MedMatch - Towards Domain Specific Semantic Matching

Jetendr Shamdasani, Peter Bloodsworth, Kamran Munir, Hanene Boussi Rahmouni, and Richard McClatchey

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

Abstract. Ontologies are increasingly being used to address the problems of heterogeneous data sources. This has in turn often led to the challenge of heterogeneity between ontologies themselves. Semantic Matching has been seen as a potential solution to resolving ambiguities between ontologies . Whilst generic algorithms have proved successful in fields with little domain specific terminology, they have often struggled to be accurate in areas such as medicine which have their own highly specialised terminology. The MedMatch algorithm was initially created to apply semantic matching in the medical domain through the use of a domain specific background resource. This paper compares a domain specific algorithm (MedMatch) against a generic (S-Match) matching technique, before considering if MedMatch can be tailored to work with other background resources. It is concluded that this is possible, raising the prospect of domain specific semantic matching in the future.

1 Introduction

Heterogeneity between data sources has been seen as a significant barrier to the exchange and use of information. Whilst ontologies have shown the potential to address some of these issues, they have often moved the problem up a level thus leading to heterogeneity between ontologies themselves. Semantic Matching has shown some promising results in terms of resolving ambiguities between ontologies [1]. Generic algorithms whilst appearing to have been applied successfully in fields with little domain specific terminology, often perform less accurately in areas such as medicine which have their own specialised terminology and use of language. MedMatch was initially created to apply semantic matching in the medical domain through the use of a domain specific background resource. This paper briefly demonstrates the potential benefits of domain specific semantic matching and seeks to understand whether the MedMatch approach can be extended to other domains which have similar attributes. A number of domains also require a specialised terminology, examples of these include Law, Physics and the Biological Sciences. The medical domain is therefore not unique in its requirement for semantics.

Currently the MedMatch algorithm has been implemented to use the UMLS as a domain specific background resource. In order to understand if it is indeed possible and if so what is required to change from one background resource to another it is necessary to select a second medical source. The Foundational Model of Anatomy ontology (FMA)was chosen for use as a test case. It was chosen because it appeared to have adequate coverage, semantics, granularity and metainformation and its model is very different from the UMLS itself. To replace the UMLS with the FMA is clearly a non-trivial task. This is primarily due to the fact that the structure of the FMA is different to the UMLS since they were designed for different purposes. The UMLS has been designed to function as a thesaurus for the domain, whereas the FMA is designed to model human anatomy. By considering the issues associated with changing background resources in this case, we aim to reach some initial conclusions regarding the expansion of MedMatch into a domain specific semantic matching framework in which new background resources can be harnessed.

The next section considers a number of terminology specific fields and draws conclusions regarding the emerging need for domain specific semantic matching in the future. Following this we briefly consider the related work in this area and then carry out a comparison between generic and domain specific semantic matching methods in order to determine the benefits that such approaches can deliver. The concluding sections describe a criteria by which background knowledge sources can be assessed and it is then used to evaluate the FMA. An analysis of the changes that would be needed to the MedMatch algorithm is carried out and final conclusions are reached.

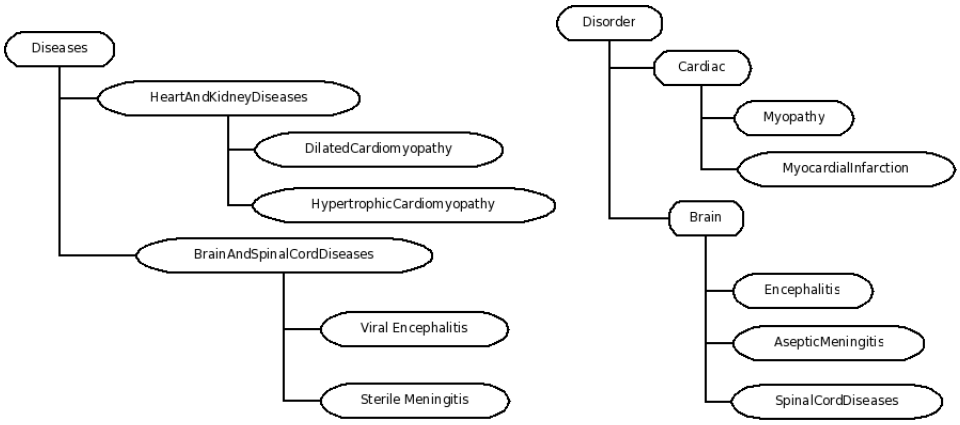
2 Domain Specific Terminology and Related Work

Many fields require a specialised terminology, examples include Law, Physics and the Biological Sciences. Creating an ontology that covers the entire legal domain is a very challenging task. Some legal ontologies appear to be well developed. These however, are often "core" ontologies which cover the most important aspects of law such as the Legal Knowledge Interchange Format (LKIF)[2]. Such resources are mainly used as a basis for creating more specific ontologies such as ones related to criminal law or other sub-domains. A number of projects are currently working on legal ontologies, at present however none of them cover the legal domain sufficiently to provide a background resource for MedMatch.

The integration of semantics within Physics experiments is developing but is still in its infancy. Once again this leads to the conclusion that an appropriate background resource for Physics is not currently available. It appears likely that this situation will improve in the medium term as existing projects mature and make their results available. In the Biological Science domain there have been attempts to classify species into categories since the early days of the field. These have been done by using what are known as biological taxonomies, or taxa for short. There are many systems that these taxa follow, the most popular of which

is the Linnaean system [1] . After some investigation it was discovered that some ontologies do exist for this domain, however, they are top-level ontologies such as BioTop [3] which are not yet sufficiently granular for the semantic matching process.

Domain specific ontologies are becoming more widely adopted and comprehensive. It is likely therefore that in the medium term a number of candidates for use as background knowledge with the algorithm will appear. The medical domain was one of the first to embrace the use of ontologies and as such, may perhaps be seen as an early indicator of how resources will develop in other domains over the next few years. It would appear likely that UMLS-like resources may well be developed. This would suggest that the need for domain specific semantic matching techniques will grow in the near future.



(a) $Ontology_1$ (b) $Ontology_2$

Fig. 1: Example Ontologies

There are many ontology alignment techniques available today [4] and research has been active in this area for quite some time mostly due to the emergent semantic web [5]. These techniques range from simple string comparisons [6] to complex machine learning approaches [7]. There have been some approaches that have used background knowledge as a dictionary [8] to more expressive techniques, for example text mining [9] or search engine distances [10]. The majority of these systems output an equivalence relationship between feature pairs with a confidence value. The focus of this work is on a subset of the ontology alignment problem known as *Semantic Matching* [11] where relationships besides equivalence are discovered between concepts in two ontologies. For example matching the concepts "BrainAndSpinalCordDiseases" from $Ontology_1$ in figure

[1] http://en.wikipedia.org/wiki/Linnaean_taxonomy

1a to "Brain" from $Ontology_2$ from figure 1b using a standard string matching approach the result of this match would be an equivalence relationship with a confidence value. In the case of semantic matching, however, the result would be that "BrainAndSpinalCordDiseases" is more general than "Brain". The purpose of the MedMatch system was to show a method of conducting semantic matching in the medical domain. The following section compares the domain specific MedMatch algorithm against the generic SMatch system in order for us to better understand the potential benefit that using domain specific background resources can deliver.

3 SMatch Comparison

A freely available version of SMatch [2] was compared with MedMatch using a set of trees which were chosen to examine some common types of alignment contexts. An automated method was used to create reference alignments. This was validated by a clinician in order to confirm that the references were accurate. It should be noted that whilst this version of SMatch was able to discover disjoint correspondences between concepts. Such mappings have been omitted from the final matching results as MedMatch presently does not discover such relationships. Table 1 shows the results of the original SMatch algorithm in comparison to the results from each of the matching tasks that were produced by MedMatch. The new algorithm performed *better* overall than the original SMatch algorithm. This was demonstrated by the higher values for f-measure in all of the tasks. Its precision was also higher than the SMatch algorithm in all of the matching tasks.

Algorithm	Expected	Overlap	Total Found	Precision	Recall	F-Measure
Test 1 – Similar Subdomains						
MedMatch	757	703	779	0.9	0.93	0.92
SMatch	757	562	763	0.736	0.742	0.739
Test 2 – Structure						
MedMatch	406	406	433	0.94	1	0.97
SMatch	406	404	532	0.759	0.995	0.861
Test 3 – Granularity						
MedMatch	1069	1003	1075	0.933	0.938	0.936
SMatch	1069	858	1149	0.746	0.802	0.773

Table 1: SMatch Results Compared To MedMatch

In relation to the first test, which was the Similar Subdomains test, the SMatch algorithm performed significantly worse than the new algorithm. The values for precision were 0.736, 0.742 recall and 0.739 f-measure. The modified algorithm produced a 0.9 precision, 0.93 recall and 0.92 f-measure. In the second experiment which was the Structure test, the new algorithm obtained a much higher value for f-measure (0.97) than the original SMatch (0.733). The reason

[2] http://semanticmatching.org/s-match.html

for this is that the current algorithm scored higher values for precision 0.94, in comparison to 0.759 and recall was also slightly higher at 1.0, when compared to 0.995. The third test, looked at the effect that two trees with different levels of granularity had on the algorithm. SMatch performed worse in terms of f-measure than the modified algorithm once again. The f-measure value for SMatch was 0.773 and this work achieved a f-measure of 0.936. This difference in f-measure is mainly because the SMatch algorithm achieved lower precision (0.746 vs 0.933) and lower recall (0.802 vs 0.938).

This experiment has indicated that MedMatch can outperform the original SMatch algorithm when matching trees which contain medical terminology. This is because MedMatch had consistently higher f-measure values than SMatch. This metric is widely seen as a good indicator of the relative performance of alignment techniques, since it considers both the metrics of precision and recall. The significance of these results is that it shows that when matching medical trees containing complex structures and terms, this version of the algorithm performs better than the SMatch algorithm. It is also of interest to note that this version of the algorithm only has a single means of matching atomic formulae, whereas the SMatch algorithm has more than just a WordNet matcher. This confirms the need for domain specific semantic matching. The following section considers what is required in terms of a background resource by the MedMatch algorithm.

4 Background Knowledge Requirements

In order for a background resource to be useable as a source of knowledge by the MedMatch algorithm, it should address the criteria of *coverage*, *semantics*, *granularity* and *metainformation* to a satisfactory degree. This means that the background resource should contain a class hierarchy from which semantics can be extracted and synonym terms to feed the semantic matching process. It is also difficult for the algorithm to adapt to sources which have a sparse number of classes since they are often incomplete and do not always contain sufficient *coverage* for the matching of ontologies in a particular domain. This means that the class hierarchy needs to cover as much of the domain of the input ontologies as possible. The class hierarchy that is present must contain *semantics* between the concepts in the class hierarchy. These need to be at the very least "broader than" and "narrower than" relationships between concepts in the chosen resource. This is so that subsumption relationships can be present between anchor points for the algorithm. Synonym relationships also need to be present so that equivalence can be determined between concepts from the input ontologies.

Granularity relates to the level of detail that the chosen background resource contains. This requirement is again tied to the number of concepts present in the background resource. There have to be anchor points present for at least some of the concepts from the input ontologies so that a relationship can exist for the final reasoning process. The more information there is present from a trusted source of background knowledge the more relationships can be induced by the final reasoning process. For the criteria of *metainformation* there must

be a string-to-concept mapping for the anchoring process to be successful. This means that labels of concepts in the input ontologies must exist in the source of background knowledge, or alternatively there must be a mapping between labels which describe the input concepts and the representation of the background knowledge source, such as is the case with the UMLS.

5 FMA evaluation

The purpose of this section is to understand what is necessary in order for other domain specific ontologies to be "plugged in" to the algorithm as background resources. This is seen as an initial step in the future generalisation of the algorithm so that it can perhaps perform semantic matching in domains other than medicine. The FMA will be assessed using the criteria that was previously described.

- **Coverage** The purpose of the FMA is to represent the anatomy of human beings. If only anatomical ontologies were to be matched, the coverage of the FMA is reasonably good. This is because the inputs to the algorithm are trees which cover the domain of anatomy. The FMA is considered to be one of the best resources that deals with the domain of anatomy. It is a very large and complete ontology describing the domain of anatomy.
- **Semantics** In the area of semantics the FMA is very expressive. It is constructed in F-Logic (Frames) and therefore it supports classes, properties and instances. The design of the FMA is based on the lateral position of parts in the human body. For example, "left arm" and "right arm" are different concepts which are both subclasses of the concept arm. The FMA also contains some synonyms for concepts with the "synonym" attribute containing alternate concept names. The FMA only contains a single source of hierarchical information for the semantic matching process.
- **Granularity** The FMA is highly granular with a high depth level for the anatomical concepts it describes. The purpose of the FMA is to describe human anatomy in great detail, therefore, it has many concepts to a high depth level.
- **Meta-information** The FMA is created in Frames using the Protege tool. Protege provides an API which provides access to the different features of the FMA such as synonyms within concepts and the FMA hierarchy.

6 Algorithm Changes Required for Use with the FMA

The MedMatch algorithm has four major steps. These are 1) String to formula conversion, 2) Context creation and Filtering 3) Atomic formula matching and 4) Reasoning. The model of the FMA is different from the UMLS, the FMA contains more specific relationships between concepts where as within the UMLS relationships taken from a source vocabulary are abstracted into higher level relationships such as "PAR" for a general parent relationship. Changes need to

be made to the algorithm to accommodate the new information that the FMA provides. This includes how concepts are organised within the FMA model such as how to extract synonyms and how to extract hierarchical information from the FMA itself. In this section the changes that are necessitated by the use of a different resource will be discussed as well as how these can be addressed. The fourth and final reasoning step requires no change since the axiom creation and reasoning scheme is identical across resources and therefore it shall not be discussed further.

6.1 Step 1 - String to Formula Conversion

The sub-steps for this part of the algorithm remain unchanged. There are multi-word concepts present in the UMLS as well, therefore the rule used in the first step for preferring multi-word concepts applies in this case as well. One important note on the anchoring scheme is that when searching through the FMA terms the synonym fields of concepts as well as the "Non English-Equivalent" field should be taken into account. This field contains the Latin equivalent of common anatomical terms such as Encephalon, which is a common synonym for Brain.

6.2 Step 2 - Context Creation and Filtering

Context is given to a node by using the background resource. This context was achieved by taking the logical formulae from the previous step then taking a conjunction from the formula of the current node to all the formulae leading to the root node. When using the FMA the context creation process is nearly identical i.e. the conjunction from the current node is still taken to its parent nodes. Where there are only single string to concept relationships present, a new filtering algorithm is required which can take into account the predicates present in the FMA to provide concepts with a more precise context.

6.3 Step 3 - Atomic Formula Matching

The UMLS hierarchies were used to match concepts attached to atomic formulae. The basic principle remains the same for using the FMA as a background resource. Semantics present in the background resource are used for the semantic matching process since these are mapped onto their propositional equivalents for the final reasoning process. These initial relationships form the background theory to seed the reasoning process from which other relationships not present in a background resource can be extracted. The rules for matching concepts attached to atomic formulae for both the UMLS hierarchies and the FMA are similar. The semantics present in both these resources are used for the creation of these rules. This would suggest that MedMatch could be modified to use the FMA as its background resource will relatively little changes being necessary to the algorithm itself. We can therefore conclude that MedMatch may well be generalisable and could be used in the future to create domain specific semantic matching systems.

7 Conclusion

MedMatch was initially created to apply semantic matching in the medical domain through the use of a domain specific background resource. This paper has shown that domain specific semantic matching can be beneficial in fields which have a specialised terminology. It has also been demonstrated that the MedMatch approach is capable of being extended to other domains which have similar attributes. These include Law, Physics and the Biological Sciences. At present the MedMatch algorithm has been implemented to use the UMLS as a domain specific background resource. In order to understand if it was possible and what was required to change from one background resource to another the Foundational Model of Anatomy ontology (FMA)was chosen for use as a test case. It was selected because it appeared to have adequate coverage, semantics, coverage and metainformation and its model is very different from the UMLS itself. The rules for matching concepts attached to atomic formulae for both the UMLS hierarchies and the FMA were found to be very similar. The semantics present in both of these resources was used to creation of these rules. This suggests that MedMatch can be modified to use the FMA as its background resource with relatively little changes being necessary to the algorithm itself. We can therefore conclude that MedMatch may well be generalisable and could be used in the future to create domain specific semantic matching systems.

References

1. F. Giunchiglia and P. Shvaiko. Semantic Matching. *The Knowledge Engineering Review*, 18(3):265–280, 2003.
2. A. Boer et al. Computable models of the law. chapter MetaLex XML and the Legal Knowledge Interchange Format, pages 21–41. Springer-Verlag, Berlin, Heidelberg, 2008.
3. E. Beisswanger et al. BioTop: An upper domain ontology for the life sciences: A description of its current structure, contents and interfaces to OBO ontologies. *Appl. Ontol.*, 3:205–212, December 2008.
4. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer Verlag, 2007.
5. T. Berners-Lee et al. The Semantic Web. *Scientific American*, 284(5):35–43, May 2001.
6. G. Stoilos et al. A String Metric for Ontology Alignment. *ISWC*, pages 624–637, 2005.
7. A. Doan et al. Ontology Matching: A Machine Learning Approach. *Handbook on Ontologies in Information Systems*, pages 385–403, 2004.
8. J. Madhavan et al. Generic schema matching using cupid. In *27th International Conference on VLDB*, pages 49–58, 2001.
9. H. Tan et al. Alignment of Biomedical Ontologies Using Life Science Literature. *KDLL*, pages 1–17, 2006.
10. G. Risto et al. Using Google Distance to Weight Approximate Ontology Matches. In *BNACI '07*, 2007.
11. F. Giunchiglia et al. Semantic Matching: Algorithms and Implementation. *Journal on Data Semantics*, 9:1–38, 2007.