Evaluating CYC: Preliminary Notes

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

July 9, 2016

1 The Problem with CYC

The problem of endowing computers with commonsense knowledge and the ability to use it in reasoning is considered a central problem for artificial intelligence (AI) but an extraordinarily difficult one (Davis and Marcus 2015). One approach toward achieving this is to have expert knowledge engineers manually encode the vast amounts of knowledge involved. The largest, best-known, and most important project of this kind is the CYC program.

CYC was initiated in 1984 by Douglas Lenat, who has let the project throughout its existence. Its initial proposed methodology was to encode the knowledge in 400 sample articles in a one-volume desk encyclopedia together with all the implicit background knowledge that a reader would need to understand the articles (hence the name) (Lenat, Prakash, & Shepherd, 1985). It was initially planned as a ten-year project, but continues to this day. In the last decade, Cycorp has released steadily increasing portions of the knowledge base for public or research use. The most recent public version, `OpenCyc 4.0,` released in June 2012 contains 239,000 concepts and 2,039,000 facts, mostly taxonomic. `ResearchCyc`, which is available to be licenced for research purposes, contains 500,000 concepts and 5,000,000 facts.

A number of successful applications of CYC to AI tasks have been reported (Conesa, Storey, & Sugumaran, 2010) including web query expansion and refinement (Conesa, Storey, & Sugumaran, 2008), and question answering (Curtis, Matthews, & Baxter, 2005), and intelligence analysis (Forbus, Birnbaum, Wagner, Baker, & Witbrock, 2005).

The dust jacket of (Lenat and Guha 1990) describes Lenat and Guha as "the authors, with distinguished colleagues, of *The Ontological Engineers Handbook*, 2nd edition (1997)." Lenant and Guha (1993) claimed that this was a joke by the publishers which came as a surprise to them, but that they believe that, indeed "something along those lines is exactly what's wanted." Sad to say, not even the first edition of the handbook had appeared by 1997, nor has it appeared to this day. Indeed, Lenat and Guha (1990) was the *last* extensive description of CYC; there has been no follow-up of any kind in the twenty-five years since. And Lenat and Guha (1990) is not very satisfying as a description of CYC; it is not very readable, and focuses on architecture rather than content.

On the whole, it is fair to say that the AI community regards CYC as a very elaborate failure. Domingos (2015 p. 51) characterizes it as "the most notorious failure in the history of AI". Domingos is a researcher in machine learning and has little use for any kind of knowledge-based methods, so his phrasing is certainly harsh, but in our experience, this opinion, more or less, is common even in

the knowledge representation (KR) community.

More importantly, when AI researchers are looking for a resource for real-world knowledge, they tend to turn to anything but CYC. There has been enough work on using Wikipedia as a source to fill a special issue of AI Journal (Hovy, Navigli, and Ponzetto, 2013). WordNet (Miller, 1995) is used ubiquitously. A number of researchers have used ConceptNet (Havasi et al. 2009) , and a few are apparently starting to use Nell (Mitchell et al. 2015). IBM in the last couple of years has gone all out to get anyone and everyone to partner with them on applications that use Watson, and apparently they have had many takers. But only a quite small number of researchers outside of CYC have used CYC. This would seem to be an extraordinary situation; it is as if Matlab and Mathematica existed, but mathematicians who wanted to use computers for symbolic mathematics were trying to apply information extraction techniques to the relevant Wikipedia articles.

This is the primary reason that I feel it is important to do the kind of evaluation and publication that I call for in this note. If the AI research community is unfairly and foolishly neglecting a valuable resource, then they should be made aware of that. If CYC could be made a valuable resource with some effort, then the CYC group, or some other group working with CYC, might want to take that on. If the needs of the AI research community and the resources offered by CYC are largely misaligned — well, that would be worth knowing too, because it would shed light, both on what those needs are and what those resources are, and one would have a clearer idea how better to satisfy those needs and to employ those resources.

The point is emphatically not to give CYC some kind of grade from A+ to D−. Nor is it to compare CYC to some competing system, because there is no other system that is competing, in any meaningful sense; the difference between using CYC and using ConceptNet or WordNet are so large and so obvious that there is no point in doing the comparison. A comparison with Watson would be a little more to the point, but not actually very valuable (and also very difficult). The point of what we are proposing is to clarify to the AI research community what CYC has and has not accomplished, in what ways they could use CYC, and what would be involved in using CYC.

The one thing that is certain about CYC from an external standpoint is that its customers must be reasonably well satisfied, since they have continued to pay the bills for 30 years. However, without information about how they are using CYC, it is hard to know what that means.

2 The Difficulties of Evaluating CYC

Evaluating an AI program is often difficult, but CYC is particularly so, for several interacting reasons.

1. CYC has no specific user-level task that it is addressed to. If one wants to evaluate Google Translate, then one has to come up for measuring the quality of translation — not, of course, an easy thing to do — and then you can test it over some samples. If you want to evaluate Google search, then you run a bunch of testmark searches and use some measure of evaluating the quality of the results. But there is nothing particular that CYC is supposed to do, in this sense, and there is no very natural class of benchmark problems.

   Moreover, the external AI research community does not have a clear idea what are the goals of CYC. Therefore, if I (for example) come up with a set of challenge problems, I really do not know which of these are "fair" tests of CYC, and which are completely irrelevant.

2. Since CYC is hand-crafted, its precision is presumably very close to 1; that is, pretty much all the assertions in CYC are true.[1] Precision is by far the easiest measure to use on a knowledge based system; e.g. if you want to evaluate NELL (Mitchell et. al, 2015) , you download a few hundred "recently learned facts" of which NELL is 99% certain, and you see how many are true, nearly true, false, meaningless, or hopelessly vague. But with CYC, what you need to measure is primarily recall, which is always much harder for general AI programs, because it is hard to know what the space is. For CYC it is especially hard, because the nature of "the space of commonsense knowledge" is exceptionally obscure.

3. By all reports, it is not easy for a newcomer to use CYC, and it is even harder to inspect the internals. For that reason, any evaluation of CYC will inherently depend on the cooperative involvement of the CYC team. By contrast, NELL and Google Translate can to a large extent be fairly evaluated by anyone with a browser.

For that reason, part of the "evaluation" that we propose here would simply be publication on the part of the CYC team; and the remainder of it, which is an evaluation in the more usual sense, would require extensive interaction and involvement of the CYC team. We will deal with these two parts in sequence.

3 What CYC should publish

There is — there has got to be — an enormous amount of knowledge about domain-specific representations, domain-independent representational issues, and meta-knowledge about representation, built into CYC, and possessed by the CYC team. It would therefore would be enormously valuable if the CYC team were to publish:

1. An overview. What is in CYC? What could easily be added to CYC? What would problems of commonsense reasoning seem to difficult to tackle?

   What has been the process of acquiring and encoding knowledge? What are the capabilities of the inference engine?

   The original goal of CYC, as stated in (Lenat, Prakash, and Shepherd, 1985) was to encode the knowledge in 400 sample articles in a one-volume desk encyclopedia, together with all the implicit background knowledge that are reader would need to understand the articles. Is that still the goal, or any part of the goal? Is it close to being achieved? Can it be estimated how much work would be needed to achieve that goal? Or is it too hopelessly ill-defined?

   How, if at all, are internal evaluations of CYC, or of parts of CYC, carried out? What are the results? What kind of testing is done when a new fact or a new feature is added, to make sure that old functionalities are not degraded? Is there any overall control of the architecture of the knowledge base? For example, Conesa, Storey, and Sugumaran (2010) reported that the microtheory taxonomy is 50 levels deep in some places; is this a feature, or is it an unavoidable consequence of something in the CYC design, or it is just an inadvertant piece of carelessness that should be cleaned up?

2. A substantial collection of interesting examples that CYC can actually handle, with variants that will help explain why CYC does not go off on the wrong track. (Having enough rules is only part of what CYC needs; the other part, certainly harder to demonstrate, and probably harder to achieve, is avoiding being distracted by the wrong rules.) The number of examples

[1] One suspects that a possible limiting factor here might be inconsistent use of symbols, leading to a situation in which one asssertion is true under one interpretation of the symbols, and another is true under a slightly different interpretation. A very careful analysis would be required to detect this kind of problem.

that have been published to date is astonishingly small, and most of these are just combinations of taxonomic inference with selectional restrictions.

3. A detailed account of one or more domains that CYC has really done well: Primitives, facts, sample inferences that CYC can carry out.

Moreover, it would be very helpful, and it would, we believe, significantly improve CYC's standing in the AI community, if the CYC team could demonstrate some specific task where CYC really visibly shines. One obvious and important type of applications where one would think that CYC would shine is in ruling out impossible interpretations of natural language text; e.g. impossible suggestions for reference resolution or impossible translations from Google Translate. If CYC could demonstrate that there this can be done to some substantial extent, then, even if it did not result in a translation or an machine reading system that beat the state of the art end-to-end, that in itself would be a noteworthy accomplishment, at least for the research community, if not for the general public.

Another application that seems like it should be a natural fit for CYC would be something like the Allen Institute Science Test Challenge (CITATION). Many of the hard problems in the science challenges are hard because they call on commonsense knowledge. The science problems in (Davis, 2016) are by design even more closely tied to commonsense reasoning.

However, if the CYC team does not like those, or does not think that those are reasonable, then let them design their own task. As Watson demonstrates, passing a self-imposed task can be impressive enough, depending on the task, and in any case it is much better than nothing. At the moment, a person who is asked, "What interesting thing has been done with CYC?" is largely at a loss for an answer.

The obvious point of contrast is Watson. Watson started from the opposite direction, with a very glitzy, publicizable accomplishment of entirely uncertain long-term significance. But it sure caught the eye of the public and of the research community, and now there are hordes of people trying to figure out what interesting or useful applications can be addressed with the technology.

4 External evaluation

Ideally, there are a number of different dimensions on which it would be useful to evaluate CYC. As I stated above, it is difficult for an outsider to know how exactly to structure an evaluation of CYC that would in fact be meaningful, so certainly the description below may well have to be modified, either in detail or in substance. In any case, one would want an external group to prepare a polished version of these kinds of suggestions. If the CYC group wants to argue that some of these suggestions are misdirected, then that's fine, time will not be wasted on those; but the explanation of why these plausible-seeming suggestions cannot be used should be published as part of the evaluation report.

4.1 Coverage

The major question has to do with coverage, actual (i.e. achieved in the current state of CYC) and potential (i.e. easily added to CYC). There are a number of different forms of coverage, and a number of different questions.

- Is the knowledge sufficient to achieve a given inference present in CYC?
- If the knowledge is present, can the CYC inference engine draw the inference? If not, is this because

a. The inference techniques are too weak.

b. The computation time required is excessive.

c. Some invalid answer ends up being preferred.

- If the knowledge is not present, how much work would be needed to add it, in a way that will not significantly degrade finding the answer to some other question, (a) by an experienced CYCist (b) by an external knowledge engineer?

There is, of course, no "representative sample of commonsense inferences" and we have no idea how to go about building such a thing. I would therefore suggest using a number of samples of different characteristics.

- Probably the best known collection is the corpus of textual entailments CITATION. This tends to involve quite shallow inference which do not draw deeply on commonsense knowledge; but it is very large, well studied and validated, and overall has large coverage.

- A while ago, I made a preliminary effort at trying to characterize commonsense inferences needed to understand some short narrative texts.
  http://www.cs.nyu.edu/faculty/davise/annotate/Tacit.html
  The project was put on hold, because there seemed to be no way either to validate or to use the results. However, these might be a starting point for assembling a collection of commonsense inferences typical of those that come up in understanding narrative.

  Another collection of commonsense inferences is the "Commonsense Problem Page"
  http://www-formal.stanford.edu/leora/commonsense
  That is a much more arbitrary collection but again might be a starting point.

- More specialized benchmark collections can be created for specific domains, either of CYC's choice (i.e. domains in which the CYC team feels that CYC is particularly strong) or of the evaluators' choice (domains that the evaluators feel are particularly fundamental). For the latter, there may be collections of problems in the literature that can be used. Preference should be given to "natural" problems rather than cute conundrums like the "Muddy Children's Problem".

4.2 Careful inspection

We suggested above that CYC should publish a detailed account of one chosen domain or a few. In a similar way, a team of outside evaluators should study one or a few specific domains, of either their choice or CYCs. On inspection, what features of the representation seem noteworthy? What gaps seem noteworthy? This analysis would produce a report, not a score.

4.3 Usability

Some evaluation should be made of how easily CYC can be used by an outsider. How easy is it to learn to use CYC for basic tasks? How easy to master the use of CYC? How easy to learn to make acceptable modification for CYC? How easy to create an interface to some application?

Our impression from talking to people is that comparatively little effort has been put into making CYC user-friendly, so this part of the evaluation might be an incentive for CYC to improve this aspect of the system.

5 Actually doing all this

Different parts of this would require very different levels of time and effort from the CYC team and from the outside evaluators. At a rough guess:

- The CYC publications we have suggested would range in length from a journal paper to a shortish monograph, and would presumably take an experienced CYC member a commensurate amount of time, i.e. some number of man-months.
- The extensions we've suggested — creating an impressive application, improving the user-interface if necessary — would need to be done by the CYC team, but it might well be possible to get a research lab, academic or industrial, to partner on it. These are hard to predict, but perhaps a man-year or so for the interface, and a substantial effort for the application.
- The effort of the external evaluation would fall mostly on the external evaluators, but it would be necessary to have a liaison in CYC who could devote serious times — perhaps a couple months of work in total, over a period of a couple of years. Ideally, this would involve a core team of four or five KR experts, each willing to devote a total of several months of work over a period of a couple years, plus perhaps an equal number of consultants with a lesser commitment. For the usability studies, one would want to have talented non-experts e.g. recent college graduates with computer science. This would all require funding; but given the importance of the undertaking, it might well be possible to raise that from government or other sources.

References

J. Conesa, V. Storey, and V. Sugumaran, (2008). "Improving web-query processing through semantic knowledge." *Data and Knowledge Engineering,* **66**(1), 18-34.

J. Conesa, V. Storey, & V. Sugumaran, (2010). "Usability of upper level ontologies: The case of ResearchCyc." *Data and Knowledge Engineering,* **69**, 343-356.

J. Curtis, G. Matthews, & D. Baxter (2005). "On the effective use of Cyc in a question answering system." *IJCAI Workshop on Knowledge and Reasoning for Answering Questions.*

E. Davis (2016). "How to Write Science Questions that are Easy for People and Hard for Computers," *AI Magazine,* Spring issue.

P. Domingos (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake Our World.* Basic Books.

K. Forbus, L. Birnbaum, E. Wagner, J. Baker, & M. Witbrock. (2005). "Analogy, Intelligent IR, and Knowledge Integration for Intelligence Analysis: Situation Tracking and the Whodunit Problem." *International Conference on Intelligence Analysis.*

R.V. Guha and D. Lenat (1993). "Re: CycLing paper reviews", *Artificial Intelligence*, **61**;1 149-174.

C. Havasi, J. Pustjekovsky, R. Speer, & H. Lieberman (2009). "Digital Intuition: Applying Common Sense Using Dimensionality Reduction." *IEEE Intelligent Systems*, **24**(4), 24-35.

E. Hovy, R. Navigli, and S.P. Ponzetto (2013). *Special Issue of Artificial Intelligence Journal on AI and Wikipedia,* Vol. 194.

D. Lenat and R.V. Guha (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Addison-Wesley.

D. Lenat, M. Prakash, and M. Shepherd (1985). "CYC: Using commonsense knowledge to overcome brittleness and knowledge acquisition bottlenecks." *AI Magazine,* **6:**(4) 65-85/

H. Levesque, E. Davis, and L. Morgenstern (2012). "The Winograd Schema Challenge." AAAI-12.

G. Miller (1995). "WordNet:A Lexical Database for English." *Communications of the ACM,* **38:**11, 39-41.

T. Mitchell et al. (2015). "Never Ending Learning" *AAAI-15.*