Evaluating Ontologies based on the Naturalness of their Preferred Terms

Soon Ae Chun
College of Staten Island
City University of New York
chun@mail.csi.cuny.edu

James Geller
Department of Computer Science
New Jersey Institute of Technology
geller@njit.edu

Abstract

The art and science of building ontologies have been developed to the point where it is not sufficient anymore to design and implement a new ontology. Rather, one needs to follow the process of building an ontology by evaluating its quality in absolute numeric terms. If another ontology in the same domain exists, then the two ontologies should be compared in a quantitative manner to determine which one of them is better. Furthermore, the quality scoring mechanism should provide clues concerning the sections of the ontology (one or both) that need improvement. Ontologies are complex structures which exist in many different variations. Even after imposing a basic structural framework and choosing a domain, two ontologies may be evaluated with respect to a number of different features. In this paper we will concentrate on one single ontology feature and assume that all other features are fixed. We have developed a mechanism to measure the quality of this ontology feature, preferred term(s) based on the concept of naturalness, and show that it agrees very well with human judgments. Thus we provide an approach towards the principled selection of the preferred terms in an ontology.

1. Introduction

Sharing knowledge and experiences among domain participants across different organizational boundaries has been facilitated by the use of domain ontologies. An ontology may be defined as a set of controlled terms or concepts, and serves as a central knowledge base in a domain. In medical informatics, the Unified Medical Language Systems (UMLS) and the SNOMED CT are two well known ontologies.

This research on ontology evaluation has been influenced by a number of sources, including object-oriented modeling of terminologies [1-3], knowledge representation formalisms such as Protégé [4] and the structure and use of the Unified Medical Language Systems (UMLS) [5-7]. An ontology consists of *concepts* as the main ingredients. Concepts are connected by binary directed hierarchical and non-hierarchical relationships. The main hierarchical relationship is referred to as *IS-A*, which defines both a generalization/specialization hierarchy and pathways for inheritance. Variations of *IS-A* exist, such as the *parent/child* relationship. Non-hierarchical relationships between pairs of concepts will be referred to as *semantic relationships*. It is also possible to attach *attributes* to concepts, which usually consist of numbers or text strings.

The concepts are also referred to as classes in some terminologies. While it is possible to work out distinctions between classes and concepts, we will take the inclusive view, just like [8], and equate the two terms in this paper. Similarly, we accept the view of [9] that ontologies and terminologies are increasingly becoming more similar by incorporating each other's features. Thus, our term *concept*, which is also used in the UMLS, will be roughly equated with the Protégé term *class*.

In [10-12], the importance of including auditing as part of the life cycle of ontologies and terminologies has been shown. Previous experience in ontology auditing and analysis has revealed that an ontology may suffer from various problems. For example, auditing medical terminologies in the UMLS [11] uncovered the following types of problems and mistakes:

- Ambiguity
- Non-uniform classification
- Classification errors
- Omissions
- Redundant classifications
- Independently listed synonyms

Ambiguity may arise when there are homonyms where one term may map to two different concepts. There is also a problem when one concept can be mapped to several synonyms, where an ontology designer needs

1530-1605/08 $25.00 © 2008 IEEE

to make a decision which term may represent the concept best. In short, the quality of ontology (QoO) is a major issue one needs to address in designing and auditing an ontology.

In the medical domain, there are often many distinct expressions for the same concept, and there is no canonical expression. For these synonyms, the designer should have a mechanism to find the term that may be the best one to describe each concept, a so-called preferred term. One may choose preferred terms randomly, but it would be better to define an intuitive, yet objective way of identifying a preferred term over the other synonymous terms. In addition, these days, ordinary users, e.g. informed patients, who are not experts in medicine, are interested in accessing medical knowledge bases, but the medical concepts are mostly described for experts and they are not indexed with intuitive terms. Thus a mechanism for choosing a preferred term from a set of synonymous medical terms should be devised to help patients with accessing medical knowledge. Towards this end, we investigate one measure of QoO using the "*Naturalness of a term*" and show how to measure naturalness of a term quantitatively in a way that is supported by an experimental study. Using a natural term to represent a concept is important, since this can facilitate different kinds of search operations, where people tend to use more natural terms as keywords in their searches, as opposed to complex or unusual terms.

This paper is organized as follows: Section 2 describes our model of the structure of an ontology and provides some background on medical ontologies, especially the SNOMED and the UMLS. In section 3, we discuss the problems of synonyms and preferred terms in medical ontologies, and present our approach based on the naturalness of a term. In section 4, we present an experimental study on measuring the naturalness by human subjects and show that it may be approximated by Google search hit counts. We also compare our results with the results of the SNOMED server at NCI and the UMLSKS server. In section 5 we present related work. Section 6 contains conclusions and Section 7 discusses suggestions for future work.

2. Medical Ontologies

The Unified Medical Language System (UMLS) is a large system that was created by the National Library of Medicine (NLM), a subdivision of the US National Institute of Health. The purpose of the UMLS is to support processing of natural language as it appears in biomedical text. Structurally, the UMLS consists of the Semantic Network, the Metathesaurus, and a set of lexical processing tools such as LVG, the Lexical Variant Generator.

The Metathesaurus is a collection of more than one hundred source terminologies, including NCI (National Cancer Institute) terminology, MeSH (Medical Subject Headings), and many others. The NLM provides quarterly updates of the UMLS to its users. Recently, the Metathesaurus has grown to over 1.4 million concepts corresponding to over 7.2 million terms. This makes the UMLS the largest publicly accessible medical terminology, and probably the largest existing terminology. As noted in the introduction, we take the inclusive view and refer to the Metathesaurus as an ontology.

The Semantic Network is a network of high level concepts which are called "semantic types." In all, there are 135 semantic types, organized as two trees. Semantic types of the Semantic Network are assigned to concepts in the Metathesaurus. As such, the semantic types are used to tie together the heterogeneous terminologies comprising the Metathesaurus.

One of the major component terminologies of the UMLS is the SNOMED CT (Standardized Nomenclature of Medicine -- Clinical Terms), designed by the College of American Pathologists (CAP). It consists of over 400,000 concepts, organized in 19 separate hierarchies, including Organism, Substance, Physical Object, etc. As opposed to the UMLS, the SNOMED CT is fairly homogeneous in its structure. One important feature of the SNOMED is that it provides for every concept a preferred term. Thus, even if several synonyms have been defined for a concept, the concept should be referred to by the preferred term. The UMLS uses a four level system of ID numbers for the contents of the Metathesaurus. A Concept Unique Identifiers (CUI) is assigned to each concept. However, a concept may be expressed by different terms. Each term has a Lexical Unique Identifier (LUI). Different terms may exist in small variants which are rarely of interest to humans, but are different to computers. Each such variant has a String Unique Identifier (SUI).

As an example of this structure, "*Colonic Neoplasms*" and "*Colonic Neoplasm*" are considered two different strings, with separate SUIs. However, they both have the same LUI. Similarly, "*Colonic tumor*" and "*Colonic Neoplasm*" have the same meaning. They are assigned two different LUIs but have the same CUI, as they express the same concept.

The AUI (Atomic Unique Identifier) describes the occurrence of a string in a terminology. Thus, "*Colonic Neoplasm*" has two different AUIs, but the same SUI, as the same string occurs both in the NCI terminology and the MSH. The UMLS is accessible at http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template/admin,user,KSS_login.vm Figure 1 shows a partial

display of the result of searching the Metathesaurus for "*Colonic Neoplasm*."

One question that arises is the following: Why does the medical community put so much effort into the development of terminologies and ontologies? In answer to that, medical researchers often need to repurpose available data. For example, a medical researcher might want to compare the incidence of infections across all hospitals of the United States. This kind of information may be available in hospital information systems in the form of electronic patient records (Electronic Medical Records, EMR). However, information in an EMR is not designed for research purposes. Therefore, a researcher would need to aggregate tens of thousands of medical records created by practitioners with very different backgrounds and conventions. Some practitioners may use abbreviations or variants of terms.

This is one place where the UMLS comes in. With its extensive repository of synonymous terms of each concept, linked together to the concept by a common CUI, it becomes possible to translate innumerable variations of terms in a medical record into a common language. If several synonyms of one concept exist, each researcher could make his own decision into which of the synonyms all terms should be translated. However, if, subsequently, another researcher runs another study on the same kind of data, he might choose to translate all synonyms into yet another term, which he might be more comfortable to him due to his education or preferences. This would recreate the problem of incompatibility again, on a different level. Thus, it would be preferable that all researchers translate synonymous terms into one specific, predetermined synonym, the *preferred term*. Preferred terms should be chosen in an objective way.

As mentioned before, Protégé is a popular tool for creating and editing medical ontologies. A large number of plug-ins has been created for Protégé, and the whole architecture of the system has been designed for making it easy to build and integrate extensions. Recently, the Protégé team at Stanford University, under the leadership of Mark Musen, celebrated the 20 year anniversary of Protégé. Due to the wide use of Protégé, we will briefly discuss its metaterminology.

The designation *IS-A relationship* is used in Protégé. However, semantic relationships and attributes are both called *slots* in Protégé, and the distinction is made by the data type that a slot refers to. A slot with a simple numeric or string data type would therefore correspond to an attribute. A slot that refers to another class would correspond to a semantic relationship. We note that axioms are not a research issue in this paper.

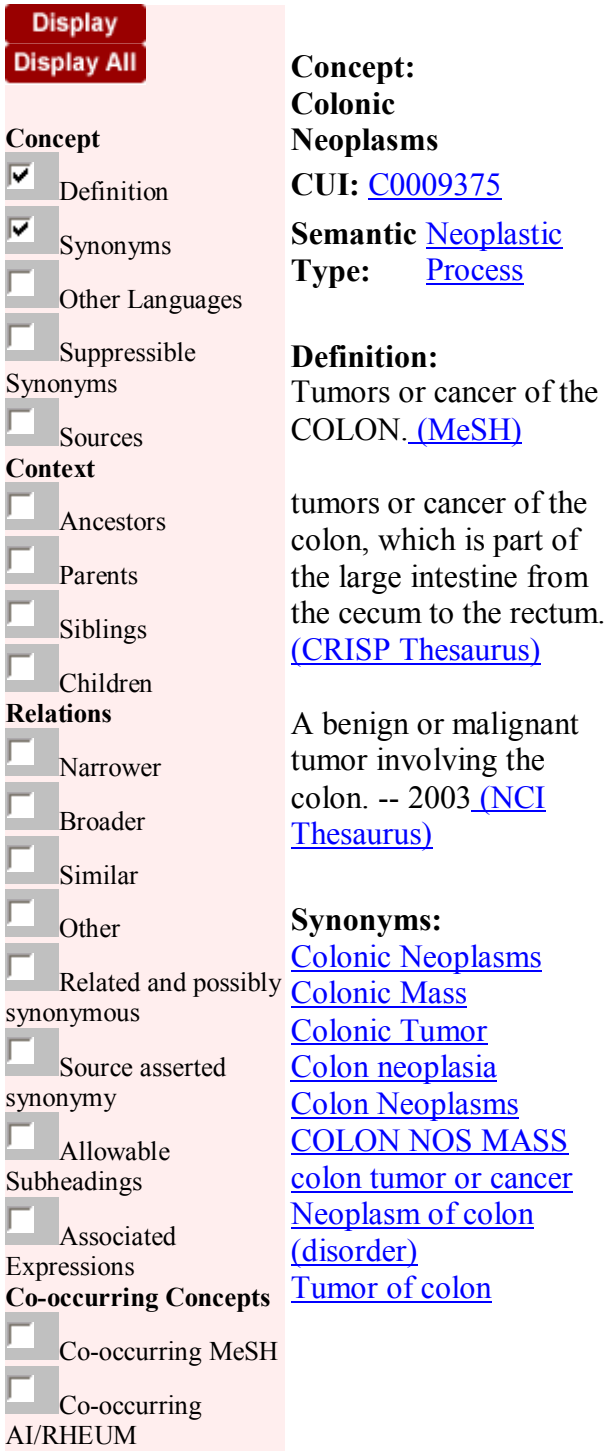

Figure 1: Partial Screen of UMLSKS search results for Colonic Neoplasm in UMLS Release 2007AA

Ontologies are commonly elucidated by visualizing them, either with a drawing produced with a general purpose editor, or with a graphical ontology display tool. Thus, Protégé supports a number of plug-ins, such as Jambalaya and OntoViz for displaying ontologies as

graphs. In such a diagram, every class (concept) appears as a box. IS-A relationships are represented as arrows. Together the boxes and arrows form a hierarchy. Attributes are visible within boxes by focusing in on them. Semantic relationships are often omitted to avoid cluttering up the diagrams. Alternative display models have been developed, such as nested box diagrams, however, we will not discuss those.

Figure 2 show an example of a small, incomplete medical ontology, as displayed with Jambalaya from within Protégé. Even though this example is simple, it shows Protégé's (and Jambalaya's) support for multiple inheritance. Thus, a Neoplastic Process is both a child of Neoplasms and a child of Pathologic Processes.

3. Synonyms and Preferred Terms

Ontologies are concept-based. However, whether one conceives of concepts as something in the mind, the intensional view point [13] or as something in the world, the extensional view point, they are not directly accessible to observation and manipulation. On a day-to-day basis, people are able to communicate, because most concepts are well described by terms. Such terms may consist of one or several words. However, there are two well-studied problems caused by the concept-to-term mapping. A term may refer to several different concepts. Such a term is considered ambiguous, and the different concepts corresponding to the term are referred to as *homonyms*.

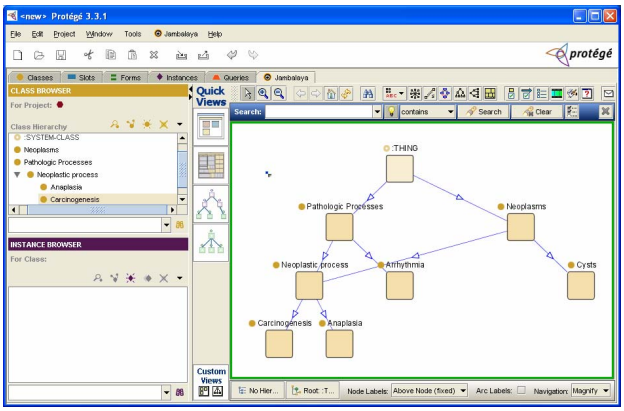

Figure 2: Medical hierarchy in Protégé, displayed with Jambalaya plug-in

Dealing with homonyms is a difficult problem that involves identification and differentiation of meanings, which usually requires human intervention. In the UMLS, numbers are assigned to the different concepts of a term, as in *baldness<1>* and *baldness<2>*.

The opposite problem, the *synonym* problem, is easier to deal with. Two terms are synonyms if they refer to the same concept. All terms describing a concept are collected and assigned to the concept. However, not all terms are considered equal in importance. The SNOMED [14], a major component of the UMLS, selects one term from all the synonyms of a concept and designates it the *preferred term*. The use of preferred terms in terminologies is explicitly sanctioned in the ISO Standard [15] which deals with principles and methods of terminology work. The preferred term of a concept is considered to be a single-valued attribute, while the list of synonyms of the preferred term may be empty, single-valued, or multi-valued. The preferred term of a concept is the term that designers and users of an ontology are encouraged to work with.

The use of preferred terms is especially important in the SNOMED CT. However, it is not clear at all how preferred terms are chosen. The SNOMED User's Guide has this to say about the selection of preferred terms:

> "Each concept has one Preferred Term meant to capture the common word or phrase used by clinicians to name that concept. For example, the concept 54987000 *Repair of common bile duct (procedure)* has the Preferred term "Choledochoplasty" to represent a common name clinicians use to describe the procedure." [16]

However, there is no specification how to determine which phrase is used in each case by clinicians. If all clinicians would always use the same term for the same concept, then there would be no need for a standardized terminology at all!

While some ontologies assign unique identifiers to concepts (such as the CUI number of the UMLS), a human who is studying, extending, or auditing the ontology, or writing an application program for it, usually cannot proceed without consulting an English language representation of the CUI. After all, C0007214 is not meaningful to a person. Thus the preferred term is of special importance, as it is the first and, in many cases, the only, representation of a concept that a user, designer, application programmer or auditor of the ontology is consulting. The preferred term is, in a way, the *handle (or index)* of the concept. Thus, choosing a good preferred term is an important task.

Natural Terms

Ceusters [8] citing [15] notes that preferred terms are designated as "*such terms, according to ISO, should also have the highest rating for acceptability in the relevant user community (though as a matter of fact*

they are often forced upon such a community with the purpose of stabilizing its terminology)."

The "acceptability" is similar to the notion of *naturalness* of terms we used in [17, 18]. Naturalness is used by [19] for expert judgments of terms, and we extended this notion to non-expert users.

The question of how to judge naturalness or acceptability is, however, a very difficult one. It is impossible to let all users of a community choose for all concepts which term they consider most acceptable. A mechanized approach is needed. To measure the naturalness, we hypothesize that the term frequency is an indication of naturalness, assuming that more users would use natural terms more frequently than unnatural synonyms. We have developed a method to quantify the naturalness of terms connected by IS-A links [18] based on the term frequency counts in results of the Google search engine [17]. In other words, a natural term as a search keyword is likely to have higher counts of occurrences in Google search results, compared to when using an unnatural synonym. Interestingly, Ceusters [8] cites the use of Google in an example as a way of rejecting a term as being a good preferred term: "Google finds only 6 hits for the term "anatomic substance" where, according to ISO, preferred terms should be those members of groups of terms which have the highest acceptability rating."

Our basic idea of naturalness deals with language *production* as opposed to language comprehension, as production is easier to observe than comprehension. We hypothesize that when people write, they will typically prefer to use one of several possible terms, such that this term is most natural to them. Thus, if an ontology with sets of synonyms (but without preferred terms) and a large corpus of English language text are available, one may check which synonym of a term is used most often in the corpus. This term should be designated the preferred term. In contrast to this approach, acceptability deals with comprehension of a term. If the ontology already provides a preferred term choice, then a search of the corpus should confirm the choice. This observation provides the key for one (of many possible) numeric measure of the quality of an ontology (QoO).

When comparing two ontologies with different preferred terms, where everything else is equal, the ontology with preferred terms that occur more often in a representative corpus (i.e., natural terms) should be considered the better ontology, than the other ontology with "unnatural terms" as its preferred terms. As the number of preferred terms of the ontology that are confirmed or rejected by the corpus is computable, it is possible to quantify the naturalness of such an ontology.

Generalizing our method to any pair of ontologies, we can compare these ontologies by computing the average number of corpus-confirmed preferred terms relative to the total number of concepts in the ontology.

Definition 1 [Quality of Ontology] *Given an ontology O as a set of terms, T(O) ={t1, t2, ...}, where PT(O) ⊂ T(O) is a set of preferred terms, and NT is a set of natural concept terms, the Quality of Ontology (QoO) is defined as the ratio of frequency of natural preferred terms to that of the total number of terms. (Below, ^ is the logical AND operator).*

$$\text{QoO(O)} = \sum_{t \in T(O)} t \in PT(O) \wedge t \in NT \Big/ \sum_{t \in T(O)} t$$

Thus, given two ontologies O and O', we can compare their quality based on the QoO values. If QoO(O) is greater than QoO(O'), then O is considered better than O'. In order to enhance the Quality of Ontology, ontology designers should choose the natural terms for preferred terms. With existing ontologies, the QoO analysis can be performed for the purpose of evaluating them.

Unfortunately, doing an exhaustive QoO analysis of even a moderately-sized ontology based on a realistic corpus (e.g. the Brown Corpus [20]) is difficult. Thus, we are following the approach in [21] of using the Web instead of a corpus. An objective and automated way of identifying naturalness of terminologies is preferable. We are using the number of hits reported by the Google search engine to measure the naturalness of a term, as hinted in [8] and implemented in [17]. This is based on the assumption that more people, collectively, would use natural terms more frequently. Thus the frequency of natural terms used as search keywords would be higher than that of synonymous other terms. In order to verify that the frequency of a term used in a Google search is in fact in alignment with a human's intuitive notion of naturalness, we performed a pilot survey to find out how good the agreement is between Google numbers and humans. The following section reports on this pilot study and how it relates Google hit numbers with human judgments.

4. Experiment of Measuring Natural Terms

In order to show that there is indeed a connection between Google numbers and human judgments of naturalness we have designed the following experiments. We have chosen 37 terms from Stedman's Concise Medical Dictionary (2nd edition). The following criteria were used when selecting terms.

Only one-word terms were selected, which had an explicitly marked one-word synonym or a one-word explanation. An attempt was made to spread the selection of words over the whole alphabet, however, not at the cost of violating the previous criterion. The resulting list of pairs of words was presented to human subjects.

Secondly, for each of the 37 defined medical terms (T) and each of its 37 synonyms (P), a Google search was performed, and the number of hits reported by Google was recorded. Then the human results and the Google hit count results were compared. Thirdly, we searched for all 74 terms with the SNOMED online server at the NCI Web site (http://nciterms.nci.nih.gov/NCIBrowser/) to determine the agreement between the SNOMED preferred terms and the preferred terms of human subjects and the high hit count Google terms. Lastly, we repeated this experiment with the UMLSKS server at http://umlsks.nlm.nih.gov/kss/.

4.1 Details of the human experiment

We created a list of 37 pairs of one-word terms, as described above. To avoid any order effects, we created two additional alternative versions of the list. In one version we swapped the T and P columns. In the other version, the order of pairs in the list was changed. These versions were randomly distributed among eight subjects. Subjects were given the following instructions:

> "Look at every row in the table below, one row at a time. Each row contains two words from the medical domain which have the same meaning. Please mark with an X which one of the two words is more *natural* to you. There is a separate space after each word for marking it. "Natural" means approximately that you understand this word better and/or that you are more likely to use it when you are speaking. If you do NOT know BOTH words, then please place an X in the rightmost column."

Importantly, there were no additional explanations given to the subjects concerning the meaning of "natural." None of the subjects had any formal medical background.

Results

Table 1 shows the alphabetized list of pairs of terms with their associated Google hit numbers. Table 2 shows judgments of humans. In Table 1, Word 1 is the word to be defined (T), and Word 2 is the synonym (P). Remarkably, in some cases the word used in the definition (P) is not simpler than the word that is being defined (T). The column Google W1 contains the hit number reported by Google for T, and similarly the column Google W2 contains the hit number for P.

Table 1 Result of judging naturalness with Google search hit numbers

Word 1 (T)	Google W1	Word 2 (P)	Google W2
abdomen	16,500,000	belly	34,500,000
aberrant	7,250,000	ectopic	5,210,000
baldness	2,440,000	alopecia	3,500,000
belching	1,030,000	eructation	121,000
calcareous	2,950,000	chalky	1,140,000
chickenpox	1,450,000	varicella	2,730,000
dehydration	10,200,000	desiccation	1,700,000
dementia	17,300,000	amentia	103,000
emmenic	653	menstrual	9,890,000
endemic	12,800,000	enzootic	401,000
feces	6,110,000	stercus	82,600
feverish	2,030,000	febrile	3,090,000
glossal	46,000	lingual	8,250,000
gravid	4,330,000	pregnant	55,600,000
gullet	1,400,000	throat	39,100,000
humpback	1,930,000	kyphosis	607,000
hyperopia	524,000	farsightedness	394,000
incubus	8,320,000	nightmare	38,100,000
insemination	3,480,000	semination	298,000
kneecap	556,000	patella	1,920,000
lactis	1,860,000	milk	103,000,000
lard	5,480,000	adeps	279,000
mammectomy	717	mastectomy	2,730,000
metamorphosis	6,030,000	transformation	108,000,000
narcolepsy	1,580,000	hypnolepsy	387
nasal	18,600,000	rhinal	142,000
odontolysis	363	erosion	34,800,000
ostein	23,200	collagen	16,700,000
pennate	211,000	feathered	4,220,000
potable	20,600,000	drinkable	1,670,000
quack	3,700,000	charlatan	1,310,000
scaly	2,090,000	squamous	6,950,000
serpigo	1,100	herpes	16,100,000
thromboplastid	253	platelet	13,400,000
tumid	105,000	swollen	12,100,000
uresis	32,400	urination	3,280,000
verticilate	180	whorled	362,000

Table 2 shows the same words as Table 1, but each word is followed in the neighboring column by how many subjects considered it more natural. The fifth column shows how many subjects did not know Word 1 and also did not know Word 2. The last column, headed with Agr, shows the agreement between Google hit numbers and human judgments. Simple majority was used for human decisions. Out of 37 pairs of words, for 29 (78%) the Google hit count agreed with human decisions, in some cases dramatically so, e.g., for *platelet*, which had a hit count five orders of magnitude greater than that for *thromboplastid*.

Table 2 Results of humans judging naturalness

Word 1(T)	W 1	Word 2(P)	W 2	Don't know both	Agr
abdomen	**3**	belly	**5**	**0**	**y**
aberrant	**4**	ectopic	**0**	**4**	**y**
baldness	**8**	alopecia	**0**	**0**	**n**
belching	**6**	eructation	**1**	**1**	**y**
calcareous	**1**	chalky	**7**	**0**	**n**
chickenpox	**7**	*varicella*	**0**	**1**	**n**
dehydration	**7**	desiccation	**1**	**0**	**y**
dementia	**5**	amentia	**1**	**2**	**y**
emmenic	**0**	menstrual	**8**	**0**	**y**
endemic	**6**	enzootic	**0**	**2**	**y**
feces	**5**	*stercus*	**0**	**3**	**y**
feverish	**6**	febrile	**1**	**1**	**n**
glossal	**0**	lingual	**6**	**2**	**y**
gravid	**0**	pregnant	**8**	**0**	**y**
gullet	**0**	throat	**8**	**0**	**y**
humpback	**6**	*kyphosis*	**0**	**2**	**y**
hyperopia	**1**	farsightedness	**7**	**0**	**n**
incubus	**0**	nightmare	**8**	**0**	**y**
insemination	**5**	*semination*	**1**	**2**	**y**
kneecap	**8**	patella	**0**	**0**	**n**
lactis	**0**	milk	**8**	**0**	**y**
lard	**5**	*adeps*	**0**	**3**	**y**
mammectomy	**0**	mastectomy	**4**	**4**	**y**
metamorphosis	**2**	transformation	**6**	**0**	**y**
narcolepsy	**3**	*hypnolepsy*	**0**	**5**	**y**
nasal	**7**	*rhinal*	**1**	**0**	**y**
odontolysis	**0**	erosion	**7**	**1**	**y**
ostein	**0**	collagen	**6**	**2**	**y**
pennate	**0**	feathered	**8**	**0**	**y**
potable	**0**	drinkable	**8**	**0**	**n**
quack	**4**	charlatan	**3**	**1**	**y**
scaly	**6**	*squamous*	**0**	**2**	**n**
serpigo	**0**	herpes	**6**	**2**	**y**
thromboplastid	**0**	platelet	**3**	**5**	**y**
tumid	**0**	swollen	**8**	**0**	**y**
uresis	**0**	urination	**8**	**0**	**y**
verticilate	**1**	whorled	**4**	**3**	**y**

For eight terms the Google hit counts contradicted the human subjects. An interesting observation was made when recording these results. Table 2 contains a number of italicized words. All those words were rejected by the Microsoft Word-2007 spelling corrector. There are 20 italicized words. Of those 20, 16 words were chosen by 0 people (nobody) as more natural. Four words were chosen by only one person as more natural. Thus the spelling corrector is in very good agreement with human naturalness judgments. The Google hit count preferred the word marked (erroneously) as wrongly spelled for three pairs. One may hypothesize that if a word is not known to MS Word, it is not a natural word for most humans, but we did not follow up on this observation.

We performed a second experiment with another set of 30 word pairs which were not medical terms. Results in this experiment were stronger than for the experiment reported on here. For 87% of term pairs (26 out of 30), human judgment was in agreement with Google hit counts. For space reasons, and due to our focus on medical terminologies in this paper, we omit further details of this second experiment.

4.2 Results of the SNOMED Experiment

In an ideal world we would hope to find most terms from Table 1 in SNOMED, and have them returned by the SNOMED online server of the National Cancer Institute. However, we found that:

- 34 out of 74 terms were not found by a search at all, i.e., the system reported 0 hits;
- 21 terms were found as the initial word of a longer term, but not by themselves;
- 19 terms were found and reported by the SNOMED as preferred terms.

There are a number of explanations for these low results. SNOMED is concept oriented, thus adjectives in the list are rare in SOMED. SNOMED was designed for the needs of the Medical and Medical Informatics communities. Thus, some of the ancillary terms from the list, even though contained in Stedman's Medical Dictionary, were not found. Lastly, it appears that the Web-based SNOMED retrieval does not report all terms. Thus, even though *Baldness* was returned as a result for the query for *Alopecia, Baldness* was not reported in return to a direct search for it. The term *kneecap* was not found at all.

It is of interest to look at the breakdown of the 19 words which appeared as preferred terms.

- For only four pairs of terms, one term was retrieved from the SNOMED as preferred term. The other one from the pair was found among the terms (synonyms) of the preferred term. This was the case for *Alopecia* (syn: *Baldness*), *Eructation* (syn: *Belching*), *Varicella* (syn: *Chickenpox*) and *Mastectomy* (syn: *Mammectomy*).
- Eight preferred terms had synonyms which were only small variations, e.g., by appending NOS (Not Otherwise Specified) to the preferred term.
- The remaining seven preferred terms had true synonyms, however these did not include the second term of the involved word pair.

As a result of this breakdown, there are only four cases where a comparison between SNOMED preferred terms and human subject preferred terms makes sense. In three cases, SNOMED disagrees with the majority

of humans. Human subjects and SNOMED agree only for *Mastectomy* that it should be the preferred term, compared to *Mammectomy*. Google hit counts, on the other hand, agree with SNOMED preferred terms on three out of these four cases. Only for SNOMED's preferred term *Eructation* does Google prefer the synonym *Belching*.

Given the small number of pairs, no general conclusions can be drawn. Furthermore, the SNOMED was designed by physicians for the medical (informatics) community, while our human subjects were, by design, laymen, non-medial experts. Thus, physicians are likely to prefer different terms than laymen. Yet, it is desirable that the designers of medical terminologies should select their preferred terms in a principled and transparent way, which does not appear to be the case for the tested system.

4.3 Result of UMLSKS Experiment

Results in this experiment were better than for the SNOMED experiment, which is expected, given that SNOMED is a subset of the UMLS. We found that:

- 15 terms were not found
- 13 terms were not found by themselves

For a number of pairs where both terms were found, they were not known as synonyms of each other. Only the following pairs had one term as preferred term (concept name) and the other term as synonym listed under this concept name:

- *Alopecia* (syn: *baldness*)
- *Eructation* (syn: *belching*)
- *Chickenpox* (syn: *varicella*)
- *Hyperopia* (syn: *farsightedness*)
- *Mastectomy* (syn: *mammectomy*)

These results are quite similar to the SNOMED results. However, *Chickenpox* is the concept name chosen, with *varicella* being the synonym. This is contrary to the SNOMED choice and in agreement with human subjects. In addition to the SNOMED results, there is also the pair of *hyperopia* and *farsightedness*, and the UMLS chose *hyperopia* as the concept name, in contrast to human subjects, but in agreement with Google hit counts. Again, the numbers are too small to draw any general conclusions.

4.4 Using Naturalness to improve an Ontology

When one ontology designates preferred terms and synonyms for each concept, the natural terms identified with our methodology can be used as preferred terms. This in turn will increase the QoO. This will work as long as the given synonyms are "synonyms in every context" [15]. Furthermore, our approach to computing naturalness may be used in defining preferred terms in the subdomains of the ontology. Thus, it is possible for a human ontology editor to evaluate the choices of preferred terms made in a subdomain, using the natural terms. The naturalness of terms will be used to audit the human choice of preferred terms, measuring the goodness of the choice, and this in turn will be used to evaluate the overall average naturalness rating of the whole ontology.

In other words, our method can be used in a constructive way to reengineer an existing ontology, even if no comparable ontology in the same domain exists. In this case, an ontological engineer would need to find new synonyms from an outside source (e.g. a thesaurus), find out the hit frequencies and then replace preferred terms of low naturalness by new terms of higher naturalness.

If two ontologies exist in the same domain, which specify different preferred terms for some concepts, then our method can be used to evaluate the quality of ontology. The poor choice of preferred terms can be replaced with more natural terms to improve the quality of ontology. Thus, our approach defines another constructive way to improve ontologies with poorly chosen preferred terms.

5. Related Work

The existence of a "name" or a preferred term for each concept in an ontology appears to be taken for granted by some authors. For example, [22] discusses the effect of using preferred terms and synonyms in queries only in the results section, as it affects the speed of processing them. According to [23], the name of a concept reflects the meaning that the ontology designer intended to encode.

In a widely cited paper defining requirements for medical terminologies [24], Cimino writes that "If a concept may have several different names, one could be chosen as the preferred name and the remainder included as synonyms." However, Cimino does not explain how the preferred name is chosen, either. Thus, our paper closes this gap by providing a quantified approach, based on Google hit numbers, how to select a preferred term among synonyms. Furthermore, our pilot study indicates that Google hit numbers are a good substitute for human subject judgments for this problem.

In [25] a set of transformations are proposed to improve the quality of an ontology. These transformations are based on schema transformation methods that were previously used to improve the quality of database schemas. Similarly, [27] presents schema level and instance level quality metrics, such as

richness in relationships, attributes and inheritance at the schema level, and class richness, cohesion, average population, importance, and fullness at the instance level. In [26] a set of ontology quality measuring dimensions are listed, such as specificity, width, coherence, maintainability, expressiveness, reusability, standards, etc. These approaches use primarily syntactic or structural analyses of ontologies to evaluate them, while our approach focuses on the semantics and usage of the terms in an ontology and proposes operational measures of the quality of an ontology.

6. Conclusions

Well designed ontologies should be measurably better than poorly designed ontologies. While there are a large number of features which could and should be compared between two ontologies, we have concentrated on the selection of good preferred terms, as a preferred term is often the main or only handle for a human user to access a concept. We are using the *naturalness* of a preferred term to quantify its goodness. We have argued that naturalness reflects the likelihood of use of a term in the production of written documents, and is thus a better measure than acceptability. Acceptability is only accessible through work-intensive and laborious experiments with large numbers of human subjects.

Practically speaking, among several potential preferred terms, we select the term with the relatively highest Google hit number as the most natural term. In this paper we reported on a small pilot experiment, which showed that in 78% of all cases human judgments of naturalness coincided with Google numbers. We also compared these results with the preferred terms of the NCI SNOMED and UMLSKS Web servers. This comparison resulted in very few matched terms, thus no results can be derived. We concluded our paper with ideas how to use naturalness to reengineer and improve existing ontologies.

7. Limitations and Future Work

The comparison of the naturalness judgments by humans with the preferred terms of SNOMED and the concept names of the UMLS is open to criticism, as the SNOMED was built with the preferred terms of clinicians and medical experts in mind. We are interested in the access of non-experts, so called informed patients, to medical information, thus we are more interested in their notion of naturalness.

Clearly there is space for methodological improvements in other aspects of this study, too. Retrieving the Google hit numbers for a homonymous term might result in skewed results, because only one of the homonyms would define a synonym of a preferred term. Thus, "court" may refer to a "courtyard" or to a "court of law." When looking for synonyms of "yard" some occurrences of "court" will actually refer to a "courtyard" which is desired. However, other occurrences would refer to a "court of law," which is not desired. In extreme cases such results would have to be excluded completely, but it would be better to use co-occurrence statistics, available in the UMLS, to select only Web pages using the correct sense of a given homonym.

It may be observed that most medical terms in the above ontologies consist of many words. Our methodology was only designed for single word terms. Extending it to longer terms is also desirable.

Lastly, our study deals only with one single feature of an ontology that is measured, the goodness of its preferred terms. As mentioned in the introduction, there are many more features which should be evaluated. Thus, for example, an omitted concept should reduce the quality of an ontology. However, even in this case, our approach of using Google hit numbers for measuring the effect of an omission is valid. Thus, a medical ontology missing the concept "Diagnosis" (Google hit number: 127,000,000) should be judged as being of much worse quality than a comparable ontology missing the concept "Cardiovascular Diagnostic Technique" (Google hit number: 1,280,000).

8. Acknowledgement

The ideas developed in this paper have greatly benefited from work done over many years with Dr. Yehoshua Perl, Dr. Michael Halper, Dr. Huanying Gu, Dr. James J. Cimino, and many students. The topic of naturalness was introduced in research done with Yoo Jung An and Dr. Yi Ta Wu over the past three years. The authors wish to thank all of them for the intellectual atmosphere that they have created.

9. References

1. Gu, H., M. Halper, J. Geller, and Y. Perl, *Benefits of an Object-oriented Database representation for Controlled Medical Terminologies.* Journal of the American Medical Informatics Association, 1999. **6**(4): p. 283-303.
2. Liu, L., M. Halper, J. Geller, and Y. Perl, *Controlled Vocabularies in OODBs: Modeling issues and implementation.* Distributed and Parallel Databases, 1999. **7**(1): p. 37-65.
3. Liu, L., M. Halper, J. Geller, and Y. Perl, *Using OODB Modeling to Partition a Vocabulary into Structurally*

and Semantically Uniform Concept Groups. IEEE Transactions on Knowledge and Data Engineering, 2002. **14**(4): p. 850-866.
4. Noy, N. F., R. W. Fergerson, and M. A. Musen, *The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility*, in *Knowledge Engineering and Knowledge Management. Methods, Models, and Tools: 12th International Conference, EKAW 2000*, R. Dieng and O. Corby, Editors. 2000, Springer: Berlin. p. 69-82.
5. Humphreys, B. L., D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett, *The Unified Medical Language System: An informatics research collaboration.* JAMIA, 1998. **5**(1): p. 1-11.
6. Lindberg, D. A. B., B. L. Humphreys, and A. T. McCray, *The Unified Medical Language System.* Methods of Information in Medicine, 1993. **32**: p. 281-291.
7. *UMLS Documentation: 2007AA Documentation.* [cited Jan. 31, 2007]; Available from: http://www.nlm.nih.gov/research/umls/umlsdoc_preface.html.
8. Ceusters, W. and B. Smith, *A Terminological and Ontological Analysis of the NCI Thesaurus.* Methods of Information in Medicine, 2005. **44**: p. 498-507.
9. Gamper, J., W. Nejdl, and M. Wolpers. *Combining Ontologies and Terminologies in Information Systems*. in *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE-99)*. 1999. Innsbruck, Austria.
10. Geller, J., H. Gu, Y. Perl, and M. Halper, *Semantic refinement and error correction in large terminological knowledge bases.* Data & Knowledge Engineering, 2003. **45**(1): p. 1-32.
11. Min, H., Y. Perl, Y. Chen, M. Halper, J. Geller, and Y. Wang, *Auditing as part of the Terminology Design Life Cycle.* Journal of the American Medical Informatics Association, 2005. **13**(6): p. 676-690.
12. Gu, H., Y. Perl, G. Elhanan, H. Min, L. Zhang, and Y. Peng, *Auditing concept categorizations in the UMLS.* Artificial Intelligence in Medicine, 2004. **31**(1): p. 29-44.
13. Shapiro, S. C. and W. J. Rapaport, *SNePS considered as a fully intensional propositional semantic network*, in *The Knowledge Frontier*, N. Cercone and G. McCalla, Editors. 1987, Springer: New York. p. 263-315.
14. *SNOMED CT.* [cited April 25, 2007]; Available from: http://www.snomed.org/snomedct/index.html.
15. *Terminology work -- Principles and methods*, in *ISO 704:2000*, International Standards Organization (ISO).
16. *SNOMED Clinical Terms User Guide.* 2007 [cited June 14, 2007]; Available from: http://www.snomed.org/snomedct/documents/snomed_ct_user_guide.pdf.
17. An, Y. J., K.-c. Huang, and J. Geller, *Naturalness of Ontology Concepts for Rating Aspects of the Semantic Web.* Communications of the IIMA, 2006. **6**(3).
18. An, Y. J., K.-c. Huang, and J. Geller. *Rating the Naturalness of Ontology Taxonomies*. in *FLAIRS-2007*. 2007. Key West, FL.
19. McCray, A. T., A. Burgun, and O. Bodenreider. *Aggregating UMLS Semantic Types for Reducing Conceptual Complexity*. in *Proceedings of Medinfo 2001*. 2001. London, UK.
20. Francis, W. N. and H. Kucera, *Frequency analysis of English language usage: Lexicon and Grammar*. 1982, Boston: Houghton Mifflin.
21. Brewster, C., F. Ciravegna, and Y. Wilks. *Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance*. in *Proceedings of the Semantic Web Workshop (SIGIR)*. 2003. Toronto, Canada.
22. Mork, P., J. F. Brinkley, and C. Rosse, *OQAFMA Querying Agent for the Foundational Model of Anatomy: a prototype for providing flexible and efficient access to large semantic networks.* Journal of Biomedical Informatics, 2003. **36**: p. 501-517.
23. Huang, J., J. Dang, M. N. Huhns, and Y. Shao, *Ontology Alignment as a Basis for Mobile Service Integration and Invocation.* Journal of Pervasive Computing and Communication, 2005. **1**(1): p. 1-11.
24. Cimino, J. J., *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.* Methods of Information in Medicine, 1998. **37**(4-5): p. 394-403.
25. Mostowfi, F. and F. Fotouhi, *Improving Quality of Ontology: An Ontology Transformation Approach,* Semantic Web and Databases Workshop (SWDB'06), Roger S. Barga, Xiaofang Zhou (Eds.): Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDE 2006, 3-7 April 2006, Atlanta, GA, USA. IEEE Computer Society 2006, p. 61.
26. Colomb. R. M., Quality of Ontologies in Interoperating Information Systems, Technical Report 18/02 ISIB-CNR, National Research Council, Institute of Biomedical Engineering, Padova, Italy, November, 2002.
27. Tartir, S., I. B. Arpinar, M. Moore, A. P. Sheth, and B. Aleman-Meza, OntoQA: Metric-Based Ontology Quality Analysis, Proceedings of IEEE ICDM 2005 Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, 2005.