Matching unstructured vocabularies using a background ontology

Zharko Aleksovski, Michel Klein, Warner ten Kate, Frank van Harmelen

Vrije Universiteit, Amsterdam and
Philips Research, Eindhoven

Podebrady, 2006

Content

- **Introduction**
- Our approach: semantic matching scheme
- Test data in our case study
- Experiments using the test data
- Evaluation agains state-of-the-art ontology matching tools
- Conclusions and future work

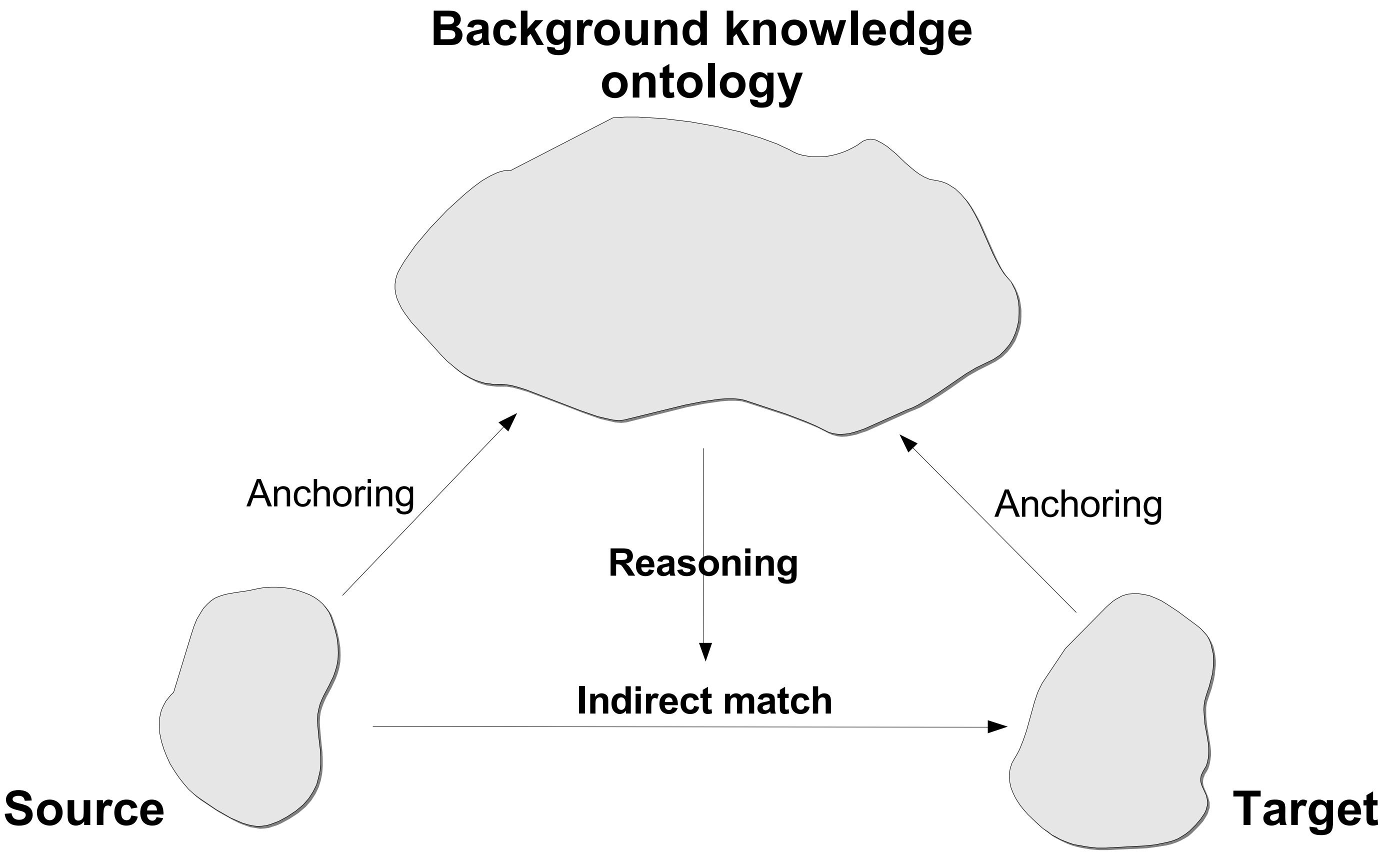
Introduction – ontology matching problem

- Semantic integration problem
  - Integrate ontologies, schemas, lists of terms
  - The goal is to find corresponding entites, semantically related
  - The problem becomes more sever because the amount of available data is rapidly growing
  - Traditional methods usually exploit lexical and structural similarity

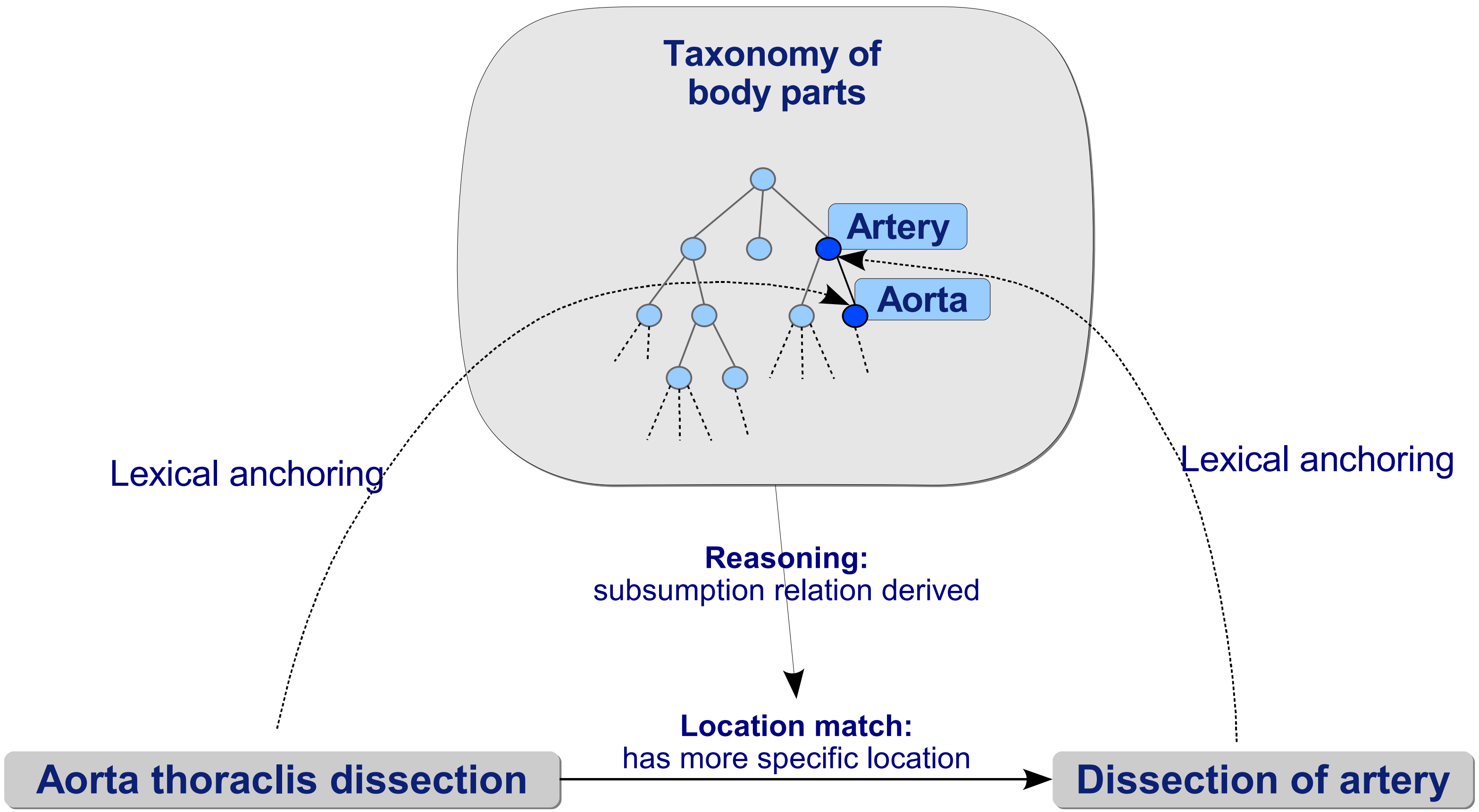
Content

- Introduction
- **Our approach: Using background knowledge ontology**
- Test data in our case study
- Experiments using the test data
- Evaluation agains state-of-the-art ontology matching tools
- Conclusions and future work

Our approach: semantic matching using background ontology



Our approach: semantic matching using background ontology



Content

- Introduction
- Our approach: Using background knowledge ontology
- **Test data in our case study**
- Experiments using the test data
- Evaluation agains state-of-the-art ontology matching tools
- Conclusions and future work

Test data in our case study

- Match two unstructured lists from intensive care domain
  - Source: OLVG list of concepts
  - Target: AMC list of concepts
- Use DICE ontology that describes the intensive care domain as a background knowledge

Case study: match OLVG to AMC list using DICE as background ontology

- Test data: Source - **OLVG** list
  - Comes from OLVG hospital in Amsterdam
  - List of reasons for admission used to register patients admitted to the Intensive care unit
  - Written in Dutch language, few in english
  - Flat list with no structure
  - 1399 concepts
  - Each concept is labeled with not more than 7 words
  - 95% of the concepts are labeled with not more than 3 words
  - Created by doctors from OLVG by adding the non-existing reasons in the system

Case study: match OLVG to AMC list using DICE as background ontology

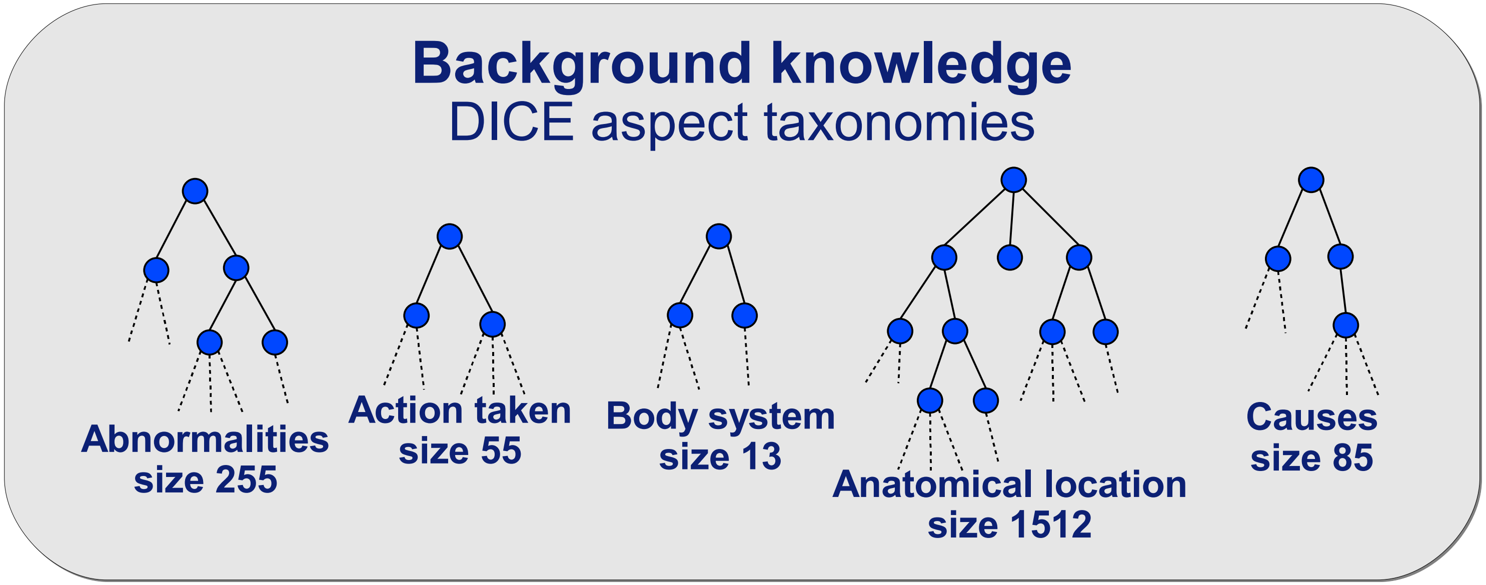
- Test data: Target - **AMC** list
  - Comes from AMC hospital in Amsterdam
  - List of reasons for admission used to register patients admitted to the Intensive care unit
  - Written in Dutch language
  - Flat list with no structure
  - 1460 concepts, labeled with synonyms as well
  - Created by medical experts from the AMC hospital

Case study: match OLVG to AMC list using DICE as background ontology

- Test data: Background ontology - **DICE** ontology
  - Describes the intensive care domain
  - In Dutch language, contains some english terms as well
  - Created by medical experts in the AMC hospital
  - In total contains 2300 concepts

Case study: match OLVG to AMC list using DICE as background ontology

- Test data: Background ontology - **DICE** ontology
  - We used five taxonomies (called aspects):



Content

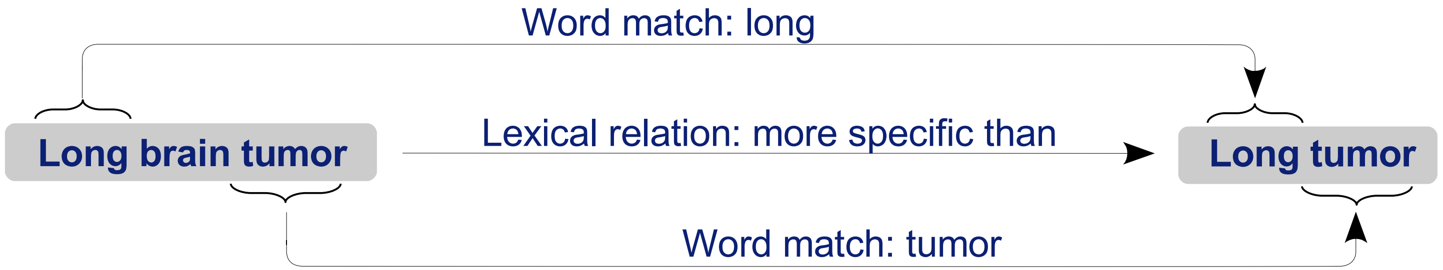
- Introduction
- Our approach: Using background knowledge ontology
- Test data in our case study
- **Experiments using the test data**
- Evaluation agains state-of-the-art ontology matching tools
- Conclusions and future work

Experiments

- Experiment 1: Lexically matching OLVG to AMC
  - Use simple lexical matching technique
  - Use state-of-the-art ontology matching tools to match OLVG to AMC
- Experiment 2: Semantic matching OLVG to AMC
  - Use DICE as background knowledge
  - Step 1: Anchor OLVG and AMC to DICE
  - Step 2: Derive matches OLVG to AMC through DICE

Experiments

- Experiment 1: Lexically matching OLVG to AMC
  - Goal: find equivalence or subsmuption relations
  - Account for lexical variations:
    - Case sensitivity
    - Word order
    - Interpunction, special characters
    - Germanic construction (*“Hersentumor”* and *“Tumor”*)
    - Use synonyms
  - Lexical heuristic for discovering subsumption relations:



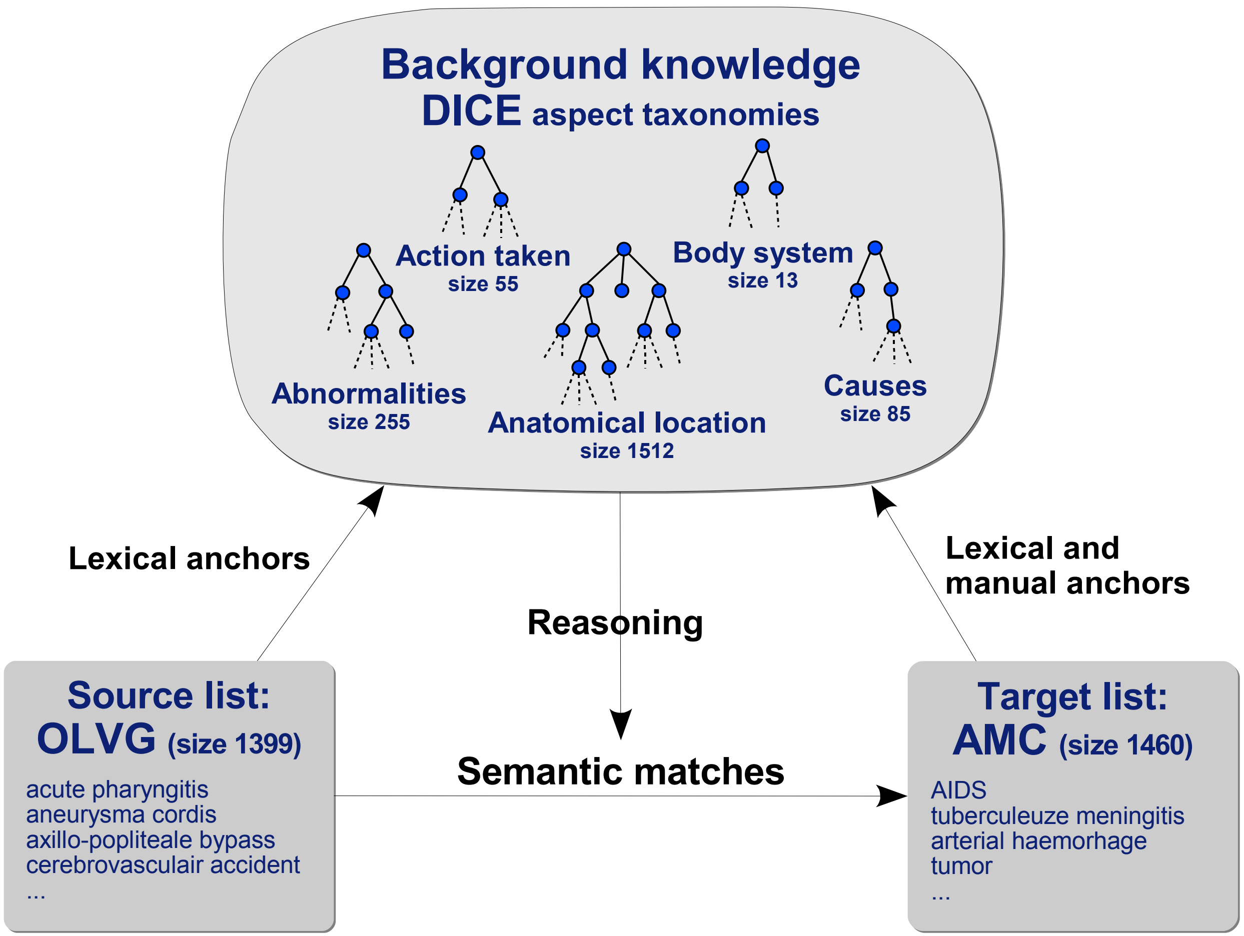
Experiments

- Experiment 1: Lexically matching OLVG to AMC
  - Perform match with state-of-the-art ontology matching tools (boils down to lexical match)
  - The results of the lexical matching are:

Method	OLVG concepts matched
Lexical matching	**582**
FOAM	**696**
Falcon-AO	**683**

Experiments

- Experiment 2: Semantic matching OLVG to AMC



Experiments

- Experiment 2: Semantic matching OLVG to AMC
- Step 1: Anchoring
  - Use the same lexical matching technique to find anchors from OLVG and AMC to DICE
  - OLVG to DICE – lexical anchoring
  - AMC to DICE – lexical + manual anchoring

Experiments

- Experiment 2: Semantic matching OLVG to AMC
- Step 1: Anchoring results:

	OLVG list	AMC list		
	Lexical	**Expert-manual**	**Additional lexical**	**Total**
Abnormality	**354**	1168	271	**1439**
Action taken	**109**	292	122	**414**
Body system	**3**	1217	2	**1219**
Location	**255**	1336	721	**2057**
Cause	**60**	555	132	**687**
	781	4568	1248	**5816**

Experiments

- Experiment 2: Semantic matching OLVG to AMC
- Step 1: Anchoring results:

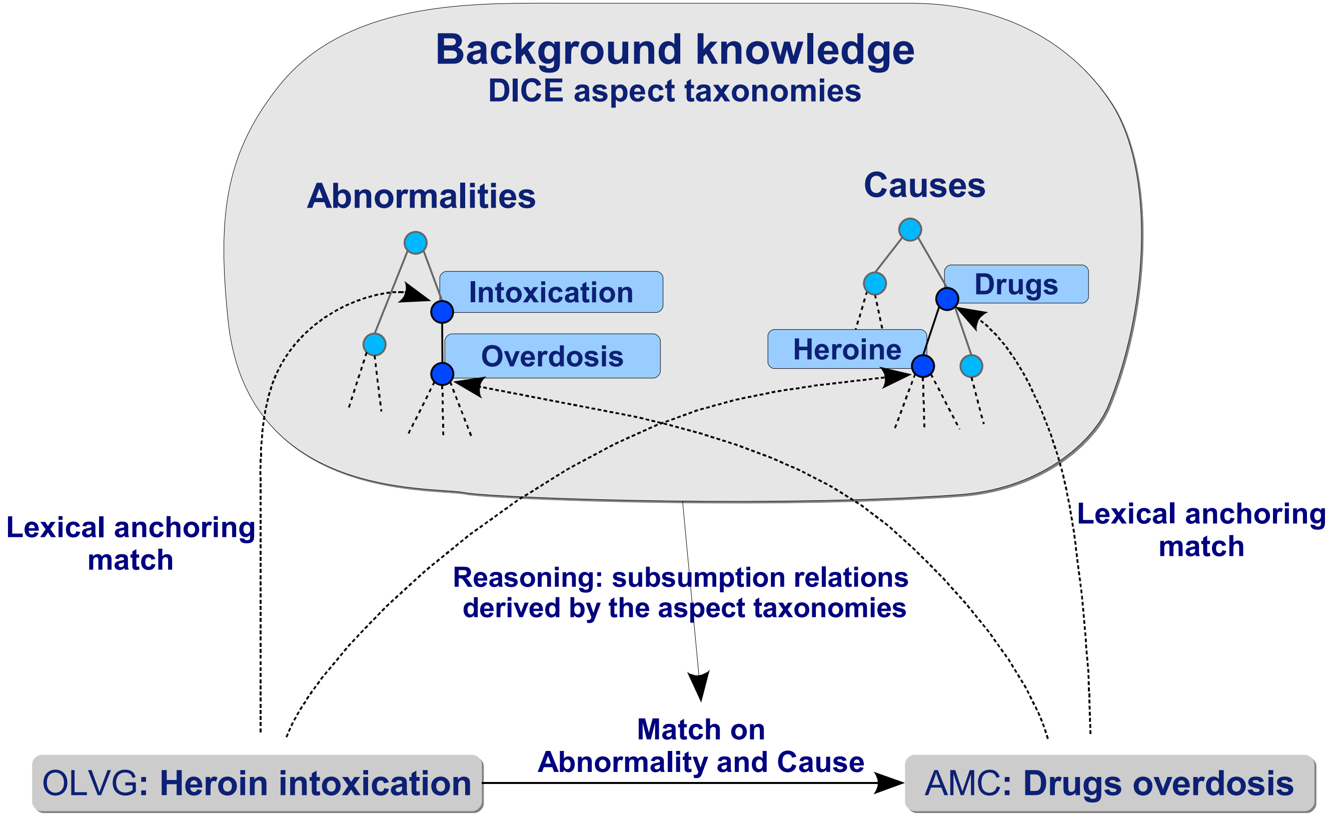
	OLVG	AMC
Anchored on 5 DICE aspects	0	2
Anchored on 4 DICE aspects	0	198
Anchored on 3 DICE aspects	4	711
Anchored on 2 DICE aspects	144	285
Anchored on 1 DICE aspects	401	208
	549	**1404**

Experiments

- Experiment 2: Semantic matching OLVG to AMC
- Step 2: Deriving semantic matches
  - Derive a match if the anchors are related
  - Multiple matches are possible for a single concept
  - Ranking the results
    - 1. Lexical equivalence is ranked highest
    - 2. Number of DICE aspects that support the match
    - 3. Favor equivalence on the DICE aspects

Experiments

- Experiment 2: Semantic matching OLVG to AMC
- Example of semantic match:



Content

- Introduction
- Our approach: Using background knowledge ontology
- Test data in our case study
- Experiments using the test data
- **Evaluation agains state-of-the-art ontology matching tools**
- Conclusions and future work

Evaluation against state-of-the-art ontology matching tools

- Evaluation against Gold Standard of size 200 concepts, created by medical expert from the AMC hospital in Amsterdam

Type of match	Semantic match	Lexical match	FOAM	Falcon-AO
Correct match found	107	73	41	28
Correct no match found	43	43	26	32
Total positive	**150 (=75%)**	**116 (=58%)**	**67 (=33%)**	**60 (=30%)**
Incorrect match found	/	5	47	78
Incorrect no match found	50	79	86	62
Total negative	**50 (=25%)**	**84 (=42%)**	**133 (=67%)**	**140 (=70%)**

Content

- Introduction
- Our approach: Using background knowledge ontology
- Test data in our case study
- Experiments using the test data
- Evaluation agains state-of-the-art ontology matching tools
- **Conclusions and future work**

Conclusions

- Semantically rich background knowledge considerably helps in matching structure poor lists of concepts
- When there is no lexical, instance and structural overlap, using background knowledge is the only way to find relatedness (*Heroin Intoxication – Drugs Overdosis* example)
- Reasoning over several hierarchies made it possible to discover matches beyond subsumption relations

Future work

- Test in different domain – experiments in music domain
- Investigate the success factors – experiments with anatomy ontologies
  - What are the properties of concepts for which background knowledge can help in matching?
  - What parts of an ontology are useful when it is used as a background knowledge?

Thank you for your attention !

Questions?