Naturalness of Ontology Concepts for Rating Aspects of the Semantic Web

Yoo Jung An, Kuo-chuan Huang and James Geller
Department of Computer Science
New Jersey Institute of Technology,
Newark, NJ 07102
ya8.kh8.geller@oak.njit.edu

ABSTRACT

The Semantic Web is expected to be the next generation of the WWW. Ontologies and agents are major ingredients of the Semantic Web. In (Lee and Geller 2005), we have argued that some existing ontologies make use of unnatural concepts. We stated that unnatural concepts make it difficult to use an ontology and they contradict the purpose of an ontology, which includes explanatory power for the purpose of sharing information. In this paper we elaborate what makes a concept unnatural. We also analyze existing ontologies to get numeric measures of how natural their concepts are.

INTRODUCTION

The Semantic Web

Recently the Semantic Web (Berners-Lee et al., 2001) has become a major research topic in several subdisciplines of Computer Science and Information Systems. The goal of the Semantic Web is to automate many tasks that humans perform with the WWW today. The vision of the Semantic Web relies on agent programs and softbots roaming the Web, finding data and services, combining them and returning them to their user. These agents will need some human-like knowledge to perform their tasks. For this purpose, ontologies will be sprinkled all over the Semantic Web.

We will now briefly summarize one possible (widely used) view of what ontologies are. Ontologies are computer implementations of human-like knowledge, for the purpose of describing domains of the world and sharing this knowledge between application programs (and also between people). Concepts are the fundamental building blocks of all ontologies. A concept is "a unit that one can think about." Concepts are typically drawn as boxes in ontology diagrams, with the concept name written inside. Concepts are connected by IS-A links (arrows) which describe generalization/specialization relationships between two concepts.

Humans appear to have additional local information about concepts. Thus, humans know, for example, the color, size, etc. of many physical concepts. We call this kind of local information "attributes." Besides the IS-A links, ontologies contain other links, e.g., part-of links. These links are called semantic relationships and are also drawn as arrows. The relationship name is written near the arrow.
Figure 1 shows an example of a (hypothetical) medical ontology applied to bacterial pneumonia and its treatment. *Drug, Pneumonia, Antibiotic* and *Erythromycin* are concepts. *Dosage* is an attribute. It is an integer number and measured in milligrams which will specify the typical dosage of a drug. The arrow next to the label "is-prescribed-for" is an example of a semantic relationship. Semantic relationships are distinct from the special purpose IS-A relationship arrows, which form the specialization/generalization hierarchy. Thus, *Erythromycin* IS-A *Antibiotic* says that *Antibiotic* is more general than *Erythromycin*. In other words, all real world instances of *Erythromycin* form a subset of all real world instances of *Antibiotics*. Figure 2 shows an analog example from an E-commerce ontology.

Experience with existing ontologies has shown that they are often not as useful and easy to use as one would hope, a phenomenon referred to as *knowledge use paradox* in (Geller et al., 2004). To quantify this problem, we are attacking the question of what it means for the concepts of an ontology to be natural, thus making the ontology (more) understandable to human users.

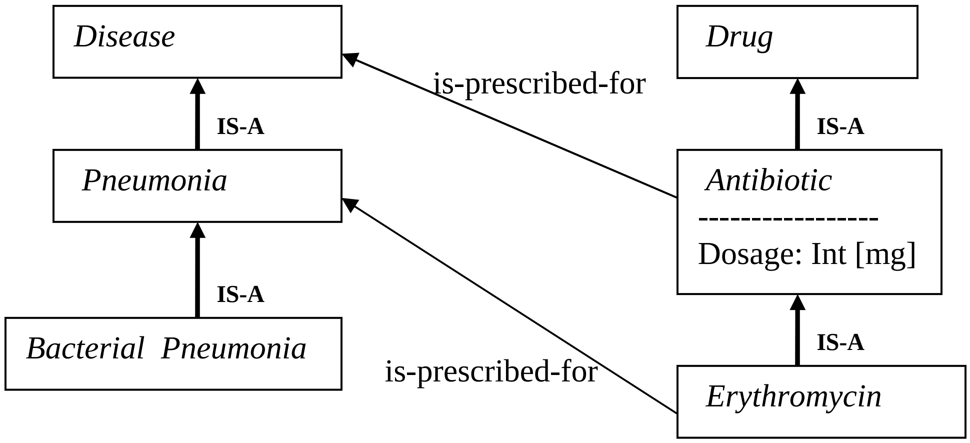


Figure 1: Example of medical ontology applied to the treatment of bacterial pneumonia

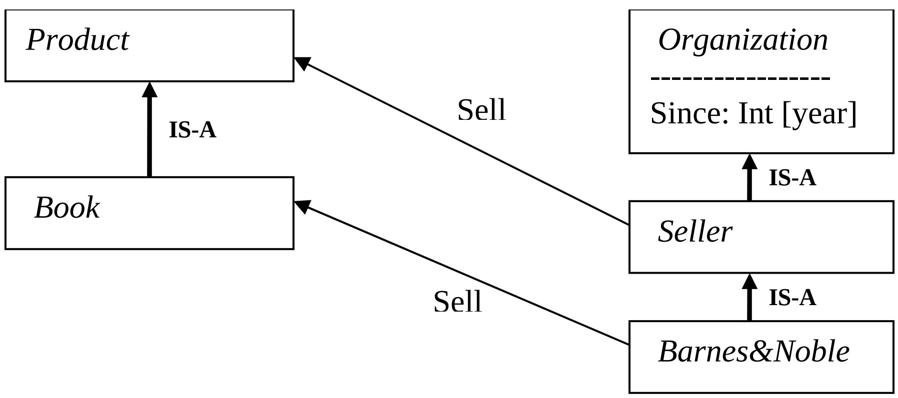


Figure 2: Analog example of an E-commerce ontology

Naturalness of an Ontology

Gruber (1995), in his original work on ontologies, stressed that ontologies are about knowledge sharing. We are raising the question whether existing ontologies are constructed so that they may succeed at this task. Specifically, we are concentrating on the concepts used in such an ontology. In this paper, we are limiting ourselves to ontologies with concepts that are labeled in English. We observe that many labels correspond to English words that could be located in a dictionary. Other concepts are labeled with complex combinations of words, sometimes written in a way such that all words after the first one have their initial letter capitalized, while all other letters are lower case. An example of such a concept label would be destinationCity. This kind of label is neither upper case nor lower case but "camel case."

A two-word lexeme or multi-word lexeme might correspond to a concept that is common enough to be listed in a dictionary or gazetteer (such as New York). In such a case, the meaning of the concept cannot be derived from the parts of the label. Thus, there is no New Massachusetts and no New Vermont. On the other hand, in many cases the meaning of a label is the compositional result of its parts. Thus, the meaning of destinationCity may be derived from the meanings of "destination" and "city."

In our review of ontologies we have found concept labels of the above kinds (single-word with successful dictionary lookup, multi-word also with successful dictionary lookup, and multi-word with compositional meaning). However, we have also found concept labels for which dictionary lookup fails to find an entry and for which the compositional meaning is difficult (if not impossible) to derive from its components in a consistent manner for different participants in an act of communication. Thus, a sender of such a label might assume a different meaning than the one understood by a receiver. For example, it is hard to judge what the exact meaning is of *Plan Action Restriction Predicate*. If a term with compositional meaning consists of many individual words, it may become increasingly more difficult to understand it.

In this paper, we attempt to quantify the idea of naturalness of ontology concepts. This paper is organized as follows. First, we review relevant literature and refine our research questions. Then our methodology is described and the results of our study are shown. Finally, conclusions are presented.

LITERATURE REVIEW

The quality of ontologies (QoO) has been investigated from two main perspectives: knowledge sharing and reusability. The first perspective is related to semantic interoperability (Ram & Park, 2004) and compatibilities of different ontologies (Burgun & Bodenreider, 2001). The importance of interoperability and compatibility of ontologies can be easily verified, as one of the main reasons why we build ontologies is the sharing of knowledge (Noy & McGuinness, 2005). Brachman (1992) presented a methodology and analysis to enhance the semantics of ontologies.

The second perspective of reusability is related to selecting an appropriate existing ontology for an application, e.g. from domain ontologies deployed for agents' interoperation on the Semantic Web. From this perspective, classification and measurement of features become research issues (Alani & Brewster, 2005; Magkanaraki et al., 2002; Supekar, Patel, & Lee, 2004). In these studies, common features, sometimes called measures of the QoO, are depth and density. These numbers are regarded as quantitative indicators of the QoO, based on the assumption that the hierarchical *structure* of an ontology can be described adequately by a set of numbers. On the other hand, a software agent might have the problem of selecting one of several ontologies that contain a certain search term. Thereby, matching the concepts of an ontology with the search term and/or evaluating semantic similarity (i.e., the minimum number of relationships a concept pair holds) are included in the framework of ranking ontologies (Alani & Brewster, 2005). In these studies, most of the ontologies experimented with are domain specific and formatted in OWL (Web Ontology Language) and/or RDF (Resource Description Framework). The sizes of these ontologies are typically relatively small and the numeric features can be obtained by processing the OWL/RDF source files.

In an important survey paper (Noy & Hafner, 1997), some existing ontologies were described and compared with regard to feature information (e.g., size, purpose, and so on), however, numeric values and classifications were discussed without being supported by a mathematical model. The survey lists three well-known ontologies, which many researchers have used and referenced, namely WordNet, OpenCyc and the UMLS. Some researchers have presented modified or enriched ontological models by adding new types and trimming some detailed relationships from existing ontologies (Stone, Wu, & Greenblatt 2004).

The perspective of evaluating the QoO in this paper is focusing on ranking large important ontologies based on certain formalized features and a mathematical model. Knowledge engineers use formal or semi-formal languages in order to build ontologies (Colomb 2005). A language provides us with interchangeable words, so the words, which are comprehensible by users, are an important factor when we evaluate ontologies (Lewis 1983).

In this paper, we attempt to circumscribe "naturalness" as a central notion of our research. An informal definition of the naturalness of the concepts used in an ontology is based on the question whether the concept labels are comprehensible to the users. As noted above, in some cases the meaning of a label can and has to be constructed from its words. In other cases this is not necessary (single-word) or not possible ("New York"). According to our assumptions described in more detail below, the naturalness of concepts will support the unambiguous communication of knowledge, which is an important factor to assess the quality of an ontology (Pisanelli, Gangemi, & Steve 1998).

It is our assumption that a term that is widely used is more likely to be natural. Note that naturalness and meaning are orthogonal. Thus DNA and Deoxyribonucleic Acid have the same meaning. Yet, DNA, commonly used in the popular press nowadays, is a more natural term to most people than Deoxyribonucleic Acid. Thus, in order to assign a numeric value to naturalness, we need to know how often a term is used, e.g. by searching a large corpus. In recent years, the Internet has become popular as an ersatz corpus, and, luckily, Google provides us with frequency information also. Thus, a Google search for DNA finds 118,000,000 hits, while a search for Deoxyribonucleic Acid finds only 1,750,000 instances.

In this paper, we propose the following measure of the naturalness of an ontology. If an ontology consists of (a majority of) labels that can be found in a dictionary lookup, then this ontology is more natural than an ontology for which this is not the case. For making this judgement, we allow both single-word and (expanded) multi-word labels. The naturalness of a concept is approximated by the numeric value which represents how many times the concept label is found by Google. Our approach is comparative. We have evaluated some of the most popular ontologies and terminologies for frequency of concept occurrence.

Research Hypothesis

As a first approximation, we propose the following measure of the concept naturalness of an ontology. If an ontology consists of (a majority of) labels that can be found in a dictionary lookup, then this ontology is more natural than an ontology for which this is not the case. For making this judgement, we allow both single-word and (expanded) multi-word labels.

Secondly, we hypothesize that an ontology with long (measured in the number of component words) multi-word labels will be less natural than an ontology with short multi-word labels. However, one may presume that the opposite is the case, and that the additional words in a label contribute to making it more clear instead of less clear. Thus, this is a research issue. Our working assumption is that an ontology with long multi-word labels will be less natural than an ontology with short multi-word labels, because the longer a phrase is, the less likely it will be that it can be located in a corpus.

METHODOLOGY

Data Sources

Three major ontologies, WordNet, UMLS and OpenCyc are investigated in this paper. WordNet is a large scale lexical reference system for the purpose of natural language processing. It contains 117,097 nouns and the average number of children for each noun is 1.027. Therefore, it can supply us with 117,097*1.207 parent-child pairs (Princeton University, 1998).

OpenCyc is the open source version of CYC, which is for general knowledge processing. CYC was built with the intention of providing common sense knowledge of concepts and reasoning rules. OpenCyc had 47,000 concepts at the initial release, and 300,000 in the release 0.9 (Cycorp, 2005). CYC vocabulary words (i.e., constants) can be conceptually divided into two parts: collections (classes) and individuals (Cycorp, 2006). A “collection” is a class of things (objects) while an “individual” is a single object, not a collection. For example, “dog” is a class while Snoopy is an individual. As this important distinction was made by the CYC designers, these conceptually different parts of CYC are separately considered in our analysis.

The UMLS is a large-scale knowledge base used in Medical Informatics. The UMLS consists of three parts of which we are interested in two, the Metathesaurus and the Semantic Network with its semantic relationships. The Metathesaurus had 900,551 concepts in the version of 2003 and the Semantic Network has 135 terms, with 612 relationships (Lister Hill National Center, 2006). As the Semantic Network is a tree with two roots, Entity and Event, there are 133 IS-A links. Entity has 99 descendants and Event has 34 descendants. In summary, we are using the following ontologies/terminologies in this research: (1) WordNet, (2) OpenCyc Class, (3) OpenCyc Individual, (4) OpenCyc (Complete), (5) UMLS Semantic Network, and (6) UMLS Metathesaurus. A partial list of concepts randomly selected for our analysis is in the Appendix.

Details of the Methodology

The flow by which our research was conducted is briefly described in Figure 3. Since the necessary data comes from different sources, several programs were implemented to gather the data.

Phase I: Extract Data

To retrieve concepts from WordNet, we used Java WordNet Library (JWNL), an API for accessing the WordNet database (Princeton University, 2006; Open Source Technology Group, 2006). We used a static method, Dictionary.getInstance().getRandomIndexWord (POS.NOUN), to get nouns at random from WordNet. Each word may have many senses. We picked the first sense as the concept.

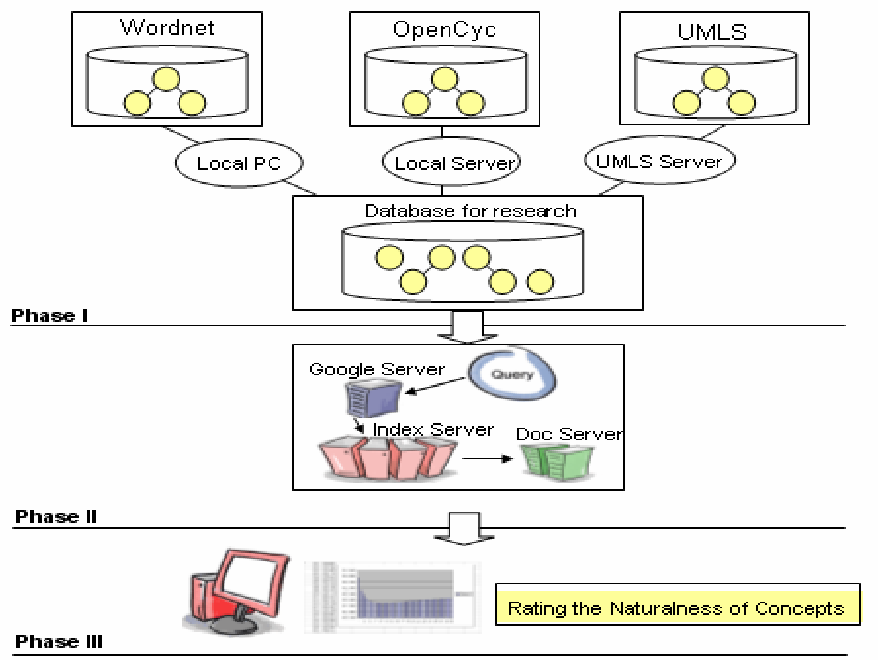


Figure 3: The flowchart of our analysis

For OpenCyc, the database server (Cycorp, 2005) is run on our local host and the OpenCyc API (Open Source Technology Group, 2006) was used to access the database. The getRandomConstant() method in the CycAccess class from the API will randomly return a constant from the database. Constants in the OpenCyc ontology are words, called concepts, that can be perceived and/or mostly understood (Cycorp, 2006).

The UMLS offers a SQL query capability in XML format to directly get records from its database, hosted at http://umlsks.nlm.nih.gov, by using Java Remote Method Invocation. The table "mrrel" in the UMLS database is used to store all IS-A records for the Metathesaurus. The XML query expression is used to retrieve all the IS-A records from the database (see Appendix). In this paper, concepts of the Metathesaurus of the UMLS are restricted to the concepts which have IS-A relationships. The set of those concepts was used in its entirety. The Semantic Network was also used in its entirety, as it is of small size. The data collection was conducted between October 16 and November 10, 2005.

Phase II: Transfer and Load Data

Our primary data collection instrument consisted of the application programs that we have developed for this research. We retrieved the number of search results returned by Google in response to a label of a concept. Concepts are sampled from the ontologies mentioned above and sent to the Google search engine (Brin and Page, 2002). Note that the number of search results of each label is the testable variable in this quantitative research.

Phase III: Analyze Data

For our statistical analysis, SAS software was used. Statistical parameters of two populations are highly unlikely to be identical, and statistical methods allow us to decide whether two different values are indeed significantly different. In this paper we use, for example, the t-statistic (pooled/equal and Sattherwaite/unequal) to determine whether significant differences exist between pairs or groups of ontology components. In order to compare more than two means of ontology components and different conditions (i.e., different label lengths), we employ the ANOVA test.

RESULTS

The Descriptive Statistics

Following are the symbols used in this paper for ontologies: “W” = WordNet, “US” = UMLS Semantic Network, “UM” = UMLS Metathesaurus, “OCC” = OpenCyc Class, “OCI” = OpenCyc Individual, and “OC” = OpenCyc (Class and Individual combined).

Table 1 shows the descriptive statistics for the naturalness of concepts. We use the following statistical measurements for our analysis: Mean, the mean value of the number of search results for a concept; Standard Deviation; Range, the difference between the minimum and the maximum. The ontology with the highest mean value, 5,059,901 is the Semantic Network and the ontology with the lowest mean value, 21,499 is the Metathesaurus of the UMLS. This agrees with our intuition, as the Metathesaurus contains many highly specialized medical terms. Note that the sample sizes are quite different. For large ontologies/terminologies, sample sizes of about 3000 were used, because processing the whole terminology was beyond our computational resources. However, we were able to process the complete UMLS Semantic Network, as it is small. Also, the part of the Metathesaurus consisting of all concepts with IS-A relationships was processed.

	W	**US**	**UM**
Sample Size	3,000	135	4,858
Mean	837,584	5,059,901	21,499
Standard Deviation	6,841,168	16,622,855	313,416
Range	195,999,948	116,999,967	16,800,000
	OC	**OCC**	**OCI**
Sample Size	3,000	1,434	1,566
Mean	540,426	771,400	328,921
Standard Deviation	5,186,753	6,983,312	2,608,801
Range	182E6	182E6	6.805584E12

Table 1: The descriptive statistics

Figure 4 shows information about the frequency of single-word labels and multi-word labels for each of our ontologies. WordNet contains relatively the most single word concepts with 45.17 percent of all the concepts. On the other hand, OpenCyc Individual contains the fewest single word concepts, with only 3.32 percent.

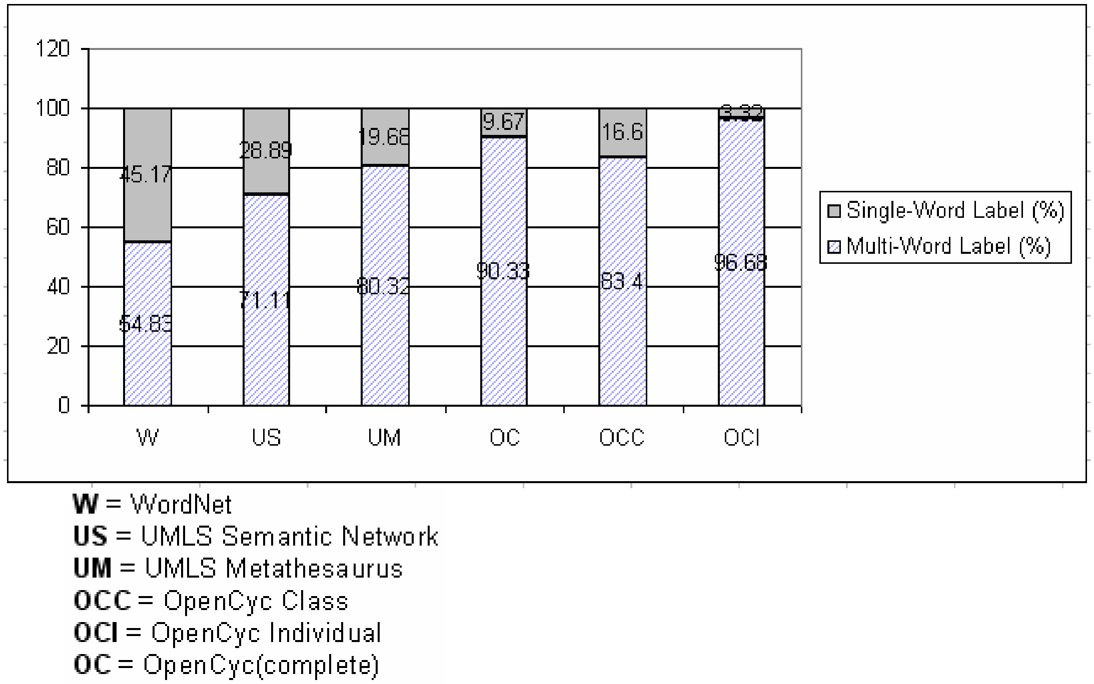


Figure 4: The Composition Ratio between Single Word Label and Multi-Word Label

Comparison of Means in Two Independent Ontologies

Table 2 shows the difference of means between different ontologies. The following statistical measurements with their corresponding symbols are used for this analysis: the degrees of freedom, “DF”, the t-statistic, “t Value”, and the probability of a t-value, “Pr>|t|”, in which the mean value of equal or greater absolute value conforms to the null hypothesis (UCLA Academic Technology Services, 2006).

1. OpenCyc vs. WordNet

The value of unequal (Satterthwaite) t is –1.90, and the p-value is 0.0580. At the 5% level, the two means are not significantly different. This is the p-value for a two sided test. Since the hypothesis is one-sided, this p-value needs to be divided by two to get the appropriate p-value. The p-value then equals 0.029. Thereby, the null hypothesis of equal mean values can be rejected at the 5% level. We conclude that the true average number of search results for WordNet is significantly higher than for OpenCyc (complete).

2. OpenCyc Class vs. WordNet

The value of unequal t is –0.30, and the p-value is 0.7644. At the 5% level, the two means are not significantly different. The null hypothesis of equal means cannot be rejected at the 5% level. We conclude that there is no significant difference between the means of WordNet and OpenCyc Class.

3. OpenCyc Individual vs. WordNet

The value of unequal t is –3.60, and the p-value is 0.0003. At the 5% level, the two means are significantly different. The null hypothesis of equal means can be rejected at the 5% level. We conclude that the true average number of search results for WordNet is significantly higher than for OpenCyc Individual.

Ontologies	Method	DF	t Value	Pr > |t|
OpenCyc vs WordNet	Pooled	5998	-1.90	0.0580
	Satterthwaite	5590	-1.90	0.0580
OpenCyc Class vs. *WordNet*	Pooled	4432	-0.30	0.7647
	Satterthwaite	2771	-0.30	0.7664
OpenCyc Individual vs. *WordNet*	Pooled	4564	-2.84	0.0046
	Satterthwaite	4268	-3.60	0.0003
UMLS Semantic Network vs. *WordNet*	Pooled	3133	6.38	0.0001
	Satterthwaite	136	2.94	0.0039
UMLS Metathesaurus vs. *WordNet*	Pooled	7856	-8.30	0.0001
	Satterthwaite	3007	-6.53	0.0001

Table 2: The t-test for the difference of means between different ontologies

4. UMLS Semantic Network vs. WordNet

The value of unequal t is 2.94, and the p-value is 0.0039. At the 5% level, the two means are significantly different. The null hypothesis of equal means can be rejected at the 5% level. We conclude that the true average number of search results for the UMLS Semantic Network is significantly higher than for WordNet.

5. UMLS Metathesaurus vs. WordNet

The value of unequal t is -6.53, and the p-value is 0.0001. At the 5% level, the two means are significantly different. The null hypothesis of equal means can be rejected at the 5% level. We conclude that the true average number of search results for WordNet is significantly higher than for the UMLS Metathesaurus.

Analysis of the Means for three Groups of Ontologies

In order to compare two or more means, we employ the ANOVA test. The function of ANOVA is to test the null hypothesis namely that the population mean is the same in all conditions or groups. Table 3 shows that the search results from Google are significantly different in the investigated ontologies. There are 10,858 labels as samples from the ontologies. The F-statistic is 32.55. Because the p-value is small, the null hypothesis of equal means for the different ontologies is rejected. The conclusion is that the numbers of search results returned by Google for concepts from different ontologies are different.

The result of Tukey comparisons is shown in Table 3. Three groups, A, B and C are shown below the means. The mean of search results for the different ontologies are significantly different. WordNet is in group A. OpenCyc is in group B. The Metathesaurus of the UMLS is in C. That is, the search results between these three groups, A, B and C are significantly different. Note that the labels from the Semantic Network of the UMLS were removed in this test since its size is very small compared to the other ontologies.

	W	**OC**	**UM**	**F**	Pr > F
Sample Size	3,000	3,000	4,858	32.55	0.0001
Mean	837,584	540,426	21,499		

Tukey Comparison	**A**	**B**	**C**		

Table 3: Results of the ANOVA test to examine the variance among the different ontologies

Analysis of the Influence of Label Length on Naturalness

Table 4 shows that the number of the search results is significantly different according to the number of words in a label. That implies there is one factor, label length, with five levels, having values 1, 2, 3, 4 and 5 which needs to be taken into consideration. We measure the number of words in a concept label as label length. There are 10,406 observations. The F-statistic for testing whether label length is significant is 56.78. Because the p-value is small, the null hypothesis of equal means for the different label lengths is rejected. The conclusion is that the results from Google are different with respect to the different lengths of the labels.

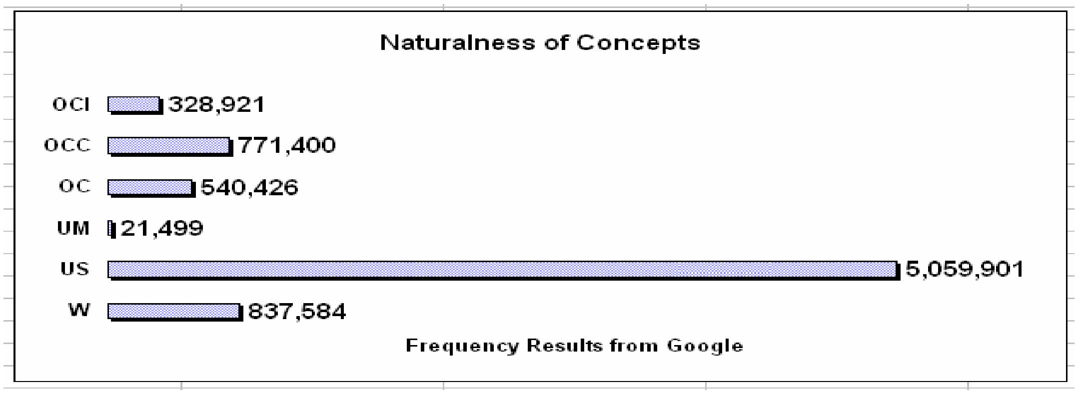
Table 4 also shows the results from the Tukey comparisons. Two groups, A and B are shown below with the means. The conclusion is that the means for the different label lengths are significantly different. The labels of length 1 are significantly different from labels with the lengths, 2, 3, 4 and 5 respectively. On the other hand, from the length 2 and up, labels are not significantly different.

	Single- word	2-word	3-word	4-word	5-word	**F**	Pr > F
Mean	1,731,209	85,815	9,283	3,330	448	56.78	<0.0001
Sample Size	2,640	3,817	2,071	1,171	707		
Tukey Comparison	**A**	**B**	**B**	**B**	**B**		

Table 4: Results of the ANOVA test to examine the variance between multi-word labels and single-word labels for Concepts

SUMMARY AND CONCLUSIONS

Figure 5 shows the mean values of concept occurrences for the different ontologies in our study. In descending order, the ontologies in Figure 5 are the UMLS Semantic Network, WordNet, OpenCyc Class, OpenCyc, OpenCyc Instance and the UMLS Metathesaurus. The rank implies that the corresponding labels of the concepts of the UMLS Semantic Network are most natural, that is they are more widely used and presumably natural. Except for the (small) UMLS Semantic Network, concepts of WordNet are the most natural by our definition of naturalness.



W = WordNet
US = UMLS Semantic Network
UM = UMLS Metathesaurus
OCC = OpenCyc Class
OCI = OpenCyc Individual
OC = OpenCyc(complete)

Figure 5: The graph to describe the naturalness of concepts

In this paper, we presented an extensive analysis of the statistical significance of the measured differences between ontologies. In summary, **(1)** the UMLS Semantic Network is the most natural followed by **(2)** WordNet **and** OpenCyc Class, **(3)** OpenCyc (Complete), **(4)** OpenCyc Individual, and **(5)** the UMLS Metathesaurus. These results are supported by t-tests and ANOVA tests.

One implication of this paper is that it is possible to formalize the quantitative measurement of aspects of the quality of ontologies (QoO). The evaluation model employed in this paper can be used to improve the QoO of an ontology by substituting concepts with low frequencies by their (more natural) synonyms. In addition, the methodology presented in this paper can provide fast on-line computation of the naturalness of concepts when the number of concepts in an ontology is small. Recently, repositories for ontologies have been established (Ding et al., 2004) where small ontologies are saved in the format of the Web Ontology Language (OWL) or Resource Description Framework (RDF). Our methodology could be adapted to on-line computation of the QoO of these small ontologies.

One may argue that our methodology will not work well for scientific domains, as Google will return relatively few hits for specialized scientific terms. There are two possible responses to this argument. On one hand, one may argue that this is exactly what *should* happen. After all, many scientific terms *are not natural* to the average person. On the other hand, different ontologies exist in a few research areas, which could potentially be compared with our method. For example, in Medical Informatics, the PubMed web site at (http://www.ncbi.nlm.nih.gov/ entrez/query.fcgi?DB=pubmed) contains over 16 million citations from medicine. However, using our method would require that all these publications were indexed by a search engine. Note that this is eventually likely to happen as there are currently projects under way of indexing whole libraries, see, e.g., Amazon's "search inside the book" feature. Thus, our approach, employing the number of search results in Google as an indicator of concept naturalness can be valid in specialist domains as well. Brewster (2003)'s experiment has a bearing on this issue. He showed that two related concepts from an ontology are less likely to be found in domain specific texts, like the journal Nature, than on the Internet. Thus, the Internet may be used for frequency analysis, even in scientific domains, until better sources have been made available and indexed.

Given the importance of ontologies for the Semantic Web, we propose to use the naturalness of concepts as the first step towards rating the quality of the Semantic Web (QoSeW). Clearly much work remains to be done in this area. Software agents are another main ingredient of the Semantic Web, and define a very large research area in their own right. The Helsinki agent bibliography, to be found at http://www.cs.helsinki.fi/u/hhelin/agents/agent-bib.pdf, lists over 3000 papers. The big vision of the Semantic Web is that communicating agents of a user will roam the Semantic Web, select useful information and combine it to satisfy the information needs of the user. Software agents will use ontologies, associated with documents and Web services, to gain an understanding of the Web resources that they will encounter in their searches. Semantic Web researchers have, to the best of our knowledge, not expressed an ambition to solve the general Artificial Intelligence (AI) problem(s). It remains to be seen how much intelligent behavior can be achieved by cooperating agents with access to ontologies, and how much Semantic Web research will contribute to solving the larger problems of AI.

REFERENCES

Alani, H., & Brewster, C. (2005). Ontology Ranking based on the Analysis of Concept Structure. *Proceedings of the Third International Conference on Knowledge Capture (K-Cap)*, Banff, Alberta, Canada, 51-58.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34-43.

Brachman, R. J. (1992). Reducing CLASSIC to practice: Knowledge Representation Theory Meets Reality. *Proceedings of the Third International Conference on the Principles of Knowledge Representation and Reasoning* (KR-92), 247-258.

Brewster, C., Ciravegna, F., & Wilks, Y. (2003). Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance. *Proceedings of the Semantic Web Workshop (SIGIR)*, n/a - n/a. Archived at http://eprints.aktors.org/307/

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Retrieved June 13, 2006, from Stanford University, InfoLab Web site: http://www-db.stanford.edu/~backrub/google.html

Burgun, A., & Bodenreider, O. (2001). Mapping the UMLS Semantic Network into General Ontologies. *Proceedings of American Medical Informatics Association (AMIA) Annual Symposium*, 81-85.

Princeton University, Princeton, Cognitive Science Laboratory. (2005). WordNet (Version 2.0) [Computer program]. Retrieved June 13, 2006, from http://wordnet.princeton.edu/oldversions

Princeton University, Princeton, Cognitive Science Laboratory. (2006). *WordNet 2.1 Database Statistics*. Retrieved June 13, 2006, from http://wordnet.princeton.edu/man/wnstats.7WN

Colomb, R. M. (2002, November). *Quality of Ontologies in Interoperating Information Systems*. Retrieved June 13, 2006, from the Institute of Cognitive Science and Technology, Laboratory for Applied Ontology Web site: http://www.loa-cnr.it/Papers/ISIB-CNR-TR-18-02.pdf

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web*. Proceedings of the 13th ACM Conference on Information and Knowledge Management*, 652-659.

Geller, J., Perl, Y., & Lee, J. (2004). Editorial: Ontology Challenges: A Thumbnail Historical Perspective. *Knowledge and Information Systems*, *6*(4), 375-379.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*(5), 907-928.

Lee, Y., & Geller, J. (2005). Semantic Enrichment for Medical Ontologies. *Journal of Biomedical Informatics*, *39*(2), 209-226.

Lewis, D. (1983). New Work for a Theory of Universals. *Australasian Journal of Philosophy*, 61, 343-377.

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine (2006). *The UMLS Semantic Network*. Retrieved June 13, 2006, from U.S. National Institutes of Health Web site: http://semanticnetwork.nlm.nih.gov/

Magkanaraki, A., Alexaki, S., Christophides, V., & Plexousakis, D. (2002). Benchmarking RDF Schemas for the Semantic Web. *Proceedings of the first International Semantic Web Conference (ISWC'02)*, 9-12.

Noy, N. F., & McGuinness, D. L. (2005). *Ontology Development 101: A Guide to Creating Your First Ontology*. Retrieved June 13, 2006, from Stanford Medical Informatics Web site: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Noy, N., & Hafner, C. (1997). The State of the Art in Ontology Design: A Survey and Comparative Review. *AI Magazine*, *18*(3), 53-74.

Noy, N. F. & McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Retrieved June 20, 2006, from http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.

Open Source Technology Group (2006). Java WordNet Library (JWNL 1.3) [Computer program]. Retrieved June 16, 2006, from http://sourceforge.net/projects/jwordnet.

Open Source Technology Group (2006). OpenCyc (Version 0.9.5.Windows NT/2K/XP) [Computer program]. Retrieved June 16, 2006, from http://sourceforge.net/project/showfiles.php?group_id=27274.

Cycorp (2005). OpenCyc (Version 0.9) [Computer program]. Retrieved June 16, 2006, from http://www.opencyc.org/releases/.

Cycorp (2006). *Cyc 101 Tutorial*. Retrieved June 16, 2006, from http://www.opencyc.org/doc/tut/?expand_all=1.

Pisanelli, D. M., Gangemi, A., & Steve, G. (1998). An Ontological Analysis of the UMLS Metathesaurus. *Journal of American Medical Informatics Association*, 5, 810-814.

Ram, S., & Park, J. (2004). Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts. *IEEE Transactions on Knowledge and Data Engineering*, *16*(2), 189-202.

Stone, J., Wu, X., & Greenblatt, M. (2004). An Intelligent Digital Library System for Biologists. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, 491-492.

Supekar, K., Patel, C., & Lee, Y. (2004). Characterizing Quality of Knowledge on Semantic Web. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Symposium Conference*, 220-228.

University of California, Los Angeles, Academic Technology Services (2006). *SAS Annotated Output*. Retrieved June 16, 2006, from http://www.ats.ucla.edu/stat/sas/output/ttest.htm.

APPENDIX 1. A PARTIAL LIST OF CONCEPTS RANDOMLY EXTRACTED FROM THE ONTOLOGIES

WordNet

subclass elasmobranchii	tribe bambuseae	enceliopsis	sea scout
kilometer	fox river	gingivitis	direct grant school
Tupi	trustingness	invalidness	film industry
drawing room car	dressing station	kea	ommatidium
ethanoyl group	menotyphla	augustinian	cortisol

OpenCyc Class

Boxed Set The Product	Inference Metric Binding Level
Olympic Mens Bantamweight Weightlifting	Military Uniform
Path Type By Surface Feature	Light Surface To Air Missile Launcher
High Wind Area	Turkey Bird
Inner Nuclear Membrane	Transitive That Clause Frame Type

OpenCyc Individual

Swarthmore College	Nootka Language
Jumping Jacks Shoes Corporation	adverb Sem Trans
Flowers Industries Corporation	corresponding Cyc Collection
N C O Group Inc	Wintrust Financial Corporation
effective Between Dates	Hotel Investors Corporation

UMLS Metathesaurus

colfosceril palmitate	Hydrocortisone
Sertraline	Pregnanetriol
17-Hydroxycorticosteroids	Tetrahydrocortisol
Cortodoxone	Tetrahydrocortisone
20-alpha-Dihydroprogesterone	17 alpha-Hydroxyprogesterone

UMLS Semantic Network

Organism	Rickettsia or Chlamydia	Amphibian	Human
Plant	Bacterium	Bird	Anatomical Structure
Alga	Animal	Fish	Embryonic Structure
Fungus	Invertebrate	Reptile	Congenital Abnormality
Virus	Vertebrate	Mammal	Acquired Abnormality

APPENDIX 2. XML QUERY FOR SAMPLING CONCEPTS FROM THE UMLS

UMLS Metathesaurus

```
<?xml version="1.0"?>
<query xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
version="1.0">
  <rowset>ConceptRelationship</rowset>
  <row>ISARelation</row>
  <distinct/>
  <dbyear>2005AB</dbyear>
  <fields>
    <fieldSet>
<table>mrrel</table><fieldNames><name>cui1</name><name>cui2</name></fieldNames>
    </fieldSet>
  </fields>
  <constraints>
    <relation>
      <operator>OR</operator>
```

```
      <constraint>
        <lhs><field>mrrel.rela</field></lhs>
        <op>=</op>
        <rhs><string>isa</string></rhs>
      </constraint>
    </relation>
  </constraints>
  <orderBy>
    <name>mrrel.cui1</name>
  </orderBy>
</query>
```

UMLS Semantic Network

```
<?xml version="1.0"?>
<query xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
version="1.0">
  <rowset>SemanticTypesRelations</rowset>
  <row>Relation</row>
  <distinct/>
  <dbyear>2005AB</dbyear>
  <fields>
    <fieldSet>
<table>srstr</table><fieldNames><name>srleft</name><name>srright</name></fieldNames>
    </fieldSet>
  </fields>
  <constraints>
    <relation>
      <operator>OR</operator>
      <constraint>
        <lhs><field>srstr.rel</field></lhs>
        <op>=</op>
        <rhs><string>isa</string></rhs>
      </constraint>
    </relation>
  </constraints>
  <orderBy>
    <name>srstr.srleft</name>
  </orderBy>
</query>
```