

Cyc: An Expert Framework for Consensus Knowledge Quantification

Kalev Hannes Leetaru

May 8th, 2002

Cyc: An Expert Framework for Consensus Knowledge Quantification

The difficulty in quantifying consensus knowledge has historically proven an insurmountable barrier to the realization of complete natural language processors and automated reasoning systems. Many ambiguities present in human discourse can only be adequately resolved through such “unspoken” shared knowledge. This knowledge allows for communication compression, as only information that deviates from the expected need be expressed. However, language compression, which is extremely beneficial to humans, is what presents the most difficulty for automated systems. Automated systems do not traditionally have access to the same reserves of consensus knowledge as humans, and thus cannot “fill in the blanks” when processing text intended for a human audience.

To address this information inequality, Douglas Lenat began the Cyc project at the Microelectronics and Computer Technology Corporation in Austin, Texas in 1984. The ultimate goal of the project is to develop a knowledgebase inclusive of all human consensus knowledge. This knowledgebase will then be capable of augmenting the domain-specific knowledgebases of natural language processors and automated reasoning systems, allowing them to incorporate external world knowledge into reasoning tasks in their restricted fields. Instead of being constrained to field-centric knowledge, systems equipped with this consensus knowledge database will be able to use general “common sense” to cope with unexpected or ambiguous inputs.

The project has garnered tremendous publicity and support within the artificial intelligence community in the 18 years since its inception. In January 1995, Cyc was spun off into a private company named Cycorp, based in Austin. Funding has been

provided by numerous private corporations including Apple, Bellcore, DEC, Interval, Kodak, and Microsoft. In addition, the Defense Advanced Research Projects Agency has provided government funding of potential military applications for the technology. To expedite the proliferation of Cyc into the marketplace, Cycorp is releasing a small portion of their knowledgebase and prepackaged reasoning engines as an open source product called OpenCyc. Their hope is to attract software developers by demonstrating the power of Cyc-powered natural language processing and automated reasoning systems and then charge for the complete knowledgebase.

At its core, Cyc is an expert system. Teams of human developers from the classic fields of reasoning, including computer science, artificial intelligence, philosophy, logic, and mathematics, have all been tasked with encoding the shared knowledge of the world into symbolic databases. These databases are primarily composed of First Order Predicate Calculus statements, but the Cyc group has expanded upon the traditional FOPC framework to provide several key extensions critical to their world representation. Two notable examples are the groups of temporal and spatial extensions created to circumvent FOPC's restrictive limitations in those areas. Rather than force their predicate rules to be recursively expanded to support ever-increasing temporal and spatial resolution, the Cyc group decided to extend the language with proprietary rulesets. By doing so, they have been able to provide a superior representation schema for their specific inference engines, and thus increase reasoning performance on the knowledgebase.

The sheer magnitude of Cyc's scope has forced the core representation schema to be redesigned several times to incorporate unforeseen large-scale phenomena. One such

phenomenon is the problem of large-scale unification. Traditional expert systems deal with an extremely restricted domain and thus have knowledgebases that typically enjoy a strong level of cohesion amongst member predicate statements. This cohesion, however, breaks down when the domain becomes sufficiently generalized. In Cyc's case, the knowledgebase reached a point where it was unable to reconcile the many conflicting rulesets it was asked to reason about. To address this, the Cyc team had to introduce the notion of logical contexts known as "microtheories" to encapsulate the cohesion in individual subdomains. Large-scale design also forced the conversion of the underlying representative foundation from a frame-based system to one based on logical assertions, which remains its present form.

While the Cyc knowledgebase is a strictly symbolic representation of its rulesets, it does contain explicitly encoded stochastic knowledge. One often-cited example is that if Cyc is asked to find "examples of males with guns", it will locate examples of soldiers, because it is aware of the assertions that soldiers are often male and that they often carry weapons. Neither of these statements are tautologies, however, as both have frequent exceptions. So, the Cyc framework has the ability to explicitly define these probabilities in its representations of entities, which gives it greater flexibility in resolving contextual ambiguities.

Due to the fact that the Cyc project is now a commercial venture, the company has understandably not published a large amount regarding the key internal aspects of their technology. Thus far they have only released a few whitepapers detailing some of the broader aspects of their extensions to First Order Predicate Calculus, along with a general overview of the Cyc knowledgebase framework. However, with the release of

the open source edition of Cyc, there should be substantially more information available in the very near future.

From the information currently published regarding Cyc's internal structure and evolution, it appears that it will share many of the same problems that afflicted the expert system revolution of the early 1980's. The expert system revolution was arguably diminished by the lack of breadth in individual knowledgebases. Each expert system was proficient in its own restricted domain, but once released "into the wild", found that it often required additional knowledge outside of its domain. Expert systems either further restricted their application areas, or else expanded their knowledgebases to include ever-larger stores of consensus knowledge. However, this approach proved to be intractable, and expert systems soon fell out of favor. The goal of Cyc is to provide a universal store of consensus knowledge that could be plugged into these expert systems' knowledgebases and give them access to external world knowledge.

The problem with this approach is that Cyc itself is an expert system and hence subject to the same fundamental limitations as the systems it promises to extend. Cyc differs from traditional expert systems in that its focus domain is extremely general, inclusive of all human consensus knowledge. However, despite its generality, the Cyc system's focus domain is still a framed one, and thus subject to localization problems. The current knowledgebase was populated by sheer brute force, an eighteen-year endeavor to capture a window of knowledge. However, this window will eventually be outdated, as new environments and cultural norms render blocks of assertions obsolete. The only solution is to continually update the Cyc knowledgebase, removing outdated assertions and adding new ones, in a never-ending process. This is not a maintainable

long-term solution, as the rapid onward march of society can easily outpace attempts to catalogue and document it. The amount of precise detail that must be captured to continually keep the Cyc knowledgebase current will prove to be impossible to obtain and record in synchrony with world changes. Although Cyc provides limited facilities for automated learning, through the use of standard inference models, it does not provide for more advanced machine learning techniques such as learning networks or augmentative inferencing. A greater focus on autonomous techniques must be pursued before it will be feasible for the Cyc system to remain current with changing trends.

Despite these limitations, the Cyc system represents a landmark breakthrough for natural language processing and automated reasoning systems. With the broad scope of Cyc's knowledgebase and its inclusion of "invisible knowledge" in the form of its consensus knowledge store, it promises to be a plug-and-play knowledge revolution. The potential applications for this technology are innumerable, but possibilities being actively pursued by the Cyc group currently include intelligent interfacing of heterogeneous databases, retrieval of textually-captioned multimedia information, integrated disparate domain-centric thesauri and syntactical references, distributed knowledge agents, and World Wide Web information retrieval. Other projects which Cycorp plans to explore in the future include online goods and services brokering, smart interfaces, character intelligence for computer games, reasoning extensions to immersive virtual environments, augmented machine translation and speech recognition systems, user modeling, and semantic data mining.

Three of these projects have already been realized in commercial products. CycSecure, CycAnswers, and Cyc Knowledge Server are all commercially available to

Cyc's sponsor corporations. CycSecure is a network intrusion detector that uses the Cyc knowledgebase to make advanced inferences regarding potential breakpoints in a network's security infrastructure. Rather than the traditional approach of brute-force scanning a network for vulnerabilities, the CycSecure system mimics a potential cracker and reproduces the steps that a human attacker would use. This is a far more powerful tactic of preventive security, as human crackers tend to use extremely clever exploitations of network topology that often cannot be captured and analyzed by a brute-force scan.

CycAnswers is a knowledge-based data warehousing application that can index large corpi of natural language texts and extract their semantic meanings. It then provides a natural language interface to this stored knowledge, permitting intuitive interaction with the stored data on a conceptual basis. The system provides the additional capability of interfacing with remote data sources and querying them using its own knowledgestores to formulate the proper queries. One instance given by Cycorp of the power of this technique is demonstrated in a query posed to a Cyc system asking for "names of individuals with advanced degrees in New England." The system is given access to a personnel database and given information about the contents of each field in the database. The Cyc system then uses its own knowledgebase to reason about the query and construct a set of SQL queries to issue to the database to retrieve the requested information. In this particular example, the system would compile a list of the states in New England and also a list of criteria that describes an individual with an advanced degree. Using this information, it can then issue simple directed SQL queries to the database. The Cyc Knowledge Server is Cycorp's development package, which contains

the Cyc knowledgebase packaged with an inferencing engine. It is designed to be incorporated into third-party products as an enabling backend.

The question of whether packaging a snapshot of human consensus knowledge into a database will prove an ultimate solution to bridging the knowledge gap between machines and humans has yet to be answered. Cyc has made admirable strides towards this goal and its ultimate success will be a strong indicator of the viability of the second generation of expert systems.