Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense

Olena Medelyan
Department of Computer Science
University of Waikato, New Zealand
olena@cs.waikato.ac.nz

Catherine Legg
Department of Philosophy and Religious Studies
University of Waikato, New Zealand
clegg@waikato.ac.nz

Abstract

Integration of ontologies begins with establishing mappings between their concept entries. We map categories from the largest manually-built ontology, Cyc, onto Wikipedia articles describing corresponding concepts. Our method draws both on Wikipedia's rich but chaotic hyperlink structure and Cyc's carefully defined taxonomic and common-sense knowledge. On 9,333 manual alignments by one person, we achieve an F-measure of 90%; on 100 alignments by six human subjects the average agreement of the method with the subject is close to their agreement with each other. We cover 62.8% of Cyc categories relating to common-sense knowledge and discuss what further information might be added to Cyc given this substantial new alignment.

1. Introduction

As information sharing became ever more sophisticated and globalized from the 1980s onwards, a new research frontier formed around the ambitious goal of developing a machine-understandable conceptual scheme (a.k.a. 'formal ontology') which would mediate information transfer in any conceivable format (Gruber, 1995). Yet current ontology research is still far from delivering such a useful product. The enormous number of concepts in human language must somehow be represented in an ontology. However, it is not enough just to index the names of concepts in some canonical list – a useful ontology needs also to capture defining facts about them, and reason about these facts. For instance, given the term 'tree', an ontology should know at least that some trees are biological organisms, and some are mathematical objects, and they are not the same.

There is an obvious trade-off between the number of concepts covered ('breadth') and the amount of information represented about each concept ('depth'), and almost all current ontology projects emphasize one without the other. For instance WordNet defines 207,000 categories but solely organizes them into a few simple relations. After initial enthusiasm for using WordNet as an ontology due to its simplicity (Mann, 2002; Niles et al., 2003), today it is still mainly appreciated as an exceptionally comprehensive linguistic resource.

Relatively sophisticated definitions of concepts in narrow domains may be found in a number of ontologies pertaining to specific subject-areas which attract research funding, e.g. the Foundation Model of Anatomy (Rosse and Mejino, 2003), and various geospatial ontologies. However, ontology integration is still an enormous challenge. Thus a cursory search for our example 'tree' on Swoogle,[1] which queries 10,000 ontologies, returns merely scattered unitary assertions (e.g. 'A Tree is a kind of LandscapeProduct', 'A TreeRing is a kind of Vegetation'), confusingly mixed with assertions concerning 'trees' as mathematical structures.

Arguably the most heroic attempt to provide breadth and depth simultaneously is the famous Cyc project (Lenat, 1995), the 'Rolls Royce of formal ontologies'. As a 20 year project, US government-funded for over 700 person-years of work, it has been able to amass 200,000 categories and provides a custom-built inference engine. Thus Cyc knows not only that `#$Tree-ThePlant` is different from `#$Tree-PathSystem`, but also assertions on the former such as, "A tree is largely made of wood", and, "If a tree is cut down, then it will be destroyed". Since 2004 sections of Cyc have been released to the public, such as OpenCyc, covering the top 40% of Cyc, and later ResearchCyc, covering over 80% (and available to research institutions). So far, however, their utilization and integration with other resources has been limited, as the combination of coding skills and philosophical nous required to understand and work with Cyc is still possessed by few.

Meanwhile vast excitement has gathered around the possibilities of leveraging websites with user-supplied content. The most general and fastest growing of these, Wikipedia, surprised many with its growth and accuracy. From its launch in early 2001 to the present it has swiftly acquired 2M concepts (indexed via 5M synonyms) and researchers soon began to mine its structure for ontology (e.g. Hepp et al., 2006; Herbelot et al., 2006). This provides a potential vast increase in concept-coverage. However if all that is

[1] http://swoogle.umbc.edu

taken from Wikipedia are names for concepts, arranged in a subsumption hierarchy via Wikipedia 'category' links, it risks becoming another WordNet – merely a 10 times bigger bag of words with no real understanding of their meaning. What is needed is some way of adding definitional information. One natural candidate for this is Cyc.

Given that Wikipedia has more concepts than Cyc, and Cyc has a richer explicitly represented knowledge framework than Wikipedia, it makes sense to integrate Wikipedia concepts into Cyc, rather than vice versa. The starting point for such integration is to establish mappings between existing Cyc terms and corresponding Wikipedia articles. To overcome terminology differences, we use rich synonymy relations in both resources. To deal with sense ambiguity, we analyze semantic similarity of possible mappings to context categories in the neighboring Cyc ontology. To bypass inconsistency in both resources, we develop a step-by-step mapping heuristic. With this strategy we can map 52,690 Cyc categories to Wikipedia articles, with a precision of 93% tested on 9,333 human alignments. Further disambiguation based on Cyc's common-sense knowledge improves the precision of 42,279 mappings to over 95%.[2]

At each mapped node in Cyc's tree, it may now be determined what new information Wikipedia can teach Cyc. So far, we have managed to identify over 145,000 possible new synonymy assertions, over 1,400 URLs, over 500,000 translations into other languages. We discuss the addition of further facts, which would produce an enlarged ontology, which may then be used for further iterative ontology alignment, including learning more facts from Wikipedia, which continues to grow and improve.

2. Related work

On the Cyc side, from its inception Cycorp sought to map existing ontologies and knowledge bases into Cyc, for instance WordNet (Reed et al. 2002). However having been automatically entered the new information required cleaning and integrating by hand, which due to limited resources was never done fully. Matuszek et al. (2005) extend Cyc by querying a search engine, parsing the results, and checking for consistency with the Cyc knowledge base. However, they required a human check before each new concept or fact was entered – thus only added 2,000 new assertions. The Cyc Foundation[3] is currently developing a user-friendly interface to integrate Cyc and Wikipedia. So far however (there are no published results yet), this appears to be merely a browser via which a human can look at Cyc and Wikipedia categories side by side, rather than any deeper integration.

On the Wikipedia side, mining for semantic relations is the prevalent research topic. Gregorowicz et al. (2006) treat Wikipedia as a semantic network, extracting hyperlinks between categories and articles, which are treated as 'semantic' but not further differentiated. Ponzetto and Strube (2006) categorize relations between Wikipedia categories by analyzing the names of concept pairs, their position in the network, as well as their occurrences in corpora, to accurately label the relation between each pair as *isa* and *not-isa*. However, there is yet no separation into further types of relations, such as *is-an-instance-of* and *has-part*. Also the approach is restricted to the 126,700 Wikipedia 'categories' (as opposed to the 2M Wikipedia articles). Thus these approaches are still far from producing full-blooded ontologies.

Several authors explicitly look at mining Wikipedia as an ontology. Hepp et al. (2006) use URIs of Wikipedia entries as identifiers for ontological concepts. This provides WordNet-style breadth without depth. Herbelot et al. (2006) extract an animal ontology from Wikipedia by parsing whole pages. Although they achieve an impressively comprehensive coverage of their subject-matter, computational demands restricted the task to only 12,200 Wikipedia pages, a tiny fraction of the total.

Suchanek et al. (2007) create a new ontology, YAGO, which unifies WordNet and Wikipedia providing 1M categories and 5M facts. Its categories are all WordNet synsets and all Wikipedia articles whose titles are not listed as common names in WordNet. It therefore misses many proper names with homonyms in WordNet—e.g. the programming language Python and the film "The Birds". Our approach differs from Yago in that we identify links between synonymous concepts in Cyc and Wikipedia using explicit semantic disambiguation, whereas Yago merely adds Wikipedia to WordNet avoiding the ambiguous items.

The DBpedia project attempts to make all structured information in Wikipedia freely available in database form (Auer et al., 2007). RDF triplets are extracted by mining formatting patterns in the text of Wikipedia articles, e.g. infoboxes, as well as categorization and other links. These authors harvest 103M facts and enable querying of their dataset via SPARQL and Linked Data. They also connect with other open datasets on the Web. But this enormous increase in data comes at a huge cost in quality. Many of the triplets' relations are not ontological but rather trivial, e.g. the most common relation in infobox triplets (over 10%) is `wikiPageUsesTemplate`. Also, amongst the relations that are ontological there are obvious redundancies not identified as such, e.g. `placeOfBirth` and `birthPlace`, `dateOfBirth` and `birthDate`.

3. Mapping of Cyc concepts to Wikipedia

The number of categories in our distribution of ResearchCyc (12/2007) is 163,317, however a significant portion of these do not represent common-sense knowledge. We therefore filtered out:

- categories describing Cyc's internal workings
- knowledge required for natural language parsing
- project-specific concepts
- microtheories
- predicates and all other instances of `#$Relation`

This leaves 83,897 categories.

[2] Available here: http://www.cs.waikato.ac.nz/~olena/cyc.html
[3] http://www.cycfoundation.org

We begin the integration of Cyc and Wikipedia by mapping Cyc concepts to Wikipedia *articles*. We do not allow mappings to Wikipedia's categories or disambiguation pages, because the former do not specifically describe concepts and the latter are inconsistent, however we do use disambiguation pages for identifying ambiguous terms (cf. Section 3.2). To Cyc terms we apply a simple cleaning algorithm to align them with Wikipedia article titles. This includes splitting the name into component words while considering the acronyms, e.g. `#$BirdOfPrey` → 'Bird of Prey'. Expressions after the dash sign in Cyc we write in brackets as this is the convention in Wikipedia, e.g. `#$Virgo-Constellation` → 'Virgo (constellation)'. We also map all but the first capitalized words to lower case, since Cyc does not distinguish between these. For example, in Wikipedia 'Optic nerve' (the nerve) and 'Optic Nerve' (the comic book) are distinct concepts; in Cyc the former is encoded as `#$OpticNerve` and the latter is missing.

Next, we differentiate between two cases: first, where a string comparison produces only one candidate Wikipedia article per Cyc term (exact mapping), and second, where it produces more than one (ambiguous mapping). For the former we propose two steps that augment each other, whereas for the latter we use two alternative approaches, which we evaluate individually.

3.1 Exact mappings

Mapping 1: We identify Cyc terms which exactly match Wikipedia article titles—or redirects, in which case the target article is retrieved. If the match is to a disambiguation page, the term is treated as ambiguous and not considered. The result is a set of possible mappings for each Cyc term. At this stage we only allow a mapping if this set contains exactly one member.

Mapping 2: If for a Cyc term Mapping 1 gives no results, we check whether its synonyms exactly match a title of a Wikipedia article, or its redirect. Again, only unitary result sets are allowed.

With this exact mapping we linked 33,481 of the chosen 83,897 Cyc terms to Wikipedia articles (40%).

3.2 Ambiguous mappings

While the above mappings ensure high accuracy (cf. Section 4.1), their coverage can be improved because many Cyc terms map to more than one Wikipedia article. Also, where no mappings were found, a less strict string comparison can improve the coverage. Therefore, before proceeding with disambiguation, we use the following conflation strategy. To each Cyc term and Wikipedia title, we apply case folding and remove brackets that specify the term's meaning (a feature used inconsistently in both resources). We do not use stemming, because most common syntactic variations are covered in either resource. We now begin to make use of links to articles on disambiguation pages as well. The set of candidate Wikipedia articles for each Cyc term now consists of:

- articles with matching titles
- articles linked from matching redirects,
- articles linked first in each disambiguation on matching disambiguation pages.

We additionally utilize anchor names (i.e. hyperlinked text) in Wikipedia as a source for synonyms (Mihalcea and Csomai, 2007). Given a search term *a*, the likelihood it will link to an article *T* is defined as

$$Commonness_{a,T} = P(T \mid a),$$

which is the number of Wikipedia articles where *a* links to *T* over the total number of articles linked from *a*. For example, the word *Jaguar* appears as a link anchor in Wikipedia 927 times. In 466 cases it links to the article *Jaguar cars*, thus the commonness of this mapping is 0.5. In 203 cases it links to the description of *Jaguar* as an animal, a commonness of 0.22.

Thus, given a Cyc term, we add to its candidate set the 5 most common Wikipedia articles, and record the most common link for each synonym of this term.

Disambiguation I: A simple disambiguation is to weight each candidate article by the number of times it has been chosen via the title of the Cyc term, or any of its synonyms. The highest weight indicates the ideal match.

Disambiguation II: Instead of relying on synonyms encoded in Cyc, this alternative disambiguation method is based on the semantic similarity of each candidate article to the *context* of the given Cyc concept. We define this context using the Cyc ontology, retrieving the categories immediately surrounding our candidate term with the following queries from Cyc's inference engine:

MIN-GENLS – direct hypernyms (collection→collection)
MAX-SPEC – direct hyponyms (collection→collection)
GENL-SIBLINGS – sister collections of a given collection.
MIN-ISA – direct hypernyms (instance→collection)
MAX-INSTANCES –direct hyponyms (collection→instance)
ISA-SIBLINGS – sister instances for a given instance.

We retrieve additional context terms via assertions on selected Cyc predicates, for instance `#$conceptually Related`, and the geographic `#$countryOfCity`.

In Cyc, specifications of a term's meaning are often provided after a dash – e.g. `#$Tool-MusicGroup`, and `#$PCS-Corporation`. If such a specification is parsed and mapped to a Wikipedia article, it serves as a context term as well. For example, 'Music group' helps mapping `#$Tool-MusicGroup` to 'Tool (band)' in Wikipedia.

Next, for each context term obtained from Cyc, we identify a corresponding Wikipedia article with Mapping I and II (Section 3.2) or ignore it if it is ambiguous.[4] Given a set of candidate Wikipedia articles and a set of related context articles, we determine the candidate that is most semantically related to a given context (Milne and Witten, 2008). For each pair, candidate article *x* and context article *y*, we retrieve the sets of hyperlinks *X* and *Y* to these articles, and compute their overlap $X \cap Y$. Given the total number *N* of articles in Wikipedia, the similarity of *x* and *y* is:

[4] Although some important information is discarded by doing so, we find that usually sufficient non-ambiguous terms are provided.

$$SIM_{x,y} = 1 - \frac{\max(\log|X|, \log|Y|) - \log|X \cap Y|}{N - \min(\log|X|, \log|Y|)}.$$

For each article in the set of possible mappings, we compute its average similarity to the context articles. If for all candidates, no similarity to the given context is observed, we return the candidate with the highest commonness weight. Otherwise, we multiply the article *T*'s average similarity to the context articles by its commonness given the n-gram *a*:

$$Score(a,T) = \frac{\sum_{c \in C} SIM_{T,c}}{|C|} \times Commonness_{a,T},$$

where c ∈ C are context articles for *T*. The article with the highest score is the best candidate.

With this method we cover an additional 19,209 Cyc terms (23%). This gives us the maximum coverage for the proposed mapping strategy, a total of 52,690 mappings, i.e. 62.8% of Cyc's common-sense knowledge. However, inaccuracies are inevitable. The following section describes how we address them.

3.3 Common-Sense Disambiguation

After mapping all Cyc terms to Wikipedia articles, we find cases where several Cyc terms map to the same article. (This is the reverse of the problem addressed by **Disambiguation I** and **II** above, where several Wikipedia articles map to the same Cyc term.) Analysis has shown that in some cases, the differentiation in Cyc is too specific, and both mappings are correct, e.g.`#$ThoracicVertebra` and `#$ThoracicVertebrae`. In other cases, one or more of the mappings are incorrect, e.g. `#$AlJazeerah-TheNewspaper`→'Al Jazeera' and `#$AlJazeera-MediaOrganizaton`→ 'Al Jazeera' – since Wikipedia describes the Al Jazeera TV network. Thus we perform two consecutive tests to further correct such mappings.

1. Similarity test.

First, we examine the semantic similarity score of each mapping. The best scoring mapping determines the minimum score for other mappings to be considered. A candidate is not considered if its score is over 30% lower than the maximum score. This helps to eliminate many unlikely mappings that were only 'found' because the Cyc concept has no equivalent Wikipedia article, or it was not located. For example, we eliminate `#$PCS-Corporation` →'Personal Computer' with a score 0.13, because it is lower than 1.58, the score of the best mapping: `#$PersonalComputer`→'Personal Computer'.

2. Disjointness test.

If the above test still leaves more than one possible mapping, we leverage Cyc's common sense knowledge about 'different kinds of things', represented in its extensive knowledge about *disjointness* of collections. We ask Cyc, whether two candidate Cyc terms (or in the case of individuals, their direct hypernyms) are disjoint. Any mapping which is disjoint with our highest scoring candidate is eliminated. All mappings for which disjointness can not be proven are retained. The following example lists Cyc terms mapped to article 'Casino' and their scores:

`#$Casino-Object`	1.1
`#$Casino-TheMovie`	1.0
`#$Casino-TheGame`	0.4
`#$Casino-Organization`	0.1

The similarity test leaves us with `#$Casino-Object` and `#$Casino-TheMovie`, where the former is more likely. But Cyc knows that a casino is a `#$SpatialThing` and a movie is an `#$AspatialThing` and the two are disjoint. Thus we only accept `#$Casino-Object`, which is the correct mapping. The philosophical purity of Cyc's ontology can produce some remarkable discriminations. For instance, Cyc distinguishes between `#$ValentinesCard` and `#$ValentinesDay` given that the former generalizes to `#$SpatialThing-NonSituational` and the latter to `#$Situation`.

Alternatively, the test allows both of these mappings:

`#$BlackPeppercorn`→'Black pepper'

`#$Pepper-TheSpice`→'Black pepper'

This is correct as the Wikipedia article, despite its title, is more general than both Cyc terms, explaining how the spice (both black and white) is produced from the peppercorns. The strategy does make some mistakes. For instance having decided that `#$Countess`→'Count' has greater semantic similarity than `#$Count-Nobleman` →'Count', the method then proceeds to reject `#$Count-Nobleman` (which would in fact be a better match) because Cyc's collections of females and males are disjoint.

With this strategy we eliminate approximately 10K mappings, which gives us a total of 42,279 – 50% of the original 83,897. Next we evaluate, whether the precision of these mappings is improved.

4. Evaluation

We evaluate the proposed methods using two data sets. The first (Testset1), kindly offered to us by the Cyc Foundation, contains 9,436 synonymous mappings between Cyc categories and Wikipedia articles – created semi-automatically by one person. Evaluation is made more difficult by the fact that at times more than one answer can be correct (e.g. `#$BabyCarrier` can be mapped to either 'Baby sling' or 'Child carrier'). Therefore we also investigate human inter-agreement on the mapping task by giving a new set (Testset2) with 100 random Cyc terms to 6 human subjects. The goal of the algorithm is to achieve as high agreement with the subjects as they with each other.

4.1 Results for Testset1

Out of 9,436 examples in the first data set, we exclude

	Found	Correct	P	R	F
Mapping I	4655	4477	96.2	48.0	64.0
Mapping I & II	6354	5969	93.9	64.0	76.1

Table 1. Results for non-ambiguous mappings in Testset1: precision (%), recall (%), F-Measure (%).

	Before common-sense disambiguation					After common-sense disambiguation				
	Found	Correct	P	R	F	Found	Correct	P	R	F
Synonym-based	8884	7958	89.6	85.3	87.4	7715	7022	91.0	72.5	82.4
Context-based	8657	8054	93.0	**86.3**	**89.5**	7763	7386	**95.1**	79.1	86.4

Table 2. Results for disambiguated mappings in Testset1: precision (%), recall (%), F-Measure (%).

those that link to Wikipedia categories or particular parts of Wikipedia articles. Tables 1 and 2 investigate mapping of the remaining 9,333. Our Mapping I alone covers 4,655 examples, out of which 4,477 are correct. Moreover, manual examination of the 'incorrect' mappings reveals that their vast majority is actually correct—our method often identified more precise mappings than the given ones, e.g.:

- `#$Plumage` → 'Plumage' instead of 'Feather'
- `#$TransportAircraft` → 'Transport aircraft' instead of 'Cargo aircraft'

By including synonyms listed in Cyc in Mapping II, we found an additional 1,699 mappings with 1,492 correct according to the test set (precision 87.8%). Here often 'incorrect' mappings occur because the meaning is too close. For example, the Cyc term `#$SacAndFoxLanguage` was mapped to 'Fox (tribe)', via Cyc's synonym *sac and fox*. which in Wikipedia means the tribe. However, in the majority of cases this strategy worked well, e.g. `#$AeolicGreekDialect` → *aeolic greek* → 'Aeolic Greek'.

The last row of Table 1 summarizes the results of Mappings I and II combined. We covered 68% of the test set with precision of almost 94%. The remaining 32% of the test set, 2,979 terms, are either difficult to find or ambiguous. With the stronger conflation strategies (cf. Section 3.2), we identify an additional bulk of terms with at least one mapping and disambiguate them to Wikipedia articles with our two methods: synonym-matching vs. context-based similarity. We additionally evaluate, whether common-sense disambiguation (Section 3.3) improved the accuracy as expected. Table 2 compares the performance of the algorithm under each setting, giving the overall results, when disambiguation is combined with Mapping I and II.

Context-based disambiguation clearly outperforms the synonym-based approach and achieves maximum precision of 95.1%, when the disjointness test is used. The best recall, 86.3%, is achieved without the common-sense disambiguation, however the precision is more than 2 points lower. There is an obvious trade-off between precision and recall, and for some applications one could be more important than the other.

Manual analysis of errors shows different reasons for incorrect mappings, e.g. inaccuracies in Wikipedia, errors in the test set, insufficient context, very close meanings, or inconsistencies in Cyc. For instance, insufficient context led to erroneous mapping `#$AnticommunistIdeology`→'Communism', because it is more common than 'Anti-Communism'. Sometimes, very similar meanings could not be differentiated, e.g. `#$CityOfKyotoJapan` and `#$Kyoto-PrefectureJapan` are both mapped to 'Kyoto'. Both pages have high similarity with their context, whereas 'Kyoto Prefecture' is less a common page. Treating specification after the dash sign as context and not as a part of the title, results in `#$Tea-Iced`→'Tea' instead of 'Iced tea.' This is an example of inconsistency in Cyc.

	Agreement with other subjects	Agreement with algorithm before final disambiguation	Agreement with algorithm after final disambiguation
Subject 1	37.6	34.0	28.0
Subject 2	40.4	41.0	31.0
Subject 3	40.8	40.0	29.0
Subject 4	40.8	41.0	30.0
Subject 5	42.4	44.0	32.0
Subject 6	37.0	35.0	28.0
Overall	39.8	39.2	29.7

Table 3. Results for the final mapping algorithm on Testset2.

4.2 Results for Testset2

We created a second test set with 100 random Cyc categories, which six human subjects independently mapped to a Wikipedia articles. The instructions were to map only if both resources define the same concept, with the aid of the Wikipedia search function.

Interestingly, the number of mapped concepts varied across the subjects. All agreed that there is no mapping in only 22 cases. On average they mapped 56 Cyc terms, ranging from 47 to 65. The algorithm was again tested with and without the common-sense disambiguation, where the former mapped 58 and the latter only 39 terms. Note that the creator of Testeset1 did not include 'mappings' where a Cyc term had no corresponding Wikipedia article, whereas Testset2 was created randomly from all common-sense terms in Cyc. This is why both humans and the algorithm have lower coverage on this set.

To compute the agreement we compared mapped concepts between each pair of human subjects, and between each human subject and our algorithm. Table 3 summarizes the results. The overall agreement between our subjects is 39.8%. The surprisingly low coverage of the algorithm, when common-sense disambiguation is applied, results in very low agreement on this data set of only 29.7%. However, without this test the algorithm performs nearly as well as the human subjects (39.2%). In fact, it outperforms Subjects 1 and 6.

Error analysis shows that in some cases the algorithm picked a more general article than the humans, e.g. `#$Crop`→'Agriculture', instead of 'Crop (agriculture)', picked by all subjects, or `#$StarTrek-GameProgram` →'Star Trek', instead of 'Star Trek Games', as identified by one subject, or 'Star Trek Generation (video game)', by another, while the others failed to produce any mapping. In a few cases, the algorithm identified a mapping where most humans failed, e.g. `#$BurmesePerson`→'Bamar'.

5. Adding new information to Cyc

Now that the alignment has been performed, could any new information be added to Cyc from Wikipedia?

Synonyms: Despite the extensive work on its natural language interface, Cyc is weak at identifying its concepts. For instance, typing "Casino" into Cyc's search function does not retrieve `#$Casino-TheMovie`. Given 42,279 more accurate mappings, we can retrieve over 154,800 synonyms from Wikipedia (≈2.6 per term), of which only 8,390 are known by Cyc.

Translations: Currently, there are over 200 different language versions of Wikipedia. 15 versions have over 100,000 articles, and 75 have at least 10,000. We have estimated that given 42,836 mappings of Cyc terms to Wikipedia articles, we can retrieve over 500,000 translations of these terms in other languages, which is about 13 per term.

Glosses: For each mapping we can retrieve the first paragraph of Wikipedia's article, which would enrich Cyc's hand-written `#$comment` on that term.

URL Resources: Using triplets in DBpedia's infobox dump, we identified 1,475 links to URLs corresponding to the Wikipedia concepts that we have mapped to Cyc.

New Relations: Many other relations in the DBpedia dataset bear a significant similarity to Cyc predicates, e.g.:

`keyPeople` ↔ `#$keyGroupMembers`
`capital` ↔ `#$capitalCity`

However manual analysis has shown that much of the dumped data is of poor quality (e.g. `keyPeople` assertions of the form "CEO" or "Bob", `capital` assertions which name districts rather than cities). Much however could be done to automatically quality-control candidate assertions using Cyc's ontological constraints on the arguments of its predicates – thus for instance as Cyc knows that the first argument to `#$capitalCity` must be a `#$City`, it can reject the claim that the capital of Bahrain is `#$AlManamahDistrict`. We will explore this in future work.

6. Conclusions

We map 52,690 Cyc terms to Wikipedia articles, with a precision of 93%. Evaluation shows that this mapping technique achieves the same agreement with 6 human subjects as they do with each other. We also show how more accurate results can be achieved using Cyc's commonsense knowledge.

Our work opens up considerable possibilities for further enriching Cyc's ontological rigor with Wikipedia's folksonomic bounty.

Acknowledgements

We thank Mark Baltzegar, Sudarshan Palliyil, Tim Wilmot-Sitwell, Glen Meyer, Kendall Lister, Brett Summers, Marios Daoutis and Cycorp Inc. for their generous help with this research.

References

Auer, S.; Bizer, C.; Kobilarov, G.; and Lehmann, C. et al. 2007. DBPedia: A nucleus for a Web of open data. Aberer, K. et al (eds.) *ISWC/ASWC 2007, LNCS 4825*. Springer-Verlag, Berlin Heidelberg, pp. 722-35.

Gregorowicz, A.; and Kramer, M.A. 2006 Mining a large-scale term-concept network from Wikipedia. *Tech. report, The MITRE Corporation*.

Gruber, T.R. 1995. Toward principles for the design of ontologies used for knowledge-sharing. *Int. Journal of Human and Computer Studies*, 43 (5/6), 907-28.

Guarino, N. (1998). Formal ontology and information systems. *Proc. FOIS-98*, Trento, Italy.

Hepp, M.; Bachlechner, D.; and Siorpaes, K. 2006. Harvesting Wiki Consensus-Using Wikipedia Entries as Ontology. *Proc. ESWC-06 Workshop on Semantic Wikis*, pp.132-46.

Herbelot, A.; and Copestake, A. 2006. Acquiring Ontological Relationships from Wikipedia Using RMRS. *Proc. ISWC-06 Workshop on Web Content Mining with Human Language.*

Legg, C. 2007. Ontologies on the Semantic Web. *Annual Review of Information Science and Technology* 41, 407-52.

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

Mann, G.S. 2002. Fine-grained Proper Noun Ontologies for Question Answering. *Proc. ICCL-02, SEMANET: Building and Using Semantic Networks*, Vol. 11, 1-7.

Matuszek, C.; Witbrock, M.; Kahlert, R.C. et al. 2005. Searching for Common Sense: Populating Cyc from the Web. *Proc. AAAI-05*. Pittsburgh, Penn., pp.1430-1435.

Milne, D.; and Witten, I.H. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *Proc. AAAI-08 Workshop Wikipedia and the AI*.

Mihalcea, R.; and Csomai, A. 2007. Wikify!: Linking documents to encyclopedic knowledge. *Proc. CIKM-07*, pp. 233-242.

Niles, I.; Pease, A. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proc. IEEE IKE-03*, pp. 412-416

Reed, S.; and Lenat, D.B. 2002. Mapping ontologies into Cyc. *Proc. AAAI Workshop Ontologies for the Semantic Web*, Edmonton, Canada.

Rosse C.; and Mejino J.V.L. 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform*. 36:478-500.

Ponzetto, S. P.; and Strube, M. 2007. Deriving a Large Scale Taxonomy from Wikipedia. *Proc. of AAAI-07*, pp. 1440-1445.