Selected Lexical/Ontology/Knowledge Base Projects

(Originally prepared and revised by Burcu Ayşen Ürgen and Bilge Say for COG 515, Wordnet added and revised by Bilge Say and Ayışığı Sevdik Çallı for COGS 523)

WordNet

"WordNet® is a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations." (see the project's web site http://wordnet.princeton.edu/) . WordNets have been developed in about 30 languages - for some languages more than one WordNet has been developed. There are ongoing project to enhance psycholinguistic and natural language processing usability of WordNet.

CYC Project

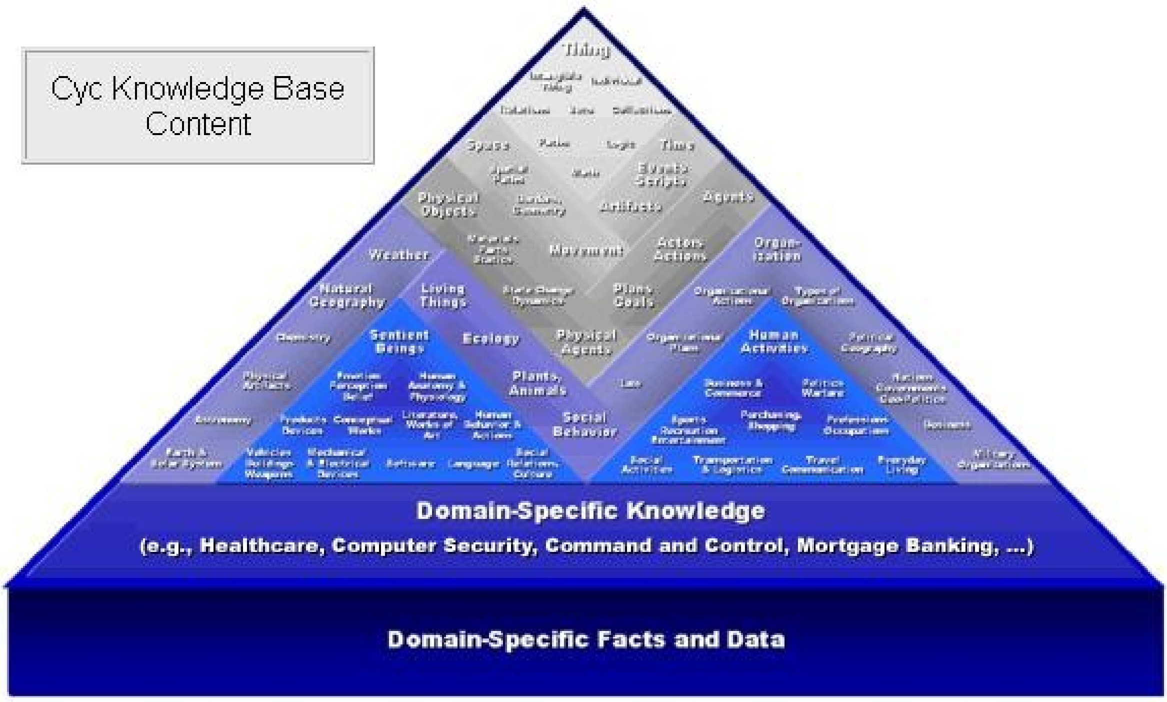
CYC is a project started by Douglas Lenat (1984) and carried out by the company Cycorp, which aims to give a formalization of human common sense knowledge. It basically consists of a knowledge base and an associated inference engine, along with several other components (see the project's web site http://www.cyc.com).

Knowledge Base (KB)

KB consists of concepts and assertions (rules) that relate those concepts, which are intended to represent fundamental human knowledge: facts, rules of thumb and heuristics in reasoning about objects and events of everyday life (such as terms or concepts like "first date" and rules of thumb like "People are more polite on their first date than they are on their nth date.").

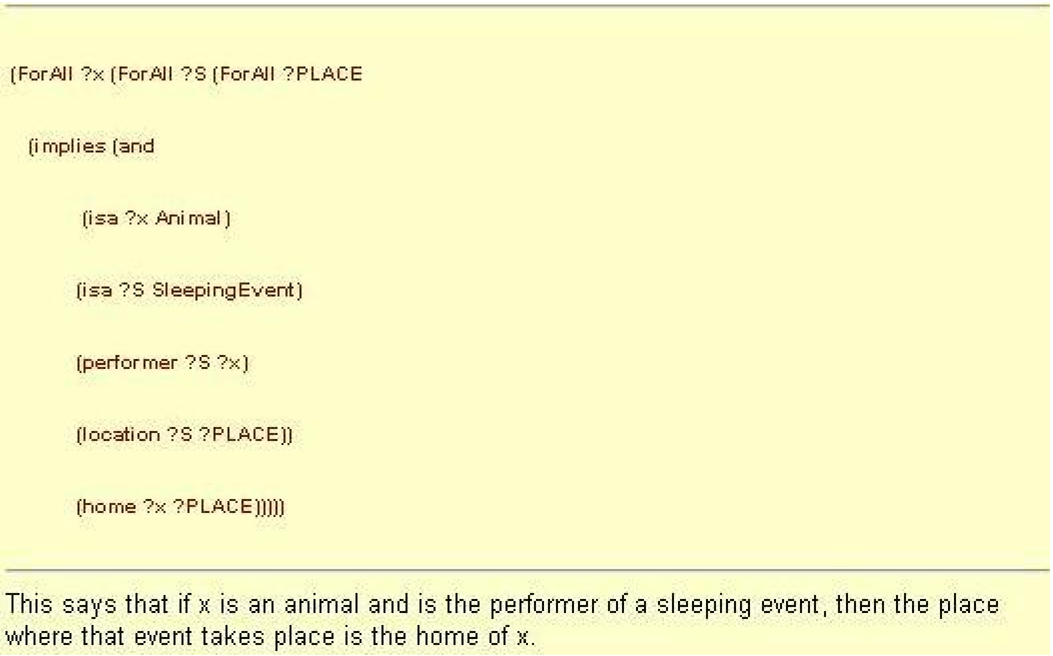
The assertions are grouped according to certain criteria such as the knowledge domain, level of detail, and time interval, which enables the system to maintain contradictory assertions.

The entire Cyc ontology currently contains hundreds of thousands of terms (more than 300, 000), along with millions of assertions relating the terms to each other, forming an upper ontology whose domain is all of human consensus reality. The graph representation of the knowledge base is as follows (Please see the address below to navigate the current knowledge base on an interactive interface,
http://www.cyc.com/cyc/technology/whatiscyc_dir/whatdoescycknow):



Currently, Thing is the concept at the top of the graph that includes everything there is. Some domain-specific concepts included currently are space, time, logic, set, human anatomy and physiology, mechanical and electrical devices, purchasing and shopping, military organizations, and many others.

For the representation of knowledge, a special language CycL, which is characterized as an extension of first-order predicate calculus is used. A knowledge representation sample from the KB is as follows:



```
(ForAll ?x (ForAll ?S (ForAll ?PLACE
  (implies (and
        (isa ?x Animal)
        (isa ?S SleepingEvent)
        (performer ?S ?x)
        (location ?S ?PLACE))
        (home ?x ?PLACE))))
```

This says that if x is an animal and is the performer of a sleeping event, then the place where that event takes place is the home of x.

Inference Engine

CYC's inference engine performs logical deduction (including modus ponens, modus tollens, universal and existential quantification), with the use of several AI techniques. It also includes several special-purpose inferencing modules, each performing a different kind of reasoning. These include equality reasoning, temporal reasoning, and mathematical reasoning.

The project is still under development, and there is an open-source version of it to be used in knowledge-intensive applications such as speech understanding, games, and email prioritizing, routing, summarization, and annotating. Its current release contains the entire Cyc ontology and the inference engine, together with the other associated tools of the project. For the open-source version see http://www.opencyc.org/.

In addition, its latest release (with the complete content of the CYC knowledge base) is made available to research community under a ResearchCyc license. See http://research.cyc.com/.

For the details and further information about the CYC project, please see the project's web site.

Open Mind Initiative

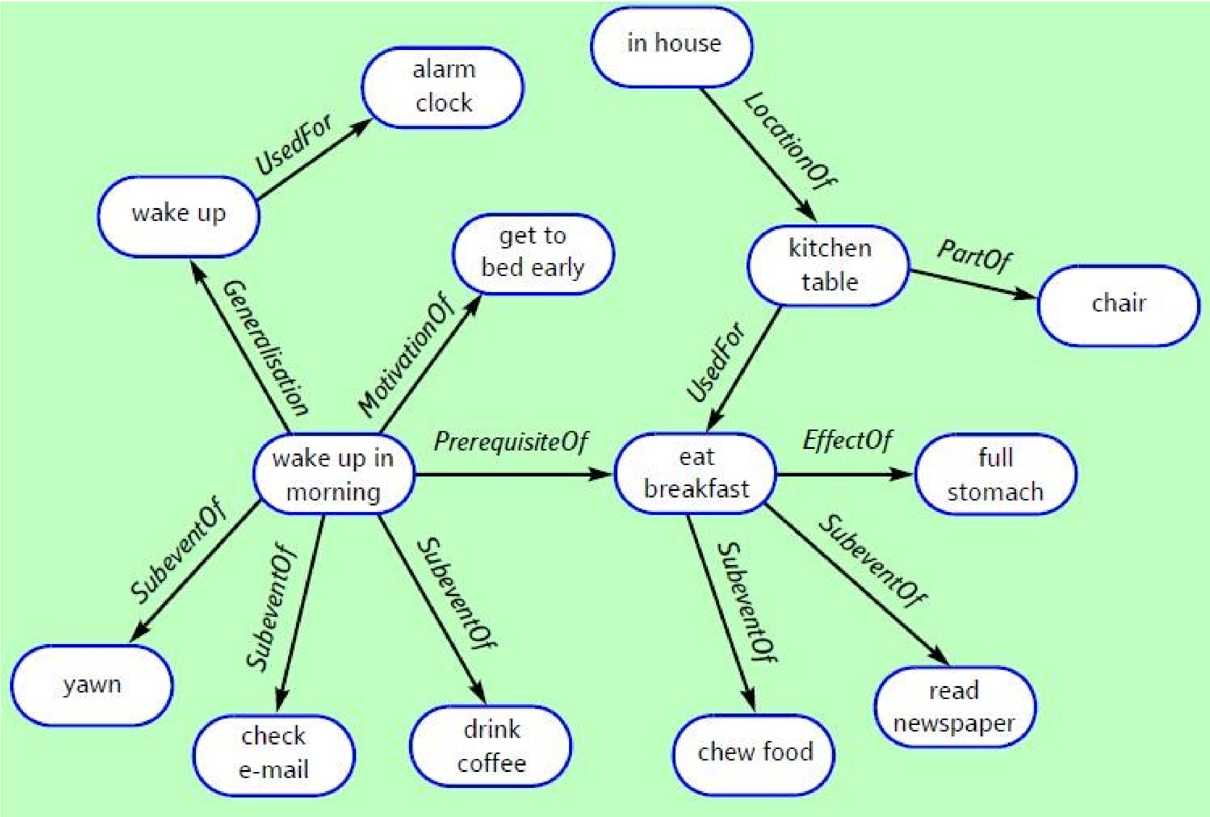
Open Mind Initiative (OMI) is a project started by Dr. David Stork to advance research in Artificial Intelligence by focusing on the aspects of humanity that aren't so easy to codify into algorithms. Currently work is carried out in the domains of the capacity to understand speech, to recognize handwriting, and to comprehend very basic, common-sense ideas. The main character of the project is that it is a collaborative framework supporting domain experts (who provide algorithms), tool developers (who provide software infrastructure and tools) and non specialist "e-citizens" (who contribute raw data) based on open source method (http://www.opensource.org/).

Open Mind Common Sense (OMCS) is one of OMI's projects being carried out in the domain of common sense reasoning. It includes a tool-kit named ConceptNet which is a freely available commonsense knowledgebase and natural-language-processing toolkit.

The ConceptNet

Knowledge Base

The knowledge base of ConceptNet is structured as a semantic network of common sense knowledge encompassing the spatial, physical, social, temporal, and psychological aspects of everyday life. It is presently available in two versions: concise (200,000 assertions) and full (1.6 million assertions). Following is an excerpt from ConceptNet's semantic network:



The full version of the semantic network at present contains 1.6 million edges (assertions) connecting more than 300 000 nodes, where nodes are semi-structured English fragments. Nodes in the semantic network can be pure lexical items (words and simple phrases with atomic meaning) or compound concepts in the sense that an action verb is combined with one or two arguments (e.g. “buy food”, “drive to store”). They are interrelated by an ontology of twenty semantic relations such as EffectOf (causality), SubeventOf (event hierarchy), CapableOf (agent’s ability), PropertyOf, LocationOf, andMotivationOf (affect).

The knowledge base is generated automatically from the English sentences of the Open Mind Common Sense corpus whose data has been collected from public, over 14 000 web-enabled users for the past six years. Any user can enter sentences of commonsense knowledge in a fill-in-the-blank fashion (e.g. “The effect of eating food is …”, “A knife is used for …”), and by the application of natural language processing and extraction rules to the semi-structured English sentences of over 700 000 in number, 300 000 concepts and 1.6 million binary-relational assertions are extracted to form the knowledge base. For example, from the sentence, ‘A lime is a sour fruit’, the assertions IsA (lime, fruit) and PropertyOf (lime, sour) are extracted. In addition, generalizations are also inferred. For example, if the majority of fruits have the property ‘sweet’, then the assertion PropertyOf (fruit, sweet) is formed.

The nature of the knowledge contained in the database is defeasible in the sense that it describes something that is often true but not always, like our commonsense knowledge (e.g. EffectOf (“fall off bicycle”, “get hurt”)

Contextual Reasoning

Apart from its knowledge base, ConceptNet also includes a mechanism of natural-language processing, operating on the knowledge base to perform various contextual-commonsense-reasoning tasks. To achieve such tasks, network-based reasoning methods like spreading activation and graph traversal are used.

For further and detailed information about the Open Mind Initiative (general information about the project and its application domains), please see http://www.openmind.org/.

For further and detailed information about the ConceptNet Project (papers, downloads, etc.), please see http://web.media.mit.edu/~hugo/conceptnet/. Particularly, you can read the article in the address http://web.media.mit.edu/~hugo/publications/papers/BTTJ-ConceptNet.pdf which is a very nice introduction to the project with information about its evolution and comparison to similar ontology/knowledge base projects.

FrameNet

FrameNet is a lexicon-building project for English, based on *frame semantics* [1], carried out by International Computer Science Institute of University of Berkeley. The product of the project is a database that documents the range of semantic and syntactic combinatory possibilities of words in each of their different senses through an annotation process performed on example sentences of a given corpus, together with a list of *frames*, and annotations of sentences containing those words. The project's recent version is its third release, and used by NLP researchers, lexicographers, language teachers and advanced language learners.

Frame: *schematic representation of a situation type (eating, spying, removing, classifying, etc.) together with lists of the kinds of participants, props, and other conceptual roles that are seen as components of such situations. The semantic arguments of a predicating word correspond to what we call the* ***frame elements(FE)*** *of the frame associated with that word* [2].

For example, Apply_Heat is a *frame* that describes a situation involving a Cook, some Food, and a Heating_Instrument (see the following example sentence), which are *frame elements* and is *evoked* by words such as bake, boil, steam, fry, etc.

[$_{Cook}$ Matilde] **fried** [$_{Food}$ the catfish] [$_{Heating\_instrument}$ in a heavy iron skillet].

The work in the project is basically carried out by first selecting words with particular meanings. Then, based on these meanings (including different senses of a single word, each word with its corresponding meaning characterized as a ***lexical unit(LU)***), underlying frames or conceptual structures are described. As a third step, the example sentences of the given corpus containing these words are examined in order to record the ways in which the components of the examined sentences express information about the frames they evoke.

Database

Currently (version 1.3), there are more than 10,000 lexical units, more than 6,000 of which are fully annotated, in more than 800 hierarchically-related semantic frames, exemplified in more than 135,000 annotated sentences in the database. Beginning with version 1.3, the quality of FrameNet data is monitored by a consistency management system. The main FrameNet corpus is 100-million words British National Corpus which is both large and balanced across genres (editorials, textbooks, advertisements, novels, sermons, etc); however it lacks many specifically American expressions. Additionally, U.S. newswire texts provided by the Linguistic Data Consortium are used. The newly released initial part of the American National Corpus has also been recently acquired to begin using soon.

Relations in the Database

There are various kinds of semantic relations between the frames in the database. These include Inheritance, Using, Subframe, SeeAlso, Causative_of and Inchoative_of. New relations have been added in the recent version.

Inheritance: Relationship in which a child frame is a more specific elaboration of the parent frame, and frame elements, subframes, and semantic types of the parent have correspondents in the child frame.

Using: Relationship that can be characterized as partial inheritance, in which a specific frame makes reference in a very general kind of way to the structure of a more abstract, schematic frame.

Subframe: Some frames (called complex frames) refer to sequences of states and transitions, each of which can itself be separately described as a frame. The separate frames (called subframes) have a Subframe relation with the complex frame.

SeeAlso: Relationship between each member of groups of frames which are similar, and should be carefully differentiated from each other.

Causative_of and Inchoative_of: Close and fairly systematic non-inheritance relationships between certain kinds of frames.

Furthermore, there is a mechanism called Semantic Type for capturing semantic facts about frames, FEs, or LUs that don't necessarily fit into developing hierarchy of frames.

It should be noted that the FrameNet database currently is not an ontology of things since it has not annotated many nouns denoting artifacts and natural kinds.

[1] See the references on Publications and Papers section of the project’s web site, http://framenet.icsi.berkeley.edu/.

[2] For detailed information about the project, please see the project's web site; particularly you can access the online version of the book of the project from http://framenet.icsi.berkeley.edu/book/book.pdf

---

For a list of ontology projects and groups worldwide, please see http://www.cs.utexas.edu/~mfkb/related.html.