Semantic Matching Using the UMLS

Jetendr Shamdasani*, Tamás Hauer, Peter Bloodsworth, Andrew Branson, Mohammed Odeh, and Richard McClatchey

CCCS Research Centre, CEMS Faculty, University of the West of England, Coldharbour Lane, Frenchay, Bristol BS16 1QY, UK
firstname.lastname@cern.ch

Abstract. Traditional ontology alignment techniques enable equivalence relationships to be established between concepts in two ontologies with some confidence value. With semantic matching, however, it is possible to identify not only equivalence ($\equiv$) relationships between concepts, but less general ($\sqsubseteq$) and more general relationships ($\sqsupseteq$). This is beneficial since more expressive relationships can be discovered between ontologies thus helping us to resolve heterogeneity between differing semantic representations at a finer level of granularity. This work concerns the application of semantic matching to the medical domain. We have extended the SMatch algorithm to function in the medical domain with the use of the UMLS metathesaurus as the background resource, hence removing its previous reliance on WordNet, which does not cover the medical domain in a satisfactory manner. We describe the steps required to extend the SMatch algorithm to the medical domain for use with UMLS. We test the accuracy of our approach on subsets of the FMA[1] and MeSH[2] ontologies, with both precision and recall showing the accuracy and coverage of different versions of our algorithm on each dataset.

1 Introduction

Semantic Matching [1] is the process of discovering set theoretic based relationships between differing concepts within two or more ontologies. It is richer than identifying equivalent concept pairs since finer relationships between concepts may be discovered, in particular *less general* ($\sqsubseteq$) and *more general* ($\sqsupseteq$). The framework is well understood, yet to our knowledge it has not yet been exploited in medical applications. Ontologies are commonly used in the medical domain for creating widely agreed terminologies, taxonomies, for annotating medical data from publications to patient records and in reformulating clinicians' queries. The value of ontology matching lies in the fact that there are numerous, largely independent terminologies, taxonomies and ontologies covering the various subdomains of medicine. In order for interoperability between these semantic models to be achieved they need to be matched. Alignment takes

* Corresponding author.

[1] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

[2] http://www.nlm.nih.gov/mesh/

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 203–217, 2009.

this a step further and allows us to describe the overlap between subdomains. As an example if we consider the case of two clinicians trying to gain access to a clinical system, each clinician has their own ontology of a subset of a domain and the clinical system has a more general view of the clinical world. For these clinicians to be able to access and share data they need to speak the same language. An alignment between these varying models is beneficial since it facilitates the sharing of information.

In this paper we present a semantic matching framework for use in the medical domain. The current work is an extension of the SMatch approach [2] and has been motivated by the requirements of the E.C. funded Health-e-Child project[3]. The SMatch framework is based on two cornerstones: firstly, it uses a predicate logic framework to unify node-wise string matching with structural matching (thus exploiting their interplay.) and secondly, it uses a background thesaurus as an anchor for tokenization, sense filtering and knowledge seeding. The direct application of the algorithm to medical ontologies is somewhat weak because of its reliance on the WordNet thesaurus [3]; this was shown in our previous work [4]. Previously we only used the UMLS Concept hierarchy for disambiguation. In this work we present results using different features of the UMLS for hierarchical disambiguation with our structural filtering implementation which differs from the original SMatch approach. We have formalised the original steps for semantic matching and created a framework built entirely on the logic behind the original algorithm.

We have chosen the UMLS metathesaurus [5] as our background resource because of its wide coverage of clinical terms and its disambiguation of these terms. We have built a prototype matcher that has been evaluated on test ontologies with some initial results. We have formalised the steps to derive a semantic match between medical ontologies; these are shown with a worked example. Section 2 presents previous work conducted in the area of semantic matching and ontology alignment. Section 3 provides a brief introduction to the UMLS. Section 4 explains the idea behind semantic matching. Section 5 describes the modifications to the original SMatch algorithm for use with the UMLS. Section 6 presents our evaluation methodology for our algorithm. Section 7 presents the results with different versions of our algorithm. Finally, conclusions and directions for further work are presented in section 8.

2 Previous Work

Previous work within the area of ontology alignment has mostly focused on discovering equivalence relationships between different concepts and roles in an ontology. Their output is normally in the form of an equivalence relationship with a confidence value between concepts within the range 0 to 1. There are many different approaches within the literature which rely on different features of an ontology (its schema, instances etc..). A diverse number of approaches have been attempted in the ontology alignment field [1] and significant progress has been

[3] `http://www.health-e-child.org (IST 2004-027749)`

made. Much of this progress has been in the form of using differing background knowledge sources. Background knowledge has been used in the form of free text [6], previous alignments [7], search engines [8] and other ontologies. In our case *Semantic Matching* uses other ontologies as background resources. There have been approaches using single or multiple resources. A standard semantic matching approach is composed of an anchoring step, deriving semantics by using background knowledge and a set of reasoning rules. Semantic matching differs from other approaches because the goal is not to look for equivalence but rather to discover other relationships.

Aleskovsi et al [9] use a single background resource to match two flat lists of medical terms using a single background ontology. Their process consists of an anchoring step where they firstly tokenize the string to be matched and then "anchor" these to concepts in the background ontology. This is done for all strings in both lists. Once this has been completed they are able to derive a semantic match between two terms using the background resource. They use a set of predefined reasoning rules to achieve this, for example, using transitive reasoning. They have shown the application of these rules in their follow-up work [10] where they matched the anatomy parts of the CRISP to MeSH and used the FMA as a background resource.

Sabou et al [11] have extended the approach by Alekovsi et al. to use multiple background ontologies using the Swoogle ontology search engine. They firstly search for the concepts they are going to match then search for ontologies which contain both of these terms. If a suitable resource cannot be found, then separate ontologies are searched for and are then matched against each other, this process is repeated until a match is found. They semi-formalise the steps involved in their framework. Their search approach has been improved in a later work [12] by using the senses available from WordNet, therefore allowing a better ontology selection process for their background knowledge. SMatch is another semantic matching approach. Instead of defining inference rules, Giunchiglia et al. cast the ontology alignment problem as a propositional satisfiability problem. Here labels of nodes in trees are tokenized and the lexical structure in addition to the tree structure are represented by logical formulae built from WordNet senses that are associated with the tokens. These formulae are fed into a standard SAT solver to check the validity of the matches based on predefined reasoning rules. One strength of this approach is that there does not necessarily have to be a complete lexical overlap between the ontologies to be matched and WordNet. SMatch and its modification to use the UMLS is explained further in section 5 with a worked example.

3 The UMLS

The UMLS is a thesaurus of biomedical knowledge. Its purpose is to integrate conceptual information from differing sources of biomedical terminology. These can be complex ontologies or simple lists, being known as source vocabularies in the schema. The UMLS contains relationships between these concepts organised into its own model and consequently it is a conceptual resource about the

biomedical domain. A concept in the UMLS is identified by a CUI (Concept Unique Identifier) which is a unique value identifying a Concept. Every Concept has a set of Terms (or Lexical Groupings) identified by an LUI (Lexical Unique Identifier) and a set of Strings identified by a SUI (String Unique Identifier). Every concept has an Atom identified by an AUI (Atomic Unique Identifier) which are the original definitions of a concept from its source vocabulary. Figure 1 shows how these relate to each other via a UML diagram.

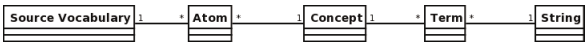


Fig. 1. A UML based representation of the UMLS

Every Concept in the UMLS is related to another concept in the UMLS hierarchy via Broader Than (RB), Narrower Than (RN), Parent (PAR), Child (CHD) and Sibling (SIB) relationships, this information being contained within the MRREL table of the UMLS. As well as relationships between concepts the UMLS also contains hierarchical information between Atoms in their original source vocabularies. Atoms are related to one another in simple taxonomic hierarchies with varying semantics for example "isa" or "part_of", this information is contained within the MRHIER table. Also for disambiguation purposes there is the MRCOC table which contains co-occurrences relationships between UMLS Concepts in text. Section 5 describes how the UMLS can be applied to semantic matching. The next section explains the idea behind semantic matching.

4 Semantic Matching Explained

Semantic Matching is the process of discovering $\equiv$, $\sqsubseteq$, $\sqsupseteq$ relationships between concepts in two ontologies. We have chosen to use the propositional reasoning based approach of the SMatch algorithm. This section will explain what semantic matching is. At the highest level, the SMatch algorithm assigns a logical formula to each node of a tree, such that the meaning of the node that is inferred from both the lexical and the tree structures which are encoded by the predicate calculus. These formulae are in turn input into reasoning rules to derive a match.

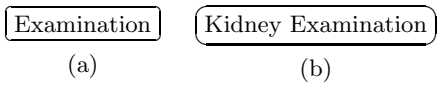


Fig. 2. String Matching Example

If we consider the matching of two strings in figure 2, these strings represent two concepts which are the concepts of "Examination" and "Kidney Examination" respectively. The expected result for this case would be that "Kidney Examination" is a $\sqsubseteq$ of "Examination". SMatch is able to discover this relationship

due to its logic based representation of strings which are converted into atomic formulae. The atomic formula for the node "Examination" would be the name of the node i.e. *examination* and the atomic formula for the node "Kidney Examination" would be the name of the node as well, however the white space character is converted to $\sqcap$ which means the meaning of the concept is the conjunction of the concepts of "Kidney" and "Examination" i.e. $kidney \sqcap examination$. The shaded region in figure 3(a) shows the meaning of this concept and figure 3(b) shows the match result.

Figure 4 shows an example of matching two trees taking their structures into account, Tree 1 (on the left) has no structure with only one node with the label "Examination" whereas Tree 2 does have structure. If we were matching the nodes "Examination" and "Midbrain" the correct result would be that "Examination" $\sqsupseteq$ "Midbrain" due to the fact that the meaning of the node "Midbrain" is influenced by its parent nodes. To achieve this SMatch takes two factors into account, firstly from our background knowledge we know that "Examination" is equivalent to "Anatomical Examination"; this is the background theory that is implicit and is used in the reasoning process. Also the node with structure ("Midbrain") is given a *context*, the manner in which this is achieved is that the logical formula of "Midbrain" is altered to take into account its relationship with its parents. This constraint is applied by an intersection with its logical formula to its path to the root, the new formula for this node is $midbrain \sqcap brain \sqcap anatomical\ examination$. The shaded area in figure 5 shows the meaning of this concept visually.

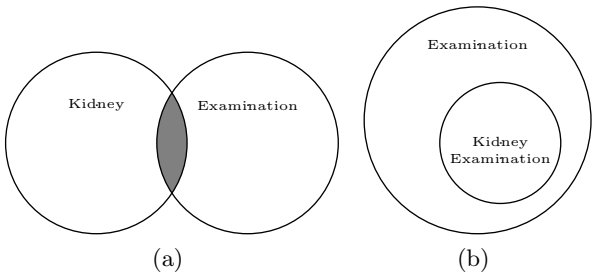


Fig. 3. Venn diagrams showing the meaning of the concept "Kidney Examination" (a) and the result of the matching process (b)

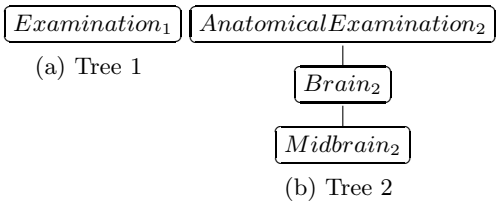


Fig. 4. Two example trees showing how SMatch takes structure into account

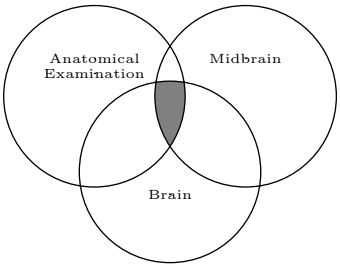
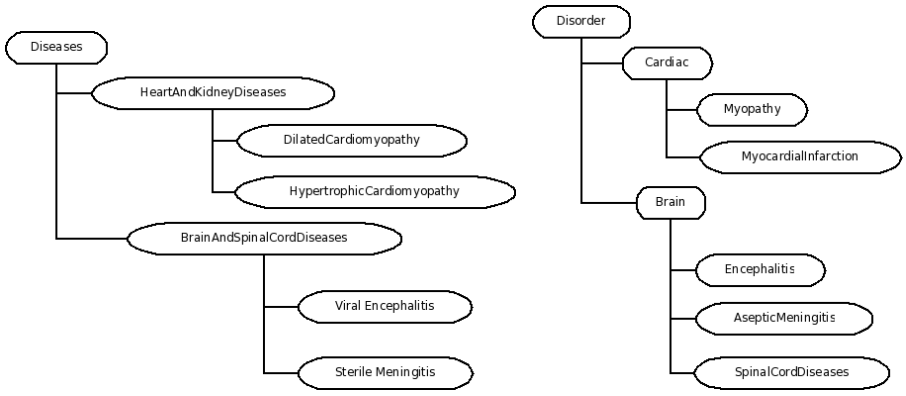


Fig. 5. The meaning of the concept Midbrain after constraining its mean via its relationship to its parents

5 The Modified SMatch Algorithm

Our modified algorithm takes two trees as input and as an output it produces a matrix of relationships between them. Figure 6 shows our input trees for this purpose. The algorithm works on the same 4-step premise as the original algorithm. At the highest level in the first step the label of a node in a tree is converted into a logical formula using specific rules with some filtering performed and concepts being attached via a background resource. Step 2 consists of creating a node formula and applying context in the form of filtering out irrelevant concepts. The third step involves the matching of atomic formulae using a background resource and the fourth and final step is the reasoning process. These will now be discussed in full.



(a) Tree 1 (b) Tree 2

Fig. 6. The two input trees for our matching process

5.1 Step 1 – String to Formula Conversion

This step, as stated previously, works exclusively on the label of a node. Firstly the label of a node is normalized then tokenized. For example we consider the

node with the label "HeartAndKidneyDiseases" from Tree 1; its tokens would be $\{heart, and, kidney, disease\}$. Once this has been completed tokens are converted to *atomic formulae* and a look up in the UMLS is performed to attach Concepts to these. Words which denote preposition, conjunctions etc. . . are ignored i.e. are not annotated with concepts from the UMLS, however they are kept for logical formula conversion. For example if we query the UMLS for the term *heart* we find that heart has a total of 5 concepts from the UMLS. Hence the formula of *heart* has a total of 5 concepts attached to it. The notation for this is $heart\{cui\#5\}$. For the node "HeartAndKidneyDiseases" we would have three atomic formulae for this node which would be $heart\{cui\#5\}, kidney\{cui\#1\}$ and $disease\{cui\#5\}$. As a final step these atomic formulae are joined with boolean connectives ($\sqcup$, $\sqcap$ and $\neg$). The rules for this are that tokens denoting disjunction (e.g. or) are converted to logical disjunction, tokens denoting conjunctive terms (e.g. for) are converted to logical conjunction and terms denoting negation (e.g. not) are converted to negation. Hence the final formula for the node "HeartAndKidneyDiseases" would be $((heart \sqcup renal) \sqcap disease)$. Notice how the term "and", despite being a connective word in the English language has been converted to disjunction, this captures the meaning of the concept. A change to this step is that before the tokenization step we check to see if a concept exists "as is" within the UMLS i.e. there is an existing match in the background resource for the entire label. For example of we consider the node labelled "MyocardialInfarction" from Tree 2, its atomic formula would be $myocardial\ infarction$ with its corresponding Concepts attached from the UMLS. This would be the single atomic formula for the node.

5.2 Step 2 – Context Creation and Filtering

The purpose of this step is to capture the *meaning* of a concept according to its relationships with its parents. The manner in which this is achieved is that the conjunction of the current node to its parent is taken, since in essence the meaning of this node is the intersection with all of its parents. For example for the node named "HyperthrophicCardiomyopathy" in Tree 2 the conjunction of the formula of this node is taken to its parent node would be $((disease) \sqcap ((heart \sqcup kidney)diesease) \sqcap (hypertrophic\ cardiomyopathy))$. This is the *context* of the node with its UMLS concepts attached to each atomic formula. Once the contextual formula has been created we can then perform a filtering of the Concepts, a necessary step in removing unrelated Concepts attached to atomic formulae and hence improving the matching process.

We filter the Concepts based on information we have available from the UMLS. The UMLS itself has three tables for disambiguation: the MRREL (Concept relationships), MRHIER (Atom relationships) and MRCOC (Co-Occurrence relationships). Our concept filtering algorithm is based on the *tree context* of the node, hence its path to the root and all of its children (not be confused with its context which is its path to root). For each concept, c, for an atomic formula a we get all the concepts attached to a in their tree context and we give a score to c if it is related to another concept in its tree context from either the

Concept, Atom or Co-Occurrences relationships to *c*. We currently retain all related concepts i.e. with a score greater than zero and disregard all unrelated concepts. Only one table at a time is used i.e. they are not used in conjunction with each other. This step is beneficial since it is able to drop concepts attached to an atomic formula which have no relationships in its current tree context. It should be noted that if an atomic formula only has one concept we retain that.

5.3 Step 3 – Atomic Formula Matching with the UMLS

Once the above steps are completed, the *concepts*, related to atomic formulae, in the two trees, are matched against each other using the UMLS hierarchy. As stated previously there is more than one source of hierarchical information in the UMLS; these are the MRREL and MRHIER tables. MRREL describes the relationships between concepts, whereas and the MRHIER describes the relationships between atoms.

If we wish to use the Atom hierarchy, the mapping rules are as follows:

- $\equiv$ **rule** - If a Concept from A contains a Concept from B then a $\equiv$ relationship is returned.
- $\sqsubseteq$ **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 returned.
- $\sqsupseteq$ **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 returned.

where A and B are both atomic formulae. If we use the Concept hierarchy of the UMLS, the mapping rules are:

- $=$ **rule** - If a Concept from A contains a Concept from B then a $\equiv$ relationship is returned.
- $\sqsubseteq$ **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 $\sqsubseteq$ relationship is returned.
- $\sqsupseteq$ **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 returned.

where A and B are both atomic formulae. As can be seen there is no $\perp$ test because unlike WordNet the UMLS does not explicitly state antonymy between concepts. We select the first relationship discovered with equivalence taking the highest level of precedence.

Tables 1 and 2 show a subset of results of matching atomic formulae from the two trees. In addition to showing equivalence and subsumption relationships there is also a "NF" relationship, this means that a relationship has not been found between atomic formulae using the UMLS. It is of interest to note that some of the relationships discovered using the differing hierarchies have presented different results e.g. *heart* is $\sqsupseteq$ *brain* in table 2, whereas this relationship does not exist in table 1. Using the atom hierarchy fewer relationships are discovered between atomic formulae, whereas using the concept hierarchy many more relationships are discovered. Whether or not these are correct is, of course another matter altogether.

Table 1. This is the output of step three of the algorithm where the labels are matched using the UMLS **Atom** hierarchy

$Tree_1$ \ $Tree_2$	$disorder$	$cardiac$	$myopathy$	$myocardial\ infarction$	$brain$
$disease$	$\equiv$	NF	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
$heart$	$\sqsubseteq$	$\equiv$	NF	NF	NF
$kidney$	$\sqsubseteq$	NF	NF	NF	NF
$dilated\ cardiomyopathy$	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF
$hypertrophic\ cardiomyopathy$	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF

Table 2. This is the output of step three of the algorithm where the labels are matched using the UMLS **Concept** hierarchy

$Tree_1$ \ $Tree_2$	$disorder$	$cardiac$	$myopathy$	$myocardial\ infarction$	$brain$
$disease$	$\equiv$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
$heart$	$\sqsubseteq$	$\equiv$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
$kidney$	$\sqsubseteq$	NF	NF	NF	NF
$dilated\ cardiomyopathy$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$
$hypertrophic\ cardiomyopathy$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$

5.4 Step 4 – Reasoning

The purpose of this step is to prove a relationship between two nodes in the trees. To achieve this equation 1 is used. If we were trying to prove a less general relationship ($\sqsubseteq$) between two nodes we would try and prove a $\rightarrow$ relationship which is its propositional equivalent and $\leftarrow$ for a more general ($\sqsupseteq$). For equivalence we would prove a $\leftrightarrow$ relationship. Before this is done however, *axioms* (eq 1) or the implicit background theory from the end of the previous step, have to be created. The relationships between atomic formulae discovered from the UMLS are converted to their propositional equivalents. Axioms are only created for the matching task at hand. For example if we were matching the node "HeartAndKidney-Diseases" from Tree 1 in figure 6 to the node "MyocardialInfarction" in Tree 2 the axioms would be $((disease \leftrightarrow disorder) \wedge (heart \leftrightarrow cardiac) \wedge (disease \leftarrow myocardial\ infarction))$ this being derived from table 1. Once this is completed the contexts from the end of the second step for each tree are fed into equation 1, rel is relationship we are trying to prove ($=$,$\sqsubseteq$,$\sqsupseteq$). Equation 3 shows a more general relationship between the two nodes. To prove this relationship we have to show that its negation is not satisfiable (Equation 2). Equation 4 shows the full propositional formula that is fed into a SAT solver before it is converted into its conjunctive normal form. We can see from equation 4 that this is not satisfiable and therefore the more general relationship between these two nodes holds. It is interesting to note that this information was not taken from the UMLS (table 1) but that this relationship was inferred. This step stays the same regardless of which features of the UMLS we use for disambiguation.

$$axioms \rightarrow rel(context_A, context_B) \tag{1}$$

$$axioms \wedge \neg rel(context_A, context_B) \tag{2}$$

$$\begin{aligned}((disease \leftrightarrow disorder) \wedge (heart \leftrightarrow \\ cardiac) \wedge (disease \leftarrow myocardial\ infarction)) \\ \rightarrow ((disease) \wedge ((heart \vee kidney) \wedge disease)) \\ \leftarrow ((disorder) \wedge (cardiac) \wedge \\ (myocardial\ infarction))\end{aligned} \tag{3}$$

$$\begin{aligned}((disease \leftrightarrow disorder) \wedge (heart \leftrightarrow \\ cardiac) \wedge (disease \leftarrow myocardial\ infarction)) \wedge \\ \neg(((disease) \wedge ((heart \vee kidney) \wedge \\ disease)) \leftarrow ((disorder) \wedge (cardiac) \wedge \\ (myocardial\ infarction)))\end{aligned} \tag{4}$$

Table 3. This is the final output of the algorithm when the Atom hierarchies are used for both filtering and matching of the concepts

$Tree_1$ \ $Tree_2$	Disorder	Cardiac	Myopathy	Myocardial Infarction	Brain	Encephalitis	Aseptic Meningitis	Spinal Cord Diseases
Diseases	$\equiv$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
HeartAndKidneyDiseases	$\sqsubseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	NF	NF	NF	NF
DilatedCardiomyopathy	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF	NF	NF	NF
HypertrophicCardiomyopathy	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF	NF	NF	NF
BrainAndSpinalCordDiseases	$\sqsubseteq$	NF	NF	NF	NF	NF	NF	NF
ViralEncephalitis	$\sqsubseteq$	NF	NF	NF	$\sqsubseteq$	$\sqsubseteq$	NF	NF
SterileMeningitis	$\sqsubseteq$	NF	NF	NF	NF	NF	NF	NF

Table 4. This is the output of the algorithm with filtering using the Atom Hierarchy of the UMLS and the Concept Matching of the third step

$Tree_1$ \ $Tree_2$	Disorder	Cardiac	Myopathy	Myocardial Infarction	Brain	Encephalitis	Aseptic Meningitis	Spinal Cord Diseases
Diseases	$\equiv$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
HeartAndKidneyDiseases	$\sqsubseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
DilatedCardiomyopathy	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF
HypertrophicCardiomyopathy	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	NF	NF	NF
BrainAndSpinalCordDiseases	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	NF	$\sqsubseteq$	NF	NF	NF
ViralEncephalitis	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	NF	$\sqsubseteq$	$\sqsubseteq$	NF	NF
SterileMeningitis	$\sqsubseteq$	$\sqsubseteq$	$\sqsubseteq$	NF	$\sqsubseteq$	NF	NF	NF

Tables 3 and 4 show the results of matching the two trees in figure 6, using the Atom and Concept hierarchies for matching atomic formulae. As can be seen the results of using either the concept hierarchy or the atom hierarchy in the third step are different, hence the axioms, or information from a background resource can have a great influence on the final result. Our evaluation focuses on using differing features of the UMLS for the filtering in the second step and the matching of atomic formulae in the third step. This will be discussed further in the next section.

6 Evaluation Methodology

Our evaluation methodology focuses on the accuracy of the differing hierarchical matching and filtering schemes used by our algorithm. We have selected subsets of the FMA and MeSH ontologies for our matching process and have created a reference alignment for our matching task. The subset of the FMA we have chosen is rooted at the concept of Brain, and follows the "regional_part_of" relationship to its leaf nodes, this contains 476 concepts in total. The subset of MeSH we have chosen is rooted at the concept of brain as well and traverses down to its leaf nodes, this contains 181 concepts in total. Both of these subsets cover the same domain, as they describe anatomical relationships with the organ "Brain" however, their level of coverage, naming of terms and models are different. All the matches are from the FMA to MeSH, i.e the FMA is the source ontology and MeSH is the target. We have used the 2008AA version of the UMLS in all of our experiments. The SMatch algorithm has been reimplemented from scratch with our modifications incorporated and we have used the SAT4J[4] library as our SAT solver.

For the evaluation of our algorithm we have chosen to use different sources for the filtering of concepts (Concepts, Atoms or Co-Occurrences, none) and for the hierarchical matching as well (Concepts or Atoms). We created our reference alignment with the aid of a medical expert, our process for creating our standard has been to select 20 random concepts from our FMA subset and 40 random concepts from our MeSH subset and manually derive a match in the semantic matching range.

7 Results

The following versions of the algorithm were run:

1. **Concept_NoFilter:** This uses the concept hierarchy for the matching of atomic formulae with no filtering of concepts.
2. **Atom_NoFilter:** This uses the atom hierarchy for the matching of atomic formulae with no filtering of concepts.
3. **Concept_ConceptFilter:** This uses the concept hierarchy of the UMLS for atomic formulae matching and the concept hierarchy for the filtering of concepts.

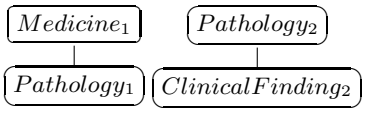
[4] http://www.sat4j.org

4. **Atom_AtomFilter:** This uses the Atom hierarchy of the UMLS for matching atomic formulae and the Atom hierarchy for filtering.
5. **Concept_AtomFilter:** This uses the Concept hierarchy for the matching of atomic formulae and the Atom hierarchy for the filtering of concepts.
6. **Atom_ConceptFilter:** This uses the Atom hierarchy of the UMLS for the matching of atomic formulae and the Concept hierarchy for the filtering of Concepts.
7. **Concept_COCFilter:** This uses the Concept hierarchy of the UMLS for the matching of atomic formulae and the Co-Occurrence relationships for filtering.
8. **Atom_COCFilter:** This uses the Atom hierarchy of the UMLS for the matching of atomic formulae and the Co-Occurrence relationships for filtering.

Table 5. The number of matches found for each version of our algorithm

Test	No. Equivalences	No. Less General	No. More General	No. Not Found
Concept_NoFilter	65	4069	937	81085
Atom_NoFilter	65	2581	726	82784
Concept_ConceptFilter	65	4069	937	81085
Atom_AtomFilter	65	2587	726	82778
Concept_AtomFilter	65	4069	937	81085
Atom_ConceptFilter	65	2587	726	82778
Concept_COCFilter	65	4069	935	81087
Atom_COCFilter	65	2581	726	82784

Table 5 shows the total number of matches found for each matching task. Overall using the Concept hierarchy for the matching of atomic formulae in the third step resulted in many more Less General and More General relationships being discovered than using the Atom hierarchy; this is because the Concept hierarchy is more expressive than the Atom hierarchy of the UMLS. The interesting thing to note is that the filtering using the Co-Occurrence relationships within the UMLS resulted in a slightly reduced number of alignments being classified as a less general relationship in the Atom hierarchy tests. Also these results are equivalent to no filtering of Concepts attached to atomic formulae. When using the UMLS Concept hierarchy for the matching of atomic formulae there seems to be no significant difference in the number of results regardless of filtering scheme employed. Our precision and recall values are identical for each version of the algorithm (both at 1), we think this is a fault with our reference alignment where there were many "NF" relationships discovered during the manual matching task by our clinical expert. Also the reference alignment did not contain any of the differing results discovered (for example the extra less general relationships discovered by the Concept matching of atomic formulae). However, with this preliminary evaluation the values are still significantly high for both precision and recall which is promising once a larger scale evaluation has been conducted.



(a) Tree 1 (b) Tree 2

Fig. 7. Two trees to show where filtering can dramatically affect the results

The filtering however is useful although this is just a special case, if we consider the example in figure 7. If there was no filtering scheme involved at the hierarchical level we would have the following relationships discovered:

- $Medicine_1 \sqsupseteq Pathology_2$
- $Medicine_1 \sqsupseteq ClinicalFinding_2$
- $Pathology_1 \equiv Pathology_2$
- $Pathology_1 \sqsupseteq ClinicalFinding_2$

However, if we use Atom filtering we would have the following relationships discovered:

- $Medicine_1 NF Pathology_2$
- $Medicine_1 NF ClinicalFinding_2$
- $Pathology_1 NF Pathology_2$
- $Pathology_1 NF ClinicalFinding_2$

This is because when we query the UMLS for the terms in Tree 1 which are "Medicine" we get the following Concepts returned:

- *C0013227:Pharmaceutical Preparations*
- *C0025118:Medicine*

For "Pathology" we have the following:

- *C0677042:Pathology processes*
- *C0334866:Medical pathologist*
- *C0205469:Pathological aspects*
- *C0919386:Pathology procedure*
- *C0030664:Pathology*

The filtering scheme retains the concept of *0025118:Medicine* and drops *C0013227:Pharmaceutical Preparations* since this concept is not related to any of the concepts of Pathology. For the node labelled "Pathology" the concept of *C0030664:Medical pathologist* is kept since it is the only concept related to any of the concepts of "Medicine". The meaning of this tree has now been constrained and given context since each node only has these concepts attached.

An identical process occurs between the nodes in tree 2 where "ClinicalFinding", only has one concept attached so it is kept which is *C0037088:Signs and Symptoms* and the node "Pathology" has the same concepts attached as

above, however, now the concept that is retained during the filtering process is *C0677042:Pathology processes*. Therefore the atomic formula for the node "Pathology" in tree 2 now has a different concept attached when compared to the node "Pathology" in tree 1. According to our rules in step 3 there is now no relationship discovered, hence the axioms are empty and no relationship can be deduced during the reasoning process. This is an ideal case where the filtering of concepts is useful, however when matching our subsets of FMA to MeSH they both cover the same domain hence the filtering (although dropping irrelevant concepts) has not been effective in the number of results discovered since identical concepts were dropped in both trees. In our future work we will match two highly unrelated trees to gauge the effectiveness of our filtering scheme.

8 Conclusion

In this work we have shown how to apply the UMLS to a semantic matching task by modifying the SMatch algorithm to its use. For this we have detailed the steps involved and how the UMLS can be used. We have found by the number of results shown that the filtering has a minimal effect on the number of results discovered, however, in some situations this is not the case. Overall the accuracy is our method is high, this is evidenced by the high precision values of our algorithm and recall is high as well (both are perfect at 1), which is to be expected on such a small reference alignment. Currently our results are preliminary, our future work will consist of a further evaluation of our approach on a much larger dataset with a complete reference alignment. This however is difficult since there are no commonly agreed upon reference alignments or datasets for semantic matching in the medical domain. We intend to submit our matching system to the OAEI [13] medical track to aid our evaluation. We will also test our algorithm on different datasets with varying degrees of overlap.

The significance of this work is that it is possible to conduct semantic matching in the medical domain, we have shown this with the use of one background resource which is the UMLS. This is beneficial because the expressivity of semantic matching is able to increase the number of alignments discovered between elements, hence increasing the number of potential matches between concepts. We will later investigate how our approach can apply to differing sources of medical knowledge and outline the steps to achieve this.

References

1. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
2. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic Matching: Algorithms and Implementation. Journal of Data Semantics, 1–38 (2007)
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
4. Shamdasani, J., Bloodsworth, P., McClatchey, R.: Semantic Matching for the Medical Domain. In: Gray, A., Jeffery, K., Shao, J. (eds.) BNCOD 2008. LNCS, vol. 5071, pp. 198–202. Springer, Heidelberg (2008)

5. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research 32 (2004)
6. Tan, H., Jakoniene, V., Lambrix, P., Aberg, J., Shahmehri, N.: Alignment of Biomedical Ontologies Using Life Science Literature. In: Bremer, E.G., Hakenberg, J., Han, E.-H.(S.), Berrar, D., Dubitzky, W. (eds.) KDLL 2006. LNCS (LNBI), vol. 3886, pp. 1–17. Springer, Heidelberg (2006)
7. Do, H., Rahm, E.: COMA a system for flexible combination of schema matching approaches. In: 28th International Conference of Very Large Databases (2002)
8. Risto, G., Aleksovski, Z., ten Kate, W., van Harmelen, F.: Using Google Distance to Weight Approximate Ontology Matches. In: BNACI 2007 (2007)
9. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching Unstructured Vocabularies using a Background Ontology. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS, vol. 4248, pp. 182–197. Springer, Heidelberg (2006)
10. Aleksovski, Z., ten Kate, W., van Harmelen, F.: Exploiting the structure of background knowledge used in ontology matching. In: Ontology Matching Workshop (ISWC 2006) (2006)
11. Sabou, M., d'Aquin, M., Motta, E.: Exploring the Semantic Web as Background Knowledge for Ontology Matching. International Journal of Data Semantics (2008)
12. Garcia, J., Lopez, V., d'Aquin, M., Sabou, M., Motta, E., Mena, E.: Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching. In: International Workshop on Ontology Matching (OM 2007) (2007)
13. Caracciolo, C., Euzenat, J., Hollink, L., Ichise, R., Isaac, A., Malais, V., Meilicke, C., Pane, J., Shvaiko, P., Stuckenschmidt, H., Svab-Zamazal, O., Svatek, V.: Results of the Ontology Alignment Evaluation Initiative 2008. In: Ontology Matching Workshop (ISWC 2008) (2008)