Crowdsourcing Ontology Verification

Jonathan M. Mortensen

Stanford Center for Biomedical Informatics Research
Stanford University, Stanford CA 94305, USA

Abstract. As the scale and complexity of ontologies increases, so too do errors and engineering challenges. It is frequently unclear, however, to what degree extralogical ontology errors negatively affect the application that the ontology underpins. For example, "Shoe SubClassOf Foot" may be correct logically, but not in a human interpretation. Indeed, such errors, not caught by reasoning, are likely to be domain-specific, and thus identifying salient ontology errors requires consideration of the domain. There are both automated and manual methods that provide ontology quality assurance. Nevertheless, these methods do not readily scale as ontology size increases, and do not necessarily identify the most salient extralogical errors. Recently, crowdsourcing has enabled solutions to complex problems that computers alone cannot solve. For instance, human workers can quickly and more accurately identify objects in images at scale. Crowdsourcing presents an opportunity to develop methods for ontology quality assurance that overcome the current limitations of scalability and applicability. In this work, I aim (1) to determine the effect of extralogical ontology errors in an example domain, (2) to develop a scalable framework for crowdsourcing ontology verification that overcomes current ontology Q/A method limitations, and (3) to apply this framework to ontologies in use. I will then evaluate the method itself and also its effect in the context of a specific domain. As an example domain, I will use biomedicine, which applies many large-scale ontologies. Thus, this work will enable scalable quality assurance for extralogical errors in biomedical ontologies.

Terminology

- Error: Extralogical ontology error (i.e., non-logical error than can only be detected by human interpretation)
- Application: A system, method, or application that uses an ontology (e.g., decision support system)
- Salient error: An error that negatively affects an application
- Verification: The process of finding errors

1 Relevance

Ontology Q/A is particularly relevant in biomedicine and healthcare, where ontologies have many applications[3]. For example, researchers in the life sciences use ontologies such as the Gene Ontology to combat the data deluge and integrate data [1]. In the clinic, ontologies such as SNOMED CT underpin Electronic Health Records (EHRs)[2]. The National Center for Biomedical Ontology's (NCBO) BioPortal has more than 350 ontologies as of this writing and many of these ontologies have thousands of concepts

H. Alani et al. (Eds.): ISWC 2013, Part II, LNCS 8219, pp. 448–455, 2013.

and tens of thousands of axioms [21]. Clearly, ontology use is becoming widespread, and accompanying that, the development, size and complexity of biomedical ontologies are increasing rapidly. As ontologies become larger, engineering and maintenance becomes more difficult and more error prone. For example, Rector, Ceusters, and others have identified significant errors in SNOMED CT and the National Cancer Institute Thesaurus (NCIt) [4,25]. Entire journal issues have focused on auditing ontologies and terminologies [7]. With the widespread and growing use of ontologies in healthcare and life sciences (where errors are costly), scalable, applicable ontology quality assurance methods are crucial.

2 Problem Statement

Many investigators believe that having a correct and error-free ontology is desirable. Though, the true effect of extralogical errors most often remains unquantified. An extralogical error might cause an application to fail, or might not even impact that application. For example, MEDLINE, a system that supports literature retrieval, may fail to retrieve meaningful results if the Medical Subject Headings (MeSH) hierarchy does not correctly describe the domain. An incorrectly specified hierarchy (i.e., one that does not correctly reflect the domain) is just one type of extralogical error that may cause an application to fail. Thus, an understanding of the type of errors and their impact on its application is essential and will guide development of methods to reduce those errors.

Current ontology quality assurance (Q/A) methods fall into two groups: (1) those that are computationally driven and (2) those that are expert driven. Computational methods provide objective, reproducible measures of ontology characteristics, like average depth, syntactic correctness, logical consistency, or fan-out, but do not necessarily find salient errors or correlate with ontology quality. Expert-driven methods usually involve application specialists manually applying a predefined guideline or set of properties (e.g., OntoClean) to evaluate an ontology. Manual Q/A, which may indeed identify errors in an ontology, is often is not scalable (in time or cost) for large ontologies such as SNOMED CT, where tagging every concept and entailment with a property is very expensive. Regardless, the errors may have little bearing on the performance of the ontology's application. Thus, in the context of biomedicine and biomedical ontologies, current Q/A methods have two main limitations: scalability and applicability.

3 Related Work

This work complements various ontology debugging methods, such reasoning and justification, which help catch and explain logical errors [29,23,15,18,13,10,28,14]. Centering on extralogical errors, there have been various related efforts to perform ontology Q/A and capture ontology quality. Researchers have developed metrics or quality criteria, including ontoQA, Ontometric, and oQual [6,17,32,8]. Such criteria, while useful for categorizing and describing ontology, do not necessarily find salient errors.

Capturing salient extralogic errors, Welty and Guarino developed OntoClean, a guideline with which experts tag concepts/classes with properties [9]. Vandrecic et al. developed AEON, an automated method that learns and predicts OntoClean properties

[33]. Even so, this method does not consider the context of how the ontology is applied. In another approach, ROMEO [35], focuses on developing a formal specification about the design on the ontology and ensuring that it meets those requirements. However, it orients itself closer to a design process and not finding salient errors.

In this work, I focus on the use of micro-task crowdsourcing, where anonymous, online workers complete small tasks for various incentives (including monetary), to solve challenges related to ontology Q/A. Crowdsourcing is a quite popular method in many fields [11,34,24,12,31]. Researchers in the Semantic Web community have begun to use crowdsourcing for a variety of tasks. For example, Sarasura and colleagues successfully applied crowdsourcing to automated ontology mapping methods to improve performance significantly [27]. Siorpaes et al. developed "games with a purpose" for semantic web engineering tasks [30]. ZenCrowd links concepts and entities using crowdsourcing [5]. Finally, demonstrating the increased research focus on crowdsourcing, the International Semantic Web Conference is holding a workshop on crowdsourcing called CrowdSem.

4 Research Questions

Considering the importance of biomedical ontologies and the need for scalable, effective ontology Q/A, this work focuses on answering three driving questions:

1. What is the effect of ontology errors on biomedical applications?
   The influence of an ontology error on its application is not well understood. Do ontology errors even affect an application? Are these errors significant or frequent? Such questions are useful to answer, particularly when the application is biomedical, a high-cost industry that directly influences health. A better understanding of this issue will motivate and direct ontology engineering and quality assurance.
2. How will we find application-relevant ontology errors at scale?
   If ontology errors do indeed affect an application, finding the salient errors is key. Thus, an effective ontology quality assurance method should find these salient errors. Current methods do not scale, and their ability to find such salient errors, specifically in large biomedical ontologies, remains unexplored.
3. How will we close the loop between method development and ontology application?
   The first two questions highlight an issue that must be addressed: Ontology research, ontology engineering, and ontology use are typically separate tasks. However, I claim that these tasks must be tightly integrated, where feedback from the application will drive both ontology engineering and the development of methods used to engineer ontologies.

5 Hypotheses

Translating the research questions, I intend to test the following hypotheses:

1. Ontology errors in biomedicine are costly (time, money, and health)

2. Crowdsourcing enables scalable detection of salient ontology errors in large biomedical ontologies
3. Quality assurance methods that improve an ontology will also improve the performance of the ontology's application

6 Approach

To address these hypotheses, I will investigate the following aims:

Aim 1 - Determine the Effect of Ontology Errors in Biomedical Applications. Currently, the cost and effect of an ontology errors are not easily characterized. It is paramount to understand how ontology quality influences the quality of whichever application uses them. To achieve this aim, I will determine which ontologies are most heavily used in biomedicine (both in clinical and life sciences). To limit the scope, this aim will not be exhaustive, but focus on archetypal ontologies and applications. For those select ontologies, I will quantify how each is used, the errors it contains, and the impact of those errors on their applications. I will determine if errors in the application are related to the ontology itself. I will use existing evaluations of these applications to guide my search in finding ontology errors and quantifying their impact.

For example, Shah and colleagues developed a novel system that leverages many biomedical ontologies and clinical notes to predict adverse drug events [16]. Furthermore, they have already developed a formal evaluation of this system. Using this evaluation, I would trace the errors in this system to the ontology it leverages, finding any salient errors. Furthermore, I will quantify the cost & effect of these salient errors. In this situation, it would be the healthcare cost (monetary units) of those affected by an adverse event between its discovery and the formal FDA withdrawl from the market.

Deliverable. Cost/error analysis of salient errors in archetypal biomedical ontologies in-use.

Aim 2 - Develop a Scalable Framework for Crowdsourcing Ontology Verification that Overcomes Current Ontology Q/A Method Limitations. Researchers have shown that crowdsourcing, wherein humans complete small tasks for a reward, is a scalable method that solves complex problems faster and more accurately than computers by leveraging the "wisdom of the crowd", where many workers solve a single task[31]. I argue that finding extralogical errors is a complex process that requires humans. Therefore, crowdsourcing is a natural candidate for ontology quality assurance methods. The types of salient errors (e.g., hierarchy errors) from the first aim will guide the specific Q/A task upon which I focus. Considering this, I propose that crowdsourcing will enable ontology verification at scale. It will assist with finding salient errors in the large scale ontologies I studied in the first aim.

I will devise a novel general framework that integrates findings from state-of-the-art crowdsourcing research and tailor this framework to the task of ontology verification. The framework will include entity selection, context provisioning, task generation, spam removal, response aggregation, and response selection. Next, I will tailor this

framework for ontology verification. To restate the above elements in terms of ontology tasks, the framework will perform axiom selection, context provisioning (e.g., provide definitions), task generation (i.e., convert axioms to natural language), spam filtering, response aggregation, and axiom selection (i.e., determine which axioms to change). Spam filtering and response aggregation are issues that span all applications of crowdsourcing, and to complete these tasks I will adapt statistical methods[26]. Figure 1 provides an overview of the framework I will develop. After I develop the framework, I will compare it with previous verification methods, including expert curation. In preliminary work, I have begun to address each of these components individually [19,22]. Furthermore, I have conducted a pilot study in which a simplified crowdsourcing framework recapitulated errors found by Rector et al. [20,25].

Deliverable. Method for crowdsourcing ontology verification focused on identifying salient errors

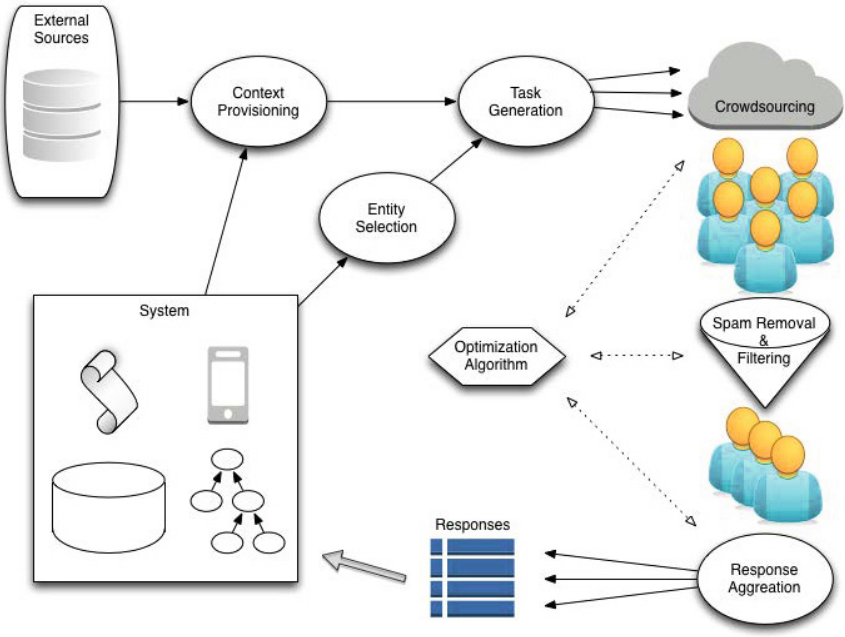

Fig. 1. An overview of the crowdsourcing workflow I propose. Suppose a system exists that leverages an ontology (Lower Left - flow clockwise). To crowdsource components of it, one first selects entities/tasks in a system to crowdsource (Upper Left) . In addition one obtains any context necessary to complete the task either from the system or external sources (Upper Left) . Next, one generates tasks using these entities, submits the tasks to a crowdsourcing platform (Upper Right), filters, and aggregates results (Right). Finally, one updates the entities in the system with those results (Bottom).

Aim 3 - Apply the Crowdsourcing Q/A Method to Biomedical Ontologies In-Use. From the first aim, I will have found archetypal ontologies that play an critical role in biomedical applications and quantified the impact of errors on the application. To understand the impact of my method from the second aim, I will apply the method I developed in the second aim to these ontologies. I will measure how the correction of errors in the ontology affects the performance of the application, comparing with initial results from the first aim. Thus, this aim will provide insight into the effectiveness of the crowdsourcing method to find salient errors and will be the first example of closing the loop between ontology research, engineering, and application.

Deliverable. Audited ontology that not only has fewer errors but also has improved its application

7 Evaluation Plan

Evaluation is a central component to demonstrating that this crowdsourcing method positively affects its application. The evaluation for this work will focus on two areas: methodological and biomedical. First, I will evaluate whether the method I developed in Aim 2 is accurate and performs as well as other techniques. To do so, I will compare this method and others with an expert curated gold standard of ontology errors. Specifically, I will work with biomedical and ontology experts to develop a reference standard of errors in SNOMED CT and the Gene Ontology (GO). Then, I will crowdsource the verification of the ontologies and compare the method's performance with other methods and experts using standard metrics such as sensitivity and specificity. While doing so, I will quantify the monetary cost and time each method requires.

Second, I will evaluate the biomedical impact of this work. In the first aim, I will show the impact of ontology errors on a small set of archetypal biomedical applications in terms of cost and health. In the third aim, I will draw a connection between the crowdsourcing method I develop, and the influence of this method on the biomedical applications in the first aim. Specifically, I will quantify how this method identifies salient errors in applications, thereby reducing cost and negative health impact. As a reference, I will compare other methods' ability to identify salient errors.

8 Reflections

This approach differs from previous ontology Q/A work in both orientation and scalability. Previous work on ontology Q/A has oriented more theoretically. In this work, I focus on the application, ensuring that the method not only finds errors but salient errors that impact the biomedical application. This is an example of the tight integration between research, engineering, and application that is necessary to develop effective methods.

In addition, current ontology Q/A generally requires costly human experts, an unscalable method. However, crowdsourcing, though it does involve humans, is readily scalable because humans need not be experts (as preliminary work shows) and cost is significantly lower. Instead, the expert knowledge arises from the wisdom of the crowd available in crowdsourcing platforms that have thousands of workers who can quickly complete a task for low cost. Therefore, this method is indeed scalable.

Risks
The first potential risk with this proposal is finding that ontology errors in biomedical ontologies have no effect on their biomedical application. This finding is is extremely unlikely given the fundamental role of ontology in biomedicine. Second, salient errors may not occur in common patterns or structures. This irregularity would render the method design in Aim 2 difficult. Based on previous work that identified general error modes, this too seems unlikely.

Acknowledgements. Many thanks to Mark Musen, Matthew Horridge, and Samson Tu for comments. This work is supported in part by Grant GM086587 from the National Institute of General Medical Sciences and by The National Center for Biomedical Ontology, supported by grant HG004028 from the National Human Genome Research Institute and the National Institutes of Health Common Fund. JMM is supported by National Library of Medicine Informatics Training Grant LM007033.

References

1. Blake, J.A., Bult, C.J.: Beyond the data deluge: data integration and bio-ontologies. J. Biomed. Inform. 39(3), 314–320 (2006)
2. Blumenthal, D., Tavenner, M.: The "meaningful use" regulation for electronic health records. New England Journal of Medicine 363(6), 501–504 (2010)
3. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Briefings in Bioinformatics 7(3), 256–274 (2006)
4. Ceusters, W., Smith, B., Goldberg, L.: A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine 44(4), 498 (2005)
5. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: 21st World Wide Web Conference WWW 2012, Lyon, France, pp. 469–478 (2012)
6. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 140–154. Springer, Heidelberg (2006)
7. Geller, J., Perl, Y., Halper, M., Cornet, R.: Special issue on auditing of terminologies. J. Biomed. Inform. 42(3), 407–411 (2009)
8. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Tech. Rep. 5–6, Knowledge Systems Laboratory, Stanford University (1993)
9. Guarino, N., Welty, C.: Evaluating Ontological Decisions with OntoClean. Communications of the ACM 45(2), 61–65 (2002)
10. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 323–338. Springer, Heidelberg (2008)
11. Howe, J.: Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business. Crown Business (2009)
12. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: Proc. of the ACM SIGKDD Workshop on Human Computation, pp. 64–67. ACM (2010)
13. Kalyanpur, A., Parsia, B., Horridge, M., Sirin, E.: Finding all justifications of OWL DL entailments. In: Aberer, K., et al. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 267–280. Springer, Heidelberg (2007)

14. Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B.: Repairing unsatisfiable concepts in OWL ontologies. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 170–184. Springer, Heidelberg (2006)
15. Kalyanpur, A., Parsia, B., Sirin, E., Hendler, J.: Debugging unsatisfiable classes in owl ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 3(4), 268–293 (2005)
16. Lependu, P., et al.: Pharmacovigilance using clinical notes. Clin. Pharmacol. Ther. 93(6), 547–555 (2013)
17. Lozano-Tello, A., Gómez-Pérez, A.: Ontometric: A method to choose the appropriate ontology. Journal of Database Management 2(15), 1–18 (2004)
18. McGuinness, D.L., Borgida, A., Alex, D.D., Borgida, E.: Explaining reasoning in description logics. Tech. rep (1996)
19. Mortensen, J.M., Alexander, P.R., Musen, M.A., Noy, N.F.: Crowdsourcing Ontology Verification. In: International Conference on Biomedical Ontologies (accepted, 2013)
20. Mortensen, J.M., Musen, M.A., Noy, N.F.: Crowdsourcing the Verification of Relationships in Biomedical Ontologies. In: AMIA Annual Symposium (submitted, 2013)
21. Musen, M.A., Noy, N.F., Shah, N.H., Whetzel, P.L., Chute, C.G., Storey, M.A., Smith, B., Team, T.N.: The National Center for Biomedical Ontology. JAMIA 19, 190–195 (2012)
22. Noy, N.F., et al.: Mechanical Turk as an Ontology Engineer? Using Microtasks as a Component of an Ontology Engineering Workflow. In: Web Science (2013)
23. Parsia, B., Sirin, E., Kalyanpur, A.: Debugging OWL ontologies. In: Proceedings of the 14th International Conference on World Wide Web, pp. 633–640. ACM (2005)
24. Quinn, A.J., Bederson, B.B.: Human computation: a survey and taxonomy of a growing field. In: Annual Conference on Human Factors in Computing Systems (CHI 2011), Vancouver, BC, pp. 1403–1412. ACM (2011)
25. Rector, A.L., Brandt, S., Schneider, T.: Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. Journal of the American Medical Informatics Association 18(4), 432–440 (2011)
26. Ruvolo, P., Whitehill, J., Movellan, J.: Exploiting structure in crowdsourcing tasks via latent factor models. Tech. rep (2010)
27. Sarasua, C., Simperl, E., Noy, N.F.: CrowdMAP: Crowdsourcing Ontology Alignment with Microtasks. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 525–541. Springer, Heidelberg (2012)
28. Schlobach, S.: Non-standard reasoning services for the debugging of description logic terminologies, pp. 355–362. Morgan Kaufmann (2003)
29. Schlobach, S., Huang, Z., Cornet, R., Van Harmelen, F.: Debugging incoherent terminologies. Journal of Automated Reasoning 39(3), 317–349 (2007)
30. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. IEEE Intelligent Systems 23(3), 50–60 (2008)
31. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
32. Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P., Aleman-meza, B.: OntoQA: Metric-based ontology quality analysis. In: IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, vol. 9 (2005)
33. Völker, J., Vrandečić, D., Sure, Y., Hotho, A.: AEON–An approach to the automatic evaluation of ontologies. Applied Ontology 3(1), 41–62 (2008)
34. Von Ahn, L., Dabbish, L.: Designing games with a purpose. Communications of the ACM 51(8), 58–67 (2008)
35. Yu, J., Thom, J.A., Tam, A.: Requirements-oriented methodology for evaluating ontologies. Information Systems 34(8), 766–791 (2009)