UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

TOWARDS EXPLAINING SEMANTIC MATCHING

Deborah L. McGuinness, Pavel Shvaiko, Fausto Giunchiglia,
and Paulo Pinheiro da Silva

February 2004

Technical Report # DIT-04-019

Also: short version in Proceedings of the International Workshop on Description Logics, 2004.

Towards Explaining Semantic Matching

Deborah L. McGuinness[1] Pavel Shvaiko[2] Fausto Giunchiglia[2]
Paulo Pinheiro da Silva[1]

[1]Stanford University, Stanford, USA
{dlm,pp}@ksl.stanford.edu.
[2]University of Trento, Povo, Trento, Italy
{pavel,fausto}@dit.unitn.it

Abstract

Interoperability among systems using different term vocabularies requires some mapping between terms in the vocabularies. Matching applications generate such mappings. When the matching process utilizes term meaning (instead of simply relying on syntax), we refer to the process as semantic matching. If users are to use the results of matching applications, they need information about the mappings. They need access to the sources that were used to determine relations between terms and potentially they need to understand any deductions performed on the information. In this paper, we present our approach to explaining semantic matching. Our initial work uses a satisfiability-based approach to determine subsumption and semantic matches and uses the Inference Web and its OWL encoding of the proof markup language to explain the mappings. The Inference Web solution also includes a registration of the OWL reasoning component of JTP, as well as other reasoner registrations, and thus provides a foundation for explaining semantic matching systems.

1 Introduction

The amount of disparate online information is increasing as is the need for interoperability. This combination increases the need for managing semantic heterogeneity (e.g., [15]). Many solutions to the problem include "matching" terms in one information source to terms in another. We will view the information sources to be graph-like structures containing terms and their inter-relationships. The *Match* operator takes two graph-like structures and produces a mapping between the nodes of the graphs that correspond to each other. We are interested in match operations that take the term meaning into account and thus produce a kind of semantic match.

We classify approaches that use syntactic similarity measures or syntax driven approaches as *syntactic matching* since, while they may use syntactic context, they do not analyze term meaning directly, e.g., [13]. We are interested in a *semantic matching* approach that generates mappings between terms (e.g., nodes of graphs) by computing *semantic relations* (for example, equivalent or subsuming elements), instead of computing coefficients rating match quality in the [0,1] range. We are also interested in determining semantic relations by analyzing meaning (concepts, not labels as in syntactic matching) captured in the ontologies.

In our effort to produce understandable systems, our goal is to be able to explain the mappings (whether they are complete, partial, or failed to be generated). In this paper, we present our approach to semantic matching as first introduced in [6], and implemented within the *S-Match* system [7][1]. An example is presented to illustrate the semantic matching approach. Then we describe the Inference Web (IW) infrastructure [9] for explanations in distributed, heterogeneous environments and its Proof Markup Language (PML)[4]. Using the matching example, we describe how the Inference Web explanations increase user understanding of semantic matching mappings, thereby increasing trust. This work is described in more detail in [14].

2 Semantic Matching

We will focus on class matching and motivate the problem by a simple catalog example shown in Figure 1. In this scenario, an agent may need to exchange documents stored according to the two class hierarchies called A1 and A2 respectively.

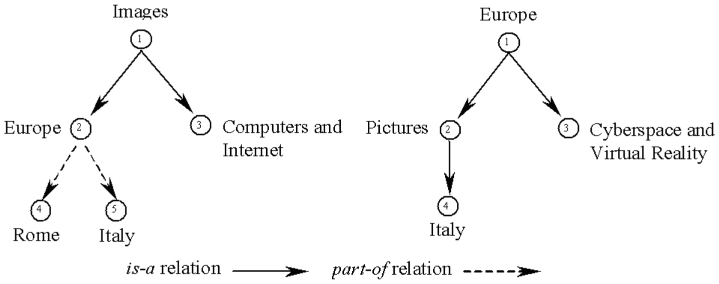


Figure 1: Simple catalog matching problem

Our semantic matching approach distinguishes the following relations between terms: *equivalence* (=, mutual subsumption); *more general* (⊒, subsumer); *less general* (⊑, subsumee); *mismatch* (⊥, disjoint); *overlapping* (⊓,

[1] The current version of *S-Match* as described in this paper is a rationalized re-implementation of the CTXmatch system [3] with a few added functionalities.

there may exist an instance of both classes). The relations form a partial order according to binding strength, with equivalence being stronger than subsumer or subsumee (which have equal binding strength) being stronger than overlapping.

The semantic relations are calculated by mapping meaning which is codified in the element descriptions and the graphs in two steps: obtaining a representation of the node meaning and by determining the meaning of the node position in the graph. In this example, we would view a node labeled *Pictures* to represent the concept "documents which are about pictures". In order to obtain some information about the node labels, our initial implementation accesses WordNet to obtain information about *senses* and subclass hierarchies. Extensions to the work would also take other DL representations of the classes as input such as full OWL ontologies. Just obtaining term information from WordNet (or any ontology) however does not account for graph position. In this example, the fact that *Pictures* is below *Europe* in A2 means that documents stored below *Pictures* are actually documents that are *both* about Europe and pictures. Thus, for A2, $\boldsymbol{C_{Pictures}} = C_{Europe} \sqcap C_{Pictures}$.

A *mapping element* is a 4-tuple $< ID_{ij}$, $n1_i$, $n2_j$, $R >$, i=1,...,N1; j=1,...,N2; where ID_{ij} is a unique identifier of the given mapping element; $n1_i$ is the i-th node of the first graph, N1 is the number of nodes in the first graph; $n2_j$ is the j-th node of the second graph, N2 is the number of nodes in the second graph; and R specifies a semantic relation which may hold between the concepts at nodes $n1_i$ and $n2_j$. *Semantic matching* can be defined as the problem: given two graphs G1, G2 compute the N1 × N2 mapping elements $< ID_{ij}$, $n1_i$, $n2_j$, $R' >$, with $n1_i \in$ G1, i=1,...,N1, $n2_j \in$ G2, j=1,...,N2 and R' the strongest semantic relation holding between the concepts of nodes $n1_i, n2_j$. We define a *mapping* as a set of mapping elements [6].

In the example, we would have a mapping element between the node labeled *Europe* in A1 and the node labeled *Pictures* in A2. Since *Europe* is below *Images* in A1, documents classified under it are about images and which are about Europe. If one knows, from WordNet or other sources, that Images and Pictures are equivalent, then we can conclude that the extension of *Europe* in A1 is the same as that for *Pictures* in A2: $< ID_{22}$, $\boldsymbol{C_{Europe}}$, $\boldsymbol{C_{Pictures}}$, $= >$

Semantic matching translates the matching problem into a validity check of the appropriate propositional formula. A translation encodes concepts at nodes using a logical propositional language where atomic formulas are atomic concepts, written as single words, and complex formulas are obtained by combining atomic concepts using the connectives of set theory and set theoretic semantics. The semantic relations are also translated into propositional connectives, namely: equivalence into equivalence, more general and less general into implication, and mismatch into negation of the conjunction.

The goal then is to prove that given a particular context or background theory, the semantic relation *rel* (translated into the propositional theory) holds between the representation of $\boldsymbol{C1_i}$ as the concept of node i in A1 and $\boldsymbol{C2_j}$ as the concept of node j in A2. We write this as: $Context \longrightarrow rel(\boldsymbol{C1_i}, \boldsymbol{C2_j})$. The background theory is the conjunction of all the relations between concepts of the labels mentioned in either graph.

From the example, trying to prove that *Europe* from A1 is the same as *Pictures* in A2, requires constructing:

$$((C1_{Images} \leftrightarrow C2_{Pictures}) \wedge (C1_{Europe} \leftrightarrow C2_{Europe})) \rightarrow$$
$$((C1_{Images} \wedge C1_{Europe}) \leftrightarrow (C2_{Europe} \wedge C2_{Pictures}))$$

The algorithm then checks for sentence validity by proving that its negation is unsatisfiable. Our implementation uses the JSAT SAT reasoner. In this example, the negated sentence is false, thus the equivalence relation holds between the nodes. Since this is the strongest relationship, no additional checks need to be made and the *S-Match* algorithm terminates and concludes that documents stored under *Pictures* in A2 are an appropriate match for documents stored under *Europe* in A1.

3 Explaining Matching

Inference Web enables applications to generate portable and distributed explanations for answers. In order to explain semantic matching and thereby increase the trust level of its users, we need to provide information about background theories (initially Wordnet), the JSAT manipulations of sentences, and the semantic match translations of graphs into propositional sentences. This paper addresses the first two topics.

In order to use Inference Web to provide explanations, question answering systems need to have their reasoners produce proofs of their answers in the Proof Markup Language, publish those proofs on the web, and provide Inference Web with a pointer to the last step in the proof. Inference Web also has a registry[10] of meta-data that is populated with information about objects used in the proof (ontologies, inference engines and their rules, etc.). In the example case, it contains meta information about JSAT and WordNet.

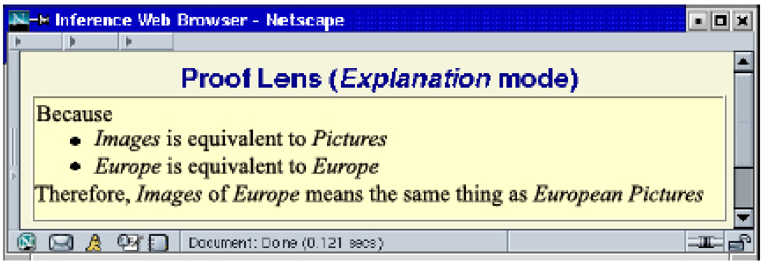
Proof and explanation documents are represented in PML and are composed of PML *node sets*. Each node set represents a step in a proof whose conclusion is justified by a set of inference steps associated with a node set. This representation could be viewed as the web-ized distributed OWL version of one author's previous work on explaining description logics [11].

The *IW Browser* is used to present proofs and explanations. Exploiting PML properties, meaningful fragments of *S-Match* proofs can be loaded on demand. Users can browse an entire proof or they can limit their view and refer only to specific, relevant parts of proofs since each node set has its own URI that can be used as an entry point for proofs, proof fragments.

4 Producing Explanations using Inference Web

Users may need different types of explanations. For example, if agents are familiar with each other, and trust each other's information sources, explanations should focus on the *S-Match* manipulations. If on the other hand, the sources may be suspect, explanations should focus on meta information about sources. We will provide a few descriptions of explanations available from our work.

In a simple case, consider the explanation in Figure 2 for why *S-Match* mapped *Europe* in A1 to *Pictures* in A2 (and thus returned documents labeled with A1:Europe as results to the query find "European pictures").



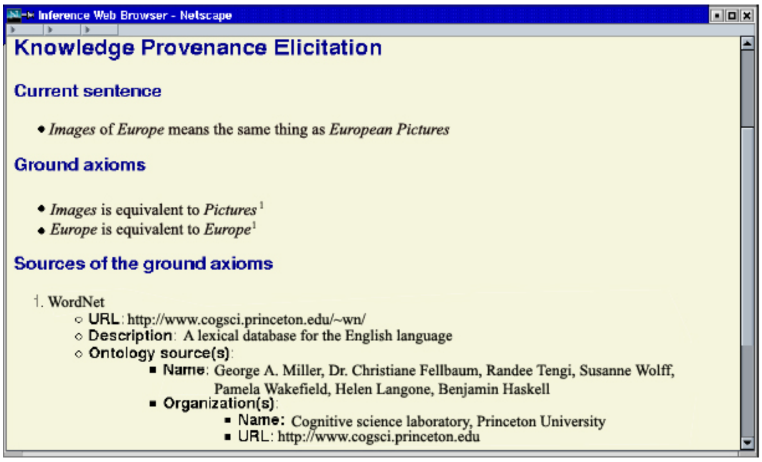
Proof Lens (*Explanation* mode)

Because
- *Images* is equivalent to *Pictures*
- *Europe* is equivalent to *Europe*

Therefore, *Images* of *Europe* means the same thing as *European Pictures*

Figure 2: An explanation in English

Users then see that *Images* in A1 and *Pictures* in A2 are equivalent words (and similarly in this unpruned version, that *Europe* in A1 denotes the same concept as *Europe (European)* in A2). If the user needed to see information about the sources of the information used, they could also obtain an explanation that reveals the provenance information about equivalences displayed. Figure 3 shows that all the information used in the *S-Match* proof came from WordNet



Knowledge Provenance Elicitation

Current sentence

- *Images* of *Europe* means the same thing as *European Pictures*

Ground axioms

- *Images* is equivalent to *Pictures*[1]
- *Europe* is equivalent to *Europe*[1]

Sources of the ground axioms

1. WordNet
   - URL: http://www.cogsci.princeton.edu/~wn/
   - Description: A lexical database for the English language
   - Ontology source(s):
     - Name: George A. Miller, Dr. Christiane Fellbaum, Randee Tengi, Susanne Wolff, Pamela Wakefield, Helen Langone, Benjamin Haskell
     - Organization(s):
       - Name: Cognitive science laboratory, Princeton University
       - URL: http://www.cogsci.princeton.edu

Figure 3: Source metadata information

(which included that the first sense of pictures is a synonym for the second sense of images). The meta information about WordNet from the IW Registry is included in this presentation.

If a user wants an explanation of the inference engine(s) embedded in a matching system, a more complete explanation may be required. Our current version of *S-Match* uses JSAT, and in particular the Davis-Putnam-Longemann-Loveland (DPLL) procedure [5].

SAT engines build an assignment $\mu \in \{\top, \bot\}$ to atoms of a propositional formula φ such that it evaluates φ to *true*. Then, φ is *satisfiable* iff $\mu \models \varphi$ for some μ. The basic DPLL procedure recursively implements the three rules: *unit resolution*, *pure literal* and *split* [5]. Our implementation works on an unoptimized version of DPLL for simplicity of the explanation effort. In the interest of space, we focus on the unit resolution rule.

Let l be a literal, φ - a propositional formula in CNF. A *unit clause* has exactly one unassigned literal. Unit resolution rule is an application of *resolution*, where one clause is a unit clause.

$$\textit{unit resolution} : \frac{\varphi \wedge \{l\}}{\varphi[l \mid \top]}$$

Let us consider the propositional formula standing for the problem of testing if the concept at node 2 in A1 is less general then the concept at node 2 in A2. To simplify the presentation we use in the following a label as a placeholder of a concept the given label denotes. DPLL procedure as implemented in JSAT handles only CNF formulas. Thus, the propositional formula and its equivalent in CNF (see Figure 4) is input into the DPLL procedure:

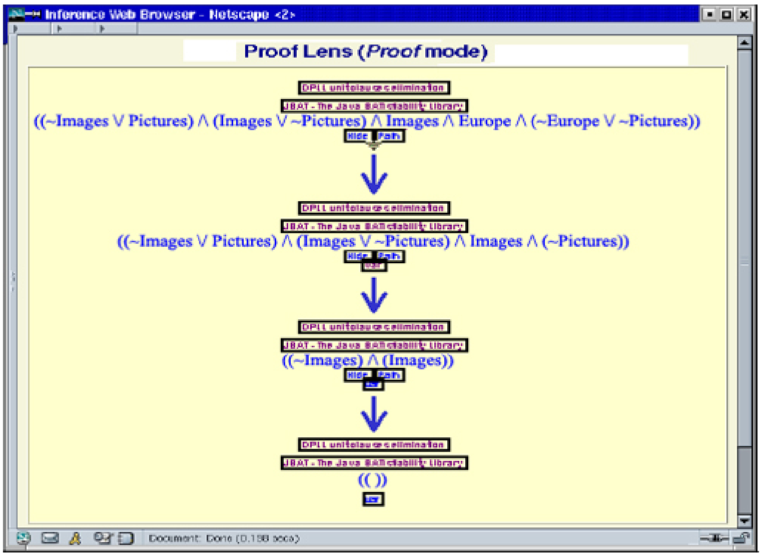
$$((Images \leftrightarrow Pictures) \wedge (Europe \leftrightarrow Europe)) \wedge$$
$$\neg((Europe \wedge Images) \rightarrow (Pictures \wedge Europe))$$

An intuitive reading of the SAT problem is "is there any situation such that the concept *Images of Europe* is less general then the concept *European Pictures* assuming that *Images* and *Pictures* denote the same concept?". The IW proof defending the negative answer is shown in Figure 4.

5 Discussion

While there are a number of other efforts in semi-automated ontology matching (see surveys in [13, 15]), we are not aware that any provide explanations. The DPLL procedure implemented in our approach, while unoptimized, includes the essence of the state of the art SAT engines such as Chaff [12], etc. Thus, one could consider using another optimized SAT reasoner that may be chosen for particular matching problems and using explanations generated by our system.

Recently there has been work on verifying SAT solvers. A direct solution is provided in [1]. They introduce a proof-producing infrastructure based on natural deduction for SAT engines (e.g., Chaff). Another approach uses explanations


Inference Web Browser - Netscape <2>
DPLL unitclause elimination
JSAT - The Java Satisfiability Library
Hide Path
Document: Done (0.158 secs)

Proof Lens (*Proof* mode)

$((\sim Images \vee Pictures) \wedge (Images \vee \sim Pictures) \wedge Images \wedge Europe \wedge (\sim Europe \vee \sim Pictures))$

$((\sim Images \vee Pictures) \wedge (Images \vee \sim Pictures) \wedge Images \wedge (\sim Pictures))$

$((\sim Images) \wedge (Images))$

$(( ))$

Figure 4: A graphical explanation

in terms of "equivalent" inference systems [2] in order to provide explanations of potentially alternative deductive paths. The key distinctions of *S-Match* proofs are:

- They are produced by a modified version of JSAT in *S-Match* that implements the Barrett and Berezin approach for generating proofs;
- They are formatted in PML and consequently they are designed for use in a distributed Web environment;
- Their sentence propositions are mapped into meaningful terms rather than numbers in sentences using the DIMACS format;
- They are supported by the Inference Web tools for explanation and interactive proof presentation.

6 Conclusion

In this paper we presented an approach for explaining answers for the semantic heterogeneity problem. By extending *S-Match* to use the Inference Web infrastructure, we demonstrated our approach for explaining matching systems that use background ontological information and reasoning engines. We presented DPLL-based IW explanations of the SAT engine used in the *S-Match* system. The explanations can be presented in different styles allowing users to understand the mappings and consequently to make informed decisions about them. The paper also demonstrates that *S-Match* users can leverage the Inference Web tools, for example, for sharing, combining, browsing proofs, and supporting proof meta-information including knowledge provenance information. Future work includes using more expressive background ontologies and other SAT engines as well as other non-SAT DPLL-based inference engines, e.g., DLP, FaCT [8] and an evaluation effort.

References

[1] C. Barrett and S. Berezin. A proof-producing boolean search engine. In *Proceedings of PDPAR'03*, 2003.

[2] A. Borgida, E. Franconi, I. Horrocks, D. McGuinness, and P. Patel-Schneider. Explaining alc subsumption. In *Proceedings of DL-99*, 1999.

[3] P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: A new approach and an application. In *Proceedings of ISWC'03*, pages 130–145, 2003.

[4] P. Pinheiro da Silva, D. L. McGuinness, and R. Fikes. A proof markup language for semantic web services. TR KSL-04-01, Stanford University, 2004.

[5] M. Davis and H. Putnam. A computing procedure for quantification theory. In *Journal of the ACM*, number 7, pages 201–215, 1960.

[6] F. Giunchiglia and P. Shvaiko. Semantic matching. In *The Knowledge Engineering Review journal*, number 18(3), 2004.

[7] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: an algorithm and an implementation of semantic matching. In *Proceedings of ESWS' 04*, 2004.

[8] I. Horrocks and P. F. Patel-Schneider. Fact and dlp. In *Automated Reasoning with Analytic Tableaux and Related Methods: Tableaux'98*, pages 27–30, 1998.

[9] D. L. McGuinness and P. Pinheiro da Silva. Infrastructure for web explanations. In *Proceedings of ISWC'03*, pages 113–129, 2003.

[10] D. L. McGuinness and Pinheiro da Silva P. Registry-based support for information integration. In *Proceedings of IJCAI'03 Workshop on IIWeb-03*, 2003.

[11] D.L. McGuinness. *Explaining reasoning in description logics.* PhD thesis, Rutgers University, 1996.

[12] M. Moskewicz, C. Madigan, Y. Zhaod, L. Zhang, and S. Malik. Chaff: Engineering an efficient sat solver. In *Proceedings of DAC'01*, 2001.

[13] E. Rahm and P. Bernstein. A survey of approaches to automatic schema matching. In *VLDB Journal*, number 10(4), pages 334–350, 2001.

[14] P. Shvaiko, F. Giunchiglia, P. Pinheiro da Silva, and D. L. McGuinness. Web explanations for semantic heterogeneity discovery. TR KSL-04-02, Stanford University, 2004.

[15] H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Huebner. Ontology-based integration of information - a survey of existing approaches. In *Proceedings of IJCAI'01 workshop on OIS*, pages 108–117, 2001.