Evaluation of Intelligent Systems: The High Performance Knowledge Bases and IEEE Standard Upper Ontology Projects

Adam Pease
Teknowledge
2 September 2001

We can consider two ways in which intelligent systems can be analyzed; with respect to a particular task and *a priori*. In this paper we discuss a particular knowledge based system and its performance on a task, as well as the *a priori* metrics which may be applied to ontologies.

The DARPA HPKB (Cohen et al, 1998) project was a large (>$30M) effort to develop large knowledge based systems that would be significantly more competent on a wider range of tasks than the expert systems of the past. In order to motivate rapid development, the program was arranged as a competition between sets of developers. Three *challenge problems* were developed as tests of performance for the systems that were created. These were *battlefield engineering* which produced reasoners which constructed plans for repairing infrastructure such as roads and bridges, *course of action analysis* which produced reasoners which critiques Army plans, and *crisis management* which produced reasoners that gave advice on aspects of international crisis situations.

In each challenge problem, the participants were presented with a set of background knowledge, expressed in English, and a set of test questions that were either expressed in English in the case of battlefield engineering and course of action analysis or in a structured language in the case of crisis management. The participants were provided the opportunity to translate background knowledge by hand or semi-automatically over several months prior to the tests as well as taking sample tests that were "graded" by human experts. The actual tests were conducted in several phases over 2-4 weeks with the results again graded by humans.

While performance was a primary metric that was assessed by number of questions answered correctly, there were additional measures that included the amount of effort expended both before and during the test by project personnel (person-hours). In (Cohen et al, 1999) an analysis was conducted after the fact to see how much knowledge based content was reused from one test to the next. This was a critical measure since one of the purported advantages of knowledge-based systems is reuse of knowledge across tasks. We believe that there is modest support for this assertion. It was found that broadly 1/3 of the most general-purpose upper level content was reused. One-third of the reuse was of "middle-level" content. This is content that addresses a particular area of knowledge such as human social interaction or common-sense knowledge about vehicles, but can be applicable across many domains. One-third of the knowledge needed to answer any particular test question was created at the time of the test.

While one might have expected to have greater reuse, these measures are somewhat conservative since they consider only the appearance of terms or axioms in the trace of the solution to a particular test question. They do not consider the considerable benefit to the knowledge engineer from having a large ontology present that aids in placing, organizing and defining brand new concepts. The authors of (Cohen, et al, 1999) discussed possible metrics for knowledge support but did not reach a set of metrics suitable for publication. More research is needed.

One key aspect of knowledge base performance is speed. The TPTP (Sutcliffe & Suttner) suite is a set of general-purpose theorem prover tests that assess both speed and expressiveness of inference systems. A compromise must often be made on creating expressive knowledge representations in order to reach acceptable speed of inference. Description logic is one class of logics that have good theoretical performance aspects that are traded off for a language that is more limited in expressiveness than full first order logic.

We will now consider *a priori* metrics and guidelines. Some guidelines were introduced in (Pease et al, 2000). A balance must be achieved as to the *fan-out* of concepts. Either the extreme of a deep and narrow or shallow and broad ontology should be avoided. A deep and narrow ontology is likely to have many unnecessary distinctions that could be better represented as properties. A shallow ontology is likely to miss important intermediate concepts that enhance the reusability of an ontology.

Another key attribute of a good ontology is the *compositionality* of concepts. The more that complex notions can be expressed as combinations of functional application and properties, instead of being *compiled* into a single concepts which lacks explicit logical definition, the more reusable the knowledge base is likely to be.

A good ontology for practical computation should also take advantage of the lessons learned from analytical philosophy. Some of those lessons are addressed in (Guarino & Welty, 2001) and include that: all instances of any sub-class are necessarily instances of the super-class, some properties (rigid properties) are ascribed to objects throughout their lifetimes, and some properties (non-rigid properties) are not permanently ascribed to objects, and that the conditions for membership in a class of the ontology should be specified as fully as possible. The IEEE Standard Upper Ontology effort (IEEE, 2001) is attempting to include these lessons in the construction of a general-purpose upper ontology. The challenge of this project is that direct measures of usefulness are not possible since no one particular application is the focus of the effort. The determination of *a priori* metrics is all the more critical. The IEEE SUO currently has two "starter documents" which are described in (Niles & Pease, 2001:1, 2001:2) and (Kent, 2001)

Cohen, P, V. Chaudhri, A. Pease, and R. Schrag (1999). "Does Prior Knowledge Facilitate the Development of Knowledge Based Systems", Proceedings of AAAI-99.

Cohen, P., R. Schrag, Jones, A. Pease, Lin, Starr, Gunning, and Burke (1998). "The DARPA High Performance Knowledge Bases Project", AI Magazine, Vol. 19 No.4, Winter.

IEEE, (2001). IEEE Standard Upper Ontology web site: http://suo.ieee.org.

Kent, R., (2001). A KIF Formalization for the IFF Category Theory Ontology, *in Proceedings of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*.

Niles, I. and Pease, A., (2001). Origins of The IEEE Standard Upper Ontology, *in Proceedings of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*.

Niles, I. and Pease, A., (2001). Towards A Standard Upper Ontology, *in Proceedings of the second conference on Formal Ontology in Information Systems*, to appear.

Sutcliffe, G., and Suttner, C., (2001). The TPTP Problem Library for Automated Theorem Proving web site: http://www.cs.miami.edu/~tptp/

Welty, C., and Guarino, N., (2001). Evaluating Ontological Decisions with OntoClean. Unpublished.