A method for recommending ontology alignment strategies

He Tan, Patrick Lambrix
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

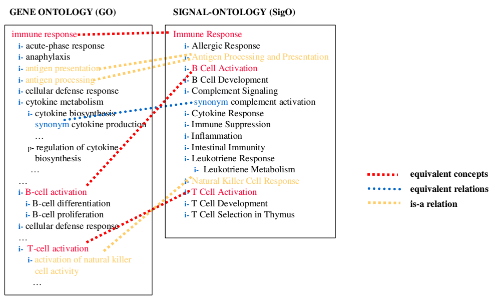
GENE ONTOLOGY (GO)

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)

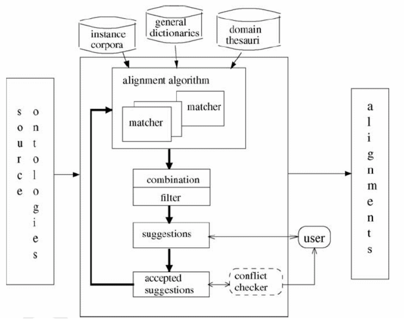
immune response
i- acute-phase response
i- anaphylaxis
i- antigen presentation
i- antigen processing
i- cellular defense response
i- cytokine metabolism
i- cytokine biosynthesis
synonym cytokine production
…
p- regulation of cytokine biosynthesis
…
…
i- B-cell activation
i- B-cell differentiation
i-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

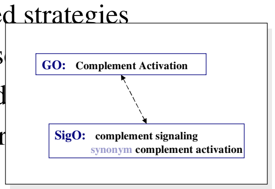
define the relationships between the terms in different ontologies

An Alignment Framework



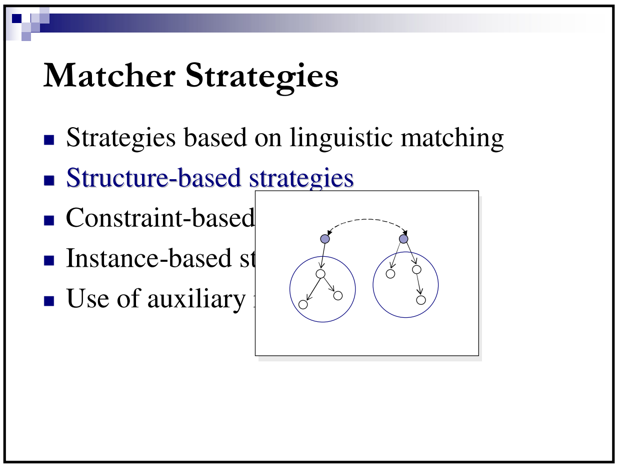
Matcher Strategies

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



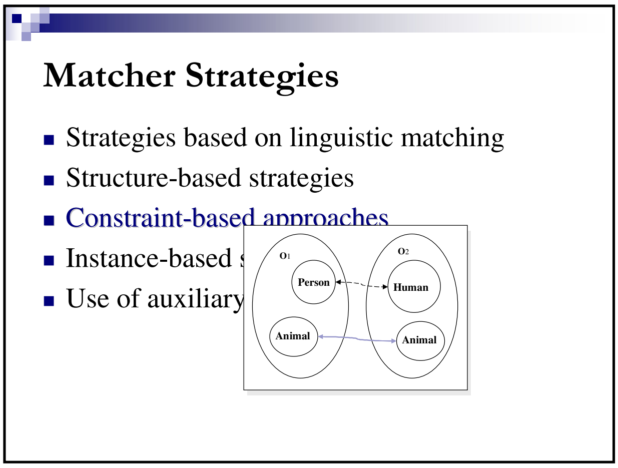
GO: **Complement Activation**

SigO: **complement signaling**
synonym complement activation



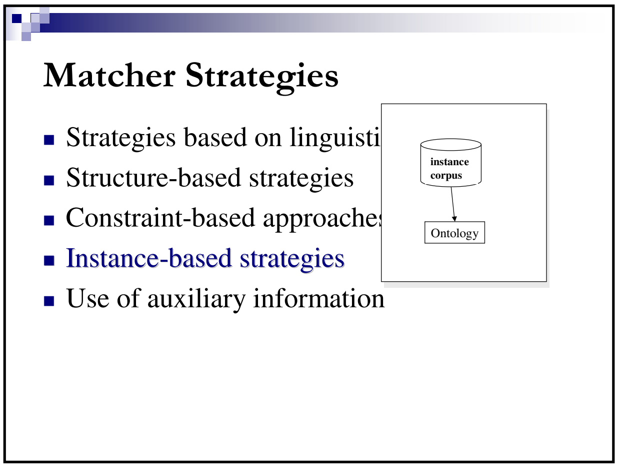
Matcher Strategies

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



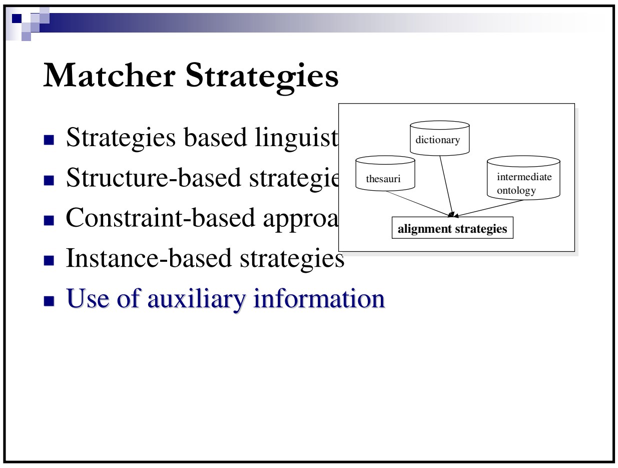
Matcher Strategies

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



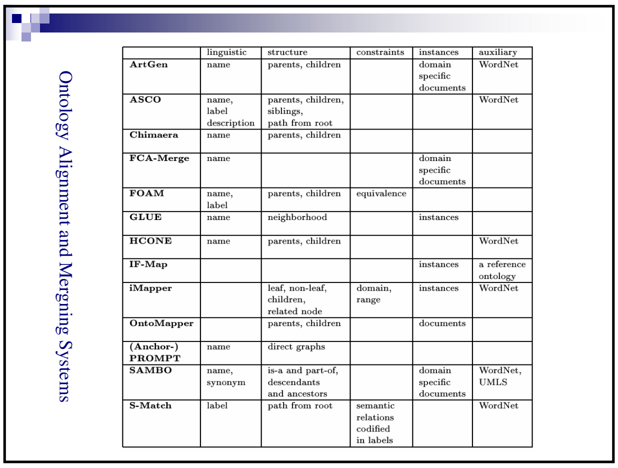
Matcher Strategies

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



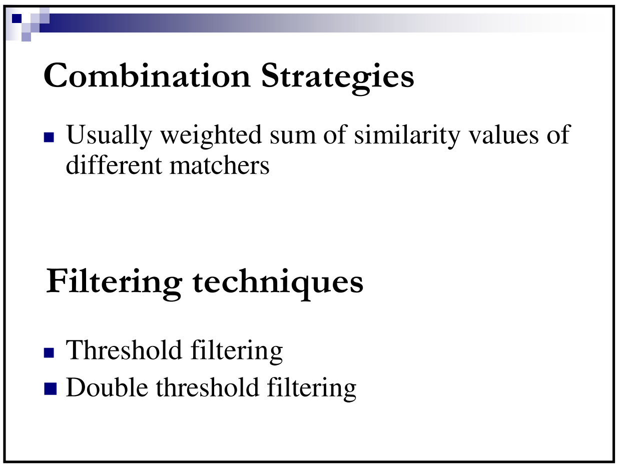
Matcher Strategies

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



Ontology Alignment and Merging 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



Combination Strategies

- Usually weighted sum of similarity values of different matchers

Filtering techniques

- Threshold filtering
- Double threshold filtering

Alignment Strategies

- Many alignment strategies (matchers, combinations, filters) available.
- Question: For a given alignment task, how to choose a strategy?

Recommending alignment strategies

Recommending strategies

- Use knowledge about previous use of alignment strategies (Mochol, Jentzsch, Euzenat 2006)
  - □ Not so much knowledge available
  - □ OAEI
- Parameters for ontologies, similarity assessment, matchers, combinations and filters + optimize parameters (Ehrig, Staab, Sure 2005)
  - □ Based on validation of alignment suggestions by users

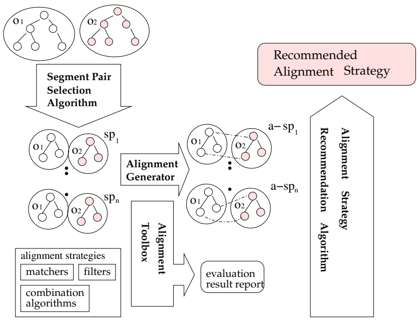
Our approach - idea

- Use the actual ontologies to align to find good candidate alignment strategies
- User/oracle with minimal alignment work
- Complementary to the other approaches

Idea

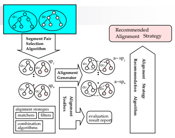
- Select small segments of the ontologies
- Generate alignments for the segments (expert/oracle)
- Use and evaluate available alignment algorithms on the segments
- Recommend alignment algorithm based on evaluation on the segments

Framework



Feasibility test

Experiment case - Ontologies

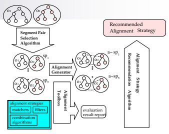


- NCI thesaurus
  - □ National Cancer Institute, Center for Bioinformatics
  - □ Anatomy: 3495 terms
- MeSH
  - □ National Library of Medicine
  - □ Anatomy: 1391 terms

Experiment case - Oracle

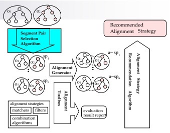
- UMLS
  - □ Library of Medicine
  - □ Metathesaurus contains > 100 vocabularies
  - □ NCI thesaurus and MeSH included in UMLS
  - □ Used as approximation for expert knowledge
  - □ 919 expected alignments according to UMLS

Experiment case – alignment strategies

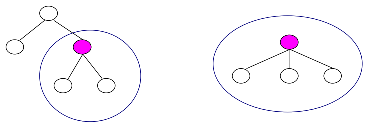


- Matchers and combinations
  - □ N-gram (NG)
  - □ Edit Distance (ED)
  - □ Word List + stemming (WL)
  - □ Word List + stemming + WordNet (WN)
  - □ NG+ED+WL, weights 1/3 (C1)
  - □ NG+ED+WN, weights 1/3 (C2)
- Threshold filter
  - □ thresholds 0.4, 0.5, 0.6, 0.7, 0.8

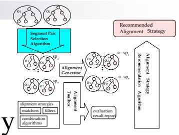
Segment pair selection algorithms



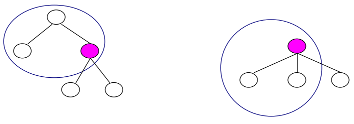
- SubG
  - □ Candidate segment pair = sub-graphs according to is-a/part-of with roots with same name; between 1 and 60 terms in segment
  - □ Segment pairs randomly chosen from candidate segment pairs such that segment pairs are disjoint



Segment pair selection algorithms



- Clust - Cluster terms in ontology
  - □ Candidate segment pair is pair of clusters containing terms with the same name; at least 5 terms in clusters
  - □ Segment pairs randomly chosen from candidate segment pairs

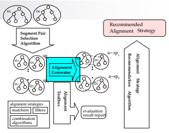


Segment pair selection algorithms

- For each trial, 3 segment pair sets with 5 segment pairs were generated

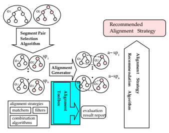
- SubG: A1, A2, A3
  - □ 2 to 34 terms in segment
  - □ level of is-a/part-of ranges from 2 to 6
  - □ max expected alignments in segment pair is 23
- Clust: B1, B2, B3
  - □ 5 to 14 terms in segment
  - □ level of is-a/part-of is 2 or 3
  - □ max expected alignments in segment pair is 4

Segment pair alignment generator



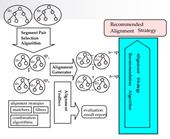
- Used UMLS as oracle

Alignment toolbox



- Used KitAMO as toolbox
- Generates reports on similarity values produced by different matchers, execution times, number of correct, wrong, redundant suggestions

Recommendation algorithm



- Recommendation scores: F, F+E, 10F+E

F: quality of the alignment suggestions

- average f-measure value for the segment pairs

E: average execution time over segment pairs, normalized with respect to number of term pairs

- Algorithm gives ranking of alignment strategies based on recommendation scores on segment pairs

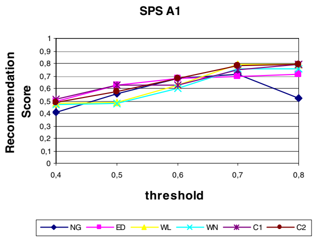
Expected recommendations for F

- Best strategies for the whole ontologies and measure F:

1. (WL,0.8)
2. (C1,0.8)
3. (C2,0.8)

Results

SubG, F, SPS A1

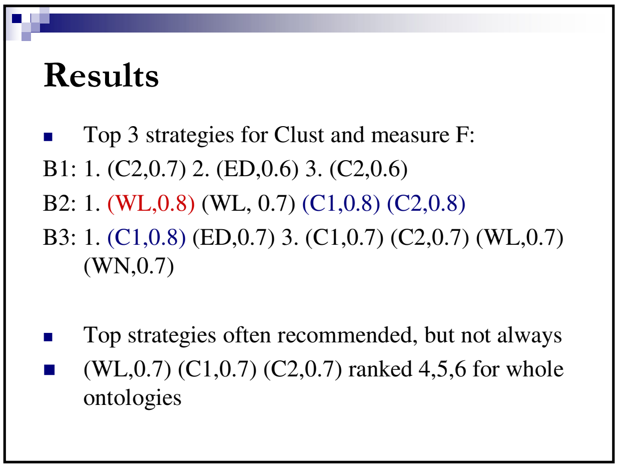


Results

- Top 3 strategies for SubG and measure F:

A1: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)
A2: 1. (WL,0.8) 2. (WL,0.7) 3. (WN,0.7)
A3: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)

- Best strategy always recommended first
- Top 3 strategies often recommended
- (WL,0.7) has rank 4 for whole ontologies



Results

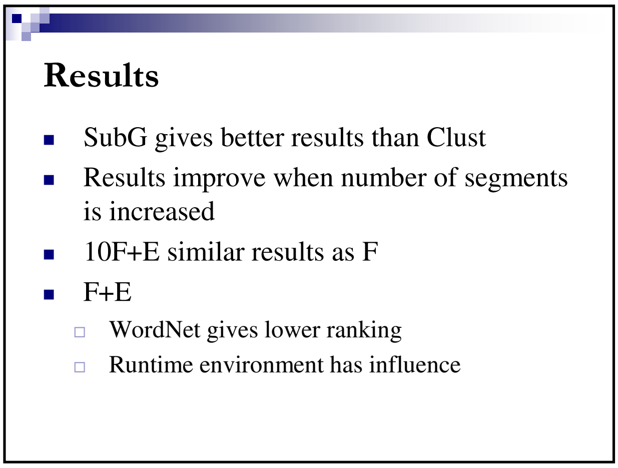
- Top 3 strategies for Clust and measure F:

B1: 1. (C2,0.7) 2. (ED,0.6) 3. (C2,0.6)

B2: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)

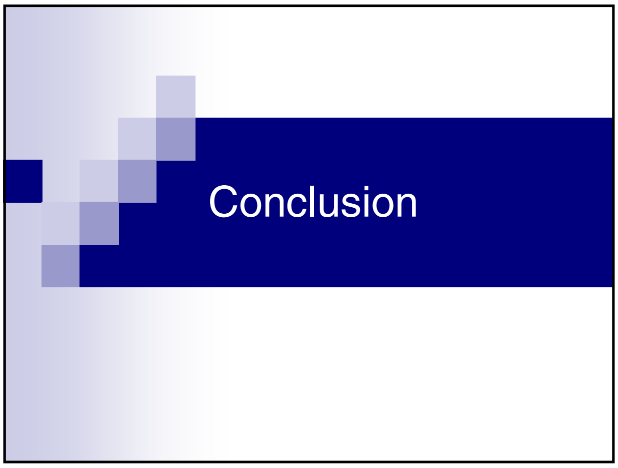
B3: 1. (C1,0.8) (ED,0.7) 3. (C1,0.7) (C2,0.7) (WL,0.7) (WN,0.7)

- Top strategies often recommended, but not always
- (WL,0.7) (C1,0.7) (C2,0.7) ranked 4,5,6 for whole ontologies

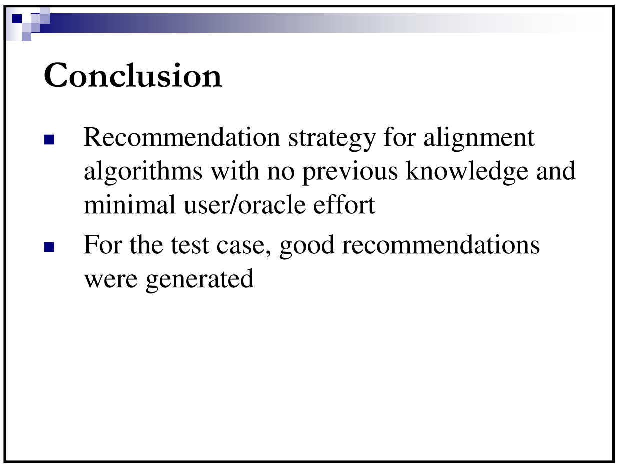


Results

- SubG gives better results than Clust
- Results improve when number of segments is increased
- 10F+E similar results as F
- F+E
  - WordNet gives lower ranking
  - Runtime environment has influence

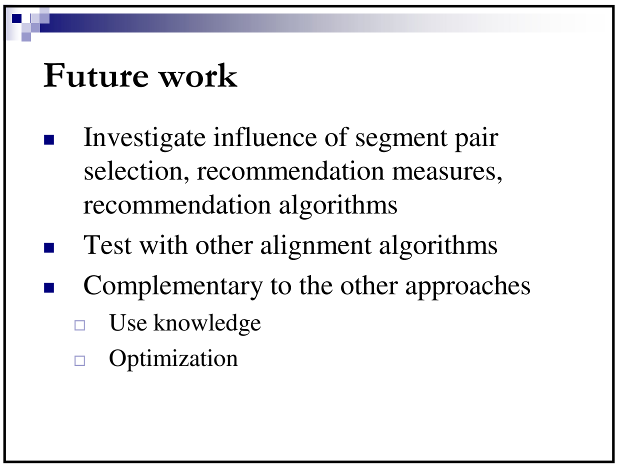


Conclusion



Conclusion

- Recommendation strategy for alignment algorithms with no previous knowledge and minimal user/oracle effort
- For the test case, good recommendations were generated



Future work

- Investigate influence of segment pair selection, recommendation measures, recommendation algorithms
- Test with other alignment algorithms
- Complementary to the other approaches
  - Use knowledge
  - Optimization