Using Partial Reference Alignments to Align Ontologies

Patrick Lambrix, Qiang Liu

Linköpings Universitet

Ontology Alignment

- Many ontologies have been developed

  Many of them have overlapping information

- Use of multiple ontologies

  e.g. custom-specific ontology + standard ontology

- Bottom-up creation of ontologies

  experts can focus on their domain of expertise

  Important to know the inter-ontology relationships

Ontology Alignment

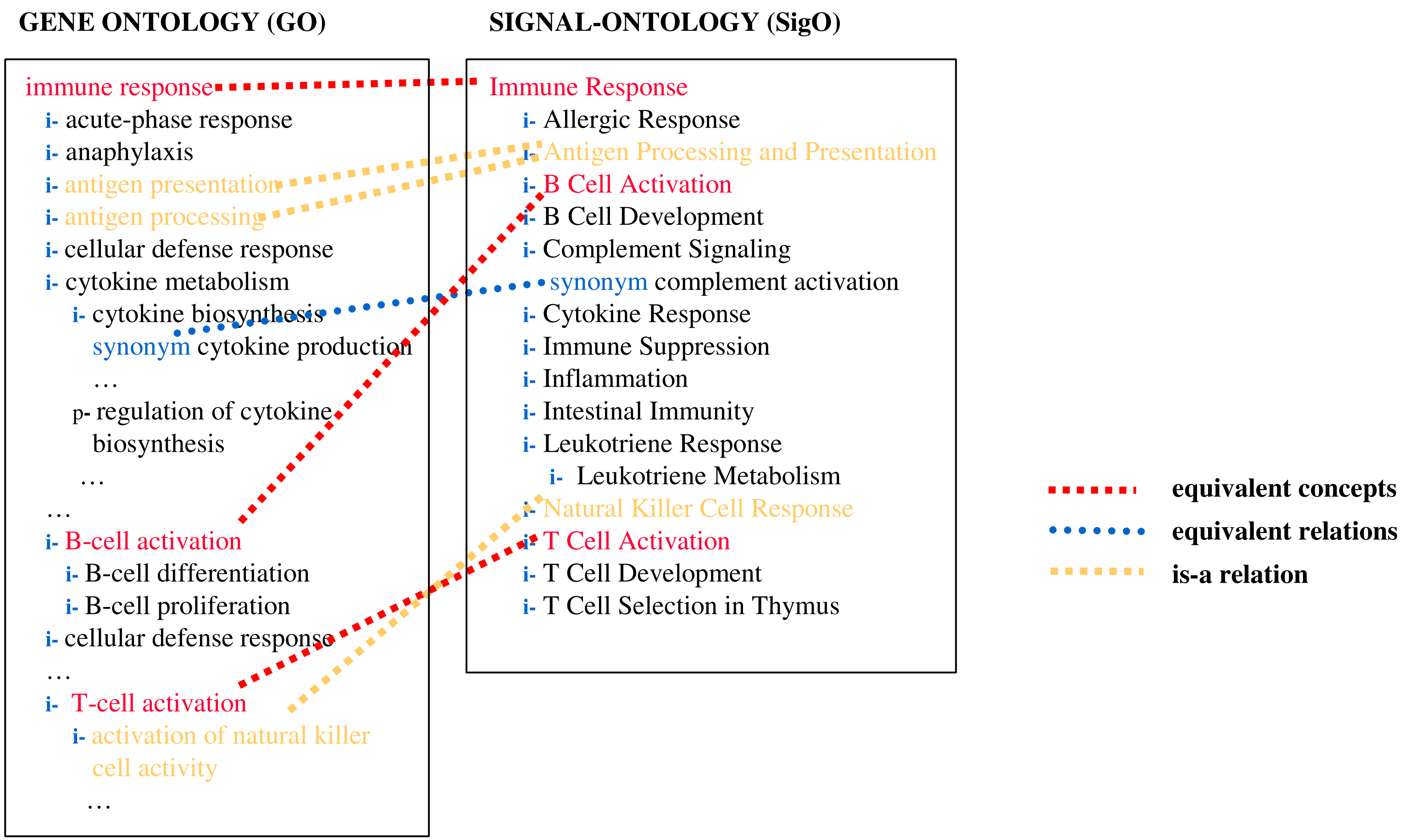
GENE ONTOLOGY (GO)

```
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



GENE ONTOLOGY (GO)

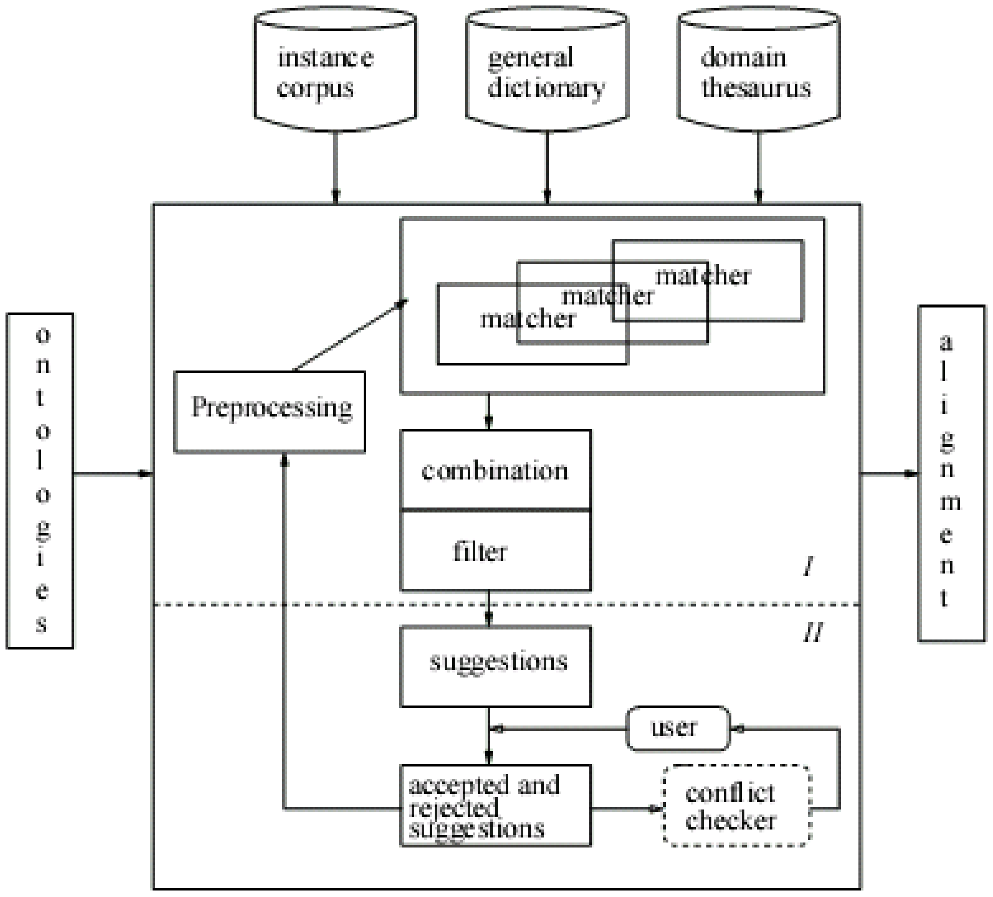
immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  …
…
- i- B-cell activation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

determine the correspondences between terms in different ontologies

Ontology Alignment Framework



Partial Reference Alignment

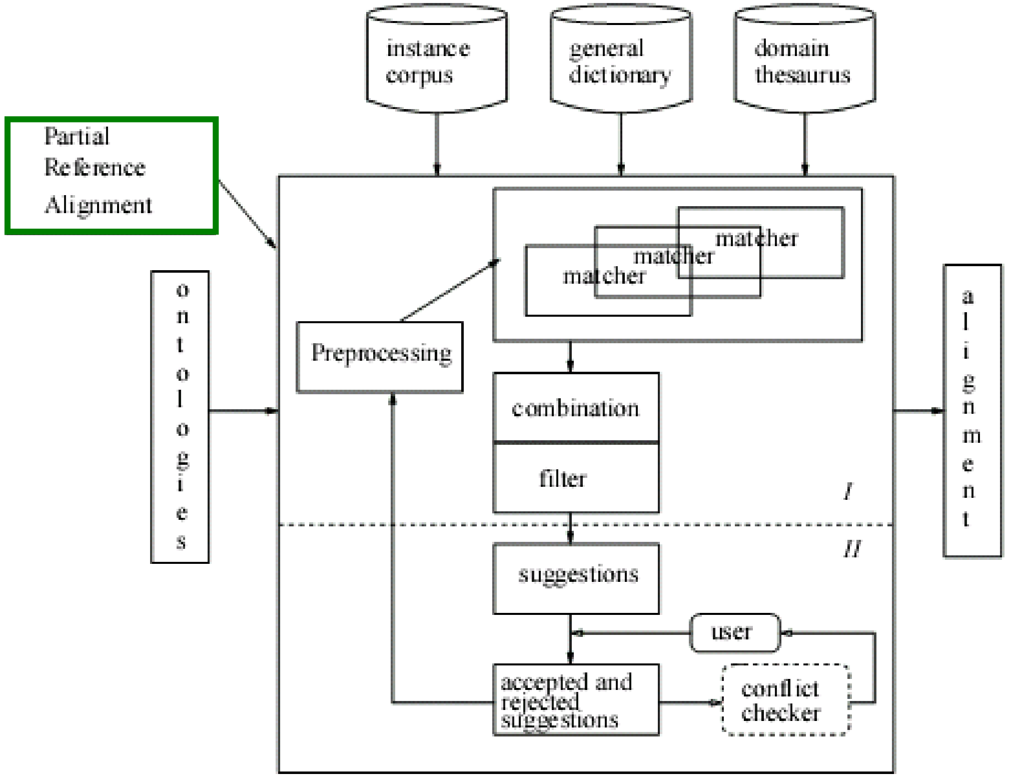
- New setting for ontology alignment:
  - Portal with mappings (e.g. BioPortal)
  - Iterative ontology alignment
  - Anatomy track, task 4 in OAEI 2008

  In all these cases some correct mappings between terms in different ontologies are given or have been obtained.
- A partial reference alignment (PRA) is a subset of all correct mappings.

Partial Reference Alignment

- Research Problem:

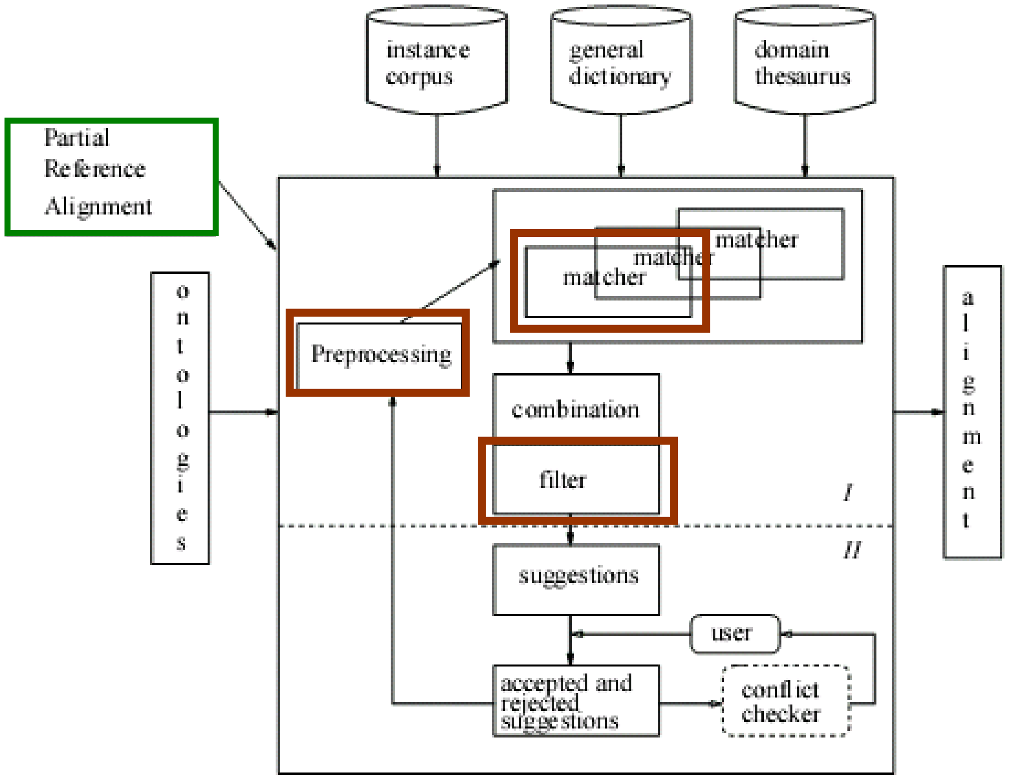
  Can we use PRAs to obtain higher quality mapping suggestions in ontology alignment?



Partial Reference Alignment

- Research Problem:

  Can we use PRAs in the different parts of the framework to obtain higher quality mapping suggestions in ontology alignment?



Outline

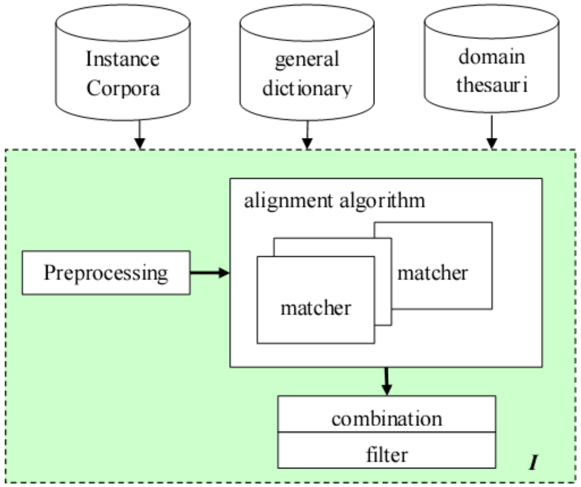
- Background and Evaluation setup
  - SAMBO and SAMBOdtf
  - Test cases and Evaluation measures
- Algorithms and evaluations
  - Use of PRA in the preprocessing step
  - Use of PRA in the matcher
  - Use of PRA in the filter step
  - Influence of size of PRA
- Conclusion & Future Work

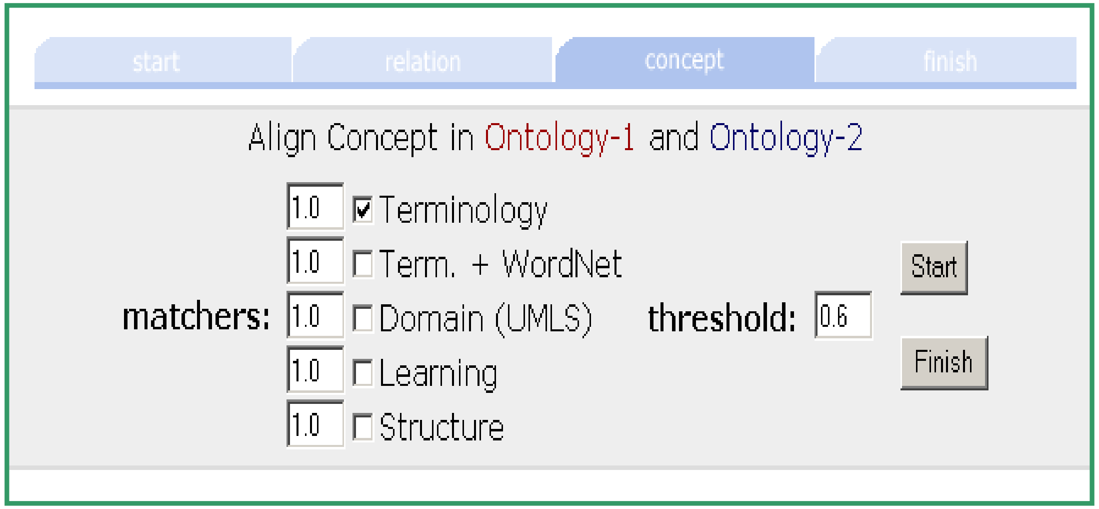
Outline

- **Background and Evaluation setup**
    - SAMBO and SAMBOdtf
    - Test cases and Evaluation measures
- Algorithms and evaluations
    - Use of PRA in the preprocessing step
    - Use of PRA in the matcher
    - Use of PRA in the filter step
    - Influence of size of PRA
- Conclusion & Future Work

SAMBO (1)

- SAMBO (System for Aligning and Merging Biomedical Ontologies)
  - Phase I
    - Matchers
    - Weighted sum combination of matcher results
    - Single threshold filtering





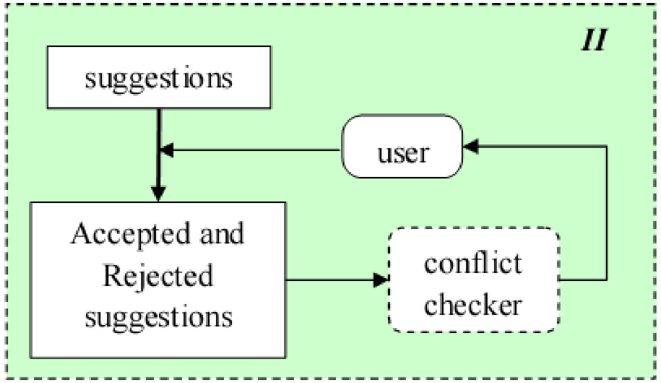
Align Concept in Ontology-1 and Ontology-2

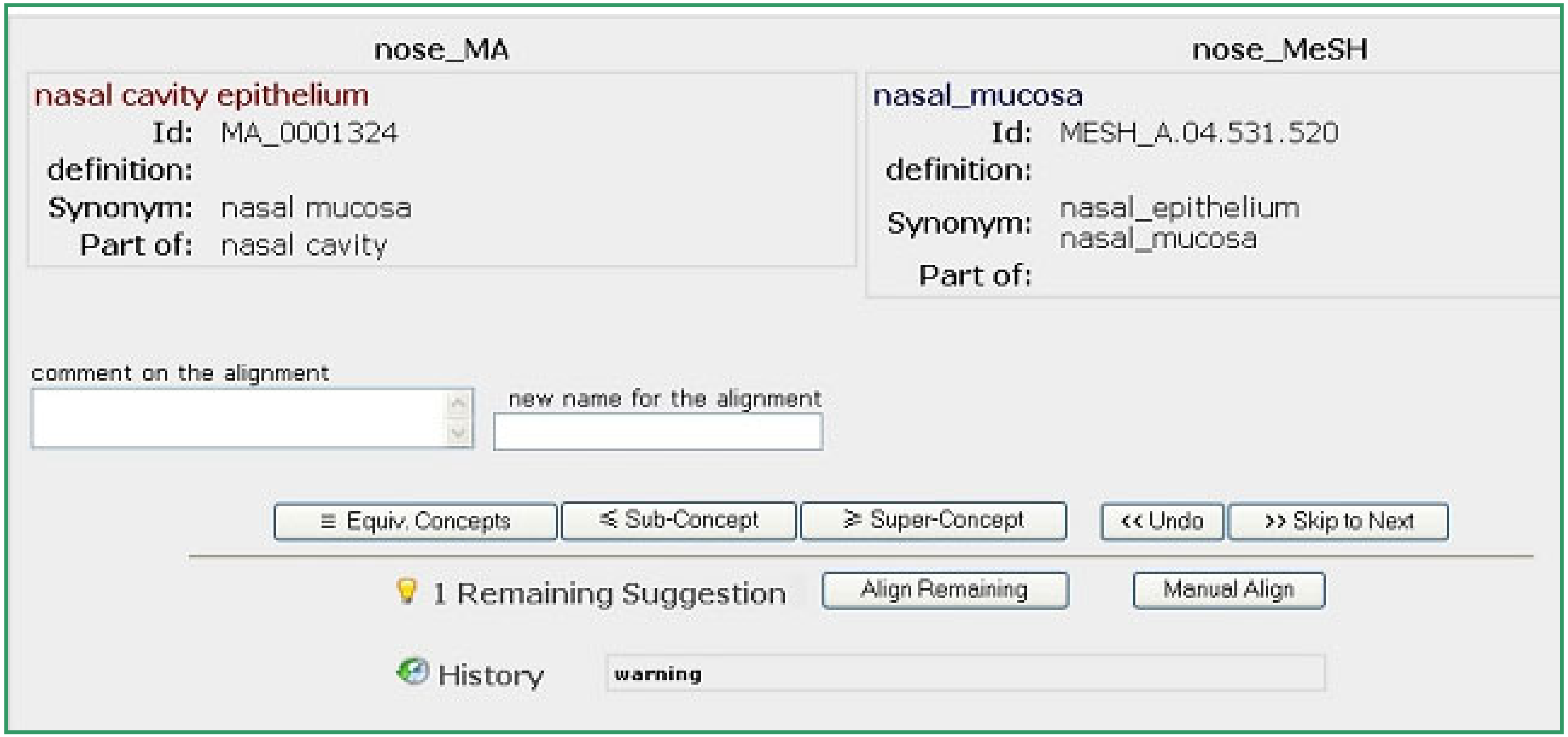
matchers:
1.0 ☑ Terminology
1.0 □ Term. + WordNet
1.0 □ Domain (UMLS)
1.0 □ Learning
1.0 □ Structure

threshold: 0.6

SAMBO (2)

- Phase II:





nasal cavity epithelium
Id: MA_0001324
definition:
Synonym: nasal mucosa
Part of: nasal cavity

nasal_mucosa
Id: MESH_A.04.531.520
definition:
Synonym: nasal_epithelium
nasal_mucosa
Part of:

comment on the alignment

new name for the alignment

SAMBOdtf (1)

- What is SAMBOdtf?

  SAMBO with **Double Threshold Filtering**

- Observation:

  For single threshold filtering,

  the higher the threshold,

  - the higher the precision
  - the lower the recall

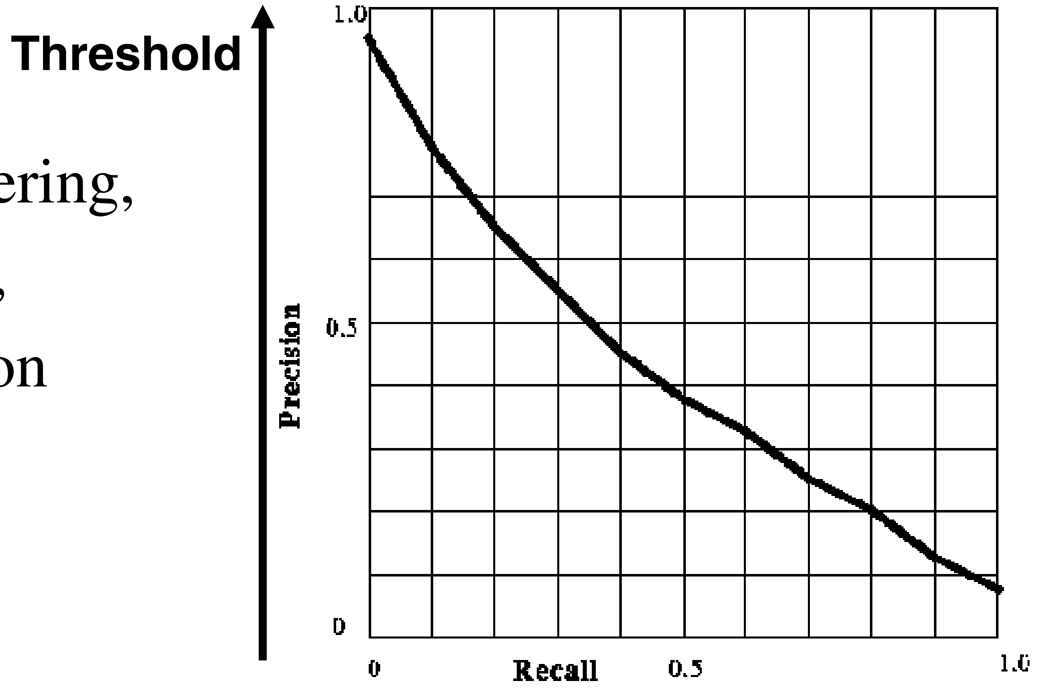


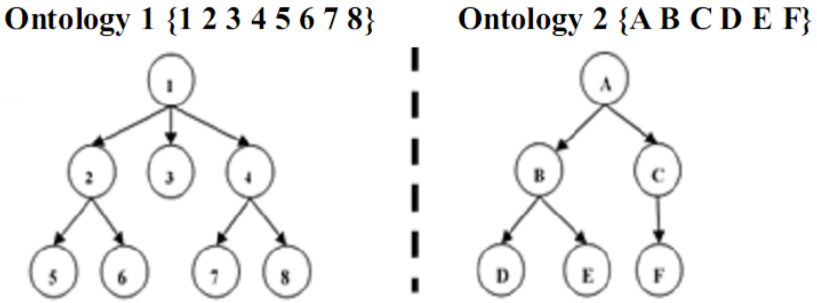
Fig. 1. A typical precision-recall graph

SAMBOdtf (2)

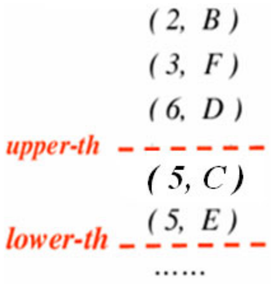
- Idea:
  - Use two thresholds
    - (i) Pairs with similarity value equal to or higher than upper threshold are retained as mapping suggestions. (Thus, upper threshold has a similar role as the threshold in single threshold filtering.)
    - (ii) Pairs with similarity value beween lower and upper threshold are retained as suggestions only if they are ’reasonable’ with respect to the structure of the ontologies and the mapping suggestions retained in step (i). Otherwise they are discarded.
    - (iii) Pairs with similarity value lower than the lower threshold are discarded.

SAMBOdtf (3)

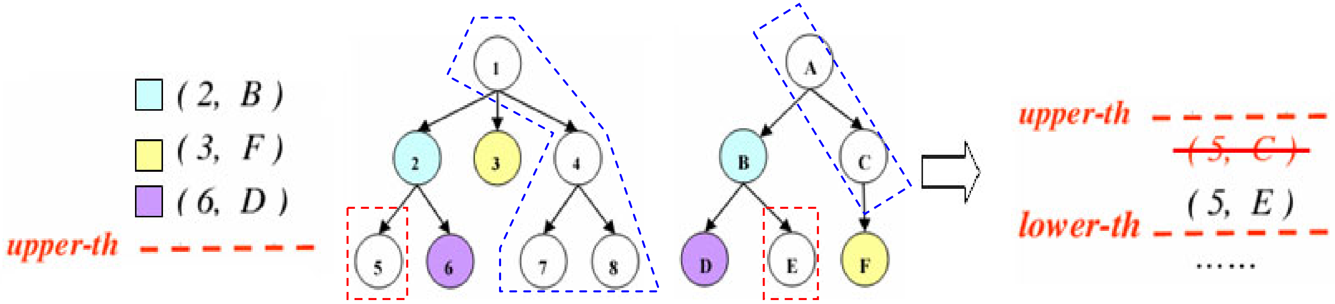
1. Given two ontologies.



2. Calculate similarity values between their concepts.



3. Use suggestions above upper threshold to partition the ontologies into ***mappable groups***, using is-a. *(For mapping suggestions (A,A') and (B,B'): A is-a B iff A' is-a B')*



4. Final mapping suggestions consist of
   1) pairs with similarity value above upper threshold and
   2) pairs of concepts with similarity value between the two thresholds for which the concepts belong to related *mappable groups.*

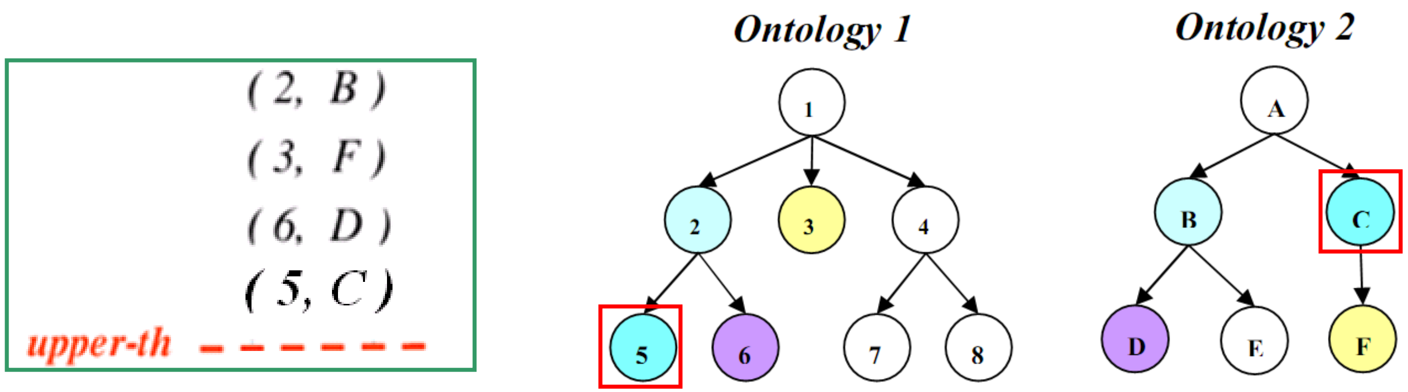
SAMBOdtf (4)

Sometimes, we cannot use *all* the suggestions with similarity values higher than or equal to the upper threshold to partition ontologies.

Example:

Suggestion *(5, C)* does not conform to structure with *(2, B)* and *(3, F)*

- 5 is-a 2, but not C is-a B
- F is-a C, but not 3 is-a 5

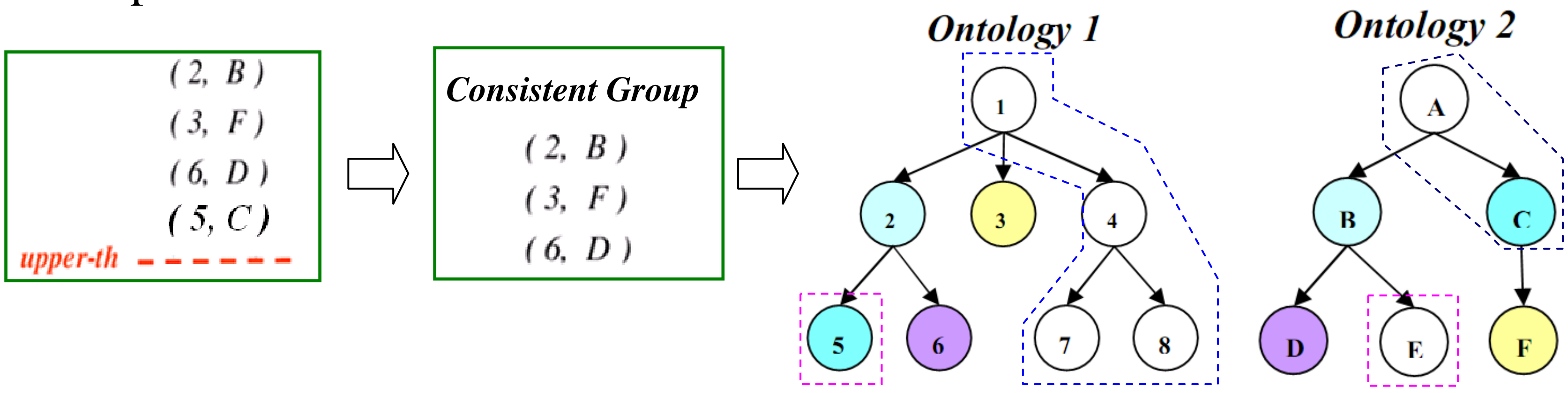


SAMBOdtf (5)

- Solution:

  In such case, we need find a ***consistent group***, in which for each pair of suggestions (A, A’) and (B, B’):   A is-a B iff A’ is-a B’

Example:

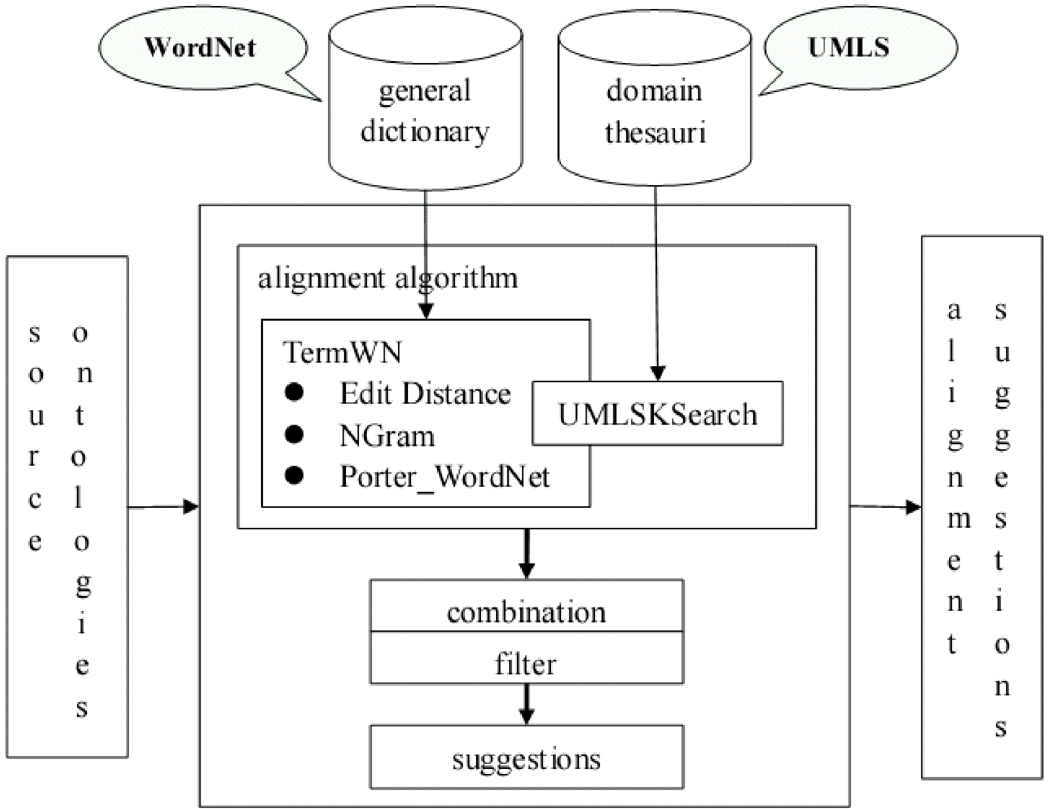


(2, B)
(3, F)
(6, D)
(5, C)
upper-th

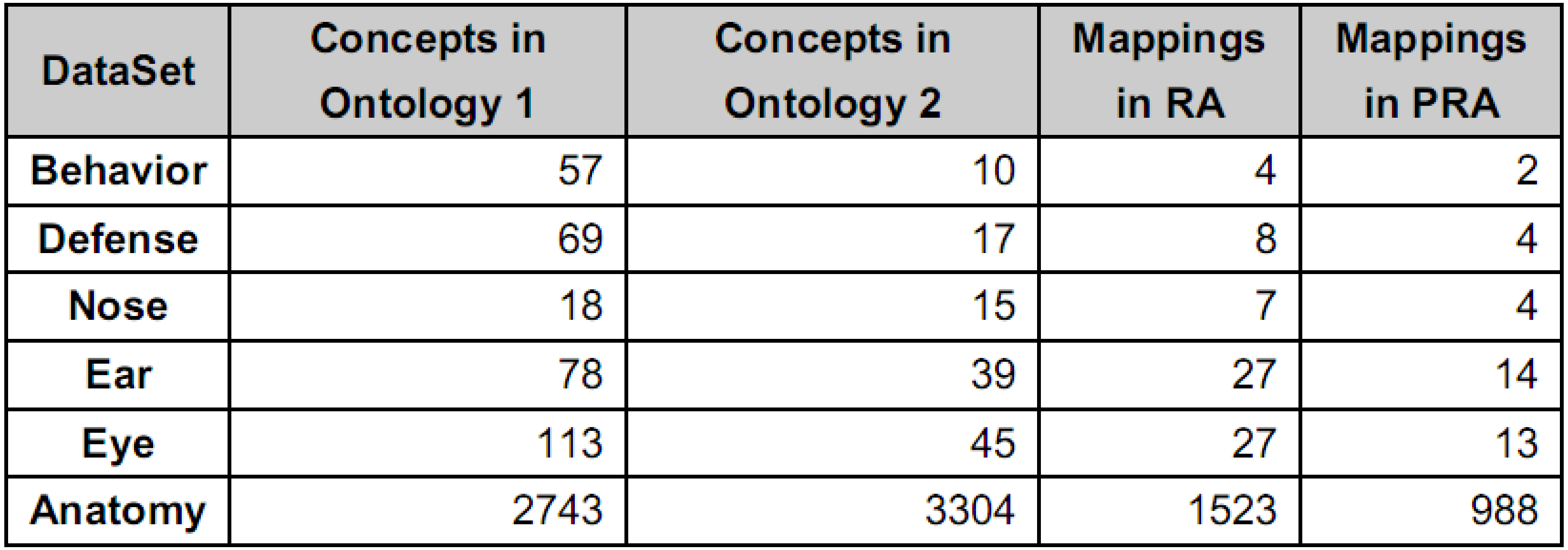
Consistent Group
(2, B)
(3, F)
(6, D)

Baseline Systems
(SAMBO and SAMBOdtf for OAEI 2008)

- Removal of Phase ***II*** – no user involvement
- As there is no user to choose between different suggestions regarding a specific term, a term appears in at most one mapping suggestion.
- Matchers:
  - TermWN
    String Matching with WordNet
  - UMLSKSearch
    Uses UMLS
- Combination
  - Maximum-based strategy
- Filters
  - Single /Double threshold filtering



Test cases



DataSet	Concepts in Ontology 1	Concepts in Ontology 2	Mappings in RA	Mappings in PRA
Behavior	57	10	4	2
Defense	69	17	8	4
Nose	18	15	7	4
Ear	78	39	27	14
Eye	113	45	27	13
Anatomy	2743	3304	1523	988

- Behavior, Defense: Gene Ontology – Signal Ontology
- Nose, Ear, Eye: Adult Mouse anatomy - MESH
- Anatomy: Adult Mouse Anatomy – NCI anatomy

Evaluation

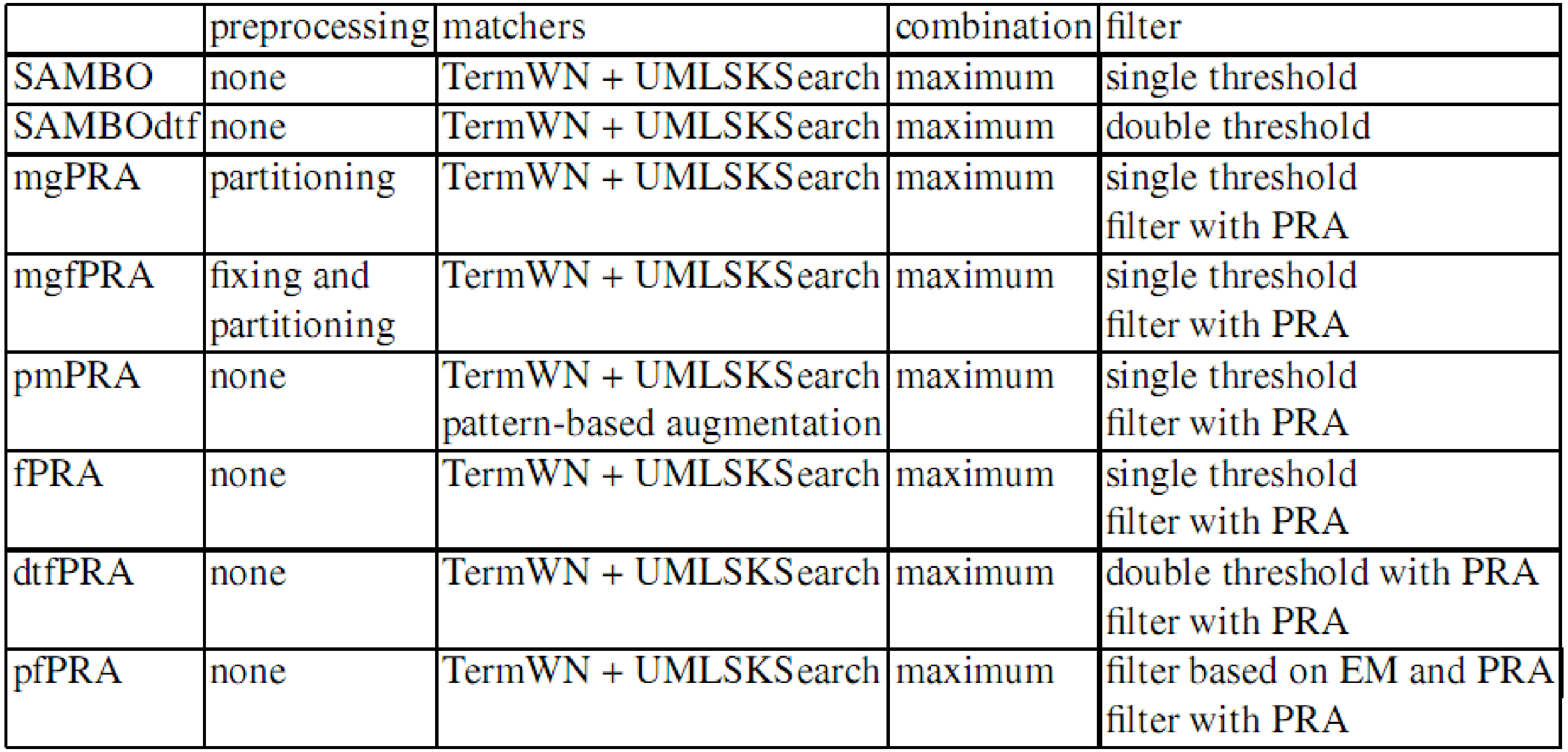
- *Precision*: number of correct suggestions divided by number of suggestions
- *Recall*: number of correct suggestions divided by number of correct mappings
- *Recall-PRA*: number of correct suggestions not in PRA divided by number of correct mappings not in PRA
- *F-measure*: harmonic mean of precision and recall

Outline

- Background and Evaluation setup
  - SAMBO and SAMBOdtf
  - Test cases and Evaluation measures
- **Algorithms and evaluations**
  - Use of PRA in the preprocessing step
  - Use of PRA in the matcher
  - Use of PRA in the filter step
  - Influence of size of PRA
- Conclusion & Future Work

Algorithms

Table 1. Alignment strategies



	preprocessing	matchers	combination	filter
SAMBO	none	TermWN + UMLSKSearch	maximum	single threshold
SAMBOdtf	none	TermWN + UMLSKSearch	maximum	double threshold
mgPRA	partitioning	TermWN + UMLSKSearch	maximum	single threshold filter with PRA
mgfPRA	fixing and partitioning	TermWN + UMLSKSearch	maximum	single threshold filter with PRA
pmPRA	none	TermWN + UMLSKSearch pattern-based augmentation	maximum	single threshold filter with PRA
fPRA	none	TermWN + UMLSKSearch	maximum	single threshold filter with PRA
dtfPRA	none	TermWN + UMLSKSearch	maximum	double threshold with PRA filter with PRA
pfPRA	none	TermWN + UMLSKSearch	maximum	filter based on EM and PRA filter with PRA

1. Use of PRA in the preprocessing step

Use of PRA in the preprocessing step

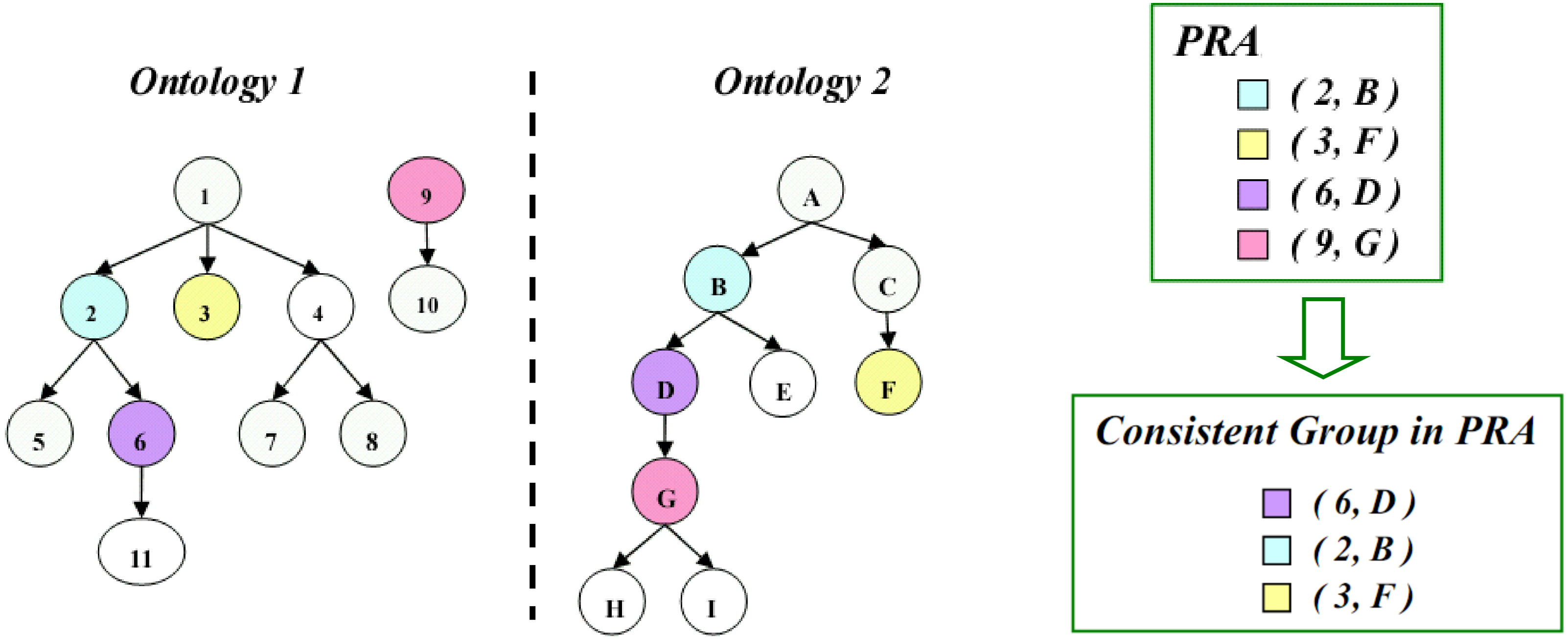
- Intuition

  During the preprocessing step, use mappings in PRA to partition the ontologies into mappable groups.

- Methods
  - mgPRA
  - mgfPRA

Use of PRA in the preprocessing step

- mgPRA (Mappable Groups with PRA)
  - Strategy
    - Find consistent group in PRA
    - Partition ontologies into mappable groups before aligning
  - Example:



PRA

- ▢ (2, B)
- ▢ (3, F)
- ▢ (6, D)
- ▢ (9, G)

Consistent Group in PRA

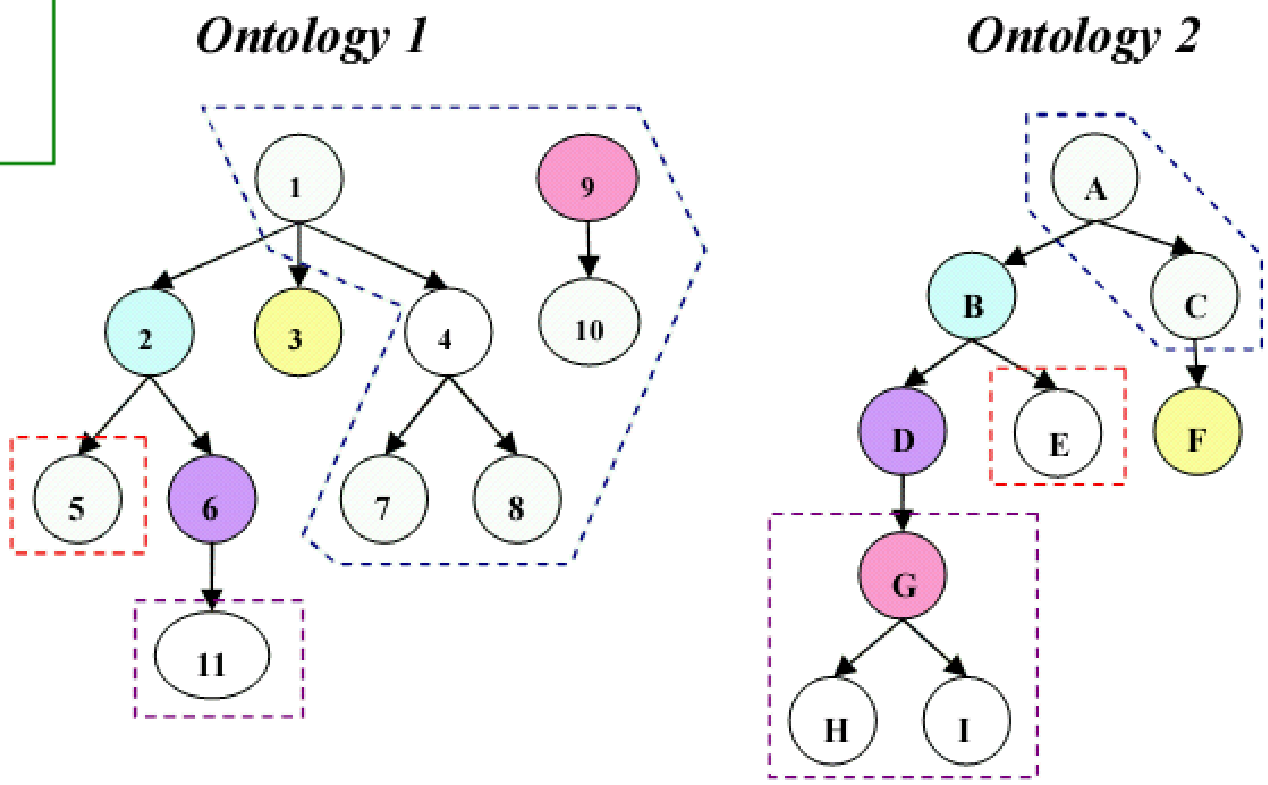
- ▢ (6, D)
- ▢ (2, B)
- ▢ (3, F)

Use of PRA in the preprocessing step

- Partition Results

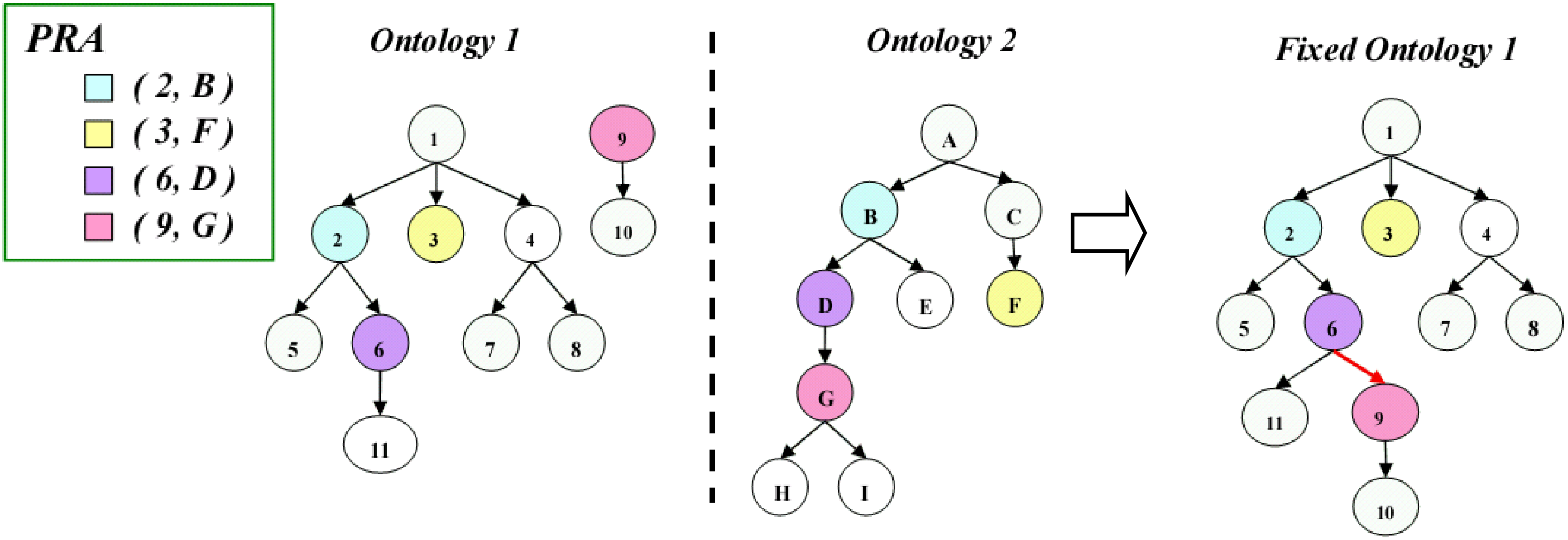
Consistent Group in PRA

🟪 *(6, D)*
🟦 *(2, B)*
🟨 *(3, F)*



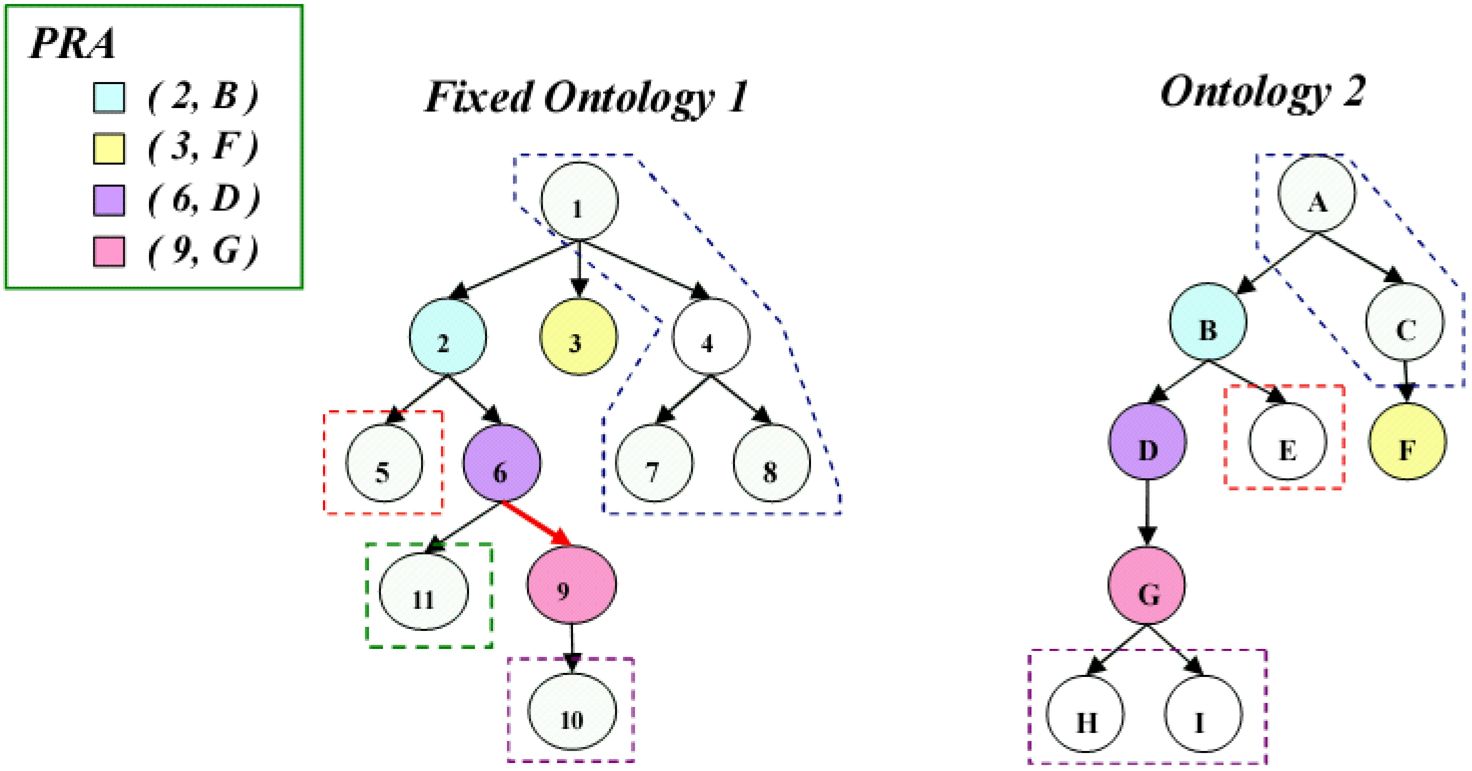
Use of PRA in the preprocessing step

- mgfPRA (Mappable Groups and Fixing with PRA)
  - Strategy
    - ‘Fix’ the missing structural relationships, making the whole PRA a consistent group
    - Then, partition ontologies into mappable groups
  - Example:



Use of PRA in the preprocessing step

- Partition Results



Use of PRA in the preprocessing step

- Result Analysis
  - For threshold 0.4, there are no conclusive results.
  - For thresholds 0.6 and 0.8,
    - mgPRA and mgfPRA almost always have equal or higher precision than SAMBO.
    - mgPRA almost always has equal or higher recall than SAMBO.
    - mgfPRA almost always has equal or lower recall than SAMBO and mgPRA.

Use of PRA in the preprocessing step

n Why does mgfPRA perform worse than mgPRA?

Incorrect use of the structural relation.

For instance, in dataset **nose**, one source ontology uses the structural relation to define both is-a and part-of.

‘Fixing’ the ontology may therefore be wrong.

For instance, the mapping (nose, nose) may lead to introducing is-a relations between nose and its parts.

2. Use of PRA in the matcher

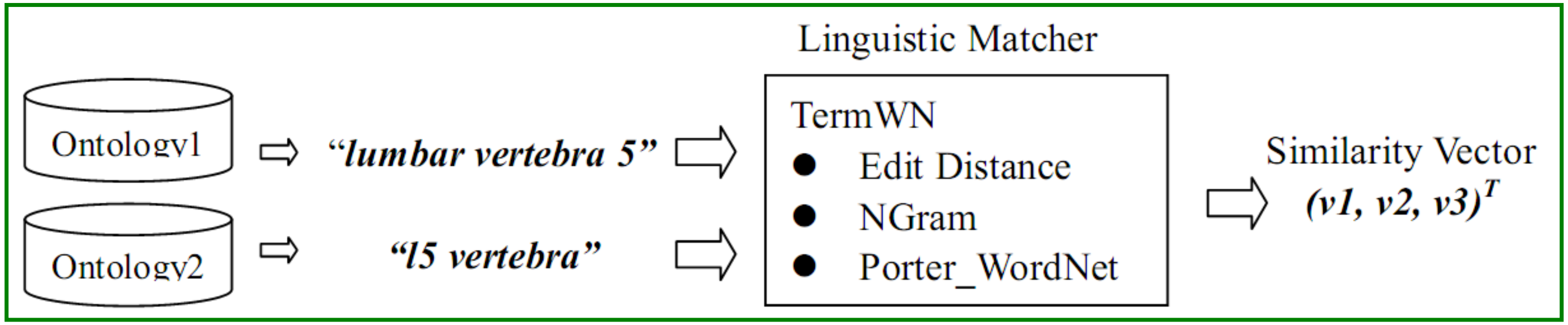
Use of PRA in a matcher

- Observation

  Some correct suggestions share a similar linguistic pattern.

  Examples from PRA of Anatomy

  - (*lumbar vertebra 5*, *l5 vertebra*) and (*thoracic vertebra 11*, *t11 vertebra*)
  - (*forebrain*, *fore brain*) and (*gallbladder*, *gall bladder*)
  - (*stomach body*, *body stomach*) and (*stomach fundus*, *fundus stomach*)



Intuition: mappings sharing a linguistic pattern have similar similarity vectors.

Use of PRA in a matcher

- Intuition

  Mapping suggestions with a similarity vector close to the similarity vector of a PRA mapping are more likely to be correct suggestions.

- pmPRA (Pattern Matcher with PRA)
  - Strategy
    - Compute a similarity vector for each PRA mapping.
    - For each mapping suggestion, we **augment** its similarity value according to the number of PRA mappings within its **neighborhood**.

Use of PRA in a matcher

- Result Analysis
  - For the small datasets, the correct suggested mappings already had high similarity values, and the missed correct mappings had no shared linguistic pattern with PRA mappings.
  - For the Anatomy dataset, the pmPRA has lower or equal precision. Recall increased for high thresholds and decreased for low thresholds.
    - New correct mappings were found.
    - For low thresholds also new wrong mappings were found.

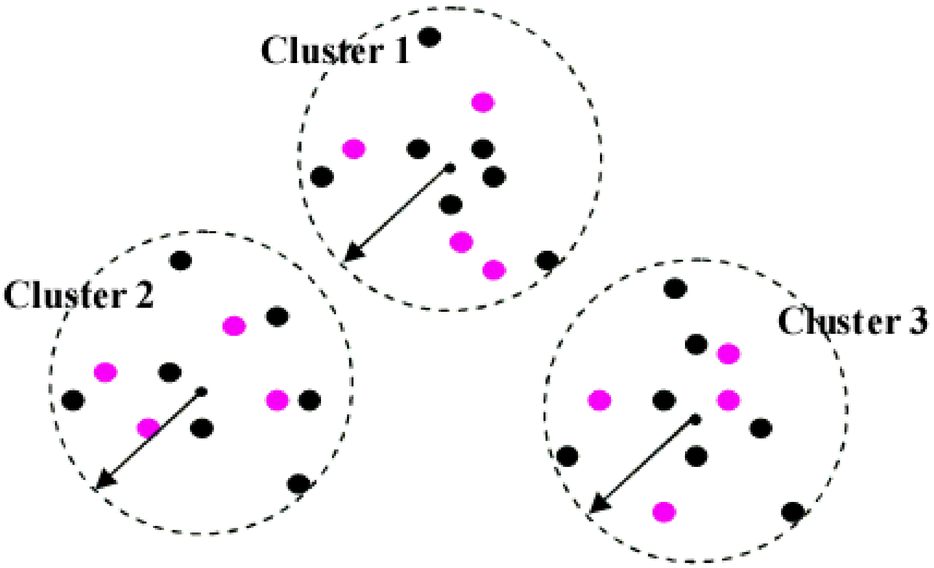
3. Use of PRA in the filter step

Use of PRA in the filter step

- fPRA (Filter with PRA)
  - Strategy
    - Implant PRA mappings in the final result. Any suggestion contradicting with PRA mappings will be filtered out.
- dtfPRA (Double Threshold Filter with PRA)
  - Strategy
    - Similar to SAMBOdtf. Use a consistent group in the PRA to filter the suggestions between upper threshold and low threshold.

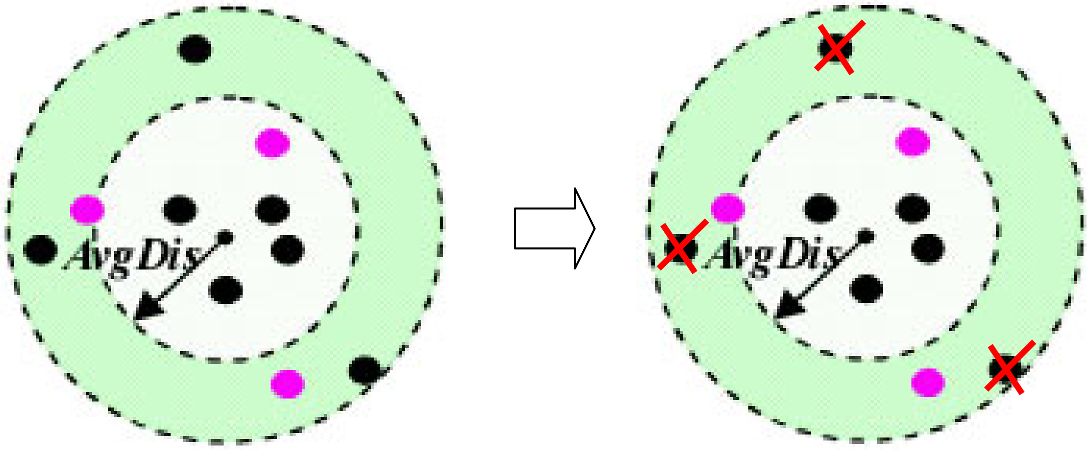
Use of PRA in the filter step

- pfPRA (Pattern Filter with PRA)
  - Strategy
    1. Cluster all suggestions according to their linguistic similarity vectors using expectation-maximization algorithm.
    2. Assign every PRA mapping to the cluster with the nearest cluster center.



Use of PRA in the filter step

- Strategy (continued..)
  3. For each cluster, calculate the ***<u>average distance</u>*** (*AvgDis*) of PRA mappings to their cluster center.
  4. Finally, only suggestions with distance to the cluster center smaller or equal than *AvgDis* will be kept. Otherwise, discarded.



Use of PRA in the filter step (1)

- Result Analysis
  - fPRA always has equal or higher precision and recall than SAMBO.
  - pfPRA always has equal or higher precision than fPRA.
  - pfPRA always has equal or lower recall than SAMBO.
    - Some correct suggestions are filtered out because they have no similar linguistic pattern to PRA mappings.

Use of PRA in the filter step (2)

- Result Analysis
  - dtfPRA always has equal or higher recall than SAMBOdtf.
  - For lower threshold 0.6, dtfPRA always has equal or higher precision than SAMBOdtf.
  - For lower threshold 0.4, dtfPRA always has equal or higher precision than SAMBOdtf, except for dataset **ear** and **eye**.
    - For dataset **ear** and **eye,** the consistent group of dtfPRA is much smaller than the consistent group of SAMBOdtf.

4. Influence of size of PRA

Use of PRA-Full vs PRA-Half

- Result Analysis

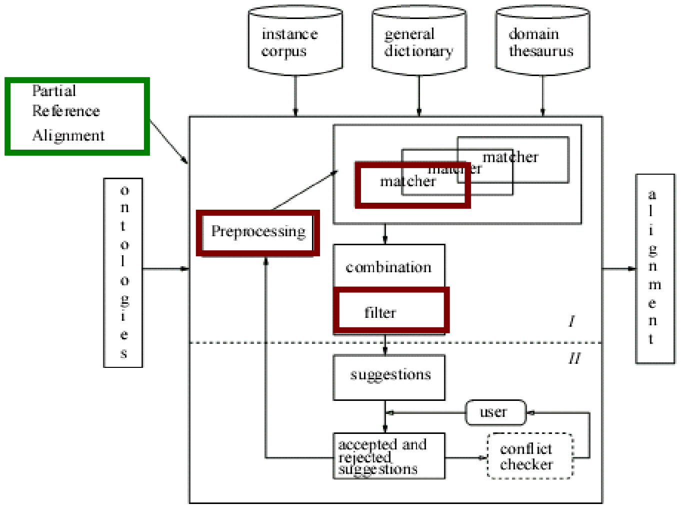
  For larger PRA

  - For all strategies, the recall is higher.
  - For the preprocessing strategies and pmPRA
    - When threshold is low, the precision is lower.
    - When threshold is high, the precision is higher.
  - For the filtering strategies
    - The precision is always equal or higher.

Outline

- Background and Evaluation setup
  - SAMBO and SAMBOdtf
  - Test cases and Evaluation measures
- Algorithms and evaluations
  - Use of PRA in the preprocessing step
  - Use of PRA in the matcher
  - Use of PRA in the filter step
  - Influence of size of PRA
- **Conclusion & Future Work**

Lessons learned



- PRA in preprocessing leads to fewer suggestions, in most cases to an improvement in precision and in some cases to an improvement in recall.

- Use the linguistic pattern matcher mainly to find new suggestions.

- Always use filter with PRA. The other filter approaches work well when the structure of the source ontologies is well-defined and complete.

- Not so large difference between PRA-based algorithms and SAMBO/SAMBOdtf
  - SAMBO/SAMBOdtf already do well on test cases
  - Anatomy case: all new correct mappings are non-trivial

Future Work

- Improve current strategies, and test on other ontologies.
- Investigate combinations and interactions of these strategies.
- Develop an iterative ontology alignment framework.