Plan for INFO 202 Lecture #11

Introduction to ontology

A vocabulary for {lexical, conceptual} relationships

Thesauri, codes, and synonym rings

WordNet
Making Sense [1]

I saw a:
  Man
  Star
  Molecule

with a:
  Telescope
  Microscope
  Binoculars

How many combinations make sense?
Making Sense [2]

"Bob saw the plane flying over Denver"
"Bob saw the mountains flying over Denver"

- What does "flying" refer to in each sentence?
"How much is that doggy in the window?"

- Who is asking the question?
- What unit of measurement does "how much" refer to?
- Is the dog really "in" the window?
Language and Meaning

Words and sentence structure only hint at meaning

Meaning is constructed from all the clues or cues in the context of use -- common knowledge, assumptions, previous discourse, the present situation, and inferences from all of these

How much "context" and "common knowledge" must be represented / understood to make sense of what meaning is intended?
Two Solutions to the "Vocabulary Problem"

Furnas et al's solution (for people) was ...

The Artificial Intelligence solution (for computers) is to give an information system all the knowledge -- including "commonsense" -- that is needed to interpret every user's expressions in every context.

A great deal of work in AI has been dedicated to building knowledge bases to support language understanding, reasoning, problem solving applications.

The most famous / infamous effort is the Cyc project (http://www.cyc.com)
Cyc -- "Formalized Commonsense Knowledge"

Cyc knows about 200,000 basic concepts and a few million human-entered assertions about the world -- "facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life"

The Cyc knowledge base consists of terms -- which constitute the vocabulary of CycL -- and assertions that relate those terms

This kind of common sense is a pre-requisite for computers to achieve anything approaching human competence on natural language processing tasks (once you get outside of narrow, constrained domains)

The Cyc KB is divided into many (currently thousands of) "microtheories", each of which is essentially a bundle of assertions that share a common set of assumptions
What Cyc Knows About [1]

Map of High-Level Cyc Topics
What Cyc Knows About [2]
Cyc Examples

Cyc can find the match between a user's query for "pictures of strong, adventurous people" and an image whose caption reads simply "a man climbing a cliff"

Cyc can notice if an annual salary and an hourly salary are inadvertently being added together in a spreadsheet

Cyc can combine information from multiple databases to guess which physicians in practice together had been classmates in medical school
Cyc Assertions About "Dog"

[Def] "A BiologicalSpecies (scientific name 'Canis familiaris') that is a specialization of CanineAnimal

Each instance of Dog is a canine animal that has either been bred to be a domestic pet (see DomesticatedAnimal) or is a wild canine animal that is not an instance of Wolf, Fox, or any other non-dog specialization of CanineAnimal

Note that although Dog and Wolf are considered distinct BiologicalSpecies, instances of the two can and do interbreed successfully. This species classification is therefore unusual, and in some circles, controversial."
What is An Ontology?

An ontology defines the terms used to describe and represent an area of knowledge.

Ontologies are used by people, databases, and applications that need to share domain information.

Ontologies include computer-usable DEFINITIONS of basic concepts in the domain and the RELATIONSHIPS among them.

They encode knowledge in a domain and also knowledge that spans domains to make that knowledge reusable.

Cyc attempts to be a "foundation" or "upper" ontology, because it includes general concepts common to all domains, but is primarily a "domain" or "lower" ontology because most of its concepts are quite specific.
That's A Very Broad Definition

The word ontology has been used to describe artifacts with different degrees of structure that differ:

- ... according to how precisely the terms are defined
- ... according to how precisely the relationships among them are expressed

So the simplest ontology is a dictionary

A thesaurus is a somewhat more complex ontology

More complete ontologies are expressed using formal logic-based language
Why Create an Ontology?

To share common understanding of the structure of information

To enable reuse of domain knowledge

To make domain assumptions explicit

To separate domain knowledge from operational knowledge

To analyze domain knowledge
Words and Concepts

A prototypical word is the minimal "meaning bearing" element of language.

Words express concepts, but not all concepts are "lexicalized".

These "lexical gaps" differ from language to language.

Whereas "conceptual gaps" -- the things we can't think of -- may be innate and universal.
Relations Among Words

Polysemy
Synonymy
Antonymy
Relations Among Concepts

Metonymy

Hyponymy/Hypernymy

Meronymy/Holonomy
Polysemy

Many "word forms" (particular spelling patterns) are polysemous with multiple senses -- they are semantically ambiguous

- That dog has floppy ears
- She has a good ear for jazz.

These senses are established in the language and stored in a person's memory, and not be just possible uses

"bank" (financial) has related senses:

- a building (the bank on Shattuck)
- a specific financial firm (Washington Mutual)
- where money is kept (abstract notion)
Polysemy vs Metaphor

A polyseme is a word with multiple senses, but in which all the senses are related.

Metaphor is a kind of nonlinear or figurative polysemy, where a sense is related but perhaps only on one or a few of the facets of a concept.

- *swallow* a pill
- *swallow* an argument
Polysemy vs Homophony

A homophone is a word with multiple senses, but for which the different senses are not conceptually related

- bank (financial sense)
- bank (river sense)

But what appears to be homophony may be polysemy from a historical perspective...and native speakers sometimes disagree about whether two senses are polysemous or homophonous

- mole (a small animal that digs underground)
- mole (a spy who infiltrates an organization)
Synonymy

Synonyms are different word forms that can express the same concept

- cat, feline, Siamese cat

Absolute synonyms that can be substitutable for each other in every conceivable context probably don't exist

- \{weep, sob, cry\}—differ in scale or degree
- "brave" implies physical, "courageous" implies moral

Propositional synonyms are more common - substitutability entails the same truth conditions

- She plays the \{violin, fiddle\}
Antonymy

Antonyms are lexical opposites

Some are "true antonyms" because they are inherently binary
  - dead / alive, true / false, on / off

Others are "graded"
  - long / short, hot / cold

Markedness: if one member of a pair is more restricted in its contexts it can stand out psychologically
  - long is unmarked, short is marked
Metonymy

Using one aspect of something to stand for the whole

- "Bank" as building stands for the institution of the "bank"
- "The White House released new budget figures today"
- (busboy says to waiter) "The ham sandwich wants his check"
Hyponymy/Hyperonymy

The IS-A relationship -- a relationship between concepts that organizes the word nouns into a "lexical hierarchy"

Often used to situate "basic categories" with respect to superordinate and subordinate categories

- A robin is a hyponym of bird
- An emu is a hyponym of bird
- A bird is a hyponym of animal
- An animal is a hypernym of bird

A is a hyponym of B if A is a type of B

Co-hyponyms are mutually exclusive categories

A is a hypernym of B if B is a type of A
A Formula for Definitions

hyponym = {adjective+} hypernym {distinguishing clause+}

Robin = Migratory BIRD with clear melodious song, a reddish breast, gray or black upper plumage

Doesn't mention every characteristic of hyponym, only those needed to distinguish from other hyponyms
Meronymy/Holonomy

Meronymy defines Part/Whole relations
- Beak is a meronym of Bird
- Bark is a meronym of Tree

Holonyms are (approximately) the inverse of meronyms
- Tree is a holonym of Bark

Meronymy is transitive conceptually but not lexically
- The Knob is part of the Door
- The Door is part of the House
- but sounds odd to say "The Knob is part of the House"
A THESAURUS is a tool for leading cataloguers or searchers to the "right" or "good" terms of a controlled vocabulary. It is a collection of (usually single) vocabulary terms annotated with lexical relationships to indicate terms that are:

- Preferred (UF "used for")
- Broader (BT "broader term")
- Narrower (NT "narrower term")
- Related (RT "related term" or "see also")

USE in a thesaurus refers the reader from a variant term to a preferred term; the inverse of UF.
Thesaurus Examples (from "Boxes and Arrows")
Thesaurus Examples (from "Boxes and Arrows")

<table>
<thead>
<tr>
<th>Women's Pants</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT Pants</td>
</tr>
<tr>
<td>NT Casual Pants</td>
</tr>
<tr>
<td>NT Dress Pants</td>
</tr>
<tr>
<td>Jeans</td>
</tr>
<tr>
<td>BT Pants</td>
</tr>
<tr>
<td>NT Levis</td>
</tr>
<tr>
<td>NT Wranglers</td>
</tr>
<tr>
<td>NT Sports Pants</td>
</tr>
<tr>
<td>UF Waist Overalls</td>
</tr>
<tr>
<td>RT Denim</td>
</tr>
<tr>
<td>RT Overalls</td>
</tr>
</tbody>
</table>
http://www.eric.ed.gov/

-- Education Resources Information Center (ERIC) is a digital library of education-related resources, sponsored by the Institute of Education Sciences of the U.S. Department of Education.
**Thesaurus Example - ERIC [2]**

![Thesaurus Example - ERIC][1]


---

**Competency Based Education**

**Record Type:** Main

**Scope Note:** Educational system that emphasizes the specification, learning, and demonstration of those competencies (knowledge, skills, behaviors) that are of central importance to a given task, activity, or career

**Category:** The Educational Process: Societal Perspectives

**Broader Terms:** Education

**Narrower Terms:** Competency Based Teacher Education

**Related Terms:** Academic Standards; Accountability; Back to Basics; Behavioral Objectives; Competence; Individualized Instruction; Minimum Competencies; Minimum Competency Testing; Outcome Based Education; Performance; Performance Based Assessment; Proficiency Based Education; Proficiency Based Education (1974-1980); Student Certification

**Used For:** Competency Based Education; Criterion Referenced Education; Output Oriented Education; Performance Based Education

**Use Term:** n/a

**Use And:** n/a

**Add Date:** 03/10/1980
The W3C's "Datatype Thesaurus"

Rick Jellife's interactive datatype hierarchy -- each type is formally related to those around it by restriction relationships (http://www.xml.com/lpt/a/2000/11/29/schemas/dataref.html)
"If you need to know about cow farming, you're probably also searching for cattle ranching, beef (or dairy) production, and Kuhbauernhof, whether you know it or not." (Tim Bray)

A synonym ring connects a series of terms together and treats them all as equivalent for search purposes.

It is a weaker mechanism of vocabulary control than an authority file or thesaurus because it doesn't designate a term as the preferred or normative form.
Recommended Types of Synonyms for Rings

Scientific terms versus popular use terms: acetylsalicylic acid, aspirin; lilioceris, lily beetle

Variant spellings: cancelled, canceled; honor, honour

Abbreviations
Types of Abbreviations

INITIALISMS - abbreviations composed from the initial letters of a series of words related to the "spelled-out" words (BBC -> British Broadcasting Company, MPG -> miles per gallon)

An ACRONYM is a special type of initialism in which the resulting abbreviation is pronounceable (NATO, ISO)

An APOCOPATION is an abbreviation formed by truncating a longer word (bicycle to bike, University of California to Cal)

A NICKNAME is often an apocopation designed to be clever, cute, derogatory, or otherwise memorable or marketable (T-bird for Thunderbird, Bud for Budweiser, Shaq for Shaquille)
WordNet

Another "ontological resource" -- a "semantic dictionary" - started in 1985 by George Miller and others at Princeton's Cognitive Science program.

Instead of using spelling as the primary organizing principle for words, it uses their semantic properties and relationships.

Designed as a network to capture the idea that words and concepts are an interrelated system.
The Structure of WordNet

The building block is the SYNSET, the set of words that are interchangeable in some contexts.

Synsets are connected by semantic relationships to represent conceptual structure.

The words are related to represent lexical structure.

This separation of conceptual and lexical relationships let you navigate in either "space"
The WordNet Lexical Hierarchy

Linking words to their hypernyms creates a lexical hierarchy.

Properties are associated with the highest level in the hierarchy to which they apply, which enables inheritance.

You can measure conceptual distance between any pair of words by traversing the lexical tree through the nodes that join them.
Using WordNet

The WordNet database is freely downloadable from
http://wordnet.princeton.edu/~wn

http://www.synonym.com/ is easier to work with

The Visual Thesaurus
is a commercial product that uses WordNet
(http://www.visualthesaurus.com)
Uses of WordNet

WordNet is very commonly used in natural language processing research and applications.

Most important automated use of WordNet is in sense disambiguation. Pick out a sense based on contexts in which it occurs.

You look up word in WordNet, and the words that appear with it (on either side of it in some window of text).

Determine how far apart the words are in "wordnet synset distance" by following the lexical hierarchy links.

The smallest distance identifies the sense of the polysemous word.
Readings for INFO Lecture #12

"The Semantic Web" Tim Berners-Lee, James Hendler, and Ora Lassila
Scientific American (May 2001)

Knowledge@Wharton, "What Is the Next Big Thing on the Web? It May
Be a Small, Simple Thing -- Microformats"


Catherine Marshall & Frank Shipman, "Which Semantic Web?" ACM
conference on Hypertext and Hypermedia (2003)