CALCULATING SEMANTIC SIMILARITY BETWEEN COMPUTER-UNDERSTANDABLE DESCRIPTORS OF SCIENTIFIC RESEARCH

Steven B. Kraines and Weisen Guo
University of Tokyo, 5-1-5 Kashiwa-No-Ha, Kashiwa-Shi, Chiba Prefecture, Japan

Keywords: Knowledge Representation, Semantic Matching, Semantic Similarity, Logic Inference.

Abstract: If researchers created computer-understandable descriptors as part of the process of authoring journal articles and other expert knowledge resources, intelligent computer-aided matching and searching applications that are critical for addressing complex and large-scale problems in society could be realized. The EKOSS system enables knowledge experts to create computer-understandable descriptors of their knowledge resources using description logics ontologies as formal knowledge representation languages. The descriptors, called semantic statements, are authored as description logic ABoxes in reference to a shared domain ontology in the form of a TBox. Reasoners using logic-based inference can then measure the semantic similarity between semantic statements, which can be applied in knowledge searching, mining and integration applications. A method for semantic matching that uses logic inference based on a DL ontology TBox to increase both the precision and recall of matching descriptors created as ABoxes is described, and the accuracy of the method compared to matching without logic inference is analyzed between a set of 15 semantic statements created using EKOSS to describe research articles related to sustainability science.

1 INTRODUCTION

Integration of knowledge from a wide range of academic and non-academic domains is needed to address the complex problems of today's society, e.g. related to achievement of sustainable societies (Takeuchi and Komiyama, 2006). However, just coming to grips with the different terminologies used in different disciplines, e.g. resolving different usages of the same term, is difficult (Allenby, 2006). If the accumulating knowledge resources remain disconnected, then soon it will be impossible to uncover the potential interrelationships and structure of all of that knowledge, which is necessary in order to solve the problems that contemporary science attempts to address (Lane and Bertuzzi, 2011).

Information technologies can be used to generate networks of expert knowledge related to specific areas such as global sustainability and bioscience (Neumann and Prusak, 2007; Cahlik, 2000). Some of these studies include semantic relationships by using automated natural language processing (NLP) techniques, such as keyword extraction (Kajikawa et al., 2007). However, even the most advanced NLP techniques today, such as relationship extraction, cannot determine meaningful relationships between keywords with high accuracy (Erhardt et al., 2006).

Technologies that enable creators of knowledge resources to provide computer interpretable descriptors of their resources themselves, rather than relying on third-party annotators or automated computer "bots", could make computer-aided knowledge sharing more effective (Gerstein et al., 2007; Uren et al., 2006; Power, 2009). In particular, technologies emerging in the context of the Semantic Web, such as ontologies, could be utilized to create an interactive knowledge sharing platform that would act as a forum for exchanging and integrating different forms of scientific knowledge related to a wide range of social issues (Allenby, 2006; Berners-Lee and Hendler, 2001; for examples see Kraines et al., 2005; Davis et al., 2009; Kumazawa et al., 2009).

If expert scientific knowledge resources, such as research articles or research project reports, were accompanied with highly accurate and semantically rich descriptors that can be interpreted by a computer, then inference and reasoning technologies

B. Kraines S. and Guo W..
CALCULATING SEMANTIC SIMILARITY BETWEEN COMPUTER-UNDERSTANDABLE DESCRIPTORS OF SCIENTIFIC RESEARCH.
DOI: 10.5220/0003637001460151
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2011), pages 146-151
ISBN: 978-989-8425-81-2

could be used to provide a wide range of knowledge processing services (Power, 2009; Alani et al., 2005; Hess and Schliedera, 2006). For example, by mapping the concepts contained in ontologies that have been constructed as knowledge models for different domains of knowledge, it should be possible to translate the concepts and relationships expressed in the descriptors between different domains of knowledge, e.g. between chemical engineering and macroeconomics.

Any third party effort, e.g. by a group of professional curators, to create such descriptors for all published research articles could not keep up with the rate of scientific publication (Attwood et al., 2009). However, knowledge processing based on computer-understandable descriptors authored by humans could be made sustainable by providing incentives for knowledge experts such as researchers and policy makers to author the descriptors for their own knowledge resources, perhaps as a part of the process of submitting research articles or project reports. Once the larger community is engaged in creating such descriptors of their expertise, then the knowledge processing based on those descriptors could be scaled up to the size of that community (DeRose et al., 2007; Ceol et al., 2008).

EKOSS is a web-based platform that supports computer-mediated sharing and integration of expert knowledge resources based on computer-understandable descriptors that are authored by human knowledge creators (Kraines et al., 2006). The TBox of a description logics ontology provides a simplified, unambiguous language for describing expert scientific knowledge with semantics that can be interpreted accurately by a computer reasoning engine. Expert knowledge is described in the form of ABoxes that instantiate the TBox. Those ABoxes, called "semantic statements" in EKOSS, are "computer-understandable" in that a computer can use logical inference and background knowledge encoded in the ontology to derive new "understanding" from a semantic statement. Each semantic statement is made of one or more triples consisting of a two ontology class instances and a typed, directed property between them.

Here, we describe work to develop and test a method for computing the semantic similarities between a set of research articles based on semantic statements that have been created for those articles. By reasoning about semantic statements with logical inference, we can identify similarities between research articles that "tell similar stories" but do not have any bibliographic evidence for similarity, such as co-citation. Thus, in comparison to conventional social networks that describe "who knows who", we aim to discover knowledge networks of "who should know who" because their work is similar in a meaningful way (Neumann and Prusak, 2007).

In Section 2, we describe the method for calculating the semantic similarity between two statements, and the process we used to create the gold standard for evaluating the calculated semantic similarities. In section 3, we present the results of the analysis of semantic similarity calculations using several different levels of inference. We discuss the results and conclude the paper in sections 4 and 5.

2 METHODS

Our hypothesis is that given a set of semantic statements of knowledge resources authored by the human creators of those resources, we can find more accurate and more "interesting" matches between knowledge resources than by using conventional matching techniques. To test this hypothesis, we examine the effectiveness of using logical inference to find which pairs of semantic statements, each of which describes the research presented in a single research article, have the highest semantic similarity. In the following sections, we describe the method for calculating the semantic similarity between a set of semantic statements, and the gold standard we have created to evaluate the matching results.

2.1 Semantic Statement Matching

In order to study the different kinds of semantic matching techniques described in the previous section, we have developed a semantic matching tool that supports three of the basic types of matching described by Guo and Kraines (2008): matching of classes only, matching of triples without logical inference, and matching using DL inference. The process flow for the matching tool is illustrated in figure 1. All matching tasks take a set of semantic statements together with a complete set of matching options as inputs, and they output a list of matching results giving the calculated matching score between each pair of semantic statements together with the specific bindings between triples or instances, depending on whether or not triples are used in the semantic matching. In addition to the matching type, the matching options include class and property generalization rules and inclusion of property inverses and symmetry for triple-based matching.

The semantic matching tool generates a search query set from each semantic statement as follows.

First, the semantic statement is decomposed into a set of atomic search queries, which are all of the classes of the instances in the statement in the case of "class-based" matching and all of the semantic triples in the statement otherwise. Because queries are matched using class and property taxonomies, queries containing general classes and properties will match with many semantic statements, but queries with specific classes and properties will often not find any matches. In order to compensate for this difference, we provide users with the option to specify class and property generalization rules. These are essentially lists of classes (properties) that are to be substituted for any subclasses (subproperties) that occur in the atomic search queries. For example, the analysis described here uses property generalization rules that include the property "has participant". Therefore, all of the properties in all search queries that are subproperties of the property "has participant", such as "produces", "consumes" and "has actor", are replaced with "has participant". The result of this process is a set of atomic search queries that represent the different semantic assertions contained in the original semantic statement at a particular level of semantic specificity that is specified by the user.

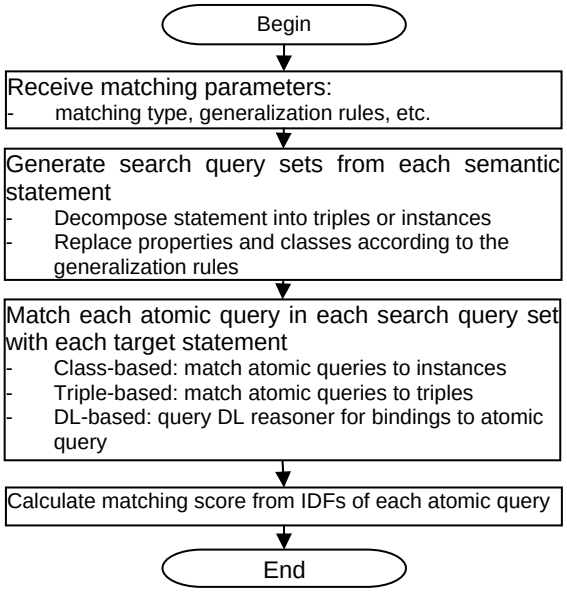

Figure 1: Flow diagram of the semantic matching algorithm.

After the search query sets are generated, each atomic search query in each search query set is matched with all of the original semantic statements, which we call "target statements". In the case of "class-based" matching, each atomic search query is just a single (possibly generalized) class; property information is ignored. Consequently, a match is recorded between the atomic search query and each instance in the target statement whose class is the same as or a subclass of the class given by the query.

For "triple-based" matching, each atomic search query is a semantic triple comprised of a directed property from a domain class to a range class. Each atomic search query is matched with each semantic triple in each target statement. A match occurs if the class of the domain instance, the class of the range instance and the property of the triple from the target statement are all equal to or subsumed by the respective classes and properties in the atomic search query. If the user has included the matching option "use property inverses" and the property in the atomic search query has an inverse, then the inverse property is substituted into the search query and the classes of the domain and range are reversed. The modified atomic search query is then matched once again with all of the triples in the target statement. Similarly, if the user has included "use property symmetry" and the property is declared in the ontology to be symmetric, then the classes of the domain and range are reversed and the modified atomic search query is matched again with all of the triples in the target statement.

In "DL-based" matching, once again each atomic search query is a semantic triple. First, all instances and properties of one target statement are loaded into the knowledge base of the DL reasoner (we use RacerPro here) as the ABox, together with the ontology which comprises the TBox. Then each atomic search query is evaluated by the reasoner against the knowledge base. If the reasoner finds an answer set to the query, then all pairs of instances in the ABox that can be bound to the class variables in the search query are recorded. The ABox of the knowledge base is then cleared, and the next target statement is loaded into the knowledge base.

In both "triple-based" and "DL-based" semantic matching, multiple pairs of instances in the target statement may match with a particular atomic search query. Also, it is possible that more than one search query may match with a particular pair of instances.

To obtain the score of a match between a search query set and a target statement, we calculate weights for each atomic search query using inverse document frequency (IDF) (Spark Jones, 1972):

$$\text{weight of atomic query} = \ln\left[\frac{\text{(total \# of statements)}}{\text{(\# of statements having at least 1 match with the atomic query)}}\right] \quad (1)$$

The matching score is then just the normalized sum of the weights of the matching atomic queries:

$$\text{score}(s,t) = \frac{\sum_{j=1}^{m}(\text{weight of matched atomic query}_j)}{\sum_{i=1}^{n}(\text{weight of atomic query}_i)} \quad (2)$$

Where n is the total number of atomic queries in the search query set s, and m is the number of atomic queries that have at least one match in target statement t.

2.2 Creating the Gold Standard

In order to evaluate the precision and recall of the matching results using each of the semantic matching techniques, we have created a gold standard that gives the "correct" matches between a set of 15 semantic statements that were created using the EKOSS system to describe research articles on topics related to sustainability science. The semantic statements were authored using the SCINTENG ontology implemented in OWL-DL, which makes extensive use of the logical constructs provided in the DL framework such as domain and range restrictions on properties, logical characteristics of properties such as being transitive or functional, and universal, existential, and cardinal restrictions on classes (Kraines and Guo, 2011). All of these constructs can be used for semantic inference.

We have chosen to focus on a small set of the semantic statements in order to be able to thoroughly investigate all of the matching results. The 15 semantic statements were selected for research articles that are recent, published in internationally recognized journals, and are representative of the coverage of the SCINTENG ontology: five articles focus on experimental studies in material science, four articles focus on modeling studies of energy devices, three articles focus on studies of natural or agrarian ecosystems, and three articles focus on analyses of economic systems. Each semantic statement contains on average about 40 semantic triples, so there are over 500 triples, not including the triples in the TBox and the triples obtained through logic inference.

We created the gold standard by examining the actual semantic similarity between pairs of semantic statements that had non-zero scores when DL inference was used for semantic matching. This set of matches necessarily includes all of the matches generated using triples, both with and without property inverses and symmetry. However, many of the matches generated using class matching will not be in this set, and these will all be treated as negative matches. We discuss this issue in section 4.

Figure 2 shows the results of matching the 15 research articles using DL inference. Each of the articles appears on both axes, so the diagonal consists of 100% matches of the semantic statement for a research article with itself. The matrix is not diagonal, however, because the semantic matching techniques are based on logical inference matches from the search query sets (on the X axis) to the target statements (on the Y axis).

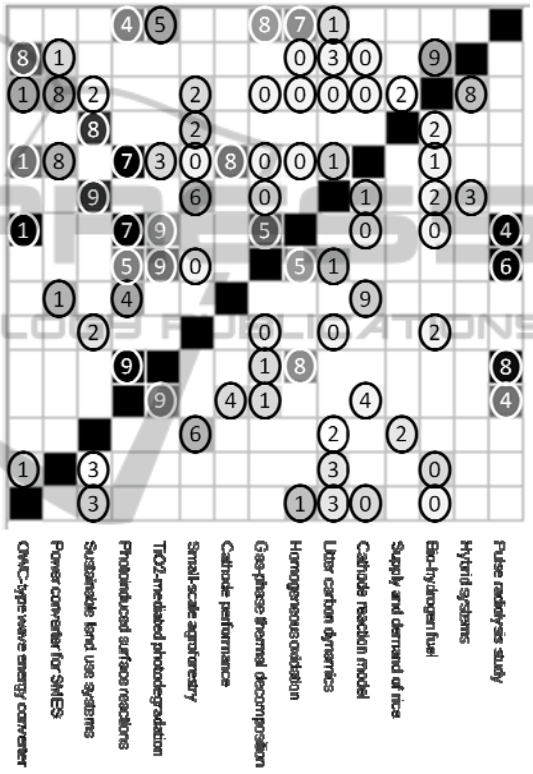

Figure 2: Results of matching semantic statements using DL inference. Labels on the X axis are short versions of the article titles. Target statements are on the Y-axis and search query sets for each of the semantic statements are on the X-axis. Shades of gray show match scores: a white cell is a zero match and a black cell is a match with a score of 20% or more. The gold standard scores are shown as circled numbers from 0 to 9 overlaying the shaded cells. Black circles are used on cells where the actual matching score using DL inference was less than 8%, and white circles are used where the match score was 8% or more.

We examined each of the 81 non-zero matches off the diagonal to evaluate the actual semantic similarity. There are 210 pairs of different semantic statements, so this gives a coverage of 39%. We manually assigned a semantic similarity score to each of the matches based on our knowledge of the

matching domain, as shown in figure 2. The occasional disagreements between the match scores and the gold standard scores, e.g. the black cell in the first column that is labelled with a "1", are indication that even matching using DL inference does not give perfectly accurate matches.

3 RESULTS

We ran the semantic matching process using class-based matching (classes), triple-based matching without inverses or symmetry (triples), triple-based matching with both inverses or symmetry (triples+), and DL inference (DL). All of the matching techniques include inference over class and property hierarchies defined in the ontology. We used a set of property generalization rules but no class generalization rules for all of the semantic matching.

To calculate precision and recall, we treat all of the matches in the gold standard having scores greater than or equal to 5 as true positives. We calculated PR curves by adjusting the cut off for the semantic matching results. Only 9 matches were greater than 15% in the DL semantic matching case, as shown in figure 2, so we have evaluated the number of true and false positives and negatives for cut off values of 1%, 2%, 5%, 8%, and 10%.

The resulting PR curves are shown in figure 3. Except for the last point for the 10% cutoff, matching with DL inference outperformed all other matching techniques. While class matching tended to have high recall, the large number of matches that did not fulfil our criteria for the gold standard (at least one matching triple) meant very low values for precision. Even at the 10% cut off, the precision for the class matching case was just 60%.

The PR value for the DL inference case at the 10% cut off is clearly a low performance result, having both lower precision and lower recall than the previous cutoff at 8%. The reason for the simultaneous decrease in precision and recall is as follows. The number of true positives decreased from 18 to 13 when the cutoff was raised from 8% to 10%. However, the cut off did not actually result in the removal of any false positives because already at 8% cutoff there were only 5 false positives. A decrease in true positives with no change in false positives resulted in a decrease in both precision and recall. In comparison, the triples+ case at 8% cutoff had one less false positive than the DL inference case at 10% cutoff with the same number of true positives, so although the recall is the same, the precision is slightly better for the triples + case.

The simultaneous decrease in precision and recall for the triples case from 5% cutoff to 10% cutoff occur for reasons similar to the DL inference case from 8% cutoff to 10% cutoff.

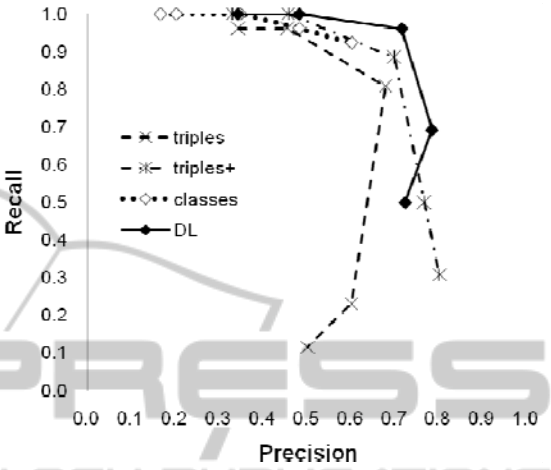

Figure 3: PR curves for the four semantic matching runs described in the text.

4 DISCUSSION

The approach we have used for creating the gold standard favors the DL matching results. Therefore, the comparison of PR curves in the previous section should be viewed as a measure of the degree to which simpler forms of semantic matching can reproduce the results given by the DL inference based semantic matching. However, we believe that this is a useful evaluation because it is reasonable to think that the manual (rather than computer-extracted) addition of rich semantic information should result in more semantically accurate matches.

Ontologies formalized in logic have well known limitations in regard to expression of fuzzy concepts and uncertainty. Furthermore, the logical formalisms provided by OWL-DL do not allow us to express the propagation of a relationship from one property to another (Horrocks et al., 2006). For example, we might want to infer that if a particular instance of activity, such as "driving", is identified to be located in a particular city "Tokyo" and have a particular participant "diesel truck", then we know that the "diesel truck" has location "Tokyo" (at least for the duration of the activity). This function is supported in the new OWL2 protocol (Grau et al., 2008). The propagation function could also be implemented by defining Horn clause rules that create new properties

every time a propagating combination is detected. For the example given above, the Horn clause would have the head "A has location B and A has participant C" and the implication "C has location B". We plan to examine the effect of adding rules such as this on semantic matching in future work.

5 CONCLUSIONS

We have described a method and tool for matching semantic statements that represent the expert knowledge reported in research articles based on a DL ontology. The precision and recall of matching using DL inference versus matching triples or classes directly without inference were measured using a gold standard prepared specifically for this study. The results indicate that the non-inference matching techniques were significantly less accurate than matching with DL inference.

ACKNOWLEDGEMENTS

This work has been supported by funding from the Japan Office of the Alliance for Global Sustainability and the Office of the President of the University of Tokyo.

REFERENCES

Alani, H., Kalfoglou, Y., O'Hara, K., Shadbolt, N., 2005. Towards a Killer App for the Semantic Web. *ISWC 2005 LNCS*, 3729, pp. 829-843.

Allenby, B., 2006. The ontologies of industrial ecology? *Progress in Industrial Ecology*, 3(1): 28-40.

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., Thorne, D., 2009. Calling International Rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, 242: 317-333.

Berners-Lee, T., Hendler, J., 2001. Publishing on the Semantic Web. *Nature*, 410: 1023—1024.

Cahlik, T., 2000. Comparison of the maps of science. *Scientometrics*, 49: 373-387.

Ceol, A., Chatr-Aryamontri, A., Licata, L., Cesareni, G., 2008. Linking Entries in Protein Interaction Database to Structured Text: the FEBS Letters Experiment. *FEBS letters*, 582(8), 1171-1177.

Davis, C., Nikolic, I., Dijkema, G. P. J., 2009. Integration of Life Cycle Assessment Into Agent-Based Modeling. *J. Industrial Ecology*, 13: 306-325.

DeRose, P., Shen, W., Chen, F., Doan, A., Ramakrishnan, R., 2007. Building structured web community portals: a top-down, compositional, and incremental approach. In *VLDB '07: Proc 33rd Intl Conf on very large data bases*, Vienna, Austria, pp. 399—410.

Erhardt, R. A-A., Schneider, R., Blaschke, C., 2006. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 11(7-8), 315-325.

Gerstein, M., Seringhaus, M., Fields, S., 2007. Structured digital abstract makes text mining easy. *Nature*, 447: 142.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U., 2008. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web* 6(4): 309—322.

Guo, W., Kraines, S. B., 2008. Explicit scientific knowledge comparison based on semantic description matching. *Annual meeting of the ASIST 2008*, Columbus, Ohio.

Hess, C., Schliedera, C., 2006. Ontology-based verification of core model conformity in conceptual modeling. *Comp, Environ Urban Sys*, 30(5):543-561.

Horrocks, I., Kutz, O., Sattler, U., 2006. The even more irresistible SROIQ. In *KR*, AAAI Press, pp: 57-67.

Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., Komiyama, H., 2007. Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, 2(2): 221—231.

Kraines, S. B., Guo, W., 2011. A system for ontology-based sharing of expert knowledge in sustainability science. *Data Science Journal*, 9: 107—123.

Kraines, S. B., Batres, R., Koyama, M., Wallace, D. R., Komiyama, H., 2005. Internet-based integrated environmental assessment: using ontologies to share computational models. *J. Industrial Ecology*, 9: 31-50.

Kraines, S. B., Guo, W., Kemper, B., Nakamura, Y., 2006. EKOSS: A knowledge-user centered approach to knowledge sharing, discovery, and integration on the Semantic Web. *ISWC 2006 LNCS*, 4273: 833—2091.

Kumazawa, T., Saito, O., Kozaki, K., Matsui, T., Mizoguchi, R., 2009. Toward knowledge structuring of Sustainability Science based on ontology engineering. *Sustainability Science*, 4(1):99—116.

Lane, J., Bertuzzi, S., 2011. Measuring the results of science investments. *Science*, 331: 678—680.

Neumann, E., Prusak, L., 2007. Knowledge networks in the age of the Semantic Web. *Briefings in Bioinfomatics*, 8 (3):141-149.

Power, R., 2009. Towards a generation-based semantic web authoring tool. *In ENLG '09: Proc. 12th European Workshop on Natural Language Generation*, Athens, Greece, pp. 9-15.

Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1): 11—21.

Takeuchi, K., Komiyama, H., 2006. Sustainability science: building a new discipline. *Sustainability Sci*, 1(1): 1-6.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F., 2006. Semantic annotation for knowledge management: requirements and a survey of the state of the art. *Journal of Web Semantics*, 4 (1): 14—28.