Semantic Matching using the UMLS

Jetendr Shamdasani,Tamas Hauer, Peter Bloodworth, Andrew Branson, Mohammed Odeh and Richard McClatchey

University of the West of England

Contact: jet@cern.ch

Semantic Matching using the UMLS

- The Medical Domain
- Semantic Matching - Definitions
- The UMLS
- The SMatch Algorithm
- Modifications
- Results

The Medical Domain

- The medical domain has very specialized terminology
- Heterogeneity occurs at both the terminological level and the conceptual level
- There are hundreds of medical ontologies available today, many of these are in use
- Integration is required for interoperability and data sharing
- Semantic Matching allows more fine grained relationships to be discovered between concepts

Semantic Matching

- We are matching two trees of Strings
- We are focused on the anchoring to a background resource
- We use a background resource to define *context*
- Discovery of Set Theoretic Relationships between concepts in two or more ontologies
- We are able to discover relationships in the ($=, \sqsubseteq, \sqsupseteq$) range
- Semantic matching in our case does not return a similarity measure

Semantic Matching

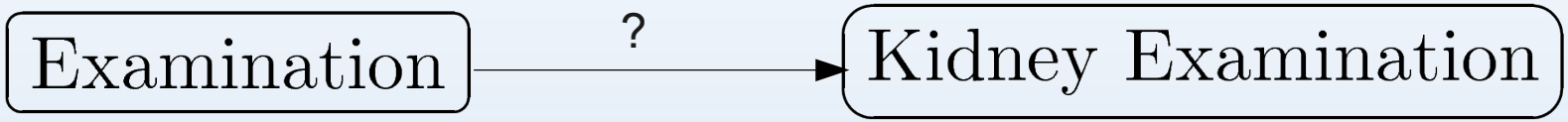
- We have extended the original SMatch algorithm [1]
- WordNet is a very general resource focusing on lexical knowledge. However, we need a background resource that is more domain specific.
- Our primary focus has been to replace its reliance on WordNet with a domain specific resource i.e. The UMLS

A story

- We have several hospitals in our project
  - Each of these are focused on the same rare diseases
  - Since these diseases are rare their data sets are small individually
  - We need to integrate this data
  - Each of them are described by their own ontology

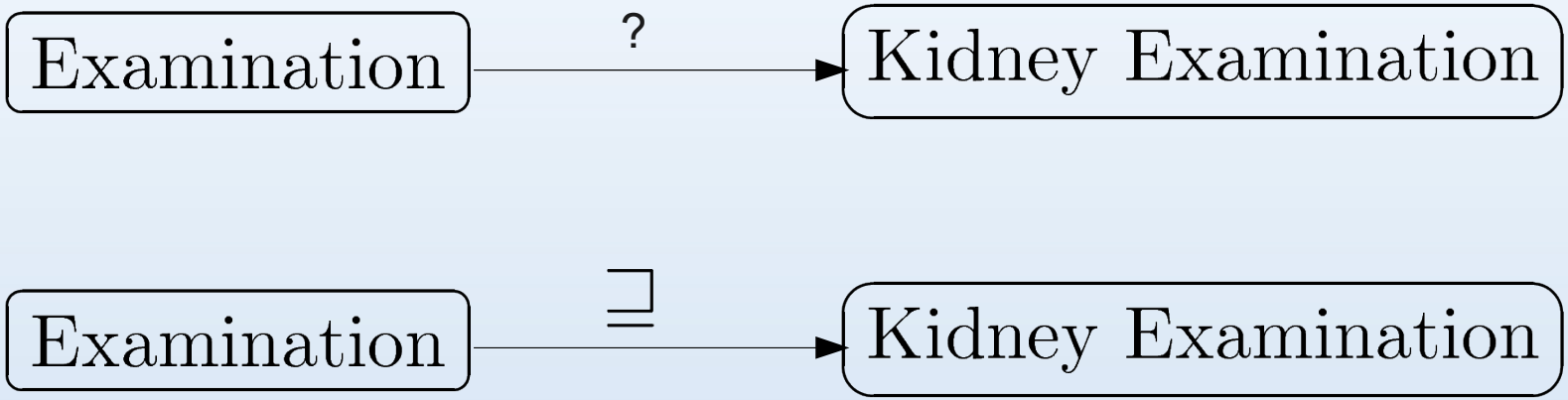
Examples

- Two concepts of “Examination” and “Kidney Examination”



Examples

- Two concepts of “Examination” and “Kidney Examination”



Examples

- The semantics of the concept “Kidney Examination” is :

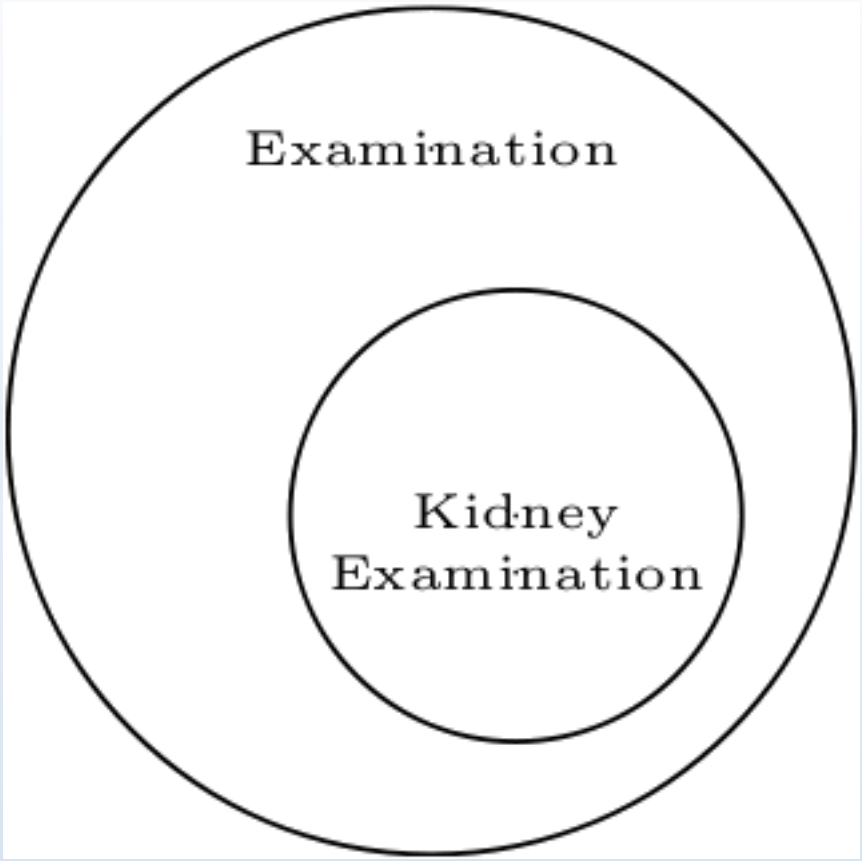
$$Kidney \sqcap Examination$$

- The semantics of the concept “Examination” is:

$$Examination$$

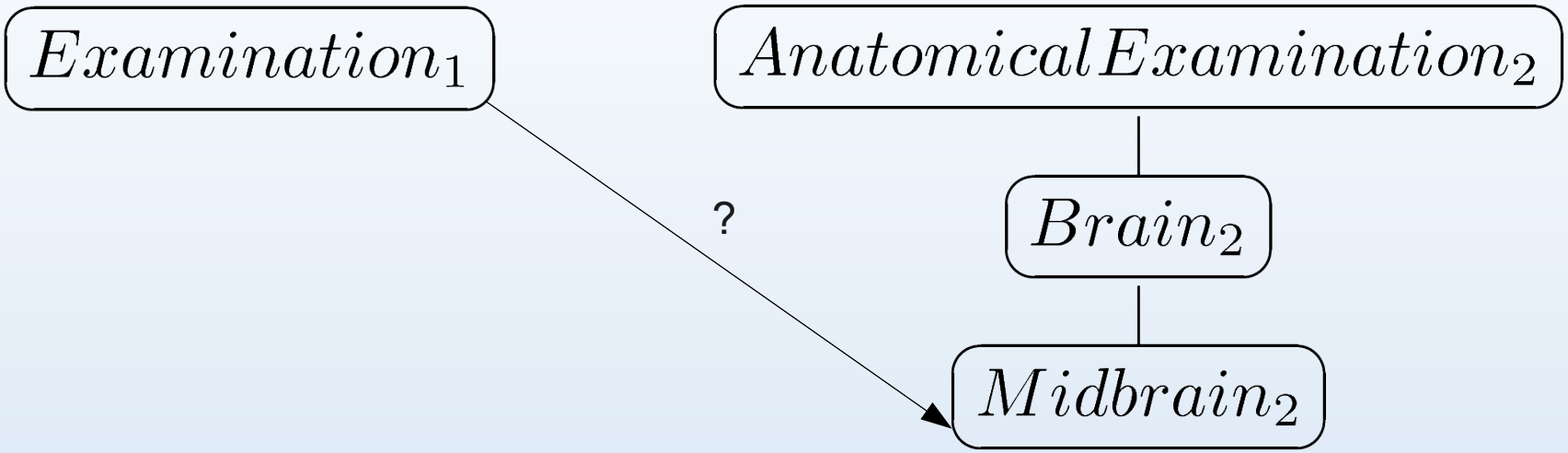
Examples

Hence visually:



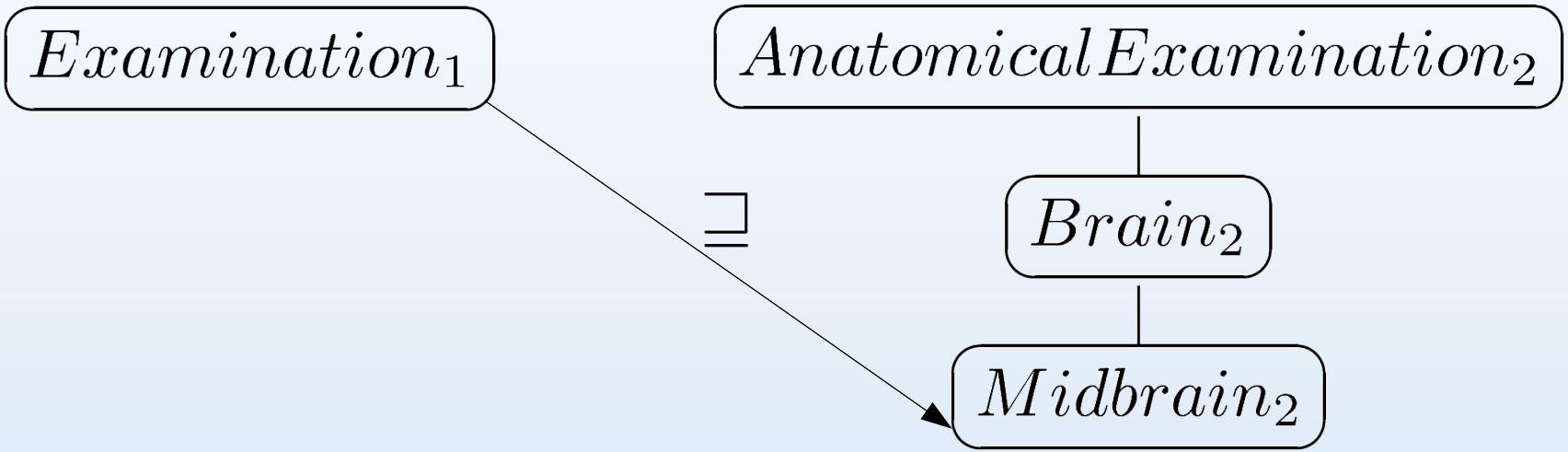
Examples with Structure

- Two input trees:

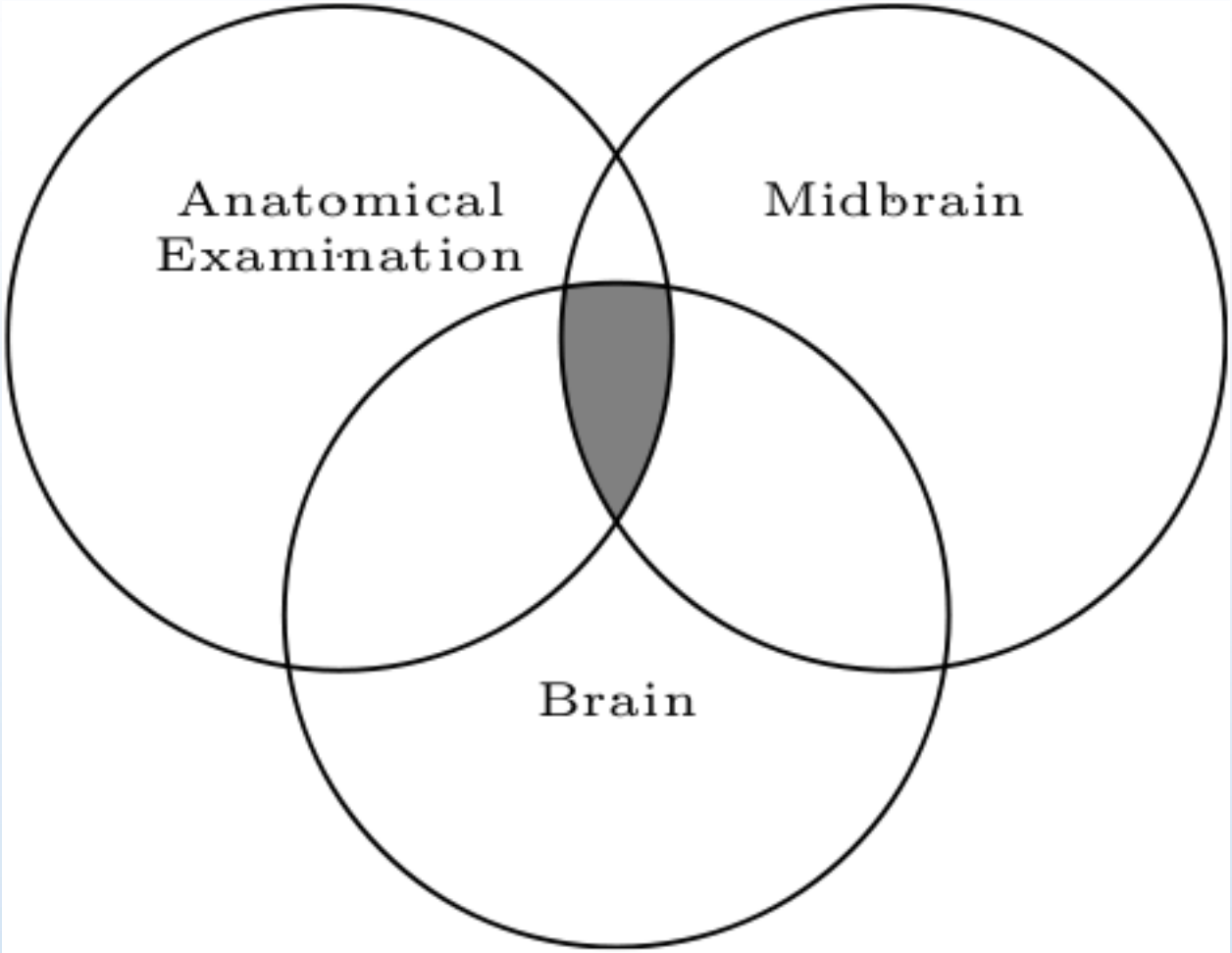


Examples with Structure

- Two input trees:



Example with Structure



$$Midbrain \sqcap Brain \sqcap Anatomical\ Examination$$

Anchoring to Background Knowledge

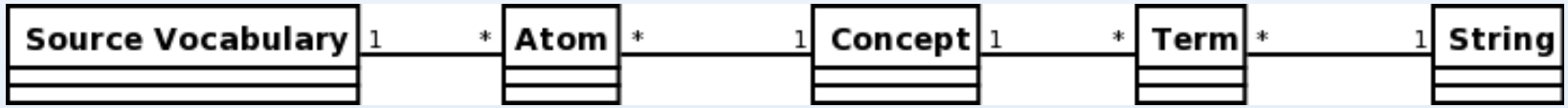
- As mentioned previously SMatch is heavily reliant on WordNet
- We have chosen to replace its reliance on WordNet with the UMLS
- This is because there is no medical WordNet available as of today
- We evaluated various forms of background knowledge and we found the UMLS to be the best fit
- The UMLS has good coverage of the medical domain and it is broad enough for the matching process

The UMLS

- The UMLS is a medical thesaurus which integrates biomedical knowledge from varying vocabularies.
- It can be considered to be a meta ontology of biomedical knowledge.
- The 2008AA version has (140) Source Vocabularies with (1553638) number of Concepts and (7781500) number of Atoms.

The UMLS

- The UMLS consists of Concepts (CUI), Atoms (AUI), Source Vocabularies (SAB), Lexical Groups (LUI) and Strings (SUI).



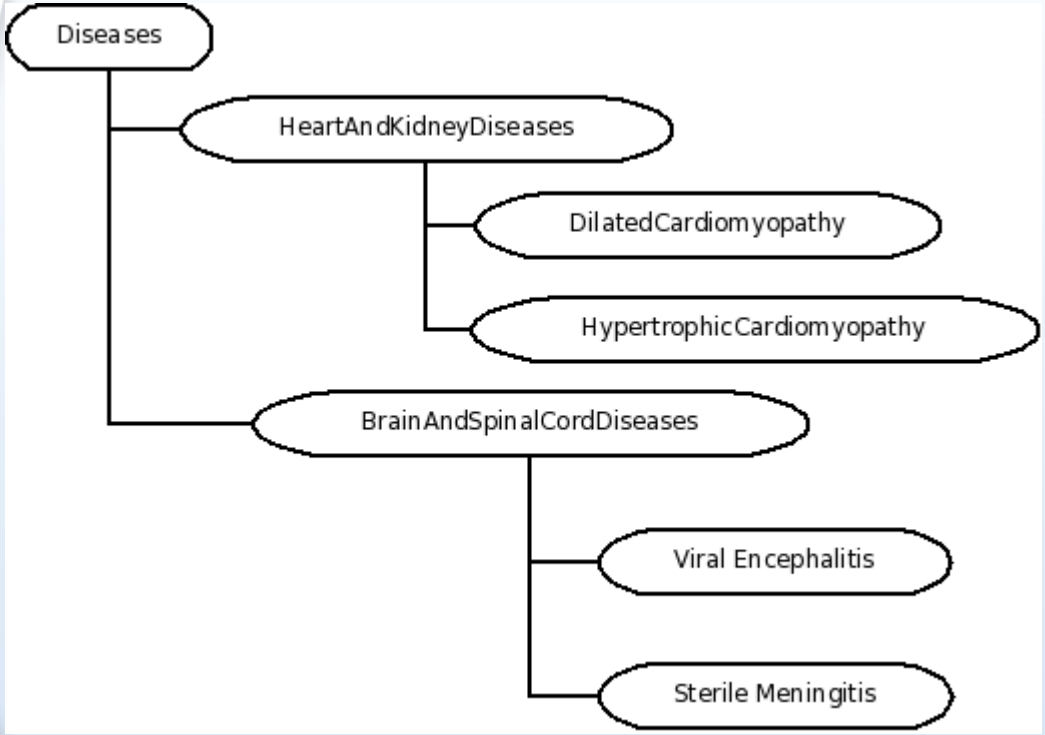
The UMLS

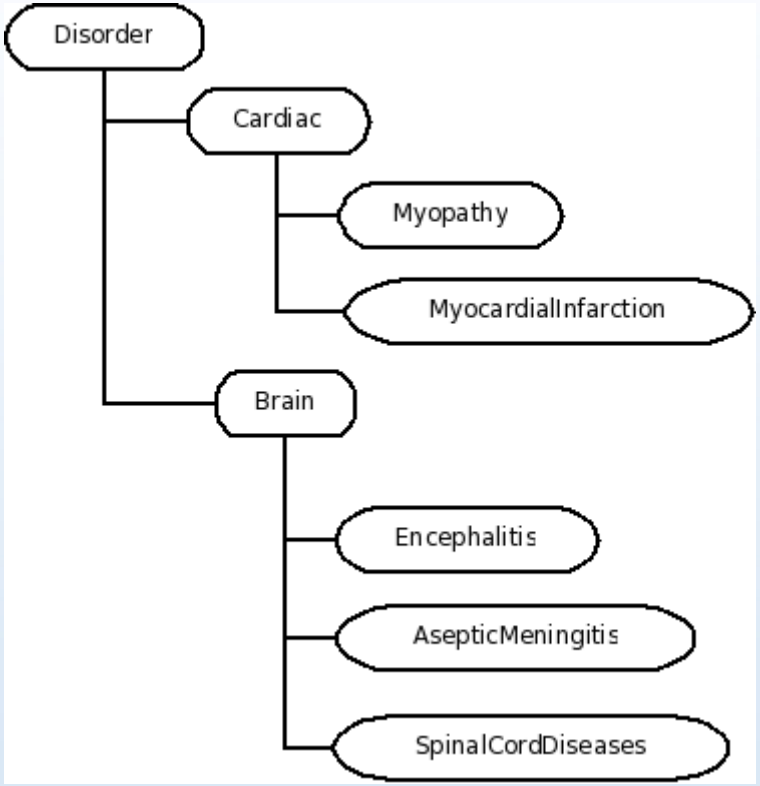
- The relationships within the UMLS are represented as a Graph of Concepts and Atoms.
- The relationships between concepts are Broader Than (RB), Narrower Than (RN), Parent (PAR), Child (CHD) and Sibling (SIB).
- The relationships between Atoms are taken from the Source Vocabularies. e.g. “part_of” and “isa”.

The Modified SMatch Algorithm

- The SMatch algorithm has 4 steps these are:
  - String to Formula Conversion
  - Context Creation and Filtering
  - Atomic Formula Matching with the UMLS
  - Reasoning

Example


Diseases
HeartAndKidneyDiseases
DilatedCardiomyopathy
HypertrophicCardiomyopathy
BrainAndSpinalCordDiseases
Viral Encephalitis
Sterile Meningitis


Disorder
Cardiac
Myopathy
MyocardialInfarction
Brain
Encephalitis
AsepticMeningitis
SpinalCordDiseases

Algorithm - String to Formula Conversion

- The purpose of this step is to convert a string to a logical formulae for the semantic matching process
- These formulae are composed of atomic formulae which have *concepts* from the UMLS attached
- For example if we take a string called “Heart” and we query the UMLS for this term we get the following concepts returned:
  - C0018787 | Heart
  - C0153500 | Malignant neoplasm of heart
  - C0153957 | Benign neoplasm of heart
  - C0795691 | Heart problem
  - C1281570 | Entire heart

Algorithm - String to Formula Conversion

- This step works exclusively on the label of a node.
- If we have a node with the label “HeartAndKidneyDiseases”
- After we tokenize the strings we attach *concepts* from the UMLS (if there is a direct string to concept match we do not tokenize the string):

$$\text{heart}\{\text{cui}\#5\},\text{kidney}\{\text{cui}\#1\},\text{disease}\{\text{cui}\#5\}$$

- The final formula is :

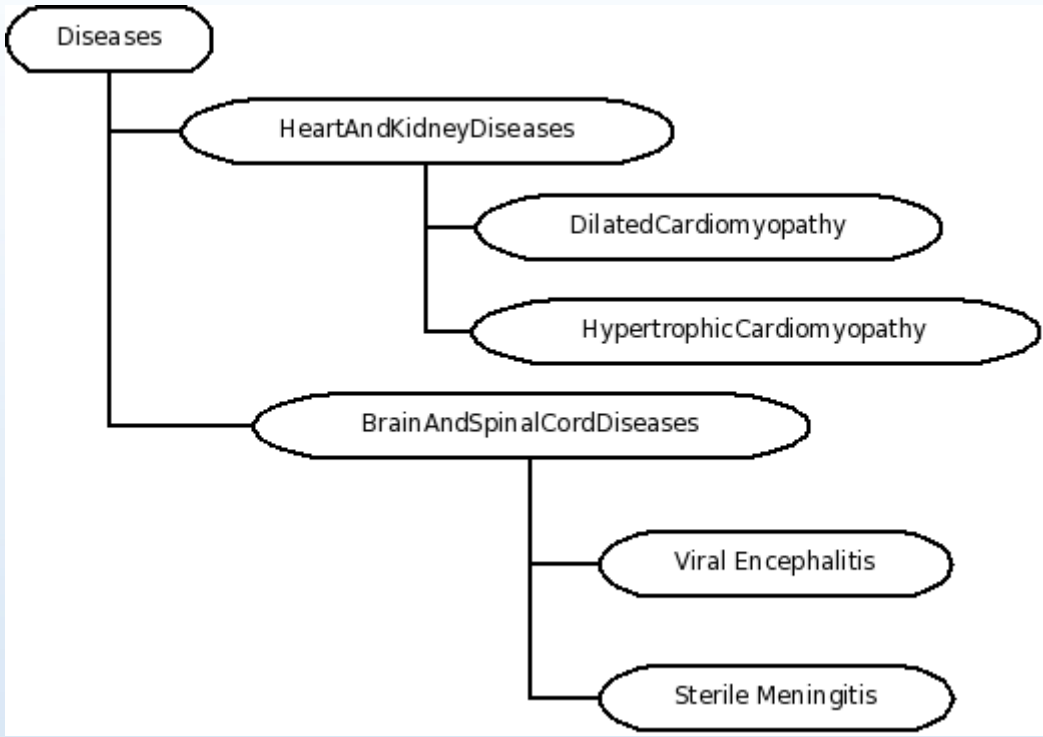
$$((heart \sqcup kidney) \sqcap disease)$$

Algorithm - Context Creation and Filtering

- The purpose of this step is to define context for a node using its formula from the previous step.
- This constrains the meaning of the node using its parents.
- The concepts from the UMLS are still attached to atomic formulae.
- We also filter concepts at the structural level as the original SMatch does.

Algorithm - Context Creation and Filtering

- Step two consists of creating a context for a given node.
- This involves taking a conjunction from the current node to its parent.
- Hence it captures the meaning of the node giving it context.



$$((disease) \sqcap ((heart \sqcup kidney) diesease) \sqcap (hypertrophic\ cardiomyopathy))$$

Algorithm - Context Creation and Filtering

- We also perform filtering of concepts for this we use the information present in the UMLS.
- There are 3 tables for disambiguation available to us these are MRREL (Concept relationships), MRHIER (Atom relationships) and MRCOC (Co-Occurrence relationships from text).
- We use all this information for our filtering process, however we only use one feature for disambiguation at a single time i.e. they are not used in conjunction with each other.

Algorithm - Atomic Formula Matching with the UMLS

- The purpose of this step is to match atomic formulae which have concepts attached using the UMLS
- For this we use the hierarchical information present in the UMLS
- We have used both the Concept and Atom hierarchies for the matching of atomic formulae.

Algorithm - Step 3

- For Atoms the rules are :

  $\equiv$ rule - If a Concept from A contains a Concept from B then a $\equiv$ relationship is declared.

  $\sqsubseteq$ rule - If an Atom from a Concept in A is a subclass of an Atom from a Concept B in a single source vocabulary then a $\sqsubseteq$ relationship is declared.

  $\sqsupseteq$ rule - If an Atom from a Concept in A is a superclass of an Atom from a Concept in B in a single source vocabulary then a $\sqsupseteq$ relationship is declared.

Algorithm - Step 3

- For Concepts the rules are :

  $\equiv$ rule - If a Concept from A contains a Concept from B then a $\equiv$ relationship is declared.

  $\sqsubseteq$ rule - If a Concept from A is a subclass of a Concept from B i.e. it is related via a RN or CHD relationship then a is $\sqsubseteq$ declared.

  $\sqsupseteq$ rule - If a Concept from A is a superclass of a Concept from B i.e. it is related via a PAR or RB relationship then a $\sqsupseteq$ relationship is declared.

Algorithm - Reasoning

- The purpose of this step is to deduce a relationship between two concepts.
- We now bring all the results from previous steps into the same logical formalism.
- As with the original SMatch we take a propositional reasoning approach.

Algorithm - Reasoning

- The equation is the following:

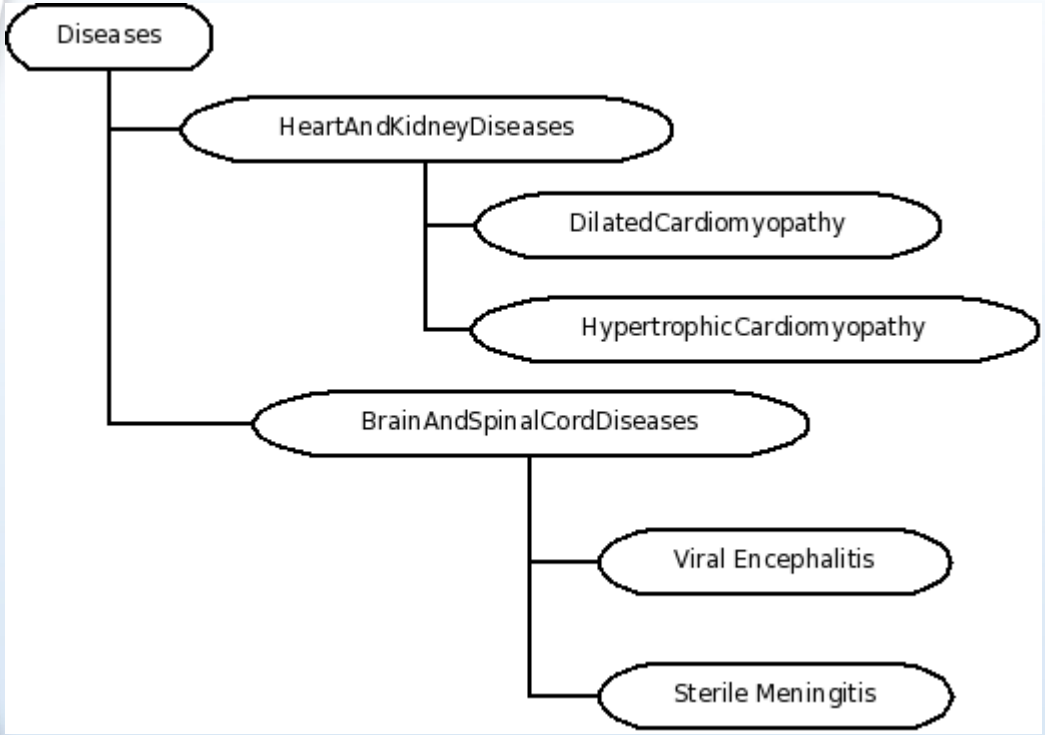
$$axioms \rightarrow rel(context_A, context_B)$$

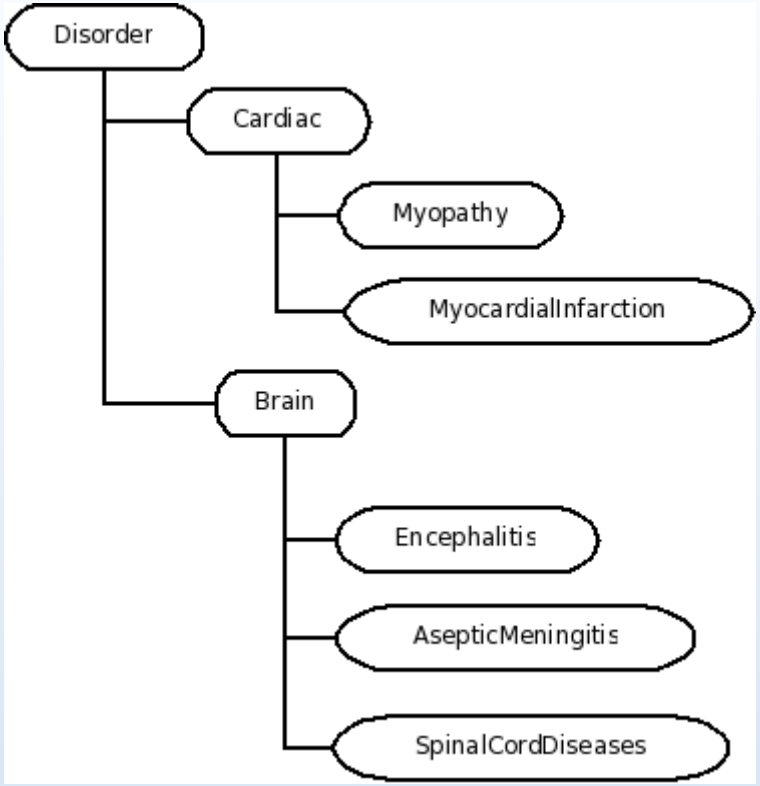
- The axioms are the background theory from the UMLS, rel is the relationship we are trying to prove. For reasoning purposes we take the negation of the above equation:

$$axioms \wedge \neg rel(context_A, context_B)$$

- Subsumption is converted to its propositional equivalent ($\rightarrow$) we try and prove both subsumption and supersumption to prove equivalence.

Example


Diseases
HeartAndKidneyDiseases
DilatedCardiomyopathy
HypertrophicCardiomyopathy
BrainAndSpinalCordDiseases
Viral Encephalitis
Sterile Meningitis


Disorder
Cardiac
Myopathy
MyocardialInfarction
Brain
Encephalitis
AsepticMeningitis
SpinalCordDiseases

Algorithm - Reasoning

- If are matching the nodes “HeartAndKidneyDiseases” from Tree1 and “MycardialInfarction” from Tree2.
- The axioms for this task would be :

$$((disease \leftrightarrow disorder) \wedge (heart \leftrightarrow cardiac) \wedge (disease \leftarrow myocardial\ infarction))$$

- The final formula if we are trying to prove a more general relationship between the two would be :

$$((disease \leftrightarrow disorder) \wedge (heart \leftrightarrow cardiac) \wedge (disease \leftarrow myocardial\ infarction)) \wedge \neg(((disease) \wedge ((heart \vee kidney) \wedge disease)) \leftarrow ((disorder) \wedge (cardiac) \wedge (myocardial\ infarction)))$$

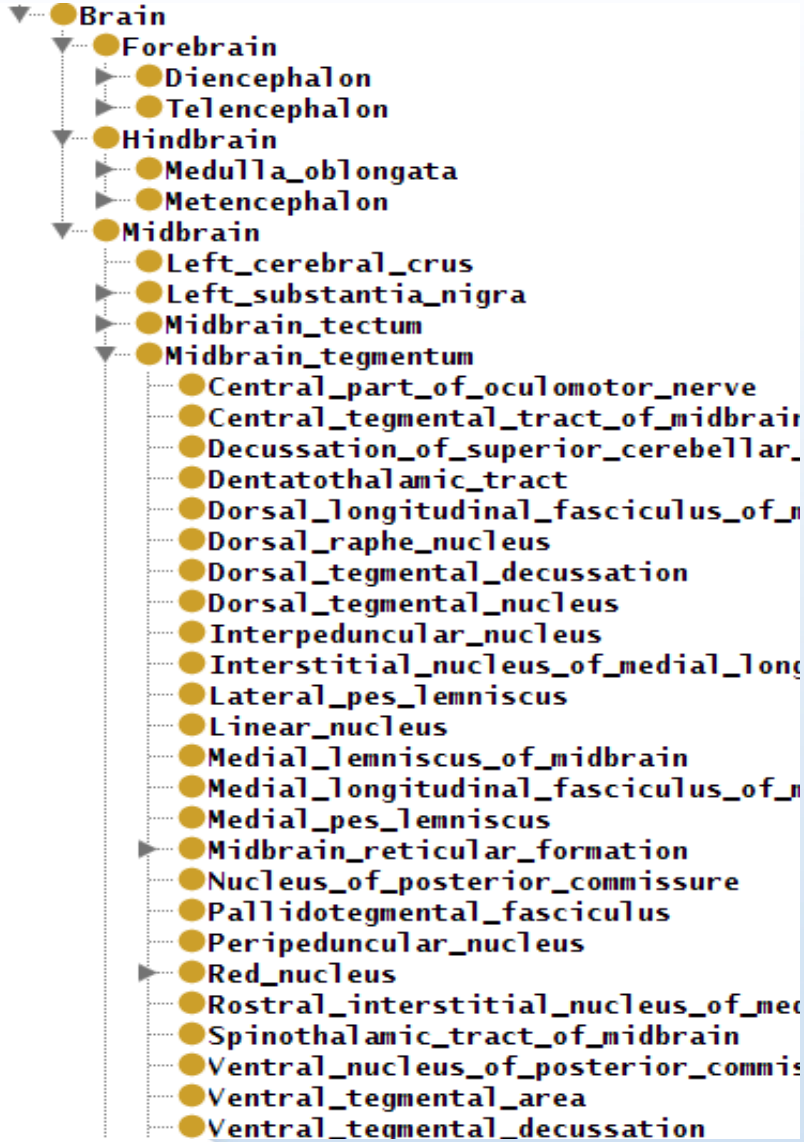
Implementation Details

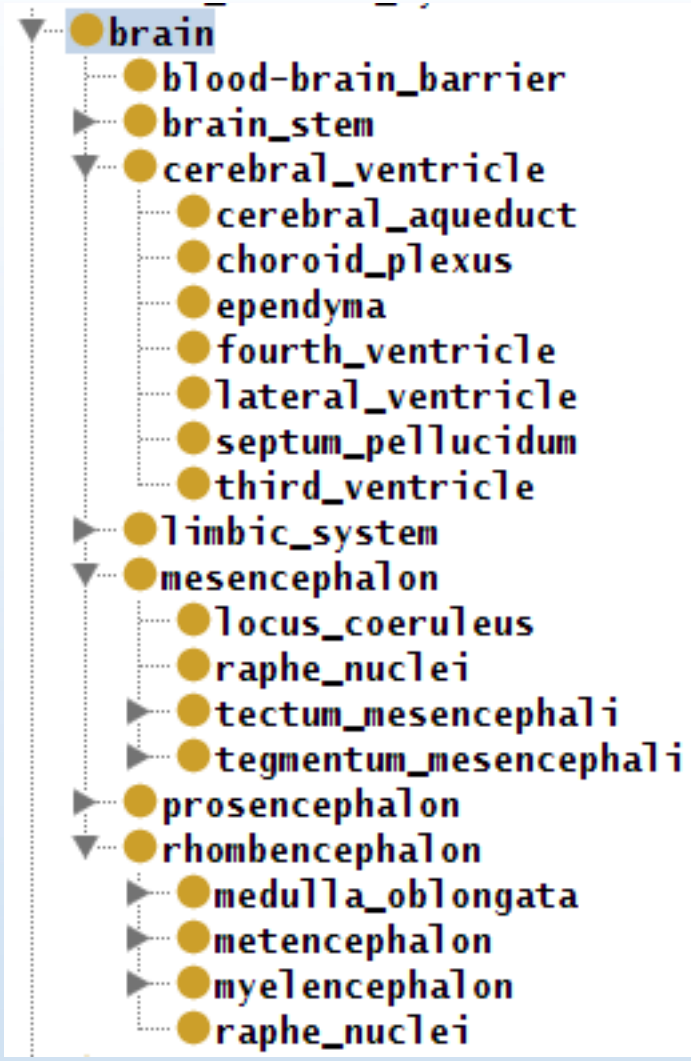
- We implemented the system from scratch
- It was developed within the framework of the Health-e-Child project (IST 2004-027749)
- We are currently preparing the implementation for an open source release

Experimental Setup

- We matched subsets of the FMA and MeSH ontologies. Our trees are rooted at the concept Brain in both Ontologies.
- To extract a tree from the FMA we followed the “regional_part_of” relationship down to its leaf nodes. (Total of 476 Concepts)
- To extract a tree from MeSH we traversed the tree down to its leaf nodes (Total of 181 Concepts)
- Our gold standard was created with the aid of a domain expert. We selected 20 random concepts from our FMA subset and 40 random concepts from our MeSH subset.
- We have used the 2008AA version of the UMLS in our experiments.

Trees : FMA vs MeSH





Results

- We ran differing versions of our algorithm using different features of the UMLS for the filtering and matching of atomic formulae.

- We found that precision and recall was unrealistically high with our gold standard. We need one with a significant size! This is future work.

Conclusions

- We have shown a means of adapting semantic matching to the medical domain using the UMLS as a background resource.
- We have shown how different features of the UMLS (Concepts, Atoms, Co-Occurrences) can be used for disambiguation.
- The filtering is only useful in cases where two ontologies may have a similar set of terms but their contexts are dissimilar.

Future Work

- We need to conduct a more thorough evaluation with a well established set of gold standards for the medical domain.
  - This is difficult since there isn't one readily available.
- We will also submit our results to this years' OAEI competition.
- We will also investigate how to adapt different background resources from different domains for the semantic matching process.

Lunch time!

- References:
  - F. Giunchiglia et al. Semantic Matching : Algorithms and Implementation. Journal of Data Semantics pp. 1-38 (2007)