Semantic Matching for the Medical Domain

Jetendr Shamdasani*, Peter Bloodsworth, and Richard McClatchey

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

Abstract. This paper proposes some modifications to the SMatch algorithm that enables the semantic matching of medical terminologies using the Unified Medical Language System (UMLS) as a source of background knowledge. Semantic Matching is the process of discovering set theoretic relationships between differing data elements. Initial results from the domain of anatomy are presented that illustrate how semantic relationships can provide greater information during the ontology alignment process than equivalence relationships alone. The paper concludes by demonstrating how this is beneficial in the medical domain.

1 Introduction and Previous Work

Semantic Matching is the process of discovering set theoretic based relationships between differing concepts within two schemas or ontologies. The power of this approach is that it is able to identify a range of expressive relationships between concepts, in particular *less general* ($\sqsubseteq$), *more general* ($\sqsupseteq$) and *disjointness* ($\perp$) relations in addition to standard equivalence ($=$). In this paper we present a modification to the SMatch [1] system to make it applicable in the medical domain. The conventional SMatch method relies heavily on the use of the WordNet (WN) thesaurus [2]. The problem with using this resource is that it is too general, with an insufficient amount of medical terminology. This often leads to few meaningful relationships being returned when two medical ontologies are matched. A source of medical terminology is therefore required to drive the alignment process in a medical context. The UMLS [3] has been chosen for this purpose because of its wide coverage of clinical terms. We have modified the SMatch method to use the UMLS during the matching of medical ontologies. A prototype has been created and experimentation comparing the results from our extension to the output from the original SMatch system in a medical context has yielded promising results.

Research in the field of schema and ontology matching is active and several different approaches have been proposed. The work contained in [4] gives a more detailed review of the current research in this area. Semantic Matching involves the use of a structured background resource to extract matches between concepts in differing ontologies. Previous work has been conducted with

* Corresponding Author: This work has been funded by the Health-e-Child project (IST 2004-027749) Thanks to Tamas Hauer, Dmitry Rogulin and Andrew Branson

use of a single background resource [5] or many background sources [6]. These approaches rely firstly on discovering terms in a background source, which they call "anchors" and then performing inferencing by using the knowledge gathered from the background knowledge by differing anchoring methods. Obviously this anchoring step relies on there being a certain amount of lexical overlap between the sources to be matched and the background resource. Our work concerns the modification of the SMatch approach by Ginunchiglia et al [1]. Section 2 presents our modifications to the SMatch system to incorporate the UMLS.

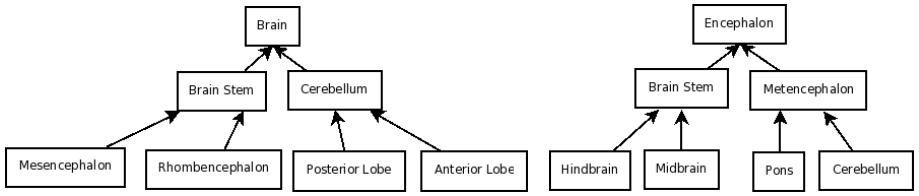
2 SMatch applied to the medical domain

WN is primarily a *lexical resource* about the English language whereas the UMLS is a *conceptual resource* about the medical domain. WN contains information about the terms in the English language. Each term has a set of senses which denote a meaning of a term. The UMLS however is a collection of many different medical ontologies. Every concept has a CUI (Concept Unique Identifier) every CUI has high level relationships linking concepts. These are PAR (Parent), CHD (Child), RB (Broader Than), RN (Narrower Than), SIB (Sibling), RO (Other), RL (Similar), SY (Source Asserted Synonymy), RQ (Possible Synonymy). Every CUI is also annotated with a top level semantic type which is a high level categorization of the CUI. A full discussion of the UMLS is beyond the scope of this paper please see [3].

SMatch takes as input two trees and outputs a set of semantic relationships between concepts. Please see [1] for a detailed discussion of SMatch. In step one of their algorithm a label of a single node is taken and converted to an atomic formula in the Description Logics (DL) sense. This string is firstly preprocessed using normalisation and tokenization to be split into its corresponding parts. Individual tokens are looked up in WN and the corresponding senses are attached to create an atomic formula. Words in the English language denoting prepositions and conjunctions are ignored and are then converted to form logical connectives. These atomic formulae are then converted to DL based formulae. For example the string "Brain Stem" would be converted to $brain \sqcap stem$. A filter is applied according the relationships in WN to remove irrelevant senses. We firstly look up the whole label to see if a term does exist in the UMLS. If this is not so, then we search for tokens then we attach *concepts* (CUIs) from the UMLS. We also filter according to the semantic types of CUIs and disregard any CUIs which do not have the same semantic type. In step two a conjunction of the logical formulae to the root node is taken from a single node. There is also structural sense filtering performed, however this has not been implemented. Hence for the node in tree 1 which is labelled "Rhombencephalon", the formula for this node to its root would be $(rhobenchepalon \sqcap brain\ stem \sqcap brain)$. Each of these formulae would have corresponding CUIs attached.

In step three a variation of their WN matcher has been implemented which is the only matcher that is able to derive semantic relationships for this step. There is a mapping from the higher level relationships between CUIs and the semantic relationships which can be derived. The mappings are the following:

(=) **Rule** - If A is connected via SY relationship to B or if A and B share the same CUI. ($\sqsupseteq$) **Rule** - If a CUI of A is a PAR or RB of a CUI of B. ($\sqsubseteq$) **Rule** - If a CUI of A CHD or a RN of B. ($\perp$) **Rule** - If a CUI of A is a SIB of B. At the end of this step a table of relationships is returned between concepts. If no relationship is found then a *null* relation is returned. In the fourth and final step we have kept their propositional reasoning approach to discover semantic relationships between different nodes. For an explanation of this step please see [1].



(a) $Tree_1$ (b) $Tree_2$

Fig. 1: The two input trees for our results comparison

3 Results and Conclusion

Our preliminary evaluation is a comparison of our approach with our own implementation of the original SMatch approach. We have used the 2.0 version of WN and 2007AB version of the UMLS. Figure 1 shows our two tree inputs which are differing conceptualisations for the parts of the brain. $Tree_1$ (1a) and $Tree_2$ (1b) contain synonyms for medical terms as well as disjointness relationships between each other for this matching task. There are more and less general relationships present as well. The top half of table 1 shows the results from using the traditional SMatch approach using WN as a source of background knowledge and the bottom half of table 1 shows the results from our approach using the UMLS as a source of background knowledge. The *null* relationship states that there was no match found between the concepts.

The most interesting result is that the WN approach has not been able to discover disjointness ($\perp$) relationships between concepts at all. Although these terms do occur in the WN thesaurus, this is due to no appropriate relationships being present between the senses for these strings (antonymy) in the WN thesaurus. However the UMLS does not explicitly state antonymy between concepts therefore this is an interesting result. Several of the results generated by the pure SMatch approach are incorrect, for example Cerebellum is not = to Encephalon instead a $\sqsubseteq$ relationship should have been returned. Our approach does return this relationship correctly; this is also true with many of the other results in table 1. The SMatch approach was able to match Encephalon to Brain correctly as they are synonymous with each other, as did our approach. But these are very general terms in the English language, for example another synonym for Brain in WN is Einstein which is incorrect for the medical domain. This clearly demonstrates that WN is a good source of lexical knowledge but not conceptual

domain knowledge which is required in the medical world. The *null* relation does occur in our approach, this is mostly because this relationship could not be found using the UMLS Metathesaurus, as the UMLS grows our approach will yield more promising results. We also found that our predicted result for this test was identical to the results presented in table 1. The results have been verified by an expert in the medical domain and he was of the opinion that our approach was correct.

In this paper we have presented a modification of the original SMatch system for use in the medical domain. We have also shown that replacing a more general source of background knowledge with a more specific resource yields greater results. For our further work we are going to extend the SMatch algorithm to utilize differing forms of background knowledge which may yield interesting results. Differing reasoning schemes will also be investigated. An extensive evaluation against real world medical ontologies will be conducted following this.

$Tree_1$ \ $Tree_2$	Encephalon	Metencephalon	Brain Stem	Midbrain	Hindbrain	Pons	Cerebellum
Brain	$=$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
Brain Stem	$\sqsubseteq$	*null*	$=$	$=$	$=$	*null*	*null*
Mesencephalon	$\sqsubseteq$	*null*	$=$	$=$	$=$	*null*	*null*
Rhombencephalon	$\sqsubseteq$	*null*	$=$	$=$	$=$	*null*	$\sqsubseteq$
Cerebellum	$=$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$=$
Posterior Lobe	$\sqsubseteq$	*null*	*null*	*null*	*null*	*null*	$\sqsupseteq$
Anterior Lobe	$\sqsubseteq$	*null*	*null*	*null*	*null*	*null*	$\sqsubseteq$
Brain	$=$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$	$\sqsupseteq$
Brain Stem	$\sqsubseteq$	*null*	$=$	$\sqsupseteq$	$\sqsupseteq$	*null*	*null*
Mesencephalon	$\sqsubseteq$	$\bot$	$\sqsupseteq$	$=$	$\sqsubseteq$	$\bot$	$\bot$
Rhombencephalon	$\sqsubseteq$	*null*	$\sqsubseteq$	$\sqsupseteq$	$=$	*null*	*null*
Cerebellum	$\sqsubseteq$	$\sqsubseteq$	*null*	$\bot$	*null*	$\bot$	$=$
Posterior Lobe	$\sqsubseteq$	$\sqsubseteq$	*null*	$\bot$	*null*	$\bot$	$\sqsubseteq$
Anterior Lobe	$\sqsubseteq$	$\sqsubseteq$	*null*	$\bot$	*null*	$\bot$	$\sqsubseteq$

Table 1: These are the comparison of our results against the SMatch approach. The top half of the table shows the results from SMatch and the bottom half shows the results using the UMLS.

References

1. F. Giunchiglia et al. Semantic Matching: Algorithms and Implementation. *Journal of Data Semantics*, pages 1–38, 2007.
2. C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
3. O. Brodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32, 2004.
4. J. Euzenat et al. State of the Art on Ontology Alignment. *Knowledge Web Deliverable*, No. 2.2.3, 2004.
5. Z. Aleksovski et al. Matching Unstructured Vocabularies using a Background Ontology. In *EKAW*. Springer-Verlag, 2006.
6. M. Sabou et al. Using the Semantic Web as Background Knowledge for Ontology Mapping. In *International Workshop on Ontology Matching (OM-2006)*, 2006.