

Building an Open Knowledge Library

Chitta Baral

Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287, USA
chitta@asu.edu

Domain-specific background knowledge is an essential component of many automated reasoning systems including general question-answering systems that reason about some aspect of the world. For example, let us consider answering questions with respect to a paragraph of text. Often it is not possible to appropriately answer the questions with respect to the text without using additional background knowledge. Consider the following text: “John, who always carries his laptop with him, took a flight from Boston to Paris on the morning of Dec 11th.”

Now suppose one were to ask the following questions:

Q1: Where is John on the evening of Dec 12th?

Q2: In which city is John’s laptop on the evening of Dec 10th?

Q3: In which city is John’s laptop on the evening of Dec 12th?

To answer the above question, one needs to use the common-sense or domain knowledge (which is unspecified in the text) that normally when one takes a flight from A to B then (s)he will be at B at the end of the flight; and if X has the physical object Y with him and X is in B then Y is also in B.

Thus if one were to represent the above story in a knowledge representation formalism, to do appropriate common-sense reasoning *he will also need encodings of the necessary background knowledge that is not quite explicitly stated in the story*. He can write those encodings himself, but what if he has access to a ‘library’ of encodings of common-sense knowledge from which he can obtain the necessary encodings with lot less work. The later will definitely make his task a lot easier. Such a library or repository is analogous to the libraries that accompany the compilers of various procedural languages. One can expect that the presence of such a knowledge library will have similar impact in building knowledge bases as program libraries in languages such as Java, C and C++ have had in software development in those languages.

In a sense such a library, CYC, developed by a company, already exists. But only small parts of CYC’s knowledge base is openly accessible, a larger subset accessible with restrictions and legalities, and its reasoning engine is not publicly available. Moreover, CYC, whose semantics is neither publicly available nor vetted by the research community outside of CYC, was a bit ahead of its time. It is only recently that the knowledge representation community has come up with a few good knowledge representation languages with a large body of support structure consisting of theoretical results, implementations and applications.

We would like to kick-start the development and compilation of a knowledge base that will be *built by the community at large, that will be openly available, that will have reasoning engines that are openly available, and that is written in a declarative*

language with precise, published and vetted (by the community) syntax and semantics. In terms of openness, the relation between CYC and our proposal can be considered analogous to the relation between Celera's genome project and the open Human genome project.

The effort to create an open repository will be similar to efforts such as wordnet, verbnet, and framenet, but unlike them our proposed repository will contain formal representations. The availability of these open repositories has had a significant impact on research in many areas including question answering, and our goal is to take this to the next level. In recent years research on knowledge representation and reasoning has come of age with projects such as Digital Aristotle (projecthalo.com) and question answering projects. One of the main bottlenecks in these efforts has been the absence of a publicly and freely available repository of background knowledge. The aim of this panel is to discuss how to go about addressing this shortcoming using the AnsProlog (Programming in logic with answer sets) framework. In particular we will discuss the theoretical problems related to the design of a repository for background knowledge and how to initiate the creation of such a repository.

In Spring 2006 a AAAI symposium on this topic¹ is being organized. In that symposium papers are requested on (a) formalizations (knowledge modules) of background knowledge in specific domains as well as (b) addressing general challenges such as formalizing background knowledge for use by multiple users on multiple reasoning tasks. The symposium is not logic or language specific. It welcomes approaches using any knowledge representation language with formal semantics and possibly an implemented querying mechanism. We hope this panel will encourage the Answer set² community to be active in the proposed open source knowledge library building effort.

Acknowledgements: Part of the above write-up is from the proposal that was submitted to the AAAI Spring Symposium organizers and the call for papers of the AAAI spring symposium. Michael Gelfond, Steve Maiorano, Vladimir Lifschitz, and Sheila McIlraith gave feedbacks and suggestions in the writing of both.

¹ The title of the AAAI symposium is: "Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering". More details on the symposium is available at <http://www.public.asu.edu/~cbaral/aaai06-ss/>.

² It seems the term 'Answer set programming' (or ASP) is understood as SAT like problem solving using logic programs with the answer set semantics. We are open to other querying approaches with respect to such programs. Hence the use of the term 'Answer set community', instead of 'ASP community'.