Ontology Engineering Using Volunteer Labor

Benjamin M Good and Mark D Wilkinson
iCAPTURE Centre for Cardiovascular and Pulmonary Research
The University of British Columbia
St. Paul's Hospital, Vancouver, BC, V6Z 1Y6 Canada
goodb@interchange.ubc.ca , mwilkinson@mrl.ubc.ca

ABSTRACT
We describe an approach designed to reduce the costs of ontology development through the use of untrained, volunteer knowledge engineers. Results are provided from an experiment in which volunteers were asked to judge the correctness of automatically inferred subsumption relationships in the biomedical domain. The experiment indicated that volunteers can be recruited fairly easily but that their attention is difficult to hold, that most do not understand the subsumption relationship without training, and that incorporating learned estimates of trust into voting systems is beneficial to aggregate performance.

Categories and Subject Descriptors
H.5.3 [Group and Organization Interfaces]: *Collaborative computing, Evaluation/methodology, Web-based interaction.*

General Terms
Design, Economics, Experimentation, Human Factors.

Keywords
ontology engineering, knowledge acquisition, semantic web

1. INTRODUCTION
Ontologies are a fundamental component of the incipient Semantic Web. To achieve its visions, ontologies need to be written in Semantic Web compatible languages such as OWL and used to annotate the resources of the Web. However, as with many previous efforts in the domain of artificial intelligence, ontology development faces the problem of the knowledge acquisition bottleneck. Given current approaches, ontology development is a slow, expertise-heavy, labor-intensive, and thus costly enterprise. The work presented here is part of a larger project that seeks to dramatically reduce the costs associated with ontology development by altering the process of knowledge acquisition such that it may be distributed across a very large number of volunteers simultaneously via the Internet.

The process starts with a seed ontology that may be generated automatically or semi-automatically; for example, from text [1], or from a translation of an existing structured resource such as a thesaurus [4]. The putative classes and relations in the inferred ontology are then validated and refined based on answers to questions about them posed to a large pool of volunteers. The simplest form of these questions ask whether or not a given ontological statement is 'true' or 'false'. Each question is posed to multiple volunteers. To make improvements to the ontology, the responses are combined using methods that attempt to incorporate estimates of trust in each volunteer.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

The goal of the work presented here is to estimate how well our system can detect errors in auto-generated statements of the subsumption relationship without any training for the volunteers. The relationships that we use are drawn from the biomedical domain. For example, how well can volunteers (individually or in aggregate) answer questions such as "is a nipple a kind of breast" or "is a lymphocyte a sub-class of a lymphatic system"?

2. Creating an OWL version of MeSH
MeSH, which stands for 'Medical Subject Headings', is the thesaurus used by the United States National Library of Medicine to index the millions of biomedical journal articles described in the MEDLINE database (http://www.nlm.nih.gov/bsd/disted/mesh/index.html).
MeSH has been automatically converted to OWL using a simple, but problematic, mapping from the 'narrower than' thesaural relation to the rdfs:subClassOf relation (http://www.berkeleybop.org/ontologies/). By our estimation, about 40% of the predicted sub-class relations are incorrect. Many are statements of meronymy, as in the nipple-breast example above, but there are many more subtle problems in the mix as well [3]. The experiment described below tests our volunteer-driven system's ability to detect these errors.

3. Experiment
Following from previous work that utilized scientific conferences as settings for focused knowledge capture efforts [2], this experiment took place at the annual meeting of a large research project directed at identifying biomarkers of allograft rejection (http://www.allomark.ubc.ca/). The setting of the meeting made it easy to identify volunteers from the biomedical domain and to provide motivation for their participation in the form of a small prize awarded to the most prolific contributor at the end of the conference.

The volunteers were asked to login to a website and answer a series of questions about subsumption relations from MeSH.owl. These questions were provided in one of two forms: "Is it true that a 'mast_cell' is also a 'connective_tissue_cell'? or "Is it true that all instances of the class 'b-lymphocyte' are also instances of the class 'antibody_producing_cell'?

3.1 Test data
To measure the performance of the volunteer-system on this task, we used a sample of 130 MeSH.owl sub-class relations which we manually labeled as either true or false. The sample relation set was generated by extracting the subgraph of the MeSH term 'immune system' which included all of its parents, all of its subclasses and all of the parents of all of its subclasses. The term 'immune system' was chosen because

the topic of the meeting where the experiment was conducted was closely related to immunology.

4. Results

Over the course of the 2 day experiment, responses from 25 volunteers were recorded. All but two of these were from the 50 attendees of the Biomarkers annual meeting, the others were from external IP addresses. As observed in previous experiments of this nature and displayed in Figure 1, the amount of labor provided per volunteer exhibited a characteristic long-tail distribution with a few volunteers contributing the large majority of the work. Overall, only 5 of the volunteers responded to more than 25% of the questions and only one volunteer responded with an assertion of 'true' or 'false' to more than 90% of the questions in the test set.

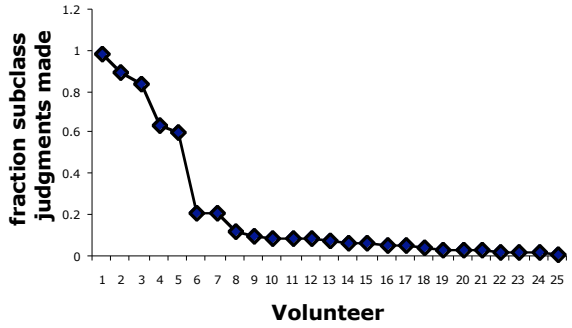

Figure 1. Volume of participation per volunteer.

4.1 Performance of aggregated responses

Five methods were tested for combining the multiple volunteer assertions about each putative MeSH sub-class relation. The simplest method was to take the majority vote for each potential sub-class. The next method weighted each vote based on the time taken between it and the previous vote. These weights on the time spent per vote provide an extremely simple estimate for the amount to 'trust' each vote based on the premise that more time spent might implies more careful thought and thus better performance. As they did not use any training, both of these aggregation methods were evaluated on the entire set of samples.

The other three methods involved machine learning algorithms (1R, Support Vector Machines, and Naive Bayes) that attempted to learn how best to combine the votes using the data collected. If, for example, one voter consistently voted correctly, these algorithms should detect that voter and weight their responses above others. Each row of training/testing data for these methods consisted of the target class (the true/false label for one sub-class relation), the votes for that relation from each volunteer who voted on it (including assertions of 'I don't know'), and the ratio of the true verse false votes gathered from all volunteers for that relation. These methods were evaluated using 10-fold cross-validation over the whole set of samples. Table 1 provides a summary of the results obtained for the various methods. It is problematic to directly compare the results of the cross-validation evaluations to those from the non-learning based approaches, but there does seem to be an advantage gained by the learning methods.

Table 1. Performance on subclass-assessment task using the different aggregation methods. The F-measure is the harmonic mean of precision P and recall R where P = tp/(fp+tp), R = tp/(fn+tp), F-measure = 2*P*R/(P+R)

Aggregation Method	% correct	F-false	F-true
A Single Volunteer	.62	.17	.75
Majority Vote (MV)	.64	.23	.77
MV weighted by time between votes	.63	.47	.71
1R	.71	.56	.78
SVM	.75	.64	.78
Naive Bayes	.75	.64	.81

5. Discussion and future work

Due to the relatively small number of volunteers and number of test cases, the work presented here should be considered as preliminary. However, it did re-iterate previous results indicating that volunteers can be found, but that this kind of task and this kind of incentive strategy are sufficient to keep the attention of only a small fraction of the recruits. It also suggested that learning algorithms can aid in forming intelligent aggregates of multiple voters on ontology evaluation tasks. Future experiments will test the effects of various training mechanisms and incentive strategies for improving individual volunteer performance and will continue to evaluate different approaches to combining the assertions of multiple volunteers of varying levels of knowledge and reliability.

6. ACKNOWLEDGMENTS

Our thanks to Robert Stevens and Andrew Gibson for advice on the design of the experiment and to the volunteers who participated in the study. BMG is supported by an award to the Better Biomarkers in Transplantation project from Genome B.C., in part through Genome Canada. MDW is supported by an award to the iCAPTURE Centre from the Michael Smith Foundation for Health Research. Core laboratory funding provided by the Natural Sciences and Engineering Research Council of Canada. Infrastructure support provided by IBM and SUN.

7. REFERENCES

[1] Cimiano, P., Hotho, A. and Staab, S. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, (2005), 24. 305-339.

[2] Good, B.M., Tranfield, E.M., Tan, P.C., Shehata, M., Singhera, G.K., Gosselink, J. and Wilkinson, M.D., Fast, cheap and out of control: A zero curation model for ontology development. in *Pacific Symposium on Biocomputing*, (Hawaii, USA, 2006), 128-139.

[3] Nelson, Stewart. *Relations in Medical Subject Headings* (2001), http://www.nlm.nih.gov/mesh/meshrels.html

[4] Van Assem, M., Menken, M., Schreiber, G., Wielemaker, J. and Wielinga, B., A Method for Converting Thesauri to RDF/OWL. in *ISWC*, (2004), 17-31.