Alignment of Biomedical Ontologies using Life Science Literature

He Tan, Vaida Jakoniene, Patrick Lambrix,
Johan Aberg, Nahid Shahmehri

Department of Computer and Information Science
Linköpings universitet

Outline

- Ontologies and ontology alignment
- Ontology alignment approaches using life science literature
- Conclusion and Future Work

Outline

- Ontologies and ontology alignment
- Ontology alignment approaches using life science literature
- Conclusion and Future Work

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.*"

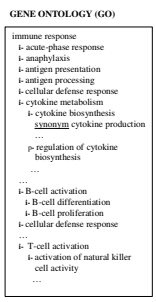
Ontologies

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

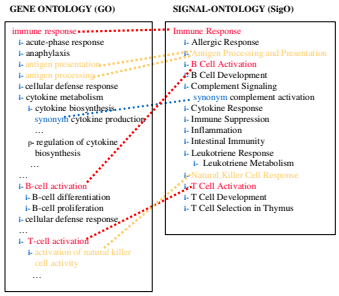
Motivation

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



Motivation

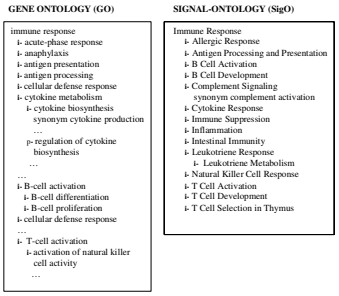
- Ontologies with overlapping information



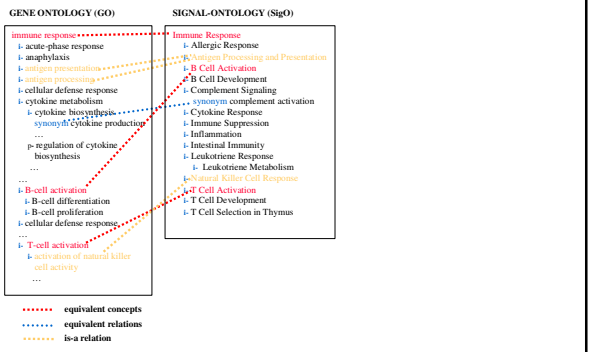
Motivation

- Use of multiple ontologies
  e.g. custom-specific ontology + standard ontology
- Bottom-up creation of ontologies
  experts can focus on their domain of expertise

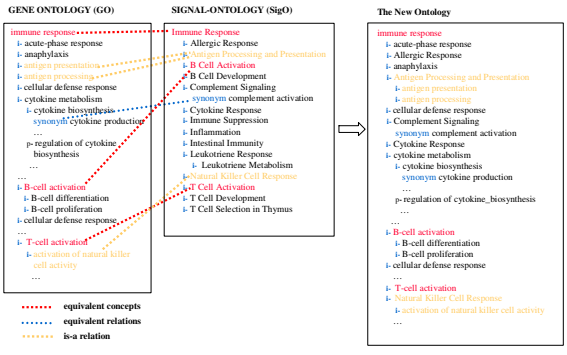
→ important to know the inter-ontology relationships



Aligning ontologies

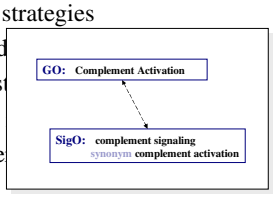


Merging ontologies



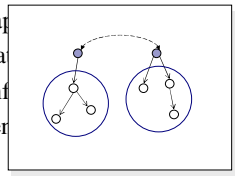
Alignment Strategies

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



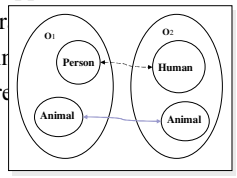
Alignment Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based ap
- Instance-based stra
- Use of auxiliary inf
- Combining differer



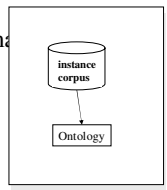
Alignment Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based str
- Use of auxiliary i
- Combining differe



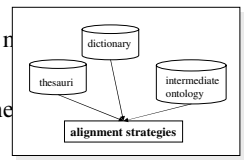
Alignment Strategies

- Strategies based on linguistic ma
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information
- Combining different approaches



Alignment Strategies

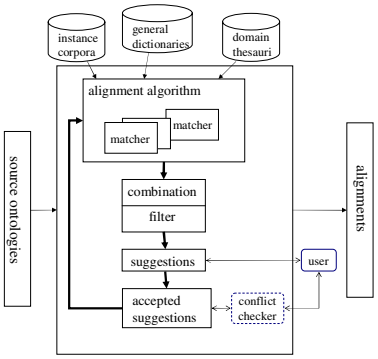
- Strategies based linguistic r
- Structure-based strategies
- Constraint-based approache
- Instance-based strategies
- Use of auxiliary information
- Combining different approaches



Alignment Strategies

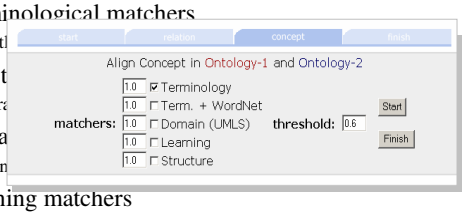
- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information
- Combining different approaches

The General Alignment Strategy



SAMBO – matchers

- Terminological matchers
  algorit
- Struct
  an itera
- Doma
  utilizin
- Learning matchers
  bayes learning algorithms based on related biomedical literature



Outline

- Introduction
  - Motivation
  - Ontologies
  - Ontology alignment
- Ontology alignment approaches using life science literature
- Conclusion and Future Work

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 classifiers
  - One classifier per ontology
- Classification
  - Abstracts related to one ontology are classified by the other ontology's classifier and vice versa
- Calculate similarities

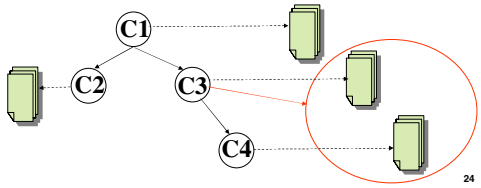
Basic learning matcher

- Generate corpora
- Generate classifiers
  - Naive Bayes classifiers
- Classification
  - Abstracts related to one ontology are classified to the concept in the other ontology with highest posterior probability
- Calculate similarities

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

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'

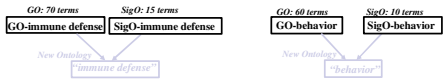
- Calculate similarities
  - Take structure of the ontologies into account when calculating similarities
  - Similarity is computed based on the classifiers applied to the concepts and their sub-concepts

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

25

Evaluation - cases

- GO vs. SigO



- MA vs. MeSH

26

Evaluation

- Matchers

  Basic, StrucCl, StrucSim, StrucClSim
  Term, TermWN, Dom

- Parameters

  Maximum number of PubMed abstracts
  Quality of suggestions: precision/recall
  Thresholds : 0.4, 0.5, 0.6, 0.7, 0.8
  Weights for combination: 1.0/1.2
  Time

27

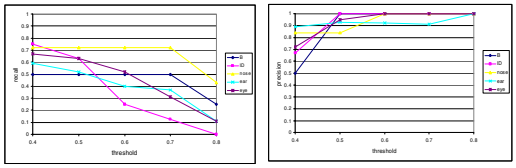
Evaluation – influence of number of PubMed abstracts

- Different results for different cases.
- Quality of results does not necessarily increase when corpora become larger.

28

Evaluation – quality of suggestions

- Basic



29

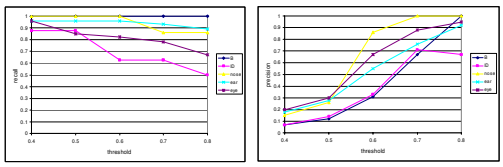
Evaluation – quality of suggestions

- Basic usually outperforms structure-based algorithms.
- No clear winner among StrucCl and StrucSim.
- StruClSim performs worse than StrucCl and StrucSim.

30

Evaluation – quality of suggestions

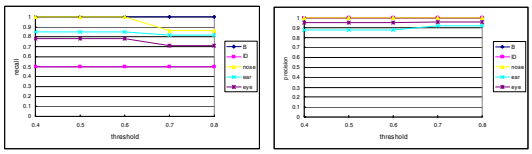
- Terminological matchers



31

Evaluation – quality of suggestions

- Domain matcher



32

Evaluation – quality of suggestions

- Comparison of the matchers

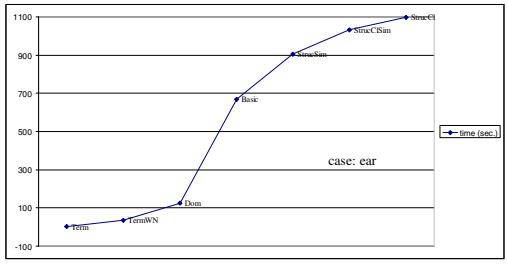
  CS_TermWN ⊇ *CS_Dom* ⊇ *CS_Basic*

- Combinations of the different matchers leads to higher quality results

33

Evaluation - time

- *T_StrucCl* ~ *T_StrucClSim* > *T_StrucSim* > *T_Basic*



34

Outline

- Introduction
  - Motivation
  - Ontologies
  - Ontology alignment
- Ontology alignment approaches using life science literature
- Conclusion and Future Work

35

Conclusion

- Instance-based algorithms for aligning ontologies using life science literature
- Evaluations of matchers
  - Basic outperforms structure-based approaches
  - Our structure-based approaches do not require previous alignments
  - Combination with other approaches gives best results

36

Future Work

- Algorithms
  - Classify abstracts to multiple concepts
  - Use of auxiliary information
  - Other classifiers
- Structure-based filtering
- Evaluation tool - KitAMO

37