UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

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

GENERATING REFERENCE MAPPINGS FOR EVALUATING ONTOLOGY MATCHING TOOLS

M. Yatskevich, P. Avesani, H. Stuckenschmidt and E. van Mulligen

May 2006

Technical Report # DIT-06-050

Generating Reference Mappings for Evaluating Ontology Matching Tools

M. Yatskevich[1], P. Avesani[2], H. Stuckenschmidt[3] and E. van Mulligen[4]

[1]University of Trento, Italy
[2]ITC-IRST, Trento
[3]University of Mannheim, Germany
[4]Erasmus University Rotterdam

Abstract. Ontology Alignment is one of the biggest challenges of semantic web research. Recently there has been some attention on automatic alignment methods and number of matching tools have been developed. One of the open problem of automatic matching currently is the evaluation of the quality of automatically generated mappings. Normally, reference mappings are used to compute the precision and recall of automatically created mappings. Often, however, such reference mappings do not exist and creating them by hand is not feasible for large ontologies. In this paper, we extend previous work on overcoming this problem by proposing a method for automatically generating highly correct reference mappings. In particular, we consider the case where we do not have shared instances for the ontologies to be aligned and report the results of an experiment in which we used automatic document classification as a basis for creating a reference mapping. We present the results of our experiment on medical datasets and show that the result provides a useful basis for comparing matching systems.

1 Introduction

A lot of work has been done in the area of databases and information systems to cope with schema heterogeneity. Recently, the problem of aligning conceptual models is gaining importance in the context of semantic web research. In this context, a number of systems that automatically try to find alignments between conceptual models have been developed [16]. In order to be able to rely on such automatic alignment methods, we need to be able to estimate the quality of the created alignments. This raises the question of how to evaluate automatically generated mappings. There are two standard approaches to this problem.

- Manual evaluation by domain experts
- Automatic evaluation against a reference mapping

In the first case, the mappings created an alignment tool are presented to a domain expert who determines which of the automatically created mappings are incorrect. This approach has some disadvantages because it has to be redone every time mappings are created. This problem can be solved by once setting up a reference alignment that is known to contain only correct mappings. This alignment does not change and can be

used repeatedly to evaluate the results of different tools and approaches. Therefore, the existence of good reference alignments is an important requirement for a successful evaluation of automatic alignment approaches. As we cannot expect that such alignments exist for relevant alignment problems – if there was an alignment there would not be a need for automatic matching – we have to think about ways to create reference alignment when needed.

In this paper we propose a method allowing (semi-) automatic construction of reference mapping set for the matching problem involving two conceptual hierarchies. The key idea of our method is to produce an incomplete set of highly correct mappings exploiting the evidence provided by documents classified under the nodes of conceptual hierarchies. Subsequently this set is verified by comparison with manually created mappings. Due to its inherent incompleteness the reference mapping set can not be exploited for evaluation of matching results precision, because correct mappings not contained in the incomplete reference mapping would be counted as false positives. However, it allows to estimate recall, what according to [15] is much more an issue for state of the art matching systems. In contrast to previous work [2] we exploit automatic classifiers in order to produce the set of the classified documents. This significantly reduces the human effort for the reference mapping acquisition and substantially extends the domain of our method applicability. We have applied our method to case study of reference mapping acquisition for two medical vocabularies (MESH and CRISP). Both vocabularies contain order of tenth thousands nodes. As a result we have obtained a reference mapping containing thousands of mappings. As from evaluation results we argue that our method allows to reduce significantly the effort for reference mapping acquisition in large scale mapping problems. We provide an empirical evidence that our methodology for reference mapping acquisition is a good approximation of a manually created reference mapping and can therefore be used to evaluate existing matching systems with respect to their completeness. Additionally we show that the reference mapping produced by our method exhibits some important properties that guarantee fairness in the evaluation of different approaches (compare [2]).

The paper is structured as follows. In Section 2 we illustrate the problem of generating and evaluating automatically generated mappings. In particular, we explains the use of reference mappings for evaluation and the difference between the problem of acquiring reference mappings and the general matching problem and present our approach to the problem. The evaluation of this approach in a real world case study is described in section 3. Section 4 concludes the paper.

2 Evaluation of Ontology Mappings

The work reported in this paper is carried out in the context of the heterogeneity workpackage of the European Research network KnowledgeWeb. The aim of this workpackage is to define a general framework for representing ontology mappings as well as to develop and test methods for automatically creating and evaluating such mappings[9]. In the following, we briefly present the notions of mapping and semantic matching used in KnowledgeWeb and explain the role of our work on generating reference mappings for evaluating matching results.

The problem of semantic matching consists of finding a set of mappings between conceptual structures that correctly represent semantic relations between elements in the two structures. This problem can be illustrated using the simple example in Figure 1 that shows small parts of the two medical terminologies used in our experiments. The task of semantic matching is to identify candidates to be merged or to have relationships under an integrated hierarchy. In Figure 1 this would mean to recognize that $Plants$ is equivalent to $plant$ and $Bacteria$ is less general than $microorganism$.

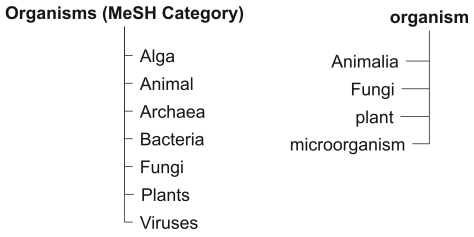

Fig. 1. Parts of MESH and CRISP conceptual hierarchies

Once these relations have been identified, they are encoded in terms of a set of mappings that are represented as a 4-tuple [9]:

$$(e, e', n, R)$$

where e and e' are elements from the different models, for example $e = Bacteria$ and $e' = microorganism$, R specifies the semantic relation that has been identified between them, in this case $R = \subseteq$ and n is a degree of confidence that R correctly describes the real sematic relation between e and e'. The idea of a reference mapping is now that it correctly describes the true semantic relations between elements in two models using expressions of the form given above.

2.1 The Evaluation Problem

Provided that a reference mapping exists, the problem of evaluating an automatically generated mapping can be reduced to the problem of comparing two sets of mappings. This comparison can be based on well known measures from information retrieval, in particular *Precision* and *Recall*. Ehrig an Euzenat describe a general framework for computing precision and recall over sets of mappings [6]. In particular, they define the notion of relaxed precision and recall in the following way, where E represents the automatically generated mapping set, R is the reference mapping set and ω is a function that returns the overlap between the two mapping sets.

$$Prec_\omega(E, R) = \frac{\omega(E, R)}{|E|} \quad Rec_\omega(E, R) = \frac{\omega(E, R)}{|R|}. \tag{1}$$

Traditionally, the overlap between between two mapping sets is computed by counting the number of mappings shared by the two sets, in particular $\omega(E, R) = E \cap R$. The situation is illustrated in figure 2.

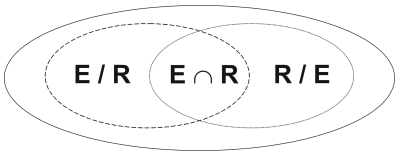

Fig. 2. Mapping comparison in the case of complete reference mapping set.

In practice, the use of this measure is complicated by two facts. First of all, not all matching systems support different kinds of semantic relations. In particular, system like COMA [12] or the approach of Euzenat and Valtchev [8] only support equivalence of elements. other relations like the inclusion relation between $Bacteria$ and $microorganism$ in our example are represented via equivalence relation and a possibly lower degree of confidence. As the reference mapping, however, is assumed to represent the true semantic relations between elements, we assume that it contains different types of semantic relations, at least the common set operations.The second problem that has already been mentioned above is the need to have a complete reference mapping set to be able to compute precision. It is clear from figure 2 and the definitions of precision and recall, that in order to compute precision, we need to know all correct mappings, because otherwise, we cannot determine the set of false positives (E/R). As a result, recent evaluation efforts such as the Ontology Alignment Evaluation Initiative [7] concentrated on rather small artificial data sets where complete reference mappings can be created manually. At the same time industrial size schemas contain up to tenth thousands of nodes. Scaling up to such real world data sets is a major problem. In fact human annotator needs to consider a quadratic number of potential relations between elements in the different models For example, in order to produce the reference mapping for the medical terminologies we used in our experiments up to $150000 \times 10000 = 1,5 \times 10^9$ potential relations have to be considered.

2.2 Our Approach

In order to overcome this problem, the idea of our work is to exploit an incomplete reference mapping set for evaluation of matching solutions. Figure 3 illustrates this idea.

Instead of computing precision and recall for automatically acquired mappings in the way described in equation 2, we compute the following measures instead:

$$Prec = \frac{|E \cap R'|)}{|E|} \quad Rec = \frac{|E \cap R'|}{|R'|}. \tag{2}$$

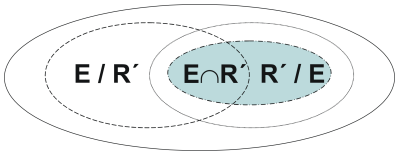

Fig. 3. Mapping comparison in the case of incomplete reference mapping set.

We can see that the definition of recall is still intact in the sense that it correctly computes the recall of E with respect to the reduced reference mapping R'. The definition of precision, however, is flawed as the expression $|E \cap R'|$ does not provide us with a correct assessment of the number of correct results as we do not know the difference between complete and incomplete reference mapping sets - the false positives (dotted and dashed-dotted areas in Figure 3). As a consequence of this we cannot estimate the precision of an automatically generated mapping. However if we assume that the incomplete reference mapping set is a good representative of the complete one we can use the recall with respect to R' as an estimate for the recall with respect to the complete reference mapping.

In [2] we define a number of criteria for deciding whether R' is a good representation of R and therefore a good basis for estimating the recall. These criteria are the following. For a more detailed motivation of these criteria, we refer to the original work.

- Correctness, namely the fact that the dataset can be a source of correct results.
- Complexity, namely the fact that the dataset is "hard" for state of the art matching systems.
- Discrimination ability, namely the fact that the dataset can discriminate among various matching approaches.

Actually creating the incomplete reference mapping set R' requires to correctly identify the semantic relations between a sufficient number of concepts in the ontologies to be aligned. In order to cope with this problem we propose to look at the pragmatic use of concepts. Often the nodes in conceptual hierarchies are used to organize documents or contents like images. Our working assumption is that the set of documents classified under a given node implicitly defines its meaning. Therefore two nodes have a similar meaning if the sets of documents classified under them have a considerable overlap. This approach has been followed by instance based matching approaches (see for example [4, 13]). In these works the interpretation of a node is approximated by a model computed through statistical learning. Of course the accuracy of the interpretation is affected by the error of the learning model. We follow a similar approach but without the statistical approximation.

The key distinction between instance-based matching and our approach is in the scope. We are focused on acquisition of relatively small number of highly precise mappings while matching approaches aim to produce complete set of mappings for the given matching task. Incomplete reference mapping (such as the one produced by our method) allows to evaluate the recall but not the precision of the matching results, where recall is defined as a ratio of reference mappings found by the system to the number of

reference mappings and precision is defined as ratio of reference mappings found by the system to the number of mappings in the result. However, as highlighted in [15], the biggest problem in state of the art matching systems is recall, while precision is much less an issue. The challenge for our evaluation methodology in this paper is to prove that reference mapping produced using our method is a good approximation of manually acquired one and that it posses desirable from matching solutions evaluation perspective properties.

3 Experimental Evaluation

The concrete problem statement underlying our work is the question of how to evaluate the results of the Ontology Alignment Evaluation Challenge[1]. In particular, the 2005 challenge featured two real world data sets in terms of ontologies to be aligned. The first data set consisted of parts of the Yahoo!, Google and Looksmart web directories. For this dataset, we successfully created a reference mapping set based on shared instances in terms of web pages classified into all directories [2]. The second data set consisted of ontologies describing human anatomy. As these ontologies do not come with instance data, the approach described in [2] cannot directly be applied to this data set. In order to also be able to provide a reference mapping for this second real world case our goal is to automatically create shared instance data for these ontologies by classifying medical documents into the two ontologies. The experiments reported in the following have been carried out as a proof of concept that it is actually possible to create shared instance data by classifying documents and successfully use it to create a reference mapping according to the method described in [2]. We did not use the actual anatomy ontologies as a basis for this proof of concept, because we wanted to compare the results of our method with a manually created reference mapping in order to be able to assess the quality of our method. For this reason, we chose to use medical terminologies from UMLS (Unified Medical Language System) as a basis for our experiments.

In a first step we extracted two conceptual hierarchies from UMLS. In the second step we created shared instance data by indexing a standardized medical document set using terms from the two hierarchies. In the third step, the we computer a reference mapping. Finally, we evaluated the generated reference mapping against manual mappings from UMLS and the quality criteria mentioned above. In the following, we provide details of this process.

3.1 Data Selection and Preparation

The UMLS metathesaurus (in the following UMLS) is a hierarchical thesaurus that integrates a number of terminological sources (such as thesauri or classifications) in the medical domain and currently contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies and classifications. The metathesaurus integrates concepts from different sources by arranging them in a common hierarchy that preserves the taxonomic relations of the source.

[1] http://oaei.ontologymatching.org/

This feature makes UMLS a perfect basis for evaluating our methodology, because we use our approach to create a reference mapping between two of the terminological sources and compare the result to the manually created mapping encoded in the metathesaurus. We selected two of the terminological sources contained in UMLS as the basis for our evaluation. The first of the sources is the MeSH (Medical Subject Headings) that contains more than 150.000 medical terms organized in a hierarchy. The second information source is the CRISP (Computer Retrieval of Information on Scientific Projects) terminology which contains more than 10.000 concepts each associated with a number of different terms. Both sources were translated into conceptual hierarchies by extracting the nodes connected by subterm/superterm relationships. In order to perform an instance-based comparison of the conceptual hierarchies, we used an existing corpus of medical documents. The OHSUMED test corpus is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The set has been used in the TREC (Text Retrieval Conferences) to assess the quality (in terms of precision and recall) of indexing software.

3.2 Step 2. Document Classification

In the second step, we classified the documents from the OHSUMED test corpus into the two conceptual hierarchies using a state of the art system for concept-based document indexing and retrieval provided by the Dutch company Collexis. The Collexis system consists of two parts: a concept-based indexing machine and a vector matching machine. The concept-based indexing machine exploits a thesaurus and basic word stemming techniques to recognize terms that identify a certain concept in a text. The indexing machine assigns to each concept found in a text a weight that indicates the (relative) importance of that concept for representing the meaning of the text. The importance of a concept with respect to a given document is determined by the normalized frequency corresponding terms occur in the document. For each document, a so-called fingerprint is created. A finger print is a vector that contains all relevant concepts with the corresponding relevance value. Based on the fingerprint, the matching engine is able to retrieve documents by comparing a query fingerprint with a collection of fingerprints for documents using various matching algorithms. The result is an ordered (descending) list of scores between the concept fingerprints in the collection and the search fingerprint. Details of the different fingerprint matching algorithms are given below.

The standard product of two vectors is used in most algorithms. It is defined as $m(q, f) = \sum_{c=1}^{n} f_c \cdot q_c$ and where f_c denotes the weight of concept c in fingerprint f. A vector f is used as a fingerprint from a collection; a vector q is used as the query fingerprint. We tested the following different concept matching algorithms to determine the relevance score (S_R) of a document with respect to a certain concept in the thesaurus.

basic: $S_R = m(f, q)$

collexis: $S_R = m(1/s_f, \delta_q)$ where s_f is a measure for the specificity of a concept (or normalized inverse frequency of a concept in a document set) and where δ_q is a vector with value 1 for concepts in q and 0 otherwise.

vector: $S_R = m(f,q)/\sqrt{m(f,f) \cdot m(q,q)}$
dice: $S_R = 2 \cdot m(f,q)/(m(f,f) + m(q,q))$
quadsum: $S_R = m(q \cdot \delta_f, q \cdot \delta_f)$
jaccard: $S_R = m(f,q)/(m(f,f) + m(q,q) - m(f,q))$

Note that for the case of classifying documents, the query fingerprint only consists of a single concept with score 1. The document was considered to be classified under the given concept if it's relevance score (S_R) exceeded a threshold taken as 0.1 by default. Hereafter we call this process classification and we refer to different concept matching algorithms as classification algorithms.

Previous experiments based on a comparable dataset showed that the system has an accuracy of about 70% [2]. We cannot expect to get much better results in this step. A basic question therefore is if this classification accuracy is sufficient for our purpose.

3.3 Hypothesis Generation

Based on the result of the document classification step, we created hypotheses for semantic relations to be included in the reference mapping set. As our goal was to create an incomplete but highly correct mapping set, we first pruned the set of concepts considered for the mapping by removing concepts with a small number of classified documents. In particular we removed all concepts from the two hierarchies that had less than five instances, because hypotheses generated based on such small sets of instances do not have enough support. In the next step, we focussed the search for semantic relations by manually pre-selecting subtrees of the two concept hierarchies that are likely to have a semantic overlap. We recognized 10 potentially overlapping pairs of subtrees. For each of these pairs we ran an exhaustive assessment between all the possible pairs of nodes in two related subtrees.

A measure of support for hypotheses about the semantic relation between two nodes S and P in a conceptual model can be derived from the F1 measure known from information retrieval [3]. In particular, the similarity of two sets of documents is defined as the ratio between the marginal sets and the shared documents:

$$Equivalence = \frac{|O_P^S|}{|M_P^S| + |M_S^P|}$$

where the set of shared documents is defined as $O_P^S = P \cap S$ and $M_P^S = S \setminus O_P^S$ is the marginal set of documents classified by S and not classified by P (similarly $M_S^P = P \setminus O_P^S$). The following equivalence applies $O_P^S = O_S^P$. Notice that "O" stands for "overlapping" and "M" stands for "Marginal set".

The *generalization* relationship holds when the first node has to be considered more general of the second node. Intuitively, it happens when the documents classified under the first nodes occur in the ancestor of the second node, or the documents classified under the second node occur in the subtree of the first node. Following this intuition we can formalize the generalization hypothesis as

$$Generalization = \frac{|O_P^S| + |O_{A_S}^P| + |O_{T_P}^S|}{|M_P^S| + |M_S^P|}$$

[2] Collexis, personal communication

where $O^P_{A_S}$ represents the set of documents resulting from the intersection between M^P_S and the set of documents classified under the concepts in the hierarchy above S (i.e. the ancestors); similarly $O^S_{T_P}$ represents the set of documents resulting from the intersection between M^S_P and the set of documents classified under the concepts in the hierarchy below P (i.e. the children).

In a similar way we can conceive the *specialization* relationship. The first node is more specific than second node when the meaning associated to the first node can be subsumed by the meaning of the second node. Intuitively, it happens when the documents classified under the first nodes occur in the subtree of the second node, or the documents classified under the second node occur in the ancestor of the first node.

$$Specialization = \frac{|O^S_P| + |O^P_{T_S}| + |O^S_{A_P}|}{|M^S_P| + |M^P_S|}$$

where $O^P_{T_S}$ represents the set of documents resulting from the intersection between M^P_S and the set of documents classified under the concepts in the hierarchy below S (i.e. the children); similarly $O^S_{A_P}$ represents the set of documents resulting from the intersection between M^S_P and the set of documents classified under the concepts in the hierarchy above P (i.e. the ancestors).

For each pair of concepts in the selected subhierarchies we computed normalized values for Equivalence, Generalization and Specialization. The corresponding semantic relation was assumed to hold between the concepts if the corresponding value was higher that 0.5. In cases where more than one of the values was above this threshold, we selected the relation with the highest support. Based on this selection the initial reference mapping set was created.

3.4 Evaluation

We carried out a detailed evaluation of the generated reference mapping set. In the course of this evaluation, we assessed the three quality criteria for a reference mapping set mentioned in section 2.2.

Correctness Correctness is the basic requirement for a reference mapping set. In order to evaluate correctness we compared them with the manually created mappings encoded in the UMLS metathesaurus, The transitive closure of the *has child* (CHD) and the *narrower meaning* (RN) relations was considered as less generality. The transitive closure of their inverses was considered as more generality. Equivalences were derived from the cases when two nodes in MESH and CRISP conceptual hierarchies were assigned to exactly the same UMLS Metathesaurus concept. We did not distinguish among different semantic relations. Therefore, for example, the mapping hypothesis $A \sqsubseteq B$ was considered to be correct if $A \equiv B$ was derived from UMLS Metathesaurus.

Additionally, we performed a manual inspection of mappings that were detected by our method but that were not contained in UMLS. For each of these mappings we determined whether the mapping was actually incorrect of whether the mapping should actually be part of UMLS.

Complexity and Discrimination Ability Assessing the additional criteria of complexity and discrimination ability requires to compare the reference mapping set with the result of different automatic matching systems in order to find out whether the mappings in the reference mapping are hard to determine and whether they do not bias a specific system. For this purpose, we applied the following state of the art matching systems to the UMLS data set and compared the results with our reference mapping set.

- S-Match [11] is a generic semantic matching tool. It takes two tree-like structures as an input and produces a set of mappings between their nodes. The matching process is based on translation of tree matching problem (assuming as a background theory context [10]) into satisfiability problem in propositional logic.
- COMA++ [1] is a generic syntactic matching tool. It takes two tree-like structures as an input and produces the mapping combining the results of several string and structure matchers.
- FOAM [5] is a syntactic OWL ontologies matching tool. It produces the mapping by aggregation of previously estimated similarity features.
- Falcon [14] is a syntactic OWL ontologies matching tool. It produces mapping exploiting linguistic and graph matchers.

In the evaluation we have exploited default settings for S-Match and COMA++. The settings for Falcon and FOAM were taken from the latest ontology mapping evaluation OAEI-2005 [7]. In order to obtain the results from ontology matching tools we have converted the dataset in OWL format exploiting the methodology of the latest ontology mapping evaluation (see [7] for more detail). Similar to reference mapping quality evaluation we did not distinguish among different semantic relations. Therefore, for example, the mapping $A \equiv B$ produced by COMA++ increased the recall of the system even if reference mapping set contained $A \sqsubseteq B$ relationship.

Since we expected a considerable overlap between the reference mapping sets and the UMLS Metathesaurus, the major requirement to the matching systems was to not exploit UMLS Metathesaurus as a knowledge source. At best of our knowledge all state of the art matching systems (including 4 we have chosen for the evaluation) comply to this requirement.

3.5 Results

Based on the six different vector matching algorithms used for classifying documents into the concepts of the two concept hierarchies involved we obtained six different reference mapping sets that we evaluated. The first observation made about the different mapping sets is the fact that while most of the classification methods produced a reference mapping set with more than 7.000 mappings, the collexis algorithms sticks out as the corresponding mapping set only contains about 1.700 mappings (compare figure 4).

Correctness Comparing the reference mappings with the manual mapping contained in UMLS we see that the small number of mappings in the set created using the collexis method also has a significant influence on the correctness of the mapping set. While the other mapping sets contain less than 50% correct mappings and are therefore not

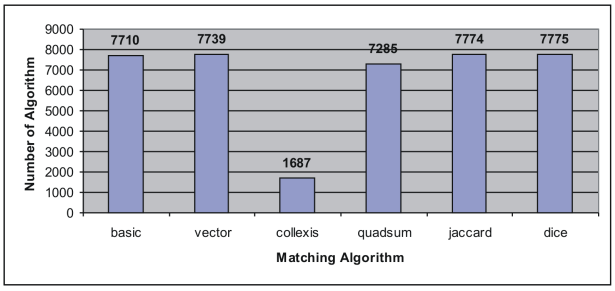

Fig. 4. Number of mappings in the reference mapping sets produced exploiting various classification algorithms

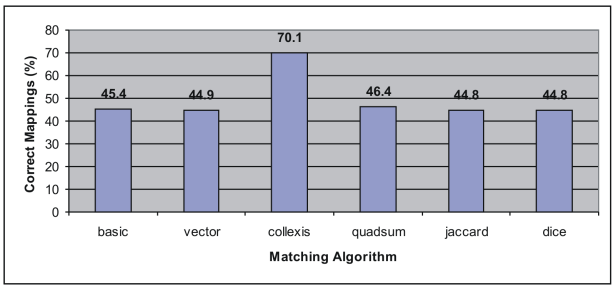

Fig. 5. Percentage of correct mappings in the reference mapping sets produced exploiting various classification algorithms. UMLS Metathesaurus is considered to be a golden standard

suitable as a reference mapping, the mapping set created using the Collexis method contains 70% correct mappings.

A manual assessment of the remaining 30% of the mappings revealed that the this mapping set is actually better than the number of 70% seems to suggest. In fact it turned out that many of these mappings are actually correct but are not included in UMLS. This can be explained by the fact that human experts have a variance of up to 20% in comparison with their own results if the mapping problem is difficult. If we take into account the number of correct mappings not in UMLS, the collexis mapping set has a correctness of 95% (compare figure 6)

We conclude that the methodology proposed in the paper allows to produce the correct reference mapping sets.

Complexity Complexity measures the difficulty of finding the mappings in the reference mapping set using automatic matching tools that try to find complete mappings. In order to determine the complexity of a mapping set, we therefore analyze the recall of different matching systems with respect to the collexis mapping set. The highest degree of completeness achieved by any of the system as slightly lower 30%. Most systems had a completeness of between 8 to 16% (compare figure 7).

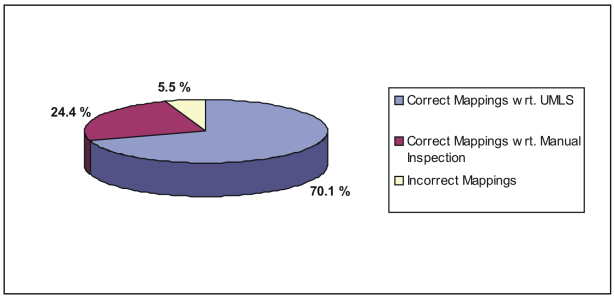

Fig. 6. Results of the Manual Assessment of the Collexis mapping set

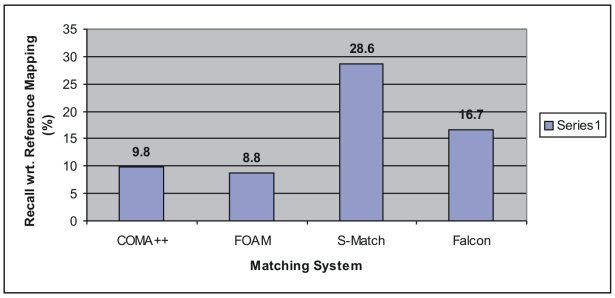

Fig. 7. Percentage of correctly determined mappings (Recall)

As previously reported recall values on artificially produced matching problems for these systems were around 60-80% we conclude that the mapping set is hard for state of the art syntactic and semantic matching systems.

Discrimination ability Discrimination ability measures the ability of the reference mapping set to clarify the strengths and weaknesses of different automatic matching systems. In order to measure this we checked for each rule in the reference mapping how many of the systems mentioned above were able to find the mapping. It turned out that only 5% of the mappings were found by all 4 systems. At the same time about 70% of the mappings were found by only one of them. This means that it contains a number of mappings which are hard to find for some of the systems and easy for others (compare figure 8).

We conclude that the collexis mapping set is highly discriminating as it enables us to analyze the weaknesses of particular systems by looking at those mappings that could not be found by that particular system.

4 Conclusion

In this paper we addressed the problem of evaluating automatic ontology matching approaches in the context of the Ontology Alignment Evaluation Initiative. In particular,

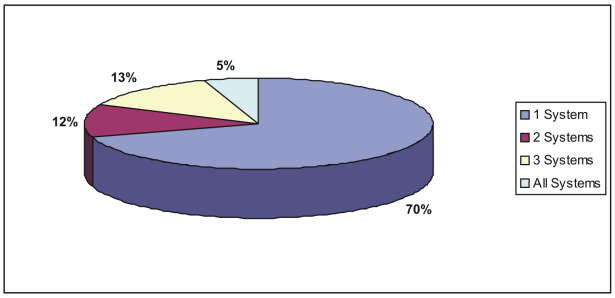

Fig. 8. Partitioning of mappings found by matching systems according to the number of systems which found them

we argued that evaluation is possible on the basis of an incomplete but highly correct reference mapping set. The specific Problem addressed in the paper was the identification of semantic relations between concepts in different ontologies based on shared instances that have been created by automatically classifying documents into the two ontologies. Our main goal was to provide a proof of concept that automatically classified instances can be used as a basis for identifying these relations. For this purpose, we classified a medical document set with respect to two concept hierarchies, computed semantic relations based on the overlap of instances and compared the result with a manually created mapping set provided by the UMLS metathesaurus.

Based on the results of our experiments we are now able to draw some conclusions about the feasibility of automatic reference set generation. Our first observation is that the approach worked quite well for the example. We were able to create a reference set with about 1.700 mappings of which 95% were correct. This result makes us optimistic that our method will also produce a useful reference mapping for the anatomy data set used in the alignment challenge. Beyond this general conclusion we gained a number of insights in the factors that influence the result. The first insight is that the correctness of the produced reference mapping set critically depends on the accuracy of the classification step. At this step it is essential to already tune the classification method towards correctness while completeness is less important. At this point, we have to rely on insights from the area of document classification to fine-tune the methods to the given data set. The second insight is that despite the promising results, it is clear that manual prost-processing is needed to ensure the correctness of the results. This step cannot be omitted, but the use of our method significantly reduces the search space for human experts that . In our case the effort for manually checking sematic relations between concepts was reduced from consideration of $150000 \times 10000 = 1,5 \times 10^9$ relations to the collexis data set, which is about 1700 mappings or 5 orders of magnitude less than the original number.

The main issue for future work is to get more insights in the generality and limitations of the approach. In particular, we need to better understand what minimal requirements ontologies and document sets must satisfy to support automatic classification of documents. A potential problem is the fact, that most existing ontologies do not con-

tain information about synonyms and often use rather complex concept names. More experiments are needed in order to find out in how far this is a problem for our method.

References

1. D. Aumueller, H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with COMA++. In *Proceedings of ACM SIGMOD 2005*, pages 906–908, New York, NY, USA.
2. P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In *Proceedings of International Semantic Web Conference (ISWC)*, 2005.
3. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
4. H. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, 50:279–301, 2003.
5. M. Ehrig and S. Staab. QOM - quick ontology mapping. In *Proceedings of the Third International Semantic Web Conference*, volume volume 3298 of LNCS, pages 683–697, Hiroshima, Japan, NOV 2004.
6. Marc Ehrig and Jerome Euzenat. Relaxed precision and recall for ontology matching. In Benjamin Ashpole, Marc Ehrig, Jrme Euzenat, and Heiner Stuckenschmidt, editors, *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies*, Banff, Canada, October 2005.
7. J. Euzenat, H. Stuckenschmidt, and M. Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Proceedings of K-CAP 2005 Workshop on Integrating Ontologies*, 2005.
8. J. Euzenat and P. Valtchev. Similarity-based ontology alignment. In *Proceedings of the European Conference on Artificial Intelligence ECAI 2004*, pages 333–337, 2004.
9. Enrico Franconi, Giorgos Stamou, Jrme Euzenat, Marc Ehrig, Markus Krtzsch, Paolo Bouquet, Pascal Hitzler, Sergio Tessaris, and York Sure. Specification of a common framework for characterizing alignment. Project Deliverable D2.2.1v2, KnowledgeWeb Network of Excellence, January 2005.
10. F. Giunchiglia. Contextual reasoning. *Epistemologia*, 16, 1993.
11. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-Match: an algorithm and an implementation of semantic matching. In *Proceedings of 1st european semantic web symposium (ESWS'04)*.
12. H.H.Do and E. Rahm. COMA - a system for flexible combination of schema matching approaches. In *Proceedings of Very Large Data Bases Conference (VLDB)*, pages 610–621, 2001.
13. R. Ichise, H. Takeda, and S. Honiden. Integrating multiple internet directories by instance-based learning. In *IJCAI*, pages 22–30, 2003.
14. N. Jian, W. Hu, G. Cheng, and Y. Qu. FalconAO: Aligning ontologies with Falcon. In *Proceedings of K-CAP 2005 Workshop on Integrating Ontologies*, 2005.
15. B. Magnini, M. Speranza, and C. Girardi. A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques. In *Proceedings of COLING-2004*, August 23 - 27, 2004.
16. Natasha Noy and Heiner Stuckenschmidt. Ontology alignment: An annotated bibliography. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum (IBFI), Schloss Dagstuhl, Germany, 2005. <http://drops.dagstuhl.de/opus/volltexte/2005/48> [date of citation: 2005-01-01].