Semantic Similarity: A Key to Ontology Alignment

Valerie Cross

Computer Science and Software Engineering Department
Miami University, Oxford, OH 45056
crossv@muohio.edu

Abstract. Many approaches to measure the similarity between concepts that exist in two different ontologies are used in the matchers of ontology alignment systems. These matchers belong to various categories depending on the context of the similarity measurement, such as lexical, structural, or extensional matchers. This paper presents a review of various forms of semantic similarity measures. Then it examines cross-ontological semantic similarity and how various OA systems have used these along with traditional semantic similarity measure on background knowledge sources. The use of mediating ontologies in ontology alignment also may incorporate the use of semantic similarity.

Keywords: Semantic similarity, ontological similarity, cross-ontological similarity, ontology alignment, information content, mediating ontology.

1 Introduction

Similarity measurement, an important notion to compare two different objects, determines how well they agree or match each other. Ambiguity exists in the meaning of the word similarity because of its diverse use in many contexts such as biology, statistics, and psychology. The term semantic similarity has been used to refer to measures between lexical words. Natural language processing applications, such as word sense disambiguation, text summarization, annotation, and information extraction and retrieval have used numerous such measures [Budanitsky, 1999]. The growth of the Semantic Web and the explosion of ontologies, the key knowledge representation model for the Semantic Web, has renewed interest in measuring similarity. In this paper, the context is ontology alignment (OA). An ontological similarity measure is a special kind of semantic similarity measure that uses the structuring relationships between concepts in an ontology to determine a degree of similarity between those concepts. Semantic similarity is used to include similarity measures that use an ontology's structure or external knowledge sources to determine similarity between entities within one ontology or between two different ontologies. They have become a key to aligning ontologies in more sophisticated domains such as biomedical. Most of the better performing current OA systems use some background knowledge source or mediating ontology and semantic similarity measures to address where simple string matching or other OA similarity measures fail at producing mappings.

2 Brief Historical Overview of Semantic Similarity

About 30 years ago, distance in a semantic network was simply the number of edges in the path between two concepts [Rada 1989]. This distance is not sensitive to the depth of the edge in the network. It assumed a weight of 1 for all edges regardless of their hierarchical depth. This weakness has been the focus in [Wu and Palmer, 1994] and [Leacock and Chodorow, 1998]. Several pieces of information about the edge are used in determining the weight: its depth, density of edges at that depth, and strength of connection between parent and child nodes. Edge weights are reduced farther down the hierarchy and in dense parts of the graph where edges represent smaller distances. Various methods are used to normalize and convert distances into semantic similarity.

Another approach is based on the insight that conceptual similarity between two ontology concepts is related to the amount of information they share [Resnik, 1995]. The more shared information, the more similar they are. Information content (IC) is a measure of how specific a concept is in a given ontology, i.e., the more specific, the higher its IC. In [Resnik, 1995], IC is calculated relative to a selected corpus and uses a logarithmic function of the probability of the concept determined by its frequency of occurrence in the corpus. Another method [Seco et al., 2004] uses the ontology structure itself as a statistical resource; it needs no external corpus. An ontology is assumed to be organized in a meaningful, structured way; the more descendants a concept has, the less information it expresses. A concept's IC is a logarithmic function of its number of descendants and the maximum number of concepts in the ontology. In [Resnick, 1995] similarity between two concepts is a function of the IC of the most specific ancestor to both concepts in the hierarchy, i.e., their shared information. In [Jiang and Conrath, 1997] [Lin, 1998] semantic similarity is determined as a function of both the IC of each individual concept with the amount of shared IC of the two concepts.

Semantic similarity measures can also be determined from a set of features for each concept. The parameterized ratio model of similarity [Tversky 1977] uses the ratio between the cardinality of the intersection of their two sets and the sum of the cardinality of this intersection and their set symmetric difference. Parameters on the set differences depend on which concept is to be emphasized as the reference concept. The Jaccard similarity measure [Jaccard, 1901] weights both set differences by 1 to produce the ratio of the intersection cardinality over the union cardinality of the two sets. A detailed discussion of semantic similarity is presented in [Cross, 2009].

3 Measuring Concept Similarity between Different Ontologies

OA research has typically focused on finding equivalences between two concepts in different ontologies. OA techniques vary greatly depending both on what features, i.e., the schema, its instances, etc. and on what background knowledge sources such as vocabularies or other ontologies, already existing alignments, free text and search engines are used to determine the mappings [Shvaiko and Euzenat, 2013].

Early on string edit distances between the concept labels were used by OA systems. Later research developed more sophisticated similarity for use in OA matchers. The foundation for many of OA matchers can be found in [Rodriguez and Egenhofer, 2003] which relies on Tversky’s parameterized ratio model of similarity. These matchers belong to various categories depending on the context of the similarity measurement, such as lexical, structural, or extensional matchers [Sabou et al., 2008]. Determining ontological similarity between entity classes *a* and *b* uses a matching process over several different sets: synonym sets (w), semantic neighborhoods (n), and distinguishing features (u). Distinguishing features are further classified into parts, functions and attributes. The similarity formula is the same for each set and given as

$$S_{\text{set-type}}(a,b)=[|\ A\cap B|]\ /\ [|A\cap B| + \alpha(a,b)\ |A\ -\ B| + (1-\alpha(a,b))\ |B\ -\ A|] \text{ for } 0 \leq \alpha \leq 1$$

where *A* and *B* are description sets for entity classes *a* and *b* and are specified by the set-type = w, u, and n. The only variation to Tversky’s parameterized ratio model is the setting of *α* which determines the importance of the non-common characteristics between *a* and *b*. The *α* parameter is simply determined from the depth of the entities within their respective ontologies. The parameter *α* is set to the ratio of the depth of *a* over the sum of the depths of *a* and *b*. Using *α* parameter gives priority to the more salient entity, i.e., the one with the greater depth. The overall similarity assessment of *a* and *b* is based on a weighted aggregation of the individual matching components S_w, S_n, and S_u. Aggregation weights depend on the assessment of the importance of each semantic component of the ontologies. Many current OA systems use this approach with their matchers for various components or features of entities and then weight the similarity of the individual matcher results either manually or through learning methods. Some OA systems employ methods to automatically weight the matchers based on an overall assessment of ontology similarity over these various kinds of sets [Pirro and Talia, 2010] [Wang et al 2010].

4 Semantic Similarity in OA Systems

OA systems have used various semantic similarity measures in a single ontology viewed as background knowledge source such as a thesaurus or mediating ontology. Concepts from the source and target ontologies are mapped into the background knowledge source. The following OA systems presented in order of their appearance in the research literature have been described in [Cross et al., 2012] in more detail. Recent OA systems use variations of semantic similarity seen in these earlier systems.

OLA [Euzenat and Valtchev, 2003] uses the lexical similarity between a pair of concept identifiers based on a set of terms for each identifier. Pairs of terms for each identifier are located in WordNet. Their term similarity is calculated using a modified Wu-Palmer measure. An aggregated similarity of proximity over all pairs of terms is calculated. **iMapper** [Su et al., 2004] increases the similarity between two concepts based on their distance in WordNet. The concepts are found in WordNet using their labels. If two terms belong to the same WordNet synset, the path distance is 1. Otherwise, the path length from each sense of one to each sense of the other is found (Rada distance). The minimum of these lengths is the semantic distance between them. If no path is found between them, they are unrelated and their similarity is not increased. **SAMBOdtf** [Lambrix et al., 2008] has the WordNet matcher that finds synonyms for concepts. If the concepts are not synonyms to the same WordNet concept, the hypernym relationships between concepts is used to determine their similarity. A domain matcher uses the UMLS. If both the source and target concepts are a synonym of the same UMLS concept, then the domain knowledge matcher sets the similarity to 0.99; otherwise the similarity is set to 0. **ASMOV** [Jean-Mary and Kabuka, 2008] checks if strings are not identical for concept labels and if available, uses WordNet or UMLS. Their lexical similarity is set to 0.99 if one label string is a synonym of the other. If one is an antonym of the other, it is set to 0. If neither and both string labels are in WordNet, it is set to Lin semantic similarity measure between the two.

CIDER [Gracia and Mena, 2008] uses a modified version of a sense semantic similarity measure to evaluate similarity between possible senses of a keyword and its synonyms to disambiguate. Semantic similarity in the filtering of mappings is adapted from the PowerMap WordNet based algorithm [Lopez et al. 2006]. The Wu-Palmer measure is used. A directional similarity is used. The validity of a mapping between concepts A and B is determined in both directions, B to A and A to B. The similarity measure is binary and is a 1 if either direction similarity is a 1 and is based on commonality between the synsets of each concept. **UFOme** [Pirro and Talia, 2010] uses a set of matchers; many have been previously developed for numerous OA systems and integrated into UFOme. The strategy predictor creates a mapping strategy by selecting and ordering the matching components. One of its matchers, the WordNet matcher, is similar to ASMOV's matcher. It uses the Lin similarity between synsets of concept terms when they do not map to the identical lexical concept in WordNet.

Using a mediating ontology is similar to using a background knowledge source. The difference if the OA system must use a simple matcher to quickly map source and target concepts to the mediating ontology, typically domain specific. Both the source and target ontologies are efficiently aligned to the mediating ontology O_M to produce a set of mappings M_{SM} and M_{TM}, respectively. In [Gross et al., 2011] [Cruz et al., 2011] a set of mediated mappings M_{ST} is created based on an exact match on the concept in O_M both the source and target concepts map to. These mediated mappings may be used when the OA process has not found direct mappings between the concepts in the two ontologies. The Uberon ontology has been used as the mediating ontology in GOMMA and AgreementMaker. An issue is both source and target concepts must map to the identical concept in the mediating ontology. The Mediating Matcher with Semantic Similarity (MMSS) was added as a new matcher to use semantic similarity measures between the mapped concepts in O_M even if no exact match exists [Cross et al., 2012].

5 Conclusions and Possible Future Directions

Semantic similarity has been reviewed and its important role in the ontology alignment task has been emphasized. Tversky's parameterized ratio model of similarity has been discussed as fundamental to developing similarity measures between concepts in different ontologies [Rodriguez and Egenhofer, 2003]. Although semantic similarity measures are essential to the OA task, more research needs to be done to determine if specific ones have better performance. The OA task should be used as a benchmark for performance evaluations on existing and new measures.

References

1. Budanitsky, A.: Lexical Semantic Relatedness and Its Application in Natural Language Processing, Computer Systems Research Group, Tech Report, University of Toronto (1999)
2. Cross, V., Silwal, P., Morell, D.: Using a Reference Ontology with Semantic Similarity in Ontology Alignment. Proc. of Int. Conf.on Biomedical Ontologies (ICBO), Graz (2012)
3. Cross, V.: Ontological Similarity. In Data Mining in Biomedicine Using Ontologies, Norwood. MA:Artech House, ISBN-13:978-1-59693-370-5, pp. 23-43 (2009)
4. Cruz, I. F., Stroe, C., Caimi, F., Fabiani, A., Pesquita, C., Couto, F. M., Palmonari, M.: Using AgreementMaker to Align Ontologies for OAEI 2011. OM Workshop, ISWC (2011)
7. Euzenat J. and Valtchev, P.: An integrative proximity measure for ontology alignment. Proc.ISWC-2003 Workshop Semantic Information integration, (FL US), pp. 33–38, (2003)
8. Gracia, J., Mena, E.: Ontology Matching with CIDER: Evaluation Report for the OAEI 2008, Proc. 3rd Ontology Matching worshop, Karlruhe (DE), pp140-146 (2008)
9. Gross, A., Hartung, M., Kirsten, T., and Rahm, E.: Mapping Composition for Matching Large Life Science Ontologies. In Proc. of Int. Conf. on Biomedical Ontologies, pp 109–116 (2011)
10. Jaccard, P., Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles, Vol. 37, pp. 547-579 (1901)
11. Jean-Mary, Y.R., Kabuka, M.R.: ASMOV: Results for OAEI 2008, OM Workshop (2008)
12. Jiang J. and Conrath D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the 10th Int. Conf. on Research on Computational Linguistics, Taiwan (1997)
13. Lambrix, P., Tan, He, Liang, Qiang: SAMBO and SAMBOdtf Results for the Ontology Alignment Evaluation Initiative, 3rd International Workshop on Ontology Matching (2008)
14. Leacock, C. and Chodorow M.: Combining local context and WordNet similarity for word sense identification. In WordNet: An Electronic Lexical Database, pp. 265-283, Cambridge, MA: The MIT Press (1998)
15. Lin. D.: An Information-theoretic Definition of Similarity. Proc. 15th International Conference on Machine Learning, Madison, Wisconsin, July 1998, pp. 296-304 (1998)
16. Lopez, V., Sabou, M., and Motta, E.: Powermap: Mapping the real semantic web on the Fly. In Proc. of 5th International Semantic Web Conference, Athens, GA (2006)
17. Pirro, G., Talia D.: UFOme: an ontology mapping system with strategy prediction capabilities. Data Knowl.Eng. 69.5, 444-471 (2010)
18. Rada R, Mili H, Bicknell E, Blettner M: Development and application of a metric on semantic nets. In: IEEE Transaction on Systems, Man, and Cybernetics. 19. pp 17–3 (1989)
19. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research Vol. 11, pp. 95-130 (1999)
20. Rodriguez, M. A., Egenhofer, M. J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering Vol 15, Issue 2(2003): p 442-456 (2003).
21. Sabou, M., d'Aquin, M., Motta, E.: Exploring the Semantic Web as Background Knowledge for Ontology Matching. J. Data Semantics 11: pp. 156-190 (2008)
22. Seco N, Veale T, Hayes J.: An intrinsic information content metric for semantic similarity in WordNet. In: ECAI. pp 1089–1090 (2004)
23. Shvaiko, P. and Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. 25(1): pp. 158-176 (2013)
24. Su, Xiaomeng: Semantic Enrichment for Ontology Mapping. Ph.D. Thesis, Dept. of Computer and Information Science, Norwegian University of Science and Technology (2004)
25. Tversky, A.: Features of Similarity. Psychological Rev., 84, pp. 327--352 (1977)
26. Wu, Z. and Palmer, M: Verb semantics and lexical selection. Proc. 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp 133-138 (1994)