# OMCSNet: A Commonsense Inference Toolkit

**Hugo Liu and Push Singh**
MIT Media Laboratory
20 Ames St., Bldg. E15
Cambridge, MA 02139 USA
{hugo, push}@media.mit.edu

## Abstract

Large, easy-to-use semantic networks of symbolic linguistic knowledge such as WordNet and MindNet have become staple resources for semantic analysis tasks from query expansion to word-sense disambiguation. However, the knowledge captured by these resources is limited to formal taxonomic relations between or dictionary definitions of lexical items. While such knowledge is sufficient for some NLP tasks, we believe that broader opportunities are afforded by databases containing more diverse kinds of world knowledge, including substantial knowledge about compound concepts like activities (e.g. "washing hair"), accompanied by a richer set of temporal, spatial, functional, and social relations between concepts.

Based on this premise, we introduce OMCSNet, a freely available, large semantic network of commonsense knowledge. Built from the Open Mind Common Sense corpus, which acquires world knowledge from a web-based community of instructors, OMCSNet is presently a semantic network of 280,000 items of common-sense knowledge, and a set of tools for making inferences using this knowledge. In this paper, we describe OMCSNet, evaluate it in the context of other semantic knowledge bases, and review how OMCSNet has been used to enable and improve various NLP tasks.

## 1 Introduction

There has been an increasing thirst for large-scale semantic knowledge bases in the AI community. Such a resource would improve many broad-coverage natural language processing tasks such as parsing, information retrieval, word-sense disambiguation, and document summarization, just to name a few. WordNet (Fellbaum, 1998) is currently the most popular semantic resource in the computational linguistics community. Its knowledge is easy to apply in linguistic applications because WordNet takes the form of a simple semantic network—there is no esoteric representation to map into and out of. In addition, WordNet and tools for using it are freely available to the community and easily obtained. As a result, WordNet has been used in hundreds of research projects throughout the computational linguistics community, running the gamut of linguistic processing tasks (see WordNet Bibliography, 2003).

However, in many ways WordNet is far from ideal. Often, the knowledge encoded by WordNet is too formal and taxonomic to be of practical value. For example, WordNet can tell us that a dog is a kind of canine which is a kind of carnivore, which is a kind of placental mammal, but it does not tell us that a dog is a kind of pet, which is something that most people would think of. Also, because it is a lexical database, WordNet only includes concepts expressable as single words. Furthermore, its ontology of relations consists of the limited set of nymic relations comprised by synonyms, is-a relations, and part-of relations.

Ideally, a semantic resource should contain knowledge not just those concepts that are lexicalized, but also about lexically compound concepts. It should be connected by an ontology of relations rich enough to encode a broad range of commonsense knowledge about objects, actions, goals, the structure of events, and so forth. In addition, it should come with tools for easily making use of that knowledge within linguistic applications. We believe that such a resource would open the door to many new innovations and improvements across the gamut of linguistic processing tasks.

Building large-scale databases of commonsense knowledge is not a trivial task. One problem is scale. It has been estimated that the scope of common sense may involve many tens of millions of pieces of knowledge. Unfortunately, common sense cannot be easily mined from dictionaries, encyclopedias, the web, or other corpora because it consists largely of knowledge obvious to a reader, and thus omitted. Indeed, it likely takes much common sense to even interpret dictionaries and encyclopedias. Until recently, it seemed that the only way to built a commonsense knowledgebase was through the expensive process of hand-coding each and every fact.

However, in recent years we have been exploring a new approach. Inspired by the success of distributed and collaborative projects on the Web, Singh et al. (2002) turned to the general public to massively distribute the problem of building a commonsense knowledgebase. They succeeded at gathering well over 500,000 simple assertions from many contributors. From this corpus of commonsense facts, we built **OMCSNet**, a semantic network of 280,000 items of commonsense knowledge. An excerpt of OMCSNet is shown in Figure 1. Our aim was to create a large-scale machine-readable resource structured as an easy-to-use semantic network representation like WordNet (Fellbaum, 1998) and MindNet (Richardson *et al.*, 1998), yet whose contents reflect the broader range of world knowledge characteristic of commonsense as in Cyc (Lenat, 1995). While far from being a perfect or complete commonsense knowledgebase, OMCSNet has nonetheless offered world knowledge on a large scale and has been employed to support and tackle a variety of linguistic processing tasks.

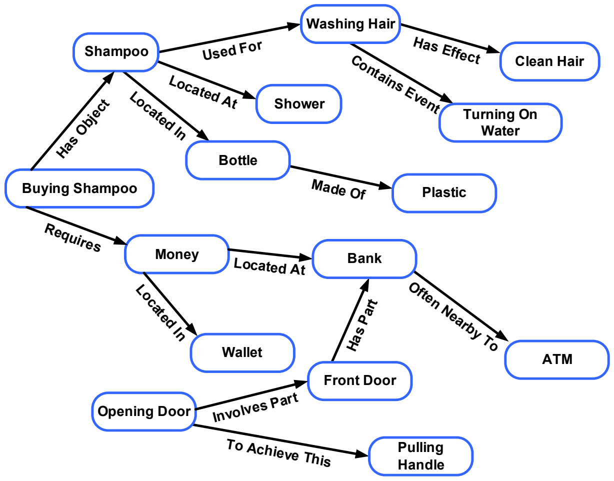


**Figure 1.** An excerpt from OMCSNet. Relation names are expanded here for illustrative purposes.

This paper is structured as follows. First, we discuss how OMCSNet was built, how it is structured, and the nature of its contents. Second, we present the OMCSNet inference toolkit distributed with the semantic network. Third, we review how OMCSNet has been applied to improve or enable several linguistic processing tasks. Fourth, we evaluate several aspects of the knowledge and the inference toolkit, and compare it to several other large-scale semantic knowledge bases. We conclude with a discussion of the potential impact of this resource on the computational linguistics community at large, and explore directions for future work.

## 2 OMCSNet

In this section, we first explain the origins of OMCSNet in the Open Mind Commonsense corpus; then we demonstrate how knowledge is extracted to produce the semantic network; and third, we describe the structure and semantic content of the network. The OMCSNet Knowledge Base, Knowledge Browser, and Inference Tool API is available for download (Liu & Singh, 2003).

### 2.1 Building OMCSNet

OMCSNet came about in a unique way. Three years ago, the Open Mind Commonsense (OMCS) web site (Singh *et al.* 2002) was built, a collection of 30 different activities, each of which elicits a different type of commonsense knowledge—simple assertions, descriptions of typical situations, stories describing ordinary activities and actions, and so forth. Since then the website has gathered nearly 500,000 items of commonsense knowledge from over 10,000 contributors from around the world, many with no special training in computer science. The OMCS corpus now consists of a tremendous range of different types of commonsense knowledge, expressed in natural language.

The earliest applications of the OMCS corpus made use of its knowledge not directly but by first extracting into semantic networks only the types of knowledge they needed. For example, the ARIA photo retrieval system (Lieberman & Liu, 2002) extracted taxonomic, spatial, functional, causal, and emotional knowledge to improve information retrieval. This suggested a new approach to building a commonsense knowledgebase. Rather than directly engineering the knowledge structures used by the reasoning system, as is done in Cyc, OMCS encourages people to provide information clearly in natural language and then extract from that more usable knowledge representations. Inspiring was the fact that there had been significant progress in the area of information extraction from text in recent years, due to improvements in broad-coverage parsing (Cardie, 1997). A number of systems are able to successfully extract facts, conceptual relations, and even complex events from text.

OMCSNet is produced by an automatic process, which applies a set of 'commonsense extraction rules' to the OMCS corpus. A pattern matching parser uses 40 mapping rules to easily parse semi-structured sentences into predicate relations and arguments which are short fragments of English. These arguments are then normalized using natural language techniques (stripped of stop words, lemmatized), and are *massaged* into one of many standard syntactic forms. To account for richer concepts which are more than words, we created three categories of concepts: Noun Phrases (things, places, people), Attributes (modifiers), and Activity Phrases (actions and actions compounded with a noun phrase or prepositional phrase, e.g.: "turn on water," "wash hair."). A small part-of-speech tag –driven grammar filters out non-compliant text fragments and massages the rest to take one of these standard syntactic forms. When all is done, the cleaned relations and arguments are linked together into the OMCSNet semantic network.

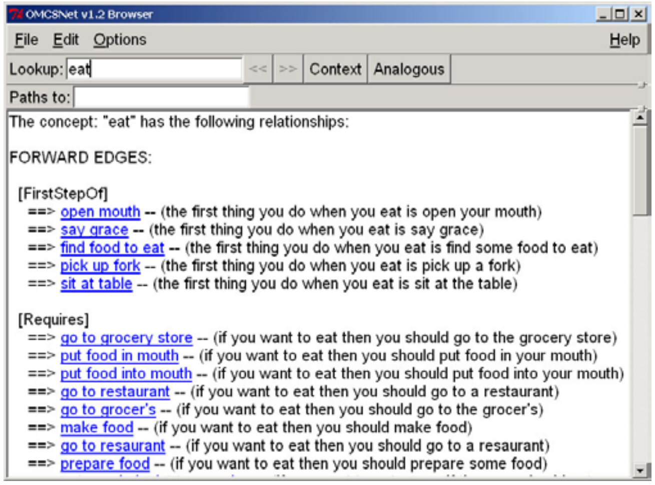
### 2.3 Contents of OMCSNet

At present OMCSNet consists of the 20 binary relations shown below in Table 1. These relations were chosen because the original OMCS corpus was built largely through its users filling in the blanks of templates like 'a hammer is for _____'. Thus the relations we chose to extract largely reflect the original choice of templates used on the OMCS web site.

**Table 1.** Semantic Relation Types currently in OMCSNet

| Category | Semantic Relations |
|---|---|
| Things | KindOf, HasProperty, PartOf, MadeOf |
| Events | SubEventOf , FirstStepOf, LastStepOf |
| Actions | Requires, HasEffect, ResultsInWant, HasAbility |
| Spatial | OftenNear, LocationOf, CityInLocality |
| Goals | DoesWant, DoesNotWant, MotivatedBy |
| Functions | UsedInLocation, HasFunction |
| Generic | ConceptuallyRelatedTo |

The OMCSNet Browser Tool can be used to browse the contents of OMCSNet by searching for concepts and following semantic links. A picture of this tool is shown in Figure 2.



```
The concept: "eat" has the following relationships:

FORWARD EDGES:

 [FirstStepOf]
   ==> open mouth -- (the first thing you do when you eat is open your mouth)
   ==> say grace -- (the first thing you do when you eat is say grace)
   ==> find food to eat -- (the first thing you do when you eat is find some food to eat)
   ==> pick up fork -- (the first thing you do when you eat is pick up a fork)
   ==> sit at table -- (the first thing you do when you eat is sit at the table)

 [Requires]
   ==> go to grocery store -- (if you want to eat then you should go to the grocery store)
   ==> put food in mouth -- (if you want to eat then you should put food in your mouth)
   ==> put food into mouth -- (if you want to eat then you should put food into your mouth)
   ==> go to restaurant -- (if you want to eat then you should go to a restaurant)
   ==> go to grocer's -- (if you want to eat then you should go to the grocer's)
   ==> make food -- (if you want to eat then you should make food)
   ==> go to resaurant -- (if you want to eat then you should go to a resaurant)
   ==> prepare food -- (if you want to eat then you should prepare some food)
```

**Figure 2.** The OMCSNet Browser Tool

## 3 OMCSNet Inference Toolkit

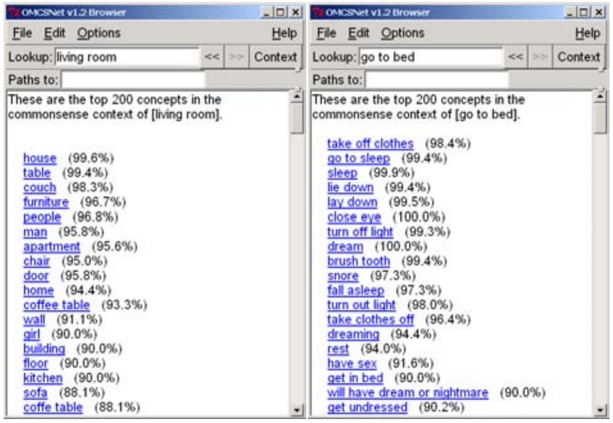
To assist in using OMCSNet in various types of inference, we built a small but growing set of tools to help researchers and application developers maintain a high-level, task-driven view of commonsense. In the following subsections, we describe some of the more basic tools.

**'Fuzzy' Inference.** So far we have presented OMCSNet as a fairly straightforward semantic network, and so one might ask the question why an inference toolkit might even be necessary when conventional semantic network graph traversal techniques should suffice. The answer lies in the structure of the nodes, and in the peculiarity of commonsense knowledge.

In the previous section we presented several types of nodes including Noun Phrases, Attributes, and Activity Phrases. These nodes can either be first-order, i.e. simple words and phrases, or second-order, such as "turn on water." Second order nodes are essentially fragments of English following a particular part-of-speech pattern. Maintaining the representation in English saves us from having to map into and out of a special ontology, which would greatly increase the complexity and difficulty-of-use of the system; it also maintains the nuances of the concept. Practically, however, we may want the concepts "buy food" and "purchase food" to be treated as the same concept.

To accomplish this, the inference mechanism accompanying OMCSNet can perform such *fuzzy* conceptual bindings using a simple semantic distance heuristic (e.g. "buy food" and "purchase food" are commensurate if a synonym relation holds between "buy" and "purchase.") Another useful approximate matching heuristic is to compare normalized morphologies produced by lemmatizing words. Using these approximate concept bindings, we can perform 'fuzzy' inference over the network.

**Context Determination.** One task useful across many natural language applications is determining the context around a concept or around the intersection of several concepts. The context determination tool enables this by performing spreading activation to discover concepts in the *semantic neighborhood*. For example, OMCSNet produced the following top concepts in the neighborhood of the noun phrase concept "living room," and the activity phrase concept "go to bed" (Figure 3). Percentages indicate confidence of overall semantic connectedness. Phrases in OMCSNet are linguistically normalized, removing plural and tense morphology (lemmatization) and filtering out determiners and possessives



```
These are the top 200 concepts in the
commonsense context of [living room].

   house   (99.6%)
   table   (99.4%)
   couch   (98.3%)
   furniture   (96.7%)
   people   (96.8%)
   man   (95.8%)
   apartment   (95.6%)
   chair   (95.0%)
   door   (95.8%)
   home   (94.4%)
   coffee table   (93.3%)
   wall   (91.1%)
   girl   (90.0%)
   building   (90.0%)
   floor   (90.0%)
   kitchen   (90.0%)
   sofa   (88.1%)
   coffe table   (88.1%)
```

```
These are the top 200 concepts in the
commonsense context of [go to bed].

   take off clothes   (98.4%)
   go to sleep   (99.4%)
   sleep   (99.9%)
   lie down   (99.4%)
   lay down   (99.5%)
   close eye   (100.0%)
   turn off light   (99.3%)
   dream   (100.0%)
   brush tooth   (99.4%)
   snore   (97.3%)
   fall asleep   (97.3%)
   turn out light   (98.0%)
   take clothes off   (96.4%)
   dreaming   (94.4%)
   rest   (94.0%)
   have sex   (91.6%)
   get in bed   (90.0%)
   will have dream or nightmare   (90.0%)
   get undressed   (90.2%)
```

**Figure 3**. Concepts in the semantic neighborhood of "living room" and "go to bed" (semantic similarity judgment based equally on all relations)

Concepts connected to "living room" through any relation were included in the context. However, we may, for example, only be interested in specific relations. If we had specified the relation "HasFunction", the context search would return results like "entertain guests," "comfortable," and "watch television." In other cases we may desire to bias the context of "living room" with another concept, e.g., "store." The output is the context of "living room" with respect to the concept "store" and returns results like "furniture," "furniture store," and "Ikea."

**Analogical Inference.** Knowledge about particular concepts is occasionally patchy. For example, the system may know "Requires(car, gas)" but not "Requires(motorcycle, gas)". Such relationships may be produced using analogical inference. For example, by employing structure-mapping methods (Getner, 1983). In the present toolkit, we are already able to make some simple conceptual analogies using structure-mapping, producing results like the following:

car is like motorcycle because both:
==[IsA]==> vehicle type
==[HasFunction]==> transportation
==[HasProperty]==> fast

## 4 NLP Applications of OMCSNet

Early versions of the OMCSNet tools are being put to use to assist a variety of NLP tasks in prototype applications, each of which uses commonsense knowledge differently. None of them actually does 'general purpose' commonsense reasoning. Below, we review some different ways that OMCSNet has supported both traditional NLP tasks, and also more niche semantic reasoning tasks.

**Semantic Type Recognition.** A very basic task in NLP is recognizing the semantic type of a word or phrase. This is similar to what is often referred to as named-entity recognition (NER). In NER, a natural language pre-processor might want to recognize a variety of entities in the text such as phone numbers, email addresses, dates, organizational names, etc. Often, syntax helps in recognition, as in the case of email addresses. Other times, naïve keyword spotting using a large domain-specific database helps, as in the case of organizational names. However, when trying to assess potential semantic roles such as everyday events, or places, there may be no obvious sources which provide laundry-lists of such knowledge. It may be easy to find a database which lists "the Rose Parade" as an event, but it may be harder to find a database that tells us a "birthday", "wedding", or "party" is an event. We argue that this is because such knowledge is often so obvious that it is never explicitly recorded. In other words, it falls within the realm of commonsense knowledge, and therefore, can be addressed by resources like OMCSNet.

Liu & Lieberman (2002) built a set of semantic agents of people, places, characteristics, events, tools, and objects for their World-Aware Language Parser (WALI) using semantic type preference knowledge from a precursor to OMCSNet. They implicitly inferred the semantic type preferences of concepts by the names of the relations that connect them. For example, from the expression, "LocationOf(A,B)", it was inferred that B can play the semantic role of PLACE. Liu & Lieberman found that while implicit semantic type preference knowledge from OMCSNet is not completely accurate by itself, it can combined with other sources of knowledge such as syntactic cues or frame semantic resources such as FrameNet (Baker et al., 1998) to produce accurate semantic recognition.

**Selectional Preferences**. One resource that is commonly used to assist in word-sense disambiguation (WSD) is a set of collocations of verbs and the arguments that they prefer. Traditional knowledge-backed approaches to selectional preferences has sought to acquire or define knowledge of semantic classes; but this has proven to be difficult because of the sheer magnitude of the knowledge engineering task. More recently, Resnik (1997), and Light & Greiff (2002) have demonstrated how selectional preferences could be calculated using corpus-derived statistical models backed by a semantic class hierarchy such as that from WordNet. However, WordNet's singular, formal, dictionary-like taxonomy does account for the large diversity of semantic class hierarchies which are practically used. OMCSNet does not maintain a single consistent hierarchy, but rather, fosters multiple and diverse hierarchies, and therefore, may prove to be more appropriate than WordNet for supporting statistical models of selectional preferences.

Selectional preferences are present in OMCSNet in two forms. First, activity concepts which are verb-object compounds (e.g. "wash hair", "drive car") provide an implicit source of selectional preferences. Because these concepts originate in a commonsense knowledgebase, they represent commonsense verb-argument usage, and therefore the resulting selectional preferences can be thought of as semantically stronger than had they be derived from a non-commonsense corpus. Second, OMCSNet contains explicit knowledge about selectional preferences, inherent in relations such as "HasAbility" and "HasFunction". In a part-of-speech tagger called MontyTagger (forthcoming), Liu uses selectional preferences from OMCSNet to correct tagging errors in post-processing. For example, in "The/DT dog/NN bit/NN the/DT mailman/NN", "bit" is incorrectly tagged as a noun. Using OMCSNet, MontyTagger performs the following inference to prefer "bit" as a verb (probabilities omitted):

mailman [IsA] person
dog [HasAbility] bite people
➔ dog [HasAbility] bite mailman

**Topic Detection and Summarization**. Eagle *et al.* (2003) are working on detecting the discourse topic given a transcript of a conversation. They are using OMCSNet to assess the semantic relatedness of concepts within the same discourse passage. First, a set of concept nodes in OMCSNet are designated as topics. By mapping concepts within the transcript to nodes in OMCSNet, the appropriate topic can be found by an aggregate nearest neighbor function or Bayesian inference.

The need for symbolic world knowledge in topic detection is further illustrated by an automatic text summarizer called SUMMARIST (Hovy & Lin, 1997). SUMMARIST uses symbolic world knowledge via WordNet and dictionaries for topic detection. For example, the presence of the words "gun", "mask", "money", "caught", and "stole" together would indicate the topic of "robbery". However, they reported that WordNet and dictionary resources were relationally too sparse for robust topic detection. We believe that OMCSNet would outperform WordNet and dictionary resources in this task because it is relationally richer and contains practical rather than dictionary-like knowledge.

**Retrieving event-subevent structure.** It is sometimes useful to collect together all the knowledge that is relevant to some particular class of activity or event. For example the Cinematic Common Sense project makes use of commonsense knowledge about event-subevent structure in OMCSNet to make suitable shot suggestions at common events like birthdays and marathons (Barry & Davenport, 2002). For the topic 'getting ready for a marathon', the subevents gathered might include: putting on your running shoes, picking up your number, and getting in your place at the starting line.

**Goal recognition and planning.** The search engines described in Singh *et al.* (2002) and Liu *et al.* (2002) exploit commonsense knowledge about typical human goals to infer the real goal of the user from their search query. For example, the search 'my cat is sick' leads to the system inferring that 'I want my cat to be healthy' because people care about their pets and they want things they care about to be healthy. Furthermore, these search engines can make use of knowledge about actions and their effects to engage in a simple form of planning. After inferring the user's true intention, they look for a way to achieve it. In this case, if you want something to be healthy you can take it to a doctor, or in the case of an animal, a veterinarian.

**Temporal projection for story generation.** The MakeBelieve storytelling system (Liu & Singh, 2002) makes use of the knowledge of temporal and causal relationships between events in order to guess what is likely to happen next. Using this knowledge it generates stories such as: *David fell off his bike. David scraped his knee. David cried like a baby. David was laughed at. David decided to get revenge. David hurt people.*

**Particular consequences of broad classes of actions.** Empathy Buddy senses the affect in passages of text (Liu *et al.*, 2003). It predicts those consequences of actions and events that have some emotional significance. This can be done by chaining backwards from knowledge about desirable and undesirable states. For example, if being out of work is undesirable, and being fired causes to be to be out of work, then the passing 'I was fired from work today' can be sensed as undesirable.

**Specific facts about particular things.** Some of OMCSNet is specific facts like "the Golden Gate Bridge is located in San Francisco", or that "a PowerBook is a kind of laptop computer." The ARIA e-mail client and photo retrieval system (Liu & Lieberman, 2002) can reason that an e-mail that mentions that "I saw the Golden Gate Bridge" meant that I was in San Francisco at the time, and proactively retrieves photos taken in San Francisco for the user to insert into the e-mail.

**Conceptual association.** OMCSNet can be used to supply associated concepts. The Globuddy program (Various Authors, 2003) uses OMCSNet to retrieve knowledge about events, actions, objects, and other concepts related to a given situation, to make a custom phrasebook of concepts you might wish to have translations for in that situation. For example, if you are arrested, it will give you a few pages translating words like 'lawyer', 'going to prison', 'find a lawyer', and so forth.

## 5 Evaluation

The original OMCS corpus was previously evaluated by Singh *et al.* (2002). Human judges evaluated a sample of the corpus and rated 75% of items as largely true, 82% as largely objective, 85% as largely making sense, and 84% as knowledge someone would have by high school.

We performed two further analyses of OMCSNet: a qualitative study (human judges) and a quantitative analysis. However, perhaps the most compelling evaluations are indirect. OMCS and OMCSNet have been used to measurably improve the behavior of intelligent language systems. In the previous section we briefly reviewed some OMCS- and OMCSNet- enabled language processing systems. For brevity, we refer the reader to each application's respective evaluations (see each application's corresponding paper).

**A Qualitative Study of OMCSNet.** We conducted an experiment with five human judges and asked each judge to rate 100 concepts in OMCSNet. 10 concepts were common to all judges (for correlational analysis), 90 were of their choice. If a concept produced no results, they were asked to duly note that and try another concept. Concepts were judged along these 2 dimensions, each on a Likert 1 (strongly disagree) to 5 (strongly agree) scale:

1) Results for this concept are fairly comprehensive.
2) Results for this concept include incorrect knowledge, nonsensical data, or non-commonsense information.

To account for inter-judge agreement, we normalized scores using the 10 common concepts, and produced the re-centered aggregate results shown below in Table 2.

**Table 2.** Measure of quality of OMCSNet.

| | Mean Score | Std. Dev. | Std. Err. |
|---|---|---|---|
| Comprehensiveness | 3.40 / 5.00 | 1.24 | 1.58 |
| Noisiness | 1.24 / 5.00 | 0.99 | 1.05 |
| % Concepts attempted, that were not in KB | 11.3% | 6.07% | 0.37% |

These results can be interpreted as follows. Judgment of comprehensiveness of knowledge in OMCSNet on average, was *several relevant concepts,* but varied significantly from *a few concepts* to *almost all of the concepts.* Noisiness was *little noise* on average, and did not vary much. % of KB misses was very consistently 11%. We consider these to be very optimistic results. Comprehensiveness was moderate but varied a lot indicating still patchy coverage, which we hope this will improve as OMCS grows. Noisiness was surprisingly low, lending support to the idea that a relatively clean KB can be elicited from public acquisition. % of KB misses was more than tolerable considering that OMCSNet has only 80,000 concepts—a fraction of that possessed by people.

**A Quantitative Analysis of OMCSNet.** 100 salient concepts already in OMCSNet were selected by the judges for each contextual "domain" as typifying that domain (for example, the domain of "everyday", concepts includes "wake up", "eat breakfast", "shower", "go to work", "prepare meal", "eat food", etc.). Concepts included appropriate distributions of concept types, i.e. people, places, things, actions, and activities. Branching factor indicates the number of relations for each node (density of knowledge). Standard deviation illustrates unevenness of knowledge. The intra-set branching factor and standard deviations indicate density and unevenness *within* each domain. Results are shown in Table 3.

**Table 3.** Coverage density and distribution in 4 domains.

| | ***Overall KB*** | ***Jobs*** | ***Family*** | ***Every-day*** | ***Trips*** |
|---|---|---|---|---|---|
| ***Branching Factor*** | 3.48 | 59.7 | 98.5 | 40.1 | 34.9 |
| ***Standard Dev.*** | 21.5 | 78.6 | 169 | 38.5 | 38.3 |
| ***Intra-set B.F.*** | | 4.06 | 8.83 | 2.2 | 1.7 |
| ***Intra-set Std. Dev.*** | | 4.17 | 9.75 | 2.35 | 2.05 |

These results show that although there is a lot of knowledge about these common domains, there is also an enormous variation of coverage. A review of the histogram of results (not shown) indicates a bimodal distribution—a concept possessed either a lot of knowledge (>100) or not much (<5). We postulate that structure of the semantic network consists of mainly dense "hub" nodes (possibly due to word-sense collision) and some outlying spoke nodes. From the intra-set results, we postulate that knowledge is not as clustered around domains as we had expected. This is an interesting result because it suggests that artificial clustering of domains prevalently practiced in AI may not work for commonsense!

## 6 Large-Scale Semantic KBs

In this section we compare OMCSNet with several other existing large-scale semantic knowledge bases.

**Cyc.** The Cyc project (Lenat 1995) is the most prominent large-scale effort to build a commonsense knowledge base. A major difference between Cyc and OMCSNet is in the choice of knowledge representation. Knowledge in Cyc is represented in a rich logical language called CycL. OMCSNet, on the other hand, explores an alternative representation grounded in structured English fragments and a limited set of predicate relations. So OMCSNet loosely resembles predicate logic over fragments of English. OMCSNet's semantic network is a much simpler and less expressive knowledge representation scheme than CycL, and as a result OMCSNet cannot represent many important types of commonsense knowledge. While not as formal as CycL, we nonetheless believe that a broad range of applications still stand to benefit from such a knowledge base.

From a practical perspective, another important difference is that the Cyc knowledge base is at present proprietary and inaccessible as a community resource, whereas both the OMCS corpus and OMCSNet are freely available resources. However, recently the developers of Cyc have released OpenCyc, a publicly available version of Cyc that includes its inference engine and Cyc's upper level ontology.

**ThoughtTreasure.** With on the order of 100,000 items of commonsense knowledge, ThoughtTreasure (TT) was built by researcher Erik Mueller to investigate the story understanding task (Mueller, 2000). TT represents commonsense knowledge in a variety of ways, including simple assertions, frames, scripts, and spatial occupancy arrays. The knowledge in TT is well-mapped onto natural language, for every concept has an associated lexical item, and the TT system itself includes a substantial natural language parsing and generation component. By comparison, knowledge in OMCSNet is completely assertional (although the OMCS corpus itself contains other types of knowledge that were not included in OMCSNet), and its representation is rooted in semi-structured English fragments.

**WordNet.** Arguably the most widely used machine-readable semantic resource in the artificial intelligence and computational linguistic communities, WordNet (Fellbaum, 1998) was not intended as a commonsense resource per se, but rather as a large lexical database of English concepts (simple words and collocations). The scope of WordNet encompasses on the order of 100,000 concepts, connected by 100,000 nymic relations of hypernymy (is-a), hyponymy (a-kind-of), synonymy, antonymy, and meronymy (part-of). It is attractive as a commonsense resource because its hierarchical system of concepts captures some basic (but limited) relationships between concepts in the everyday world, and is comprehensive enough to have wide application.

WordNet's popularity with researchers and developers illustrates the two communities' thirst for semantic knowledge bases. Its representational simplicity (all binary relations) and its being rooted in plain English (no complex representational language to map into or out of) lends it an ease of use and integration into applications that has also promoted adoption. We feel that OMCSNet, with a comparable knowledge representation but offering more diverse semantic content, will also help to address the knowledge needs of the communities and foster innovation that would not be possible otherwise.

OMCSNet differs from WordNet in a few important ways. First, concepts in OMCSNet are not sense disambiguated as in WordNet, though it is possible to introduce a statistical notion of "sense" by clustering conceptual nodes in a graph. Second, concepts in WordNet are organized into syntactic categories of nouns, verbs, adjectives, and adverbs and are usually one word or a collocation with one head word; in contrast, concepts in OMCSNet contain a variety of semantic categories like things, people, properties, actions, activities, and events, and may contain many hyperlexical concepts (e.g. “buy groceries”) in addition to lexical ones. Third, relations in WordNet are primarily hierarchical and are limited in the relationships they can express; OMCSNet presently uses 20 relations including temporal, spatial, causal, and functional relations, which are arguably more useful for commonsense reasoning problems.

**MindNet.** OMCSNet and MindNet (Richardson *et al.*, 1998) follow a very similar approach. Also based on the premise that large, useful semantic networks can be extracted from natural language text corpora, the MindNet project mines reference materials like dictionaries using broad-coverage parsing techniques to populate a semantic network with named relations. The two semantic networks have comparable numbers of named semantic relations, and go beyond basic WordNet nymic relations, which are largely hierarchical. However, there are several pointed differences.

First, MindNet is fundamentally a *lexical* knowledge base—concepts that are words. This reflects the fact that they are parsing primarily lexical resources including Longman’s Dictionary of Common English (LDOCE) and American Heritage Dictionary, 3rd edition (AHD3); in addition, imperfect broad coverage parsing over unstructured text (dictionary definitions are unstructured) makes it hard to parse relationships between entities much larger than individual words. Because of its knowledge source, OMCSNet’s concept nodes are often hyperlexical (second order nodes), including English fragments such as activity phrases (e.g. “wash hair”, “brush teeth”), and concept phrases (e.g. “automatic teller machine”). For the same reason, MindNet’s relations primarily describe lexical-level properties such as Part, Possessor, Material, Source, etc. As a result, nonlexical commonsense not resembling dictionary definitions is harder to express in the MindNet formalism, e.g. “eating a lot of food will make you less hungry.”

Second, MindNet relies on dictionary corpora, and dictionary definitions and wording are often not very representative of the *practical* and *everyday* meaning of concepts. Mined from dictionaries, MindNet will provide only one or two definitions of each concept, while OMCSNet maintain a plurality of ways of representing a concept’s meaning, and a plurality of different ways to phrase a definition. Mining of dictionaries and reference resources may be useful for acquiring a small subset of denotational, lexical commonsense, but ultimately a large part of commonsense is not written in existing references.

**Relative sizes of Knowledgebases.** Table 4 compares the sizes of these five large-scale semantic knowledgebases. The size of Cyc is on the order of 1.5 million assertions, though we caution that numbers given throughout this section are specific to each project’s knowledge representation and therefore they should be compared with caution.

**Table 4.** The relative size of knowledgebases. Adapted with permission from Mueller (1999)).

| Name | Concepts | ako/isa | part-of | Other |
|---|---|---|---|---|
| Cyc | 30,000 | ~375,000 | ~525,000 | ~600,000 |
| ThoughtTreasure | 27,093 | 28,818 | 666 | 21,821 |
| WordNet 1.6 | 99,642 | 78,446 | 19,441 | 42,700 |
| MindNet | 45,000 | 47,000 | 14,100 | 32,900 |
| OMCSNet | 81,430 | 45,382 | 5,208 | 151,692 |

## 7 Extending OMCSNet

We are presently extending OMCSNet in several directions. First, we would like to disambiguate the senses of the concepts in OMCSNet, perhaps into WordNet’s sense definitions. The Open Mind Word Expert web site (Chklovski and Mihalcea, 2002) allows users to disambiguate the senses of the words in the OMCS corpus, and we are looking into making use of the data they are collecting to build a disambiguated OMCSNet. Second, the current set of 20 relation types in OMCSNet is small compared to the wide array of assertion types that exist in the OMCS corpus. We wish to employ a broad coverage parser that can extract a wider range of knowledge from the corpus. Third, we are developing a special version of the OMCS web site that focuses specifically on further growing the OMCSNet knowledge base, including special activities for elaborating, validating, repairing items of knowledge.

## Conclusions

OMCSNet is presently the largest freely available database of commonsense knowledge. It comes with a browser and a preliminary set of tools to support basic semantic processing tasks, and is being used in a number of applications. While the contents of the knowledgebase are still patchy in comparison to what people know, our analysis has shown it to be surprisingly clean, and it has proven more than large enough to enable experimenting with entirely new ways to tackle traditional NLP tasks.

## References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998): The Berkeley FrameNet project. in *Proceedings of the COLING-ACL*, Montreal, Canada.

Barry, B. & Davenport. G. (2002). *Why Common Sense for Video Production?* (Interactive Cinema Technical Report #02-01). Media Lab, MIT.

Cardie, C. (1997). Empirical Methods in Information Extraction, *AI Magazine*, 65-79.

Chklovski, T. and Mihalcea, R. (2002). Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of the Workshop on "Word Sense Disambigua-*

*tion: Recent Successes and Future Directions", ACL 2002*.

Eagle, N., Singh, P., and Pentland, A. (2003). Using Common Sense for Discourse Topic Prediction. Forthcoming MIT Media Lab, Human Design Group Technical Report.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, pp 155-170.

Fellbaum, Christiane. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Hovy, E.H. and C-Y. Lin. (1999). Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization.* Cambridge: MIT Press, pp. 81-94.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. CACM 38(11): 33-38.

Lieberman, H., Rosenzweig E., Singh, P., (2001). Aria: An Agent For Annotating And Retrieving Images, *IEEE Computer*, July 2001, 57-61.

Lieberman, H. and Liu, H. (2002). Adaptive Linking between Text and Photos Using Common Sense Reasoning. In Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, (AH2002) Malaga, Spain. Springer-Verlag, Berlin, 2002, pp. 2-11.

Light, M., Greiff, W. (2002). Statistical models for the induction and use of selectional preferences. Cognitive Science 87 (2002) 1–13.

Liu, H., Lieberman, H. (2002). Robust photo retrieval using world semantics. *Proceedings of the LREC2002 Workshop: Creating and Using Semantics*. Las Palmas, Canary Islands.

Liu, H., Lieberman, H., Selker, T. (2002). GOOSE: A Goal-Oriented Search Engine With Commonsense. *Proceedings of AH2002.* Malaga, Spain.

Liu, H., Lieberman, H., Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. In *Proceedings of IUI 2003.* Miami, Florida.

Liu, H. (forthcoming). Using Commonsense to Improve a Brill-Based Part-of-Speech Tagger. White paper available at: web.media.mit.edu/~hugo/montytagger/

Liu, H., Singh, P. (2002). MAKEBELIEVE: Using Commonsense to Generate Stories. In *Proceedings of AAAI-02*. Edmonton, Canada.

Liu, H., Singh, P. (2003). OMCSNet v1.2. Knowledge Base, tools, and API available at: web.media.mit.edu/~hugo/omcsnet/

Mueller, E. (1999). Prospects for in-depth story understanding by computer. arXiv:cs.AI/0003003 http://www.signiform.com/erik/pubs/storyund.htm.

Mueller, E. (2000). ThoughtTreasure: A natural language/commonsense platform. Retrieved from http://www.signiform.com/tt/htm/overview.htm on 1/9/03.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ANLP-97 workshop: Tagging text with lexical semantics: Why, what, and how?* Washington, DC. 424

Richardson, S. D., Dolan, B., and Vanderwende, L. (1998). MindNet: Acquiring and structuring semantinc information from text. In *COLING-ACL'98*.

Singh, Push, Lin, Thomas, Mueller, Erik T., Lim, Grace, Perkins, Travell, & Zhu, Wan Li (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In *Proceedings of ODBASE'02*. Lecture Notes in Computer Science. Heidelberg: Springer-Verlag.

Various Authors (2003). Common Sense Reasoning for Interactive Applications Projects Page. Retrieved from http://www.media.mit.edu/~lieber/Teaching/Common-Sense-Course/Projects/Projects-Intro.html on 1/9/03.

WordNet Bibliography. (2003). Retrieved from http://engr.smu.edu/~rada/wnb/ on 2/24/2003.