Bill Gates is not a Parking Meter: Philosophical Quality Control in Automated Ontology-building

Catherine Legg, Samuel Sarjant
The University of Waikato, New Zealand

Email: clegg@waikato.ac.nz, sjs31@cs.waikato.ac.nz

Abstract

The somewhat old-fashioned concept of philosophical categories is revived and put to work in automated ontology building. We describe a project harvesting knowledge from Wikipedia's category network in which the principled ontological structure of Cyc was leveraged to furnish an extra layer of accuracy-checking over and above more usual corrections which draw on automated measures of semantic relatedness.

1. Philosophical Categories

S1: The number 8 is a very red number.

There is something clearly wrong with this statement, which seems to make it somehow 'worse than false'. For a false statement can be negated to produce a truth, but

S2: The number 8 is not a very red number.

doesn't seem right either.[1] The problem seems to be that numbers are not the kind of thing that can have colours – if someone thinks so then they don't understand *what kinds of things numbers are*[2].

The traditional philosophical term for what is wrong is that **S1** commits a *category mistake*. It mixes kinds of thing nonsensically. A traditional task of philosophy was to identify the most basic categories into which our knowledge of reality should be divided, and thereby produce principles for avoiding such statements. One of the first categorical systems was produced by Aristotle, who divided predicates into ten groups (*Substance*, *Quantity*, *Quality*, *Relation*, *Place*, *Time*, *Posture*, *State*, *Action*, and *Passion*). The differences between these predicates were assumed to reflect differences in the ontological natures of their arguments. For example, the kinds of things that are earlier and later (*Time*) are not the kinds of things that are heavy or light (*Substance*). Category lists were also produced by Kant, Peirce, and many other Western philosophers.

We believe there is a subtle but important distinction between philosophical *categories* and mere *properties*. Although both divide entities into groups, and may be represented by classes, categories arguably provide a deeper, more sortal division which enforces *constraints*, which distinctions between properties do not always do. So for instance, while we know that the same thing cannot be both a colour and a number, the same cannot be said for green and square. However, at what 'level' of an ontology categorical divisions give way to mere property divisions is frequently unclear and contested. This has led to skepticism about the worth of philosophical categories which will now be touched on.

This task of mapping out categories largely disappeared from philosophy in the twentieth century[3]. The logical positivists identified such investigations with the "speculative metaphysics" which they sought to quash, believing that the only meaningful questions could be settled by empirical observation (Schlick, 1936. Carnap, 1932).

Following this, Quine presented his famous logical criterion of ontological commitment: "to be is to be the value of a bound variable… [in our best scientific theory]" (Quine, 1953). This widely admired pronouncement may be understood as flattening all philosophical categories into one 'mode of being'. Just as there is just one existential quantifier in first-order logic, Quine claimed, ontologically speaking there is just one kind of existence, with binary values (does and does not exist). Thus there are no *degrees* of existence, nor are there *kinds* – rather there are different *kinds of objects*

---

[1] Some philosophers do take a hard line on statements such as **S2**, claiming that it is literally true, but it does at least seem to have misleading pragmatic implications.

[2] There is the phenomenon of synaesthesia. But the rare individuals capable of this feat do not seem to converge on any objective colour-number correlation.

[3] Notable exceptions: (Weiss, 1958), (Chisholm, 1996), (Lowe, 1997, 1998).

which all have the same kind of existence.

This move to a single mode of being might be thought to reopen the original problem of why certain properties are instantiated by certain kinds of objects and not others, and why statements such as **S1** seem worse than false. A popular response – common in the analytic tradition as a reply to many problems – has been to fall back on faith in an ideal language, such as modern scientific terminology (perhaps positions of atoms and molecules), which is fantasized as 'category-free'.

Be that as it may, we will now examine an IT project which recapitulated much of the last 3000 years of philosophical metaphysics in a fascinating way.

2. The Cyc Project

2.1 Goals and Basic Structure

When the field of Artificial Intelligence struggled in the early 80s with brittle reasoning and inability to understand natural language, the Cyc project was conceived as a way of blasting through these blocks by *codifying common sense*. It sought to represent in a giant knowledge base, "the millions of everyday terms, concepts, facts, and rules of thumb that comprise human consensus reality", sometimes expressed as everything a 6-year-old knows that allows her to understand natural language and start learning independently (Lenat, 1995, Lenat and Guha, 1990).

This ambitious project has lasted over 25 years, producing a taxonomic structure purporting to cover all conceivable human knowledge. It includes over 600,000 categories, and over 2 million axioms, a purpose-built inference engine, and a natural language interface. All knowledge is represented in *CycL*, which has the expressivity of higher-order logic – allowing assertions about assertions, context logic (Cyc contains 6000 "Microtheories"), and some modal statements.

The initial plan was to bring the system as quickly as possible to a point where it could begin to learn on its own, for instance by reading the newspaper (Lenat and Guha, 1990, Lenat, 1995). Doug Lenat estimated in 1986 that this would take 5 years (350 person-years) of effort and 250,000 rules, but it has still not happened, leading to widespread scepticism about the project.

2.2 Categories and Common Sense Knowledge

Nevertheless, it is worth noting that the Cyc project *did* meet some of its goals. Consider the following, chosen at random as a truth no-one would bother to teach a child, but which by the age of 6 she would know by common-sense:

S3: Bill Gates is not a parking meter.[4]

This statement has never been asserted into Cyc. Nevertheless Cyc knows it, and can justify it, as follows:

```
BillGates is known not to be an instance of
     ParkingMeter in mt WikipediaToCycDataMt.
sbhl conflict: (isa BillGates ParkingMeter) TRUE
           WikipediaToCycDataMt
     because: (isa BillGates MaleHuman)
             True-JustificationTruth
(genls MaleHuman MaleAnimal) TRUE
(genls MaleAnimal Animal) TRUE
(genls Animal AnimalBLO) TRUE
(genls AnimalBLO BiologicalLivingObject) TRUE
(disjointWith BiologicalLivingObject
     Artifact-Generic) TRUE
(genls Technology-Artifact Artifact-Generic) TRUE
(genls MechanicalDevice Technology-Artifact) TRUE
(genls ParkingMeter MechanicalDevice) TRUE
```

(Justification produced in ResearchCyc 1.0, 2009)

The crucial premise is the claim of disjointness between the classes of living things and artifacts. The Cyc system only contains several thousand explicit **disjointWith[5]** statements, but as seen above, these ramify through the knowledge hierarchy in a powerful, open-ended way.

A related feature of Cyc's common-sense knowledge is its so-called *semantic argument constraints on relations*. For example **(arg1Isa birthDate Animal)** represents that only animals have birthdays. These features of Cyc are a form of categorical knowledge. Although some of the categories invoked might seem relatively specific and trivial compared to Aristotle's, logically the constraining process is the same.

In the early days of Cyc, knowledge engineers laboured to input common-sense knowledge in the form of rules (e.g. "If people do something for recreation that puts them at risk of bodily harm, then they are adventurous"). Reasoning over such rules required inferencing of such complexity that they almost never 'fired' (were recognized as relevant), or if they did fire they positively hampered query resolution (i.e. finding the answer). By contrast Cyc's disjointness and semantic predicate-argument constraints were simple and effective, so much so that they were enforced at the knowledge-entry level. Thus returning again to **S1**, this statement could not

---

[4]   Presenting this material to research seminars it has been pointed out that there is a metaphorical yet highly meaningful sense in which Bill Gates (if not personally, then in his capacity as company director) *does* serve as a parking meter for the community of computer users. Nevertheless, in the kinds of applications discussed in this paper we must alas confine ourselves to literal truth, which is challenging enough to represent.

[5]   Terms taken from the CycL language are represented in bolded Courier through the paper

be asserted into Cyc because redness is represented as the class of red things which generalizes to spatiotemporally located things, while numbers generalizes to abstract objects, and once again these high level classes are known to be disjoint in Cyc.

We believe these constraints constitute an untapped resource for a distinctively ontological quality control for automated knowledge integration. Below we show how we put them to work in a practical project.

3. "Semantic Relatedness"

When 'good-old fashioned' rule-based AI systems such as Cyc apparently failed to render computers capable of understanding the meaning of natural language, AI researchers turned to more brute, statistical ways of measuring meaning. A key concept which emerged is *semantic relatedness*, which seeks to quantify human intuitions such as: *tree* and *flower* are closer in meaning than *tree* and *hamburger*. Simple early approaches analysed term co-occurrence in large corpora (Jiang and Conraith, 1997; Resnik, 1999). Later, more sophisticated approaches such as Latent Semantic Analysis constructed vectors around the compared terms (consisting of, for instance, word counts in paragraphs, or documents) and computed their cosine similarity.

Innovative extensions to these methods appeared following the recent explosion in free user-supplied Web content, including the astoundingly detailed and organized Wikipedia. Thus Gabrilovich and Markovitch (2007) enrich their term vectors with Wikipedia article text: an approach called Explicit Semantic Analysis. Milne and Witten (2008) develop a similar approach using only Wikipedia's internal hyperlinks. Here semantic relatedness effectively becomes a measure of likelihood that each term will be anchor text in a link to a Wikipedia article about the other.

In the background of this research lurk fascinating philosophical questions. Is closeness in meaning sensibly measured in a single numeric value? If not, how should it be measured? Can the semantic relatedness of two terms be measured overall, or does it depend on the context where they occur? Yet automated measures of semantic relatedness now have a high correlation with native human judgments (Medelyan et al, 2009).

4. Automated Ontology Building: State of the Art

Dissatisfaction with the limitations of manual ontology-building projects such as Cyc led to a lull in formal knowledge representation through the 1990s and early 2000s, but the new methods of determining semantic relatedness described above, and the free user-supplied Web content on which they draw, has recently begun a new era in *automated* ontology building.

One of the earliest projects was YAGO (Suchanek et al, 2007, 2008), which maps Wikipedia's leaf categories onto the WordNet taxonomy of synsets, adding articles belonging to those categories as new elements, then extracting further relations to augment the taxonomy. Much useful information is obtained by parsing category names, for example extracting relations such as *bornInYear* from categories such as *1879 birth*.

A much larger, but less formally structured, project is DBpedia (Auer et al. 2007, Auer and Lehmann 2007), which transforms Wikipedia's infoboxes and related features into a vast set of RDF triples (103M), to provide a giant open dataset on the web. This has since become the hub of a Linked Data Movement which boasts billions of triples (Bizer et al. 2009). Due to the lack of formal structure there is however much polysemy and many semantic relationships are obscured (e.g. there are redundant relations from different infobox templates, for instance *birth_date*, *birth* and *born*). Therefore they have also released a DBpedia Ontology generated by manually reducing the most common Wikipedia infobox templates to 170 ontology classes and the 2350 template relations to 940 ontology relations asserted onto 882,000 separate instances.

The European Media Lab Research Institute (EMLR) built an ontology from Wikipedia's category network in stages. First they identified and isolated *isA* relations from other links between categories. (Ponzetto and Strube, 2007). Then they divided *isA* relations into *isSubclassOf* and *isInstanceOf* (Zirn et al, 2008), followed by a series of more specific relations (e.g. *partOf, bornIn*) by parsing category titles and adding facts derived from articles in those categories (Nastase and Strube, 2008). The final result consists of 9M facts indexed on 2M terms in 105K categories.[6]

What is notable about these projects is that firstly, all have found it necessary to build on a manually created backbone (in the case of YAGO: Wordnet, in the case of the EMLR project: Wikipedia's category network, and even DBPedia produced its own taxonomy). Yet none of these ontologies can recognize the wrongness of `S1`. Although YAGO and EMLR's system possess rich taxonomic structure, it is property-based rather than categorical, and does not enforce the relevant constraints. A second important issue concerns evaluation. With automation, accuracy becomes a key issue. Both YAGO and DBPedia (and Linked Data) lack any formal evaluation, though EMLR did evaluate the first two stages of their project – interestingly, using

---

[6]   Downloadable at http://www.eml-research. de/english/research/nlp/download/wikirelations.php

Cyc as a gold standard – reporting precision of 86.6% and 82.4% respectively.

Therefore we wondered whether Cyc's more stringent categorical knowledge might serve as an even more effective backbone for automated ontology-building, and also whether we might improve on the accuracy measurement from EMLR. We tested these hypotheses in a practical project, which transferred knowledge automatically from Wikipedia to Cyc (ResearchCyc version 1.0).

5. Automated Ontology Building: Cyc and Wikipedia

5.1. Stage 1: Concept Mapping

Mappings were found using four stages:

Stage A: Searches for a one-to-one match between Cyc term and Wikipedia article title.

Stage B: Uses Cyc term synonyms with Wikipedia redirects to determine a single mapping.

Stage C: When multiple articles map, a 'context' set of articles (comprised of article mappings for Cyc terms linked to the current term) is used to identify the article with the highest semantic-related score using (Milne and Witten, 2008).

Stage D: Disambiguates and removes incorrect mappings by performing Stage A and B backwards (e.g. `DirectorOfOrganisation` → *Film director* → `Director-Film`, so this mapping is discarded).

5.2 Stage 2: Transferring Knowledge

Here new subclasses and instances ('children') were added to the Cyc taxonomy, as follows.

Finding possible children

Potential children were identified as articles within categories where the category had an equivalent Wikipedia article mapped to a Cyc collection (about 20% of mapped articles have equivalent categories).

Wikipedia's category structure is not as well-defined as Cyc's collection hierarchy, containing many merely associatively-related articles. For example *Dogs* includes *Fear of dogs* and *Puppy Bowl*. Blind harvesting of articles from categories as subclasses and instances of Cyc concepts was therefore inappropriate.

Identifying correct candidate children

Each article within the given category was checked to see if a mapping to it already existed from a Cyc term. If so, the Cyc term was taken as the child, and the relevant assertion of parenthood made if it did not already exist. If not, a new child term was created if verified by the following methods:

Link parsing: The first sentence of an article can identify parent candidates by parsing links from a regularly structured sentence. Each link represents a potential parent if the linked articles are already mapped to Cyc collections (in fact multiple parents were identified with this method).

The regular expression set was created from the most frequently occurring sentence structures seen in Wikipedia article first sentences. Examples included:

- **X are a Y**
  'Bloc Party are a British indie rock band…'
- **X is one of the Y**
  'Dubai is one of the seven emirates…'
- **X is a Z of Y**
  'The Basque Shepherd Dog is a breed of dog…'
- **X are the Y**
  'The Japanese people are the predominant ethnic group of Japan.'

.

Infobox pairing: If an article within a category was not found to be a child through link parsing, it was still asserted as a child if it shared the same infobox template as 90% of the children that *were* found.

Results

The project added over 35K new concepts to the lower reaches of the Cyc ontology, each with an average of 7 assertions, effectively growing it by 30%. It also added documentation assertions from the first sentence of the relevant Wikipedia article to the 50% of mapped Cyc concepts which lacked this, as illustrated below:

Collection : WrestlingRing

Bookkeeping Assertions :

(myCreationTime WrestlingRing 19920918) in BookkeepingMt

GAF Arg : 1

Mt : UniversalVocabularyMt
isa : ExistingObjectType
gens : SportsPlayingArea

Mt : WikipediaToCycDataMt
comment : "A wrestling ring is the ring stage that professional wrestlers wrestle in."
salientURL : "http://en.wikipedia.org/wiki/Wrestling_ring"

Mt : WikipediaToCycLexicalMt
(synonymousExternalConcept WrestlingRing Enwiki_20080727 "5160881")

An evaluation of these results was performed with 22 human subjects on testsets of 100 concepts each. It showed that the final mappings had 93% precision, and that the assignment of newly created concepts to their 'parent' concepts was 'correct or close' 90% of the time (Sarjant et al, 2009). This suggests a modest improvement on the EMLR results, though more extensive testing would be required to prove this. Work on an earlier version of the algorithm (Medelyan and Legg, 2008) also tested its accuracy against the inter-agreement of six human raters, measuring the latter at 39.8% and the agreement

between algorithm and humans as 39.2%.

5.3 Categorical Quality Control

During the initial mapping stage, Cyc's disjointness knowledge was put to work discriminating rival candidate matches to Cyc concepts which had near-equal scores in quantitative semantic relatedness. In such cases Cyc was queried for disjointness between ancestor categories of the rivals, and if disjointness existed, the match with the highest score was retained and others discarded. Failing that, all high-scoring matches were kept. Examples of where this worked well were the Wikipedia article *Valentine's Day*, which mapped to both `ValentinesDay` and `ValentinesCard`, but Cyc knew that a card is a spatiotemporal object and a day is a 'situation', so only the former was kept. On the other hand, the test allowed *Black Pepper* to be mapped to both `BlackPeppercorn` and `Pepper-TheSpice`, which despite appearances was correct given the content of the Wikipedia article.

During the knowledge transfer stage an interesting phenomenon occurred. Cyc was insistently 'spitting out' a given assertion and it was thought that a bug had occurred. To the researchers' surprise it was found that Cyc was ontologically correct. From that time on, the assertions Cyc was rejecting were gathered in a file for inspection. At the close of the project this file contained 4300 assertions, roughly 3% of the assertions fed to Cyc. Manual inspection suggested that 96% of these were 'true negatives,' for example:

```
(isa CallumRoberts Research)
(isa Insight-EMailClient EMailMessage)
```

This compares favourably with the evaluated precision of assertions successfully added to Cyc.

The examples above usefully highlight a clear difference between quantitative measures of semantic relatedness, and an ontological relatedness derivable from a principled category structure. Callum Roberts is a *researcher*, which is highly semantically related to *research* and Insight is an *email client*, which is highly semantically related to *email messages*. Thematically or topically these pairs are incredibly close, but ontologically speaking, they are very different kinds of thing. Thus if we state:

```
S4: Callum Roberts is a research
```

we once again hit the distinctively unsettling silliness of the traditional philosophical category mistake, and a kind of communication we wish our computers to avoid.

6. Plans for Further Feeding

Given the distinction between semantic and ontological relatedness, we may note that combining the two has powerful possibilities. In fact this observation may usefully be generalized to note that in automated information science, *overlapping independent heuristics* are a boon to accuracy, and this general principle will guide our research over the next few years.

Our first step will be to develop strategies to automatically augment Cyc's disjointness network and semantic argument constraints on relations (where Cyc's manual coding has resulted in excellent precision but many gaps) using features from Wikipedia. For instance, systematically organized infobox relations, helpfully collected in DBPedia, are a natural ground to generalize argument constraints. The Wikipedia category network will be mined – with caution – for further disjointness knowledge. This further common-sense categorical knowledge will then bootstrap further automated ontology-building.

7. Philosophical Lessons

Beyond the practical results described above, our project provides fuel for philosophical reflection. It suggests the notion of philosophical categories should be rehabilitated as it leads to measurable improvements in real-world ontology-building. Just how extensive a system of categories should be will of course require real-world testing. But now we have the tools, the computing power, and most importantly the wealth of free user-supplied data to do this.

The issue of where exactly the line should be drawn between categories proper and mere properties remains open. However, modern statistical tools raise the possibility of a quantitative treatment of ontological relatedness that is more nuanced than Aristotle's ten neat piles of predicates, yet can still recognize that `S1` is highly problematic, and why.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J, Cyganiak, R., Ives, Z. (2007). DBPedia: A Nucleus for a Web of Open Data. In Aberer, K. et al (eds.) *ISWC/ASWC 2007, LNCS 4825*. Springer, Berlin Heidelberg, 722-735.

Auer, S., and Lehmann, J. (2007). What have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In Franconi *et al.* (eds), *Proc. of European Semantic Web Conference (ESWC'07)*, LNCS 4519, Springer, 503–517.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia—A Crystallization Point for the Web of Data. *Journal of Web Semantics* **7**(3), 154-165.

Carnap, R. (1932). The Elimination of Metaphysics

through Logical Analysis of Language, *Erkenntnis* **2**, 219–41.

Chisholm, R. (1996). *A Realistic Theory of Categories : an Essay on Ontology*. Cambridge: Cambridge University Press.

Gabrilovich, G. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*, Hyderabad, India, January 2007, 1606–1611.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*. Taiwan, 1997.

Legg, C. (2007). Ontologies on the Semantic Web. *Annual Review of Information Science and Technology* **41**, 407-452.

Legg, C. (in press). Peirce, Meaning and the Semantic Web. *Semiotica*.

Lenat, D.B. (1995). Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* **38** (11).

Lenat, D.B., Guha, R.V. (1990). *Building Large Knowledge Based Systems*. Reading, Massachusetts: Addison Wesley.

Lowe, E. J. (1997). Ontological Categories and Natural Kinds. *Philosophical Papers* **26**, 29-46.

Lowe: E.J. (1998). *The Possibility of Metaphysics: Substance, Identity, and Time.* Oxford: Oxford University Press.

Matuszek, C., Witbrock, M., Kahlert, R.C., Cabral, J., Schneider, D., Shah, P. And Lenat, D. (2005). "Searching for Common Sense: Populating Cyc from the Web". *Proceedings of the 20th National Conference on Artificial Intelligence.* Pittsburgh, Penn., 1430-1435.

Medelyan, O., and Legg, C. (2008). Integrating Cyc and Wikipedia: Folksonomy meets Rigorously Defined Common-sense. *Proc., Workshop on Wikipedia and AI, AAAI08*, Chicago, USA, July 2008.

Medelyan, O., Legg, C., Milne, D., and Witten, I.H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, **67**(9), 716–754.

Milne, D., and Witten, I. H. (2008). An Effective, Low-cost measure of Semantic Relatedness Obtained from Wikipedia links. *Proc., Workshop on Wikipedia and AI, AAAI08*, Chicago, USA, July 2008.

Milne, D., Medelyan, O. and Witten, I. H. (2006). Mining Domain-specific Thesauri from Wikipedia: A Case study. *Proc. of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006)*, Hong Kong.

Nastase, V. and Strube, M. (2008). Decoding Wikipedia Categories for Knowledge Acquisition. *Proc. of AAAI'08*, Chicago, USA, July 2008.

Ponzetto, S. P. (2007). Creating a Knowledge Base from a Collaboratively-generated Encyclopedia. *Proc., Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Doctoral Consortium*. Rochester, NY, 22–27 April, 2007, 9–12.

Ponzetto, S. P., and Strube, M. (2007). Deriving a Large Scale Taxonomy from Wikipedia. *Proc. of AAAI'07*, Vancouver, Canada, July 2007, 1440-45.

Quine, W.V.O. (1953). On What There Is. *From a Logical Point of View*. Cambridge MA: Harvard University Press, 1-19.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

Sarjant, S., Legg, C., Robinson, M., Medelyan, O. (2009) 'All you can eat' Ontology-building: Feeding Wikipedia to Cyc." *Proc IEEE/WIC/ACM International Conference on Web Intelligence*, WI-09, Milan.

Schlick, M. (1936). Meaning and Verification. *The Philosophical Review*. **45** (4), 339-69.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A Core of semantic knowledge. *Proc. of 16th World Wide Web Conference, (WWW'07)*, New York, NY: ACM Press.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A Large Ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics* **6**(3), 203-217.

Weiss, P. (1958). *Modes of Being*. Carbondale, IL: Southern Illinois University Press.

Zirn, C., Nastase, V., and Strube, M. (2008). Distinguishing between Instances and Classes in the Wikipedia Taxonomy. *Proc. of The 5th European Semantic Web Conference (ESWC'08)*, Tenerife, Spain, June 2008.