Aligning Biomedical Ontologies

Patrick Lambrix
Linköpings universitet

Outline

- Ontologies and ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies

Ontologies

"*Ontologies define the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.*"

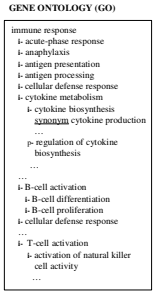
Ontologies used …

- for communication between people and organizations
- for enabling knowledge reuse and sharing
- as basis for interoperability between systems
- as repository of information
- as query model for information sources

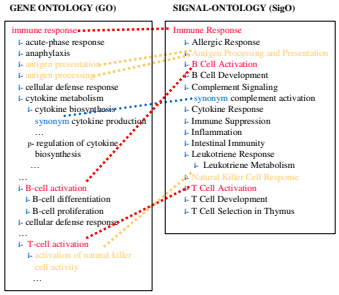
Key technology for the Semantic Web

Ontologies in biomedical research

- many biomedical ontologies
  e.g. GO, OBO, SNOMED-CT
- practical use of biomedical ontologies
  e.g. databases annotated with GO



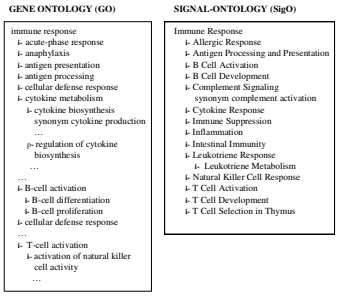
Ontologies with overlapping information



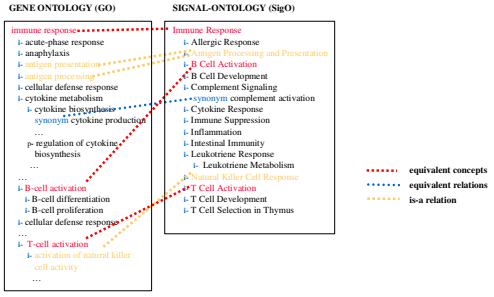
Ontologies with overlapping information

- 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



- To define the relations between the terms in different ontologies

Defining ontologies is not so easy ...

The Celestial Emporium of Benevolent Knowledge, Borges
"On those remote pages it is written that animals are divided into:
a. those that belong to the Emperor
b. embalmed ones
c. those that are trained
d. suckling pigs
e. mermaids
f. fabulous ones
g. stray dogs
h. those that are included in this classification
i. those that tremble as if they were mad
j. innumerable ones
k. those drawn with a very fine camel's hair brush
l. others
m. those that have just broken a flower vase
n. those that resemble flies from a distance"

Slide from talk by C. Goble

Defining ontologies is not so easy ...

Dyirbal classification of objects in the universe

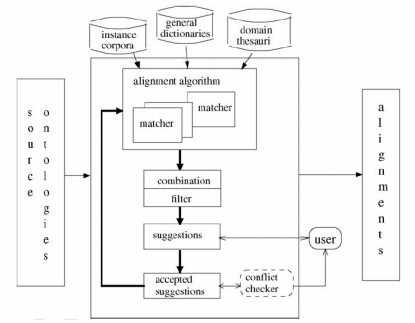
- Bayi: men, kangaroos, possums, bats, most snakes, most fishes, some birds, most insects, the moon, storms, rainbows, boomerangs, some spears, etc.
- Balan: women, anything connected with water or fire, bandicoots, dogs, platypus, echidna, some snakes, some fishes, most birds, fireflies, scorpions, crickets, the stars, shields, some spears, some trees, etc.
- Balam: all edible fruit and the plants that bear them, tubers, ferns, honey, cigarettes, wine, cake.
- Bala: parts of the body, meat, bees, wind, yamsticks, some spears, most trees, grass, mud, stones, noises, language, etc.

Slide from talk by C. Goble

Outline

- Ontologies and ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies

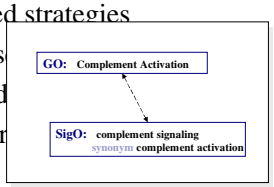
An Alignment Framework



Matchers

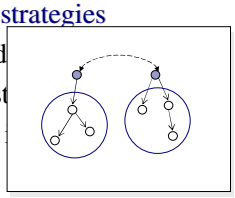
Matcher Strategies

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



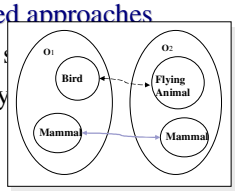
Matcher Strategies

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



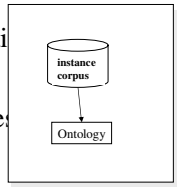
Matcher Strategies

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



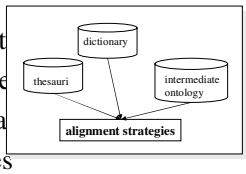
Matcher Strategies

- Strategies based on linguisti
- Structure-based strategies
- Constraint-based approache
- 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 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

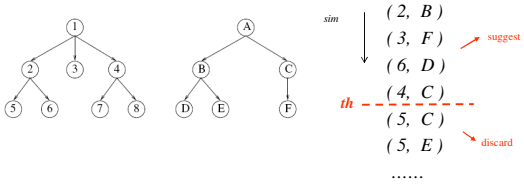
- Usually weighted sum of similarity values of different matchers

Filtering

Filtering techniques

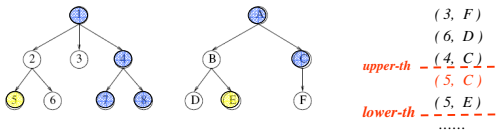
- Threshold filtering

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



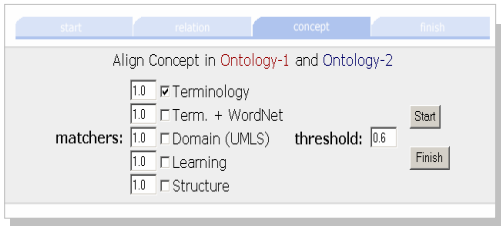
Filtering techniques

- 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)



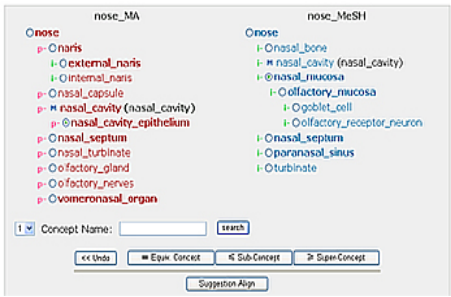
(2, B)
(3, F)
(6, D)
(4, C)
upper-th
(5, C)
lower-th
(5, E)
.....

Example alignment system SAMBO – matchers, combination, filter



Example alignment system SAMBO – suggestion mode

Example alignment system SAMBO – manual mode



Outline

- Ontologies and ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies

Evaluation measures

- Precision:

$$\frac{\#\text{ correct suggested alignments}}{\#\text{ suggested alignments}}$$

- Recall:

$$\frac{\#\text{ correct suggested alignments}}{\#\text{ correct alignments}}$$

- F-measure: combination of precision and recall

Ontology Alignment Evaluation Initiative

OAEI

- Since 2004
- Evaluation of systems
- Different tracks
  - comparison: benchmark (open)
  - expressive: anatomy (blind)
  - directories and thesauri: directory, food, environment, library (blind)
  - consensus: conference

OAEI

- Evaluation measures
  - Precision/recall/f-measure
  - recall of non-trivial alignments
  - full / partial golden standard

OAEI 2007

- 17 systems participated
  - benchmark (13)
    - ASMOV: p = 0.97, r = 0.97
  - 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 2007

- 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

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

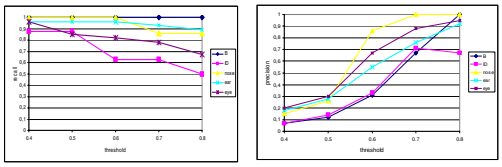
  Threshold filtering : 0.4, 0.5, 0.6, 0.7, 0.8

  Weights for combination: 1.0/1.2

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

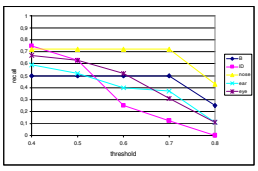
Results

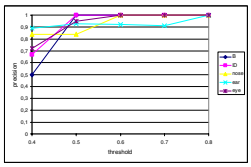
- Terminological matchers



Results

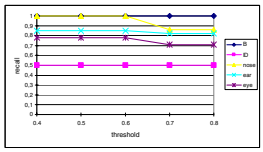
- Basic learning matcher

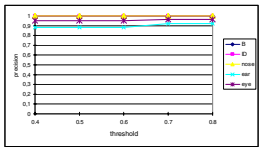




Results

- Domain matcher





Results

- Comparison of the matchers

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

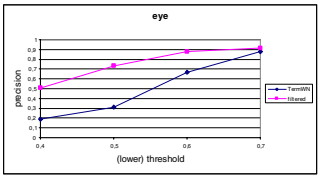
- Combinations of the different matchers
  - combinations give often better results
  - no significant difference on the quality of suggestions for different weight assignments in the combinations

- Structural matcher did not find (many) new correct alignments
  (but: good results for systems biology schemas SBML – PSI MI)

Evaluation of filtering

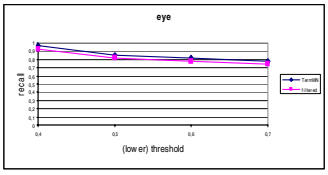
- Matchers
  - TermWN
- Parameters
  - Quality of suggestions: precision/recall
  - Double threshold filtering using structure:
    - Upper threshold: 0.8
    - Lower threshold: 0.4, 0.5, 0.6, 0.7, 0.8

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

Outline

- Ontologies and ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies

Recommending strategies - 1

- Use knowledge about previous use of alignment strategies
  - gather knowledge about input, output, use, performance, cost via questionnaires
  - Not so much knowledge available
  - OAEI

(Mochol, Jentzsch, Euzenat 2006)

Recommending strategies - 2

- Optimize
  - Parameters for ontologies, similarity assessment, matchers, combinations and filters
  - Run general alignment algorithm
  - User validates the alignment result
  - Optimize parameters based on validation

(Ehrig, Staab, Sure 2005)

Recommending strategies - 3

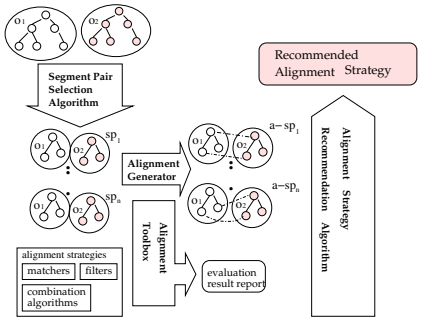
- Based on inherent knowledge
  - Use the actual ontologies to align to find good candidate alignment strategies
  - User/oracle with minimal alignment work
  - Complementary to the other approaches

(Tan, Lambrix 2007)

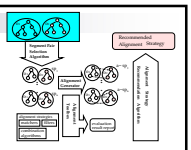
Idea

- 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



Experiment case - Ontologies

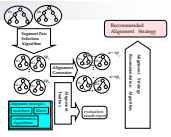


- NCI thesaurus
  - National Cancer Institute, Center for Bioinformatics
  - Anatomy: 3495 terms
- MeSH
  - National Library of Medicine
  - Anatomy: 1391 terms

Experiment case - Oracle

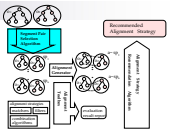
- UMLS
  - Library of Medicine
  - Metathesaurus contains > 100 vocabularies
  - NCI thesaurus and MeSH included in UMLS
  - Used as approximation for expert knowledge
  - 919 expected alignments according to UMLS

Experiment case – alignment strategies

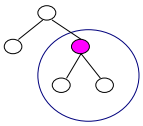


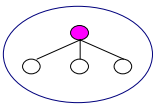
- Matchers and combinations
  - N-gram (NG)
  - Edit Distance (ED)
  - Word List + stemming (WL)
  - Word List + stemming + WordNet (WN)
  - NG+ED+WL, weights 1/3 (C1)
  - NG+ED+WN, weights 1/3 (C2)
- Threshold filter
  - thresholds 0.4, 0.5, 0.6, 0.7, 0.8

Segment pair selection algorithms



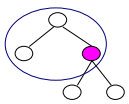
- SubG
  - Candidate segment pair = sub-graphs according to is-a/part-of with roots with same name; between 1 and 60 terms in segment
  - Segment pairs randomly chosen from candidate segment pairs such that segment pairs are disjoint

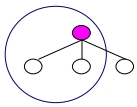




Segment pair selection algorithms

- 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

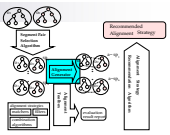




Segment pair selection algorithms

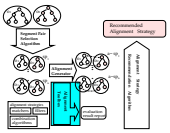
- For each trial, 3 segment pair sets with 5 segment pairs were generated
- SubG: A1, A2, A3
  - 2 to 34 terms in segment
  - level of is-a/part-of ranges from 2 to 6
  - max expected alignments in segment pair is 23
- Clust: B1, B2, B3
  - 5 to 14 terms in segment
  - level of is-a/part-of is 2 or 3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

Segment pair alignment generator



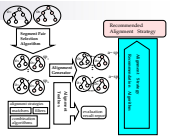
- Used UMLS as oracle

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

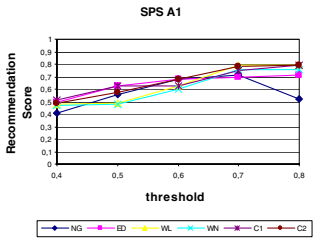
Expected recommendations for F

- 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

Current issues

- Systems and algorithms
  - Complex ontologies
  - Use of instance-based techniques
  - Alignment types (equivalence, is-a, …)
  - Complex alignments (1-n, m-n)
  - Connection ontology types – alignment strategies

Current issues

- Evaluations
  - Need for Golden standards
  - Systems available, but not always the alignment algorithms
  - Evaluation measures
- Recommending 'best' alignment strategies

Further reading

- http://www.ontologymatching.org

(plenty of references to articles and systems)

- Ontology alignment evaluation initiative: http://oaei.ontologymatching.org

(home page of the initiative)

- 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)

Further reading

- 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 H, Lambrix P, `A method for recommending ontology alignment strategies', *International Semantic Web Conference*, 494-507, 2007.

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

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

(recommendation of alignment strategies)