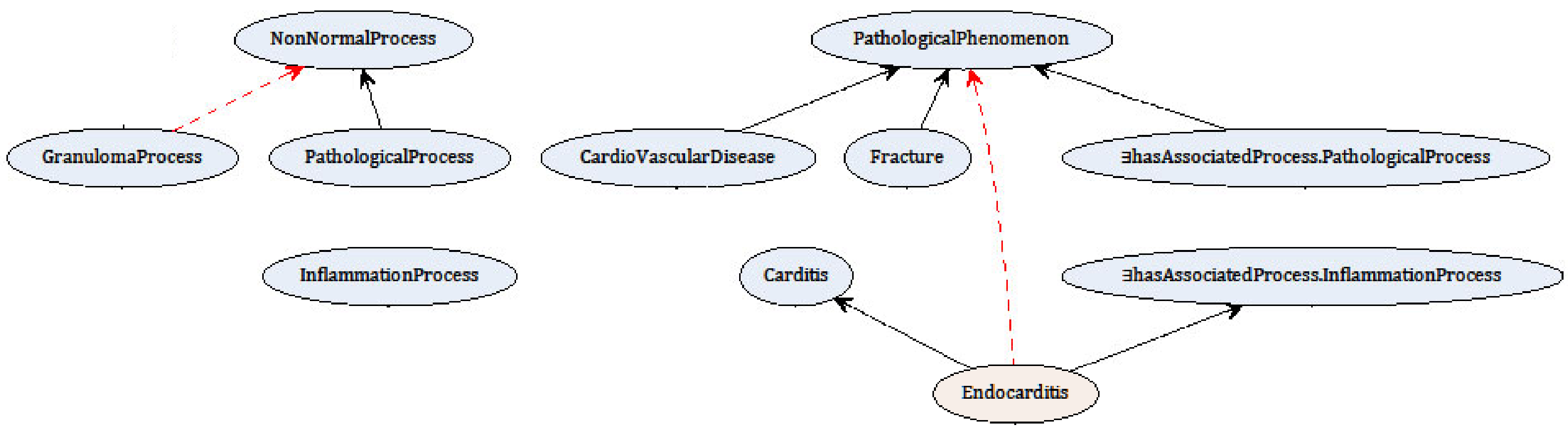
Atomic concept	A
Universal concept	$\top$
Intersection of concepts	$C \sqcap D$
Existential restriction	$\exists r.C$

- Terminological axioms: equivalence and subsumption

Generalized Tbox Abduction Problem – GTAP(**T**,**C**,Or,M)

- Given
  - **T**- a Tbox in EL
  - **C**- a set of atomic concepts in **T**
  - $M = \{A_i \sqsubseteq B_i\}_{i=1..n}$ and $\forall$ i:1..n: $A_i, B_i \in$ **C**
  - Or: $\{C_i \sqsubseteq D_i \mid C_i, D_i \in$ **C**$\} \rightarrow \{true, false\}$
- Find
  - $S = \{E_i \sqsubseteq F_i\}_{i=1..k}$ such that
    $\forall$ i:1..k: $E_i, F_i \in$ **C** and $Or(E_i \sqsubseteq F_i) = true$
    and T U S is consistent and T U S |= M

GTAP - example



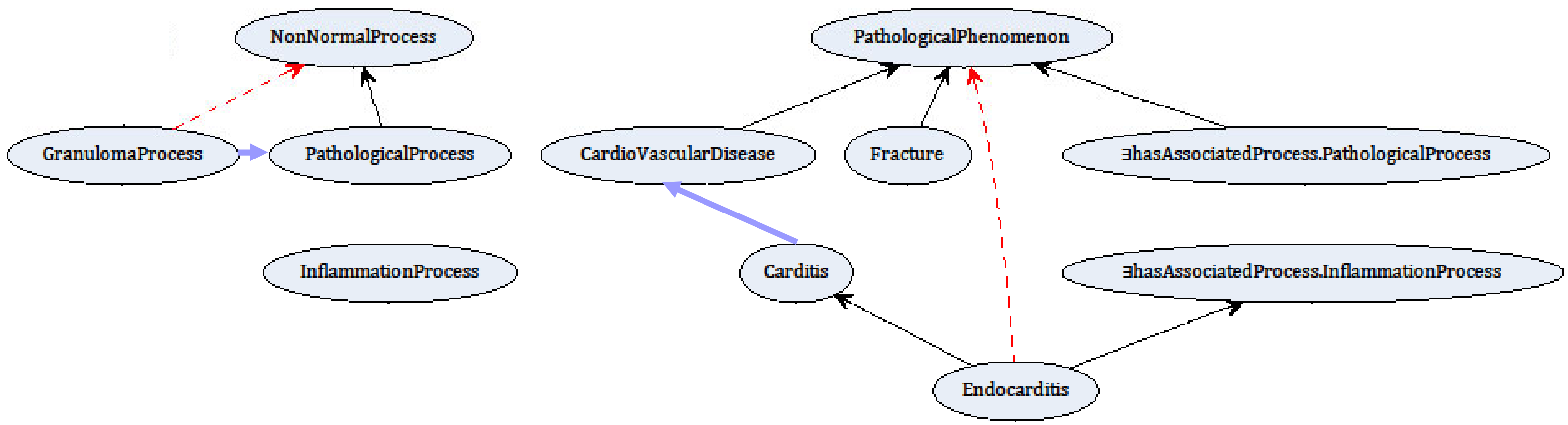
C = { GranulomaProcess, CardioVascularDisease, PathologicalPhenomenon, Fracture, Endocarditis, Carditis, InflammationProcess, PathologicalProcess, NonNormalProcess}

T = { GranulomaProcess $\dot{\sqsubseteq}$ $\top$, hasAssociatedProcess $\dot{\sqsubseteq}$ $\top \times \top$, CardioVascularDisease $\dot{\sqsubseteq}$ PathologicalPhenomenon, Fracture $\dot{\sqsubseteq}$ PathologicalPhenomenon, ∃hasAssociatedProcess.PathologicalProcess $\dot{\sqsubseteq}$ PathologicalPhenomenon, Endocarditis $\dot{\sqsubseteq}$ Carditis, Endocarditis $\dot{\sqsubseteq}$ ∃hasAssociatedProcess.InflammationProcess, PathologicalProcess $\dot{\sqsubseteq}$ NonNormalProcess }

M = { Endocarditis $\dot{\sqsubseteq}$ PathologicalPhenomenon, GranulomaProcess $\dot{\sqsubseteq}$ NonNormalProcess }

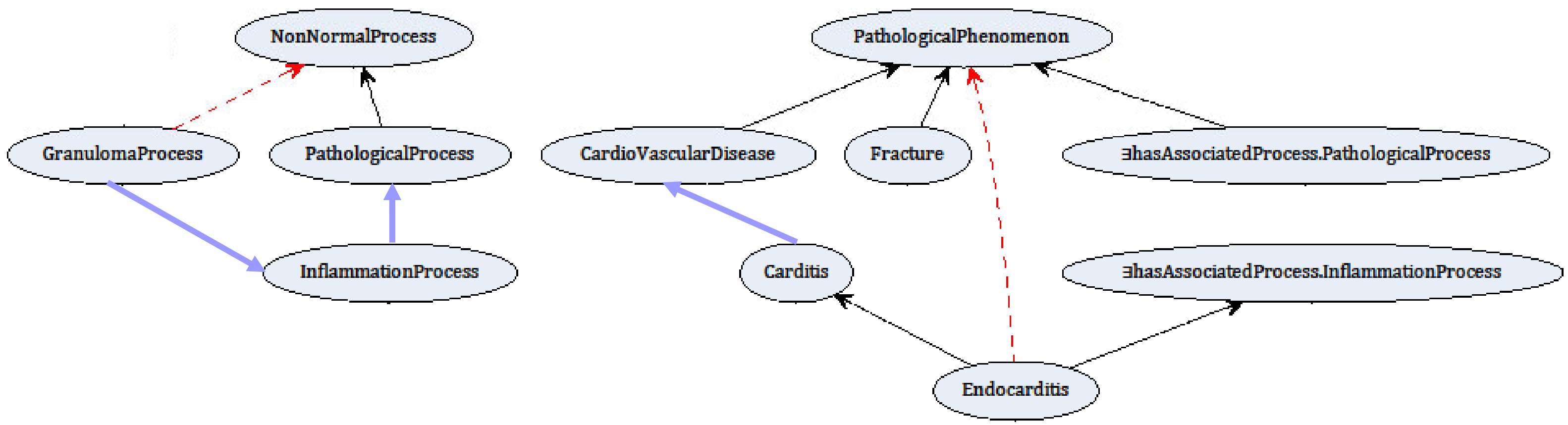
Preference criteria

- There can be many solutions for GTAP



Preference criteria

- There can be many solutions for GTAP



Not all are equally interesting.

More informative

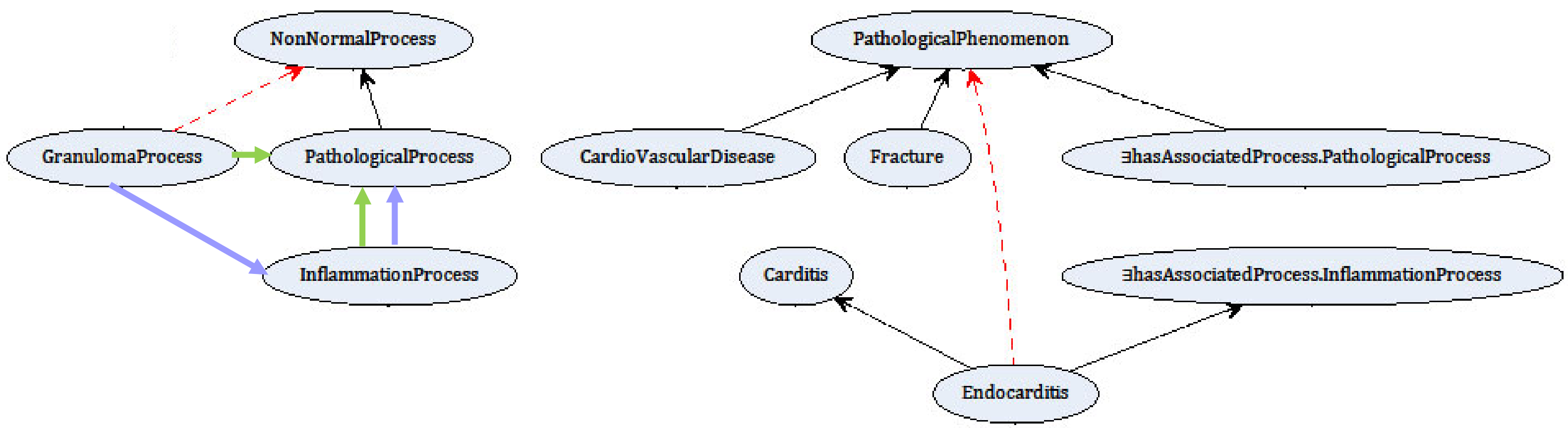
- Let *S* and *S'* be two solutions to GTAP(**T**,**C**,Or,M). Then,

 - *S* is more informative than *S'*
iff **T** U *S* |= *S'* but not **T** U *S'* |= *S*

 - *S* is equally informative as *S'*
iff **T** U *S* |= *S'* and **T** U *S'* |= *S*

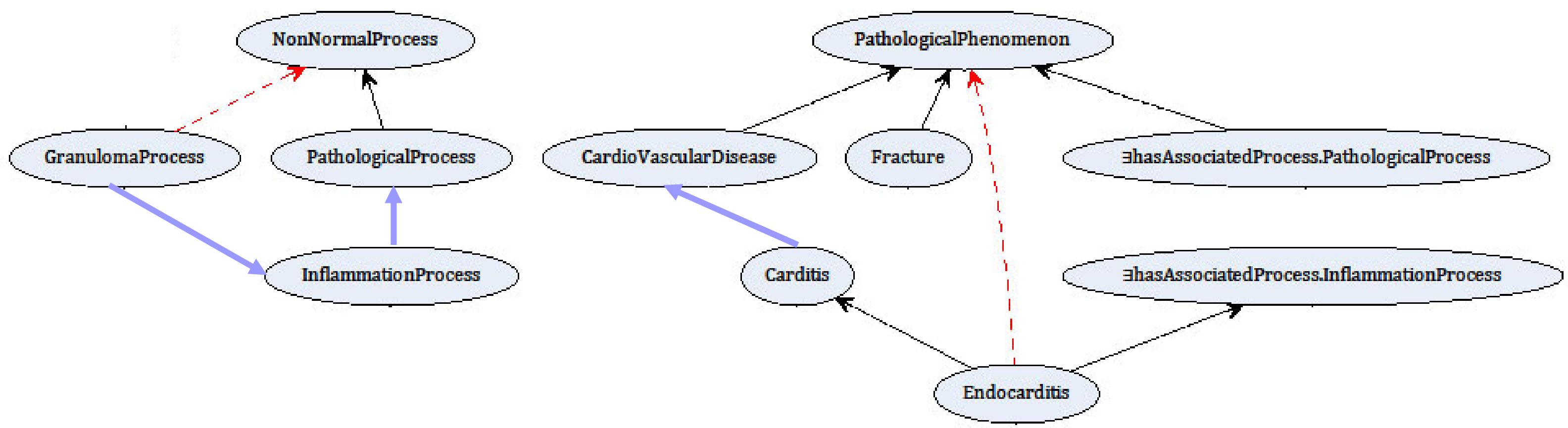
More informative

- ’Blue’ solution is more informative than ’green’ solution



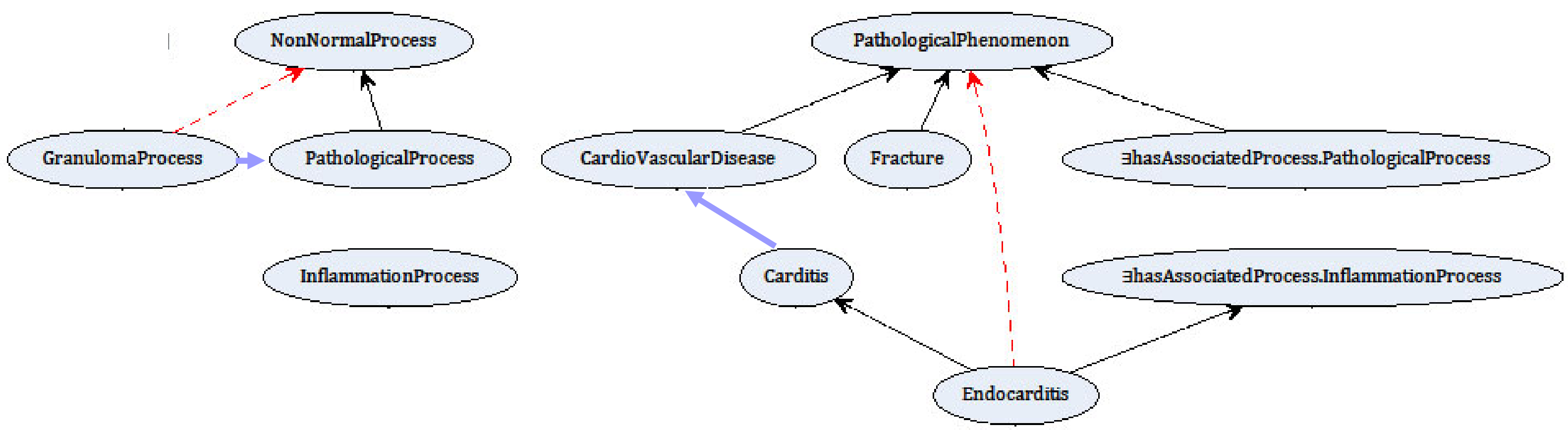
Semantic maximality

- A solution *S* to GTAP(**T**,**C**,Or,M) is semantically maximal iff there is no solution *S*' which is more informative than *S*.



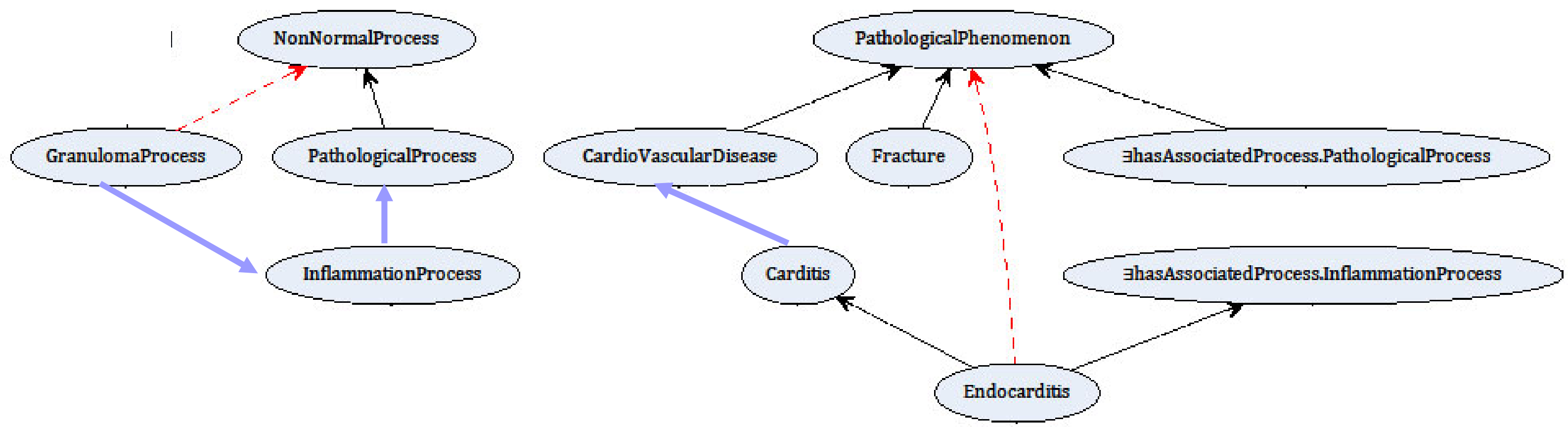
Subset minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is subset minimal iff there is no proper subset *S'* of *S* that is a solution.



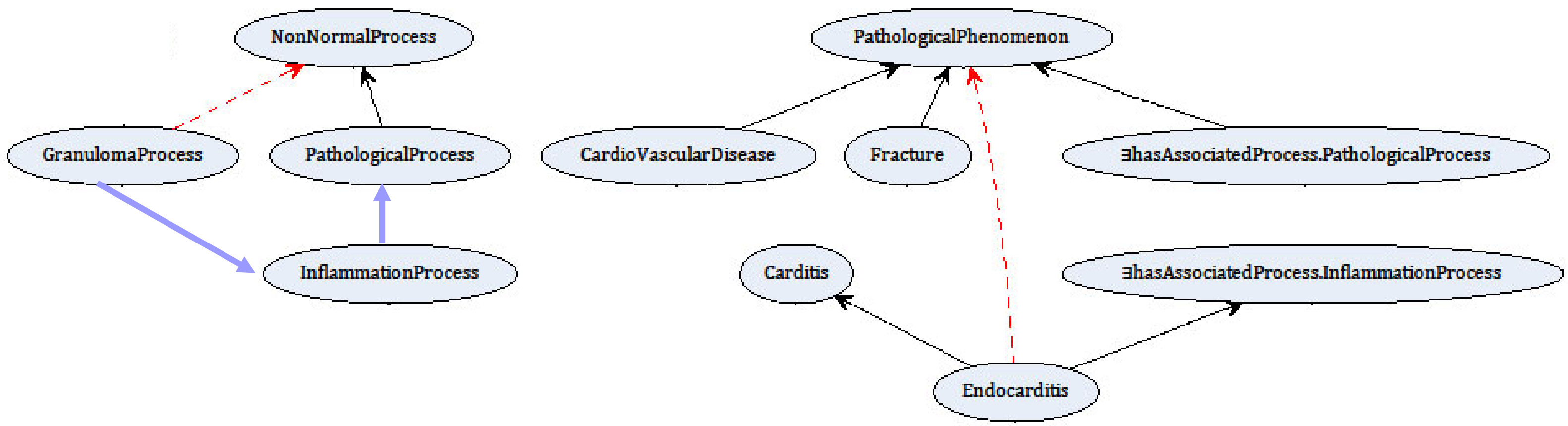
Combining with priority for semantic maximality

- A solution *S* to GTAP(**T**,**C**,Or,M) is maxmin optimal iff *S* is semantically maximal and there is no other semantically maximal solution that is a proper subset of *S*.



Combining with priority for subset minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is minmax optimal iff *S* is subset minimal and there is no other subset minimal solution that is more informative than *S*.



Combining with equal preferences

- A solution *S* to GTAP(**T**,**C**,Or,M) is skyline optimal iff there is no other solution that is a proper subset of *S* and that is equally informative than *S*.
  - All subset minimal, minmax optimal and maxmin optimal solutions are also skyline optimal solutions.
  - Semantically maximal solutions may or may not be skyline optimal.

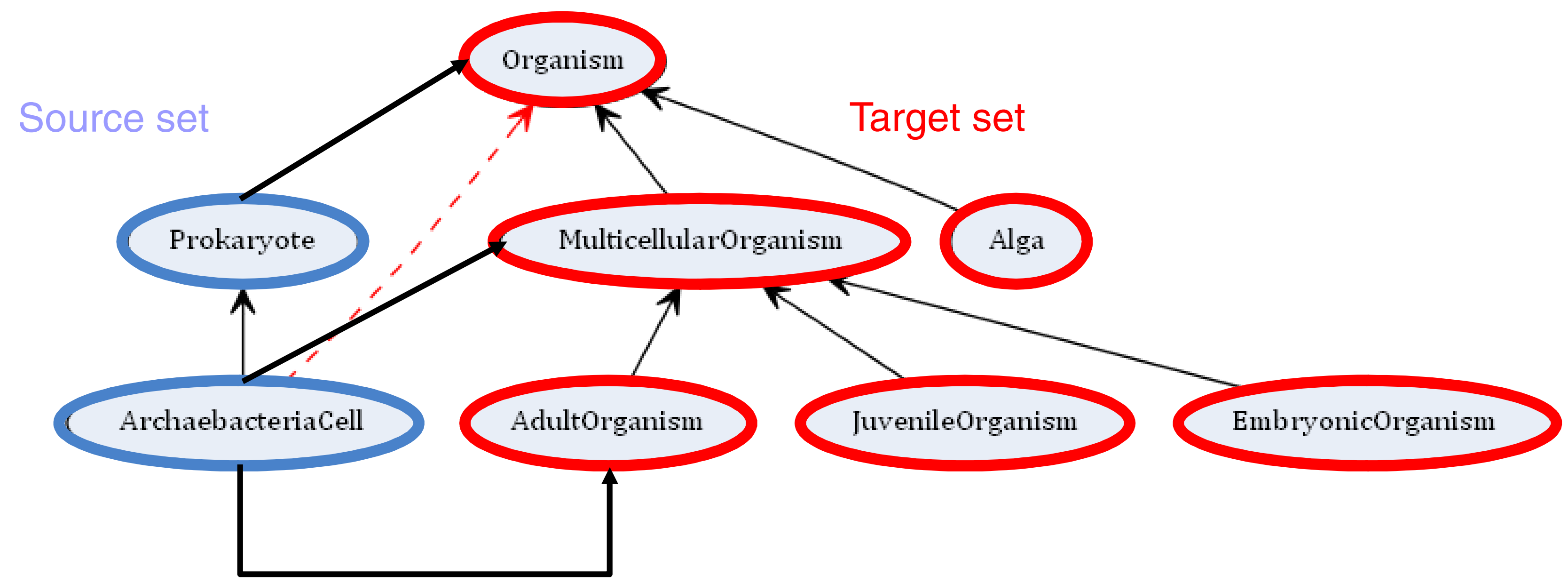
Preference criteria - conclusions

- In practice it is not clear how to generate maxmin or semantically maximal solutions (the preferred solutions)
- Skyline optimal solutions are the next best thing and are easy to generate

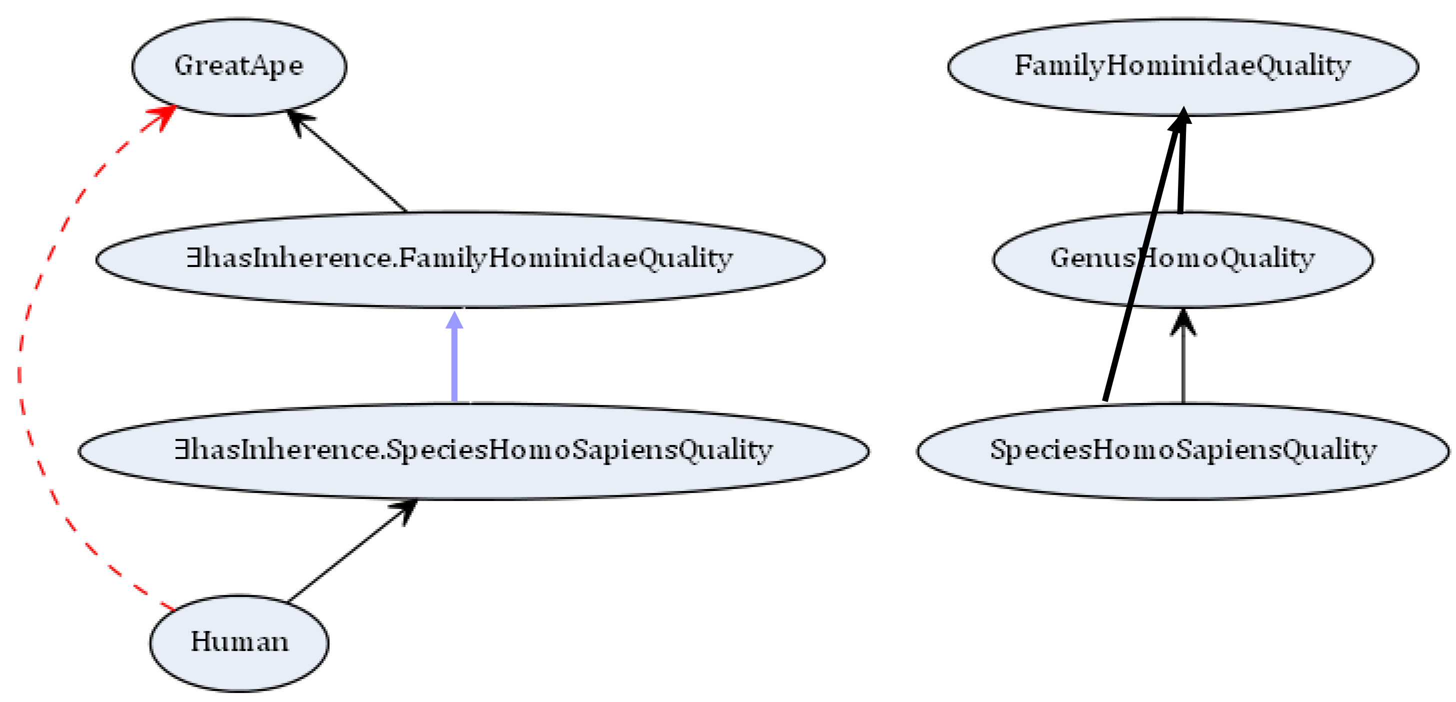
Approach

- Input
  - Normalized EL - TBox
  - Set of missing is-a relations (correct according to the domain)
- Output – a skyline-optimal solution to GTAP
- Iteration of three main steps:
  - Creating solutions for individual missing is-a relations
  - Combining individual solutions
  - Trying to improve the result by finding a solution which introduces additional new knowledge (more informative)

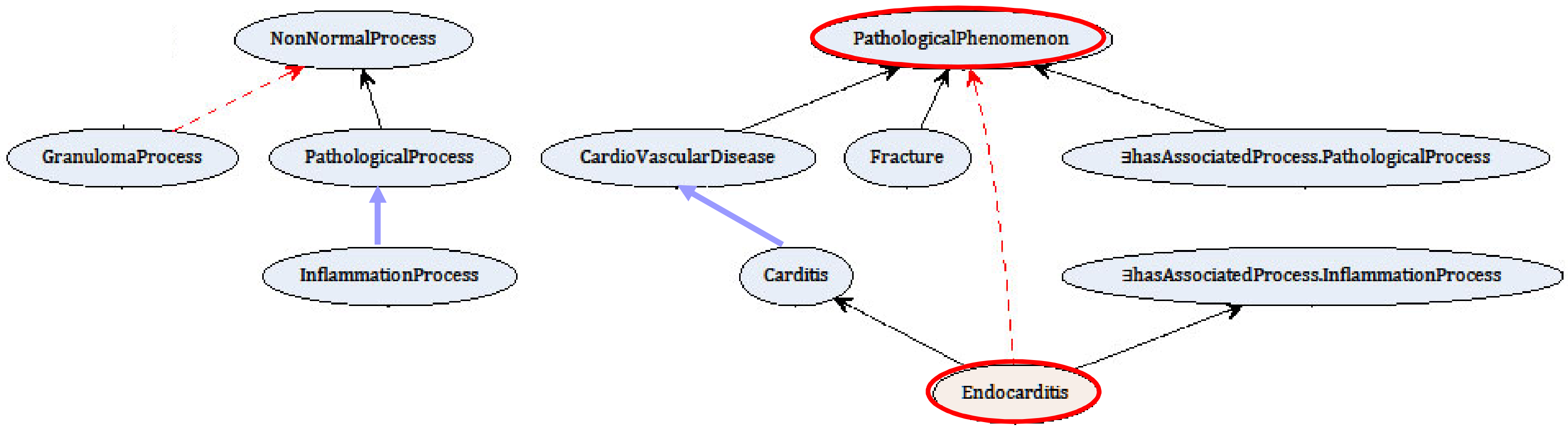
Intuition 1


Source set
Target set
Organism
Prokaryote
MulticellularOrganism
Alga
ArchaebacteriaCell
AdultOrganism
JuvenileOrganism
EmbryonicOrganism

Intuitions 2/3



Example – repairing single is–a relation



~~Endocarditis $\dot\sqsubseteq$ PathologicalPhenomenon~~

~~Endocarditis $\dot\sqsubseteq$ Fracture~~ false

~~Endocarditis $\dot\sqsubseteq$ CardioVascularDisease~~

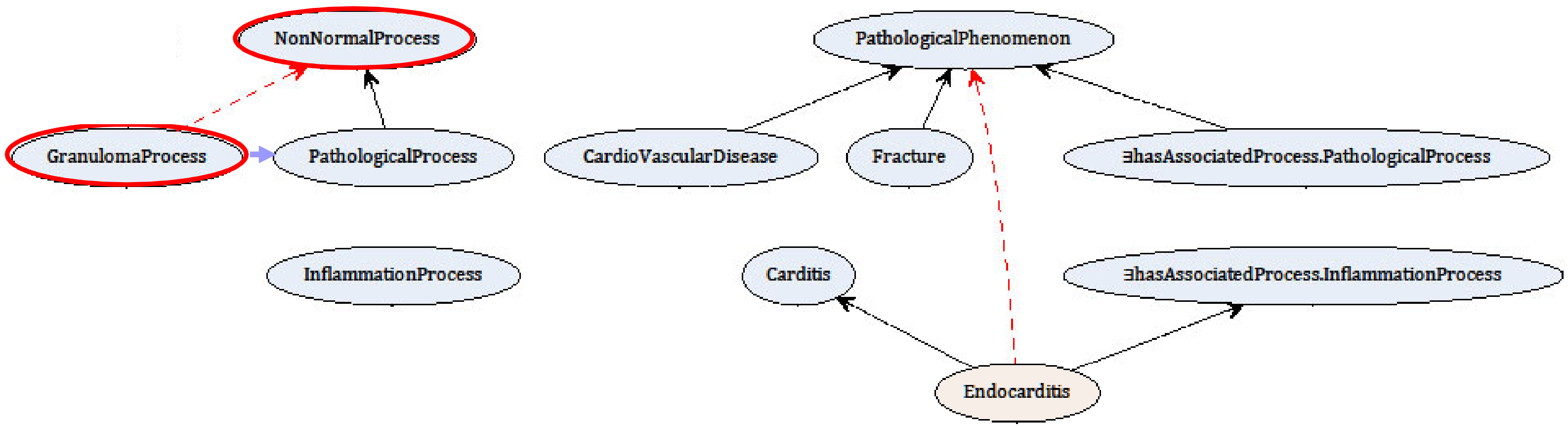
~~Carditis $\dot\sqsubseteq$ PathologicalPhenomenon~~

~~Carditis $\dot\sqsubseteq$ Fracture~~ false

Carditis $\dot\sqsubseteq$ CardioVascularDisease

InflammationProcess $\dot\sqsubseteq$ PathologicalProcess

Example – repairing single is–a relation



$\cancel{\text{GranulomaProcess} \ \dot{\sqsubseteq} \ \text{NonNormalProcess}}$

$\text{GranulomaProcess} \ \dot{\sqsubseteq} \ \text{PathologicalProcess}$

Algorithm - Repairing multiple is-a relations

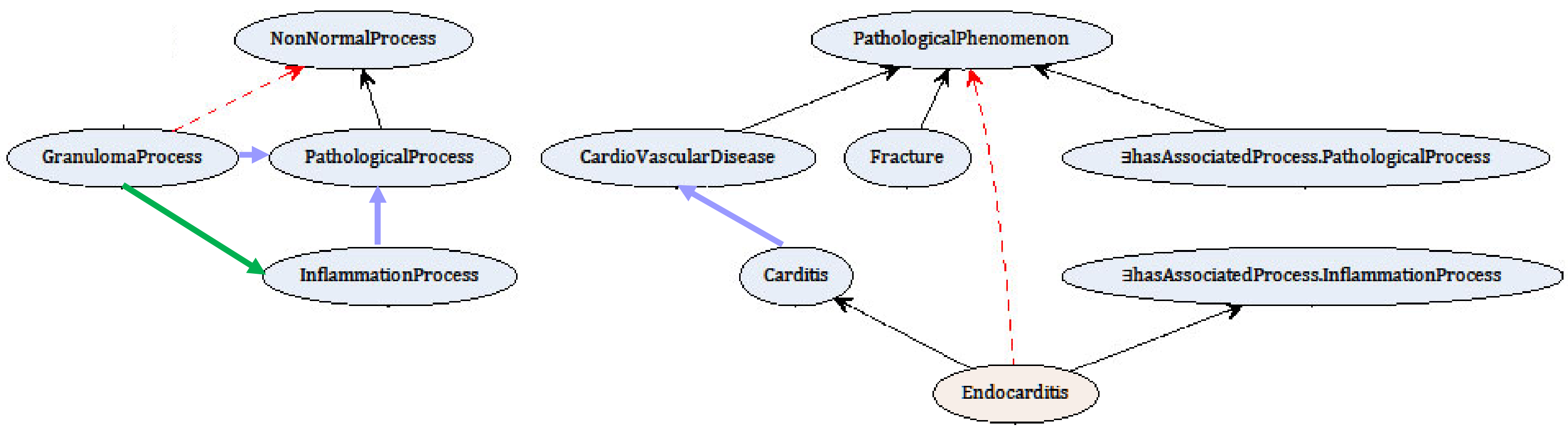
- Combine solutions for individual missing is-a relations
- Remove redundant relations while keeping the same level of informativness
- Resulting solution is a skyline optimal solution

$$
\begin{aligned}
&\{\text{InflammationProcess} \mathrel{\dot{\sqsubseteq}} \text{PathologicalProcess},\\
&\text{Carditis} \mathrel{\dot{\sqsubseteq}} \text{CardioVascularDisease},\\
&\text{GranulomaProcess} \mathrel{\dot{\sqsubseteq}} \text{PathologicalProcess}\}
\end{aligned}
$$

Algorithm – improving solution

- Solution S from previous step may contain relations which are not derivable from the ontology.
- These can be seen as new missing is-a relations.
- We can solve a new GTAP problem: GTAP(**T** U *S*, **C**, Or, *S*)

Example – improving solutions



$$\text{GranulomaProcess} \;\dot{\sqsubseteq}\; \text{InflammationProcess}$$

$$Or(GranulomaProcess \;\dot{\sqsubseteq}\; InflammationProcess) =$$

$$\{\text{InflammationProcess} \;\dot{\sqsubseteq}\; \text{PathologicalProcess},\\ \text{Carditis} \;\dot{\sqsubseteq}\; \text{CardioVascularDisease},\\ \text{GranulomaProcess} \;\dot{\sqsubseteq}\; \text{InflammationProcess}\}$$

Algorithm properties

- Sound
- Skyline optimal solutions

Experiments

Two use-cases

- Case 1: given missing is-a relations
  AMA and a fragment of NCI-A ontology – OAEI 2013
  - AMA (2744 concepts) – 94 missing is-a relations
    → 3 iterations, 101 in repairing (47 additional new knowledge)
  - NCI-A (3304 concepts) – 58 missing is-a relations
    → 3 iterations, 54 in repairing (10 additional new knowledge)

- Case 2: no given missing is-a relations
  Modified BioTop ontology
  - Biotop (280 concepts, 42 object properties)
    randomly choose is-a relations and remove them: 47 'missing'
    → 4 iterations, 41 in repairing (40 additional new knowledge)

Further reading

Starting points for further studies

Further reading ontology debugging

- http://www.ida.liu.se/~patla/DOOM/

Semantic defects

- Schlobach S, Cornet R. Non-Standard Reasoning Services for the Debugging of Description Logic Terminologies. *18th International Joint Conference on Artificial Intelligence - IJCAI03*, 355-362, 2003.
- Schlobach S. Debugging and Semantic Clarification by Pinpointing. *2nd European Semantic Web Conference - ESWC05*, LNCS 3532, 226-240, 2005.

Further reading ontology debugging

Completing ontologies

- Fang Wei-Kleiner, Zlatan Dragisic, Patrick Lambrix. Abduction Framework for Repairing Incomplete EL Ontologies: Complexity Results and Algorithms. 28th AAAI Conference on Artificial Intelligence - AAAI 2014, 1120-1127, 2014.

- Lambrix P, Ivanova V, A unified approach for debugging is-a structure and mappings in networked taxonomies, *Journal of Biomedical Semantics* 4:10, 2013.

- Lambrix P, Liu Q, Debugging the missing is-a structure within taxonomies networked by partial reference alignments, *Data & Knowledge Engineering* 86:179-205, 2013.

Further reading ontology debugging

- Lambrix P, Ivanova V, Dragisic Z, Contributions of LiU/ADIT to Debugging Ontologies and Ontology Mappings, in Lambrix, (ed), *Advances in Secure and Networked Information Systems - The ADIT Perspective*, 109-120, LiU Tryck / LiU Electronic Press, 2012. http://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A573657&dswid=4198