9. Ontology

INFO 202 - 29 September 2008

Bob Glushko

Plan for INFO 202 Lecture #9

Introduction to ontology
A vocabulary for {lexical, conceptual} relationships
Indexes, thesauri, synonym rings
Topic maps
Making Sense [1]

<table>
<thead>
<tr>
<th>I saw a:</th>
<th>with a:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>Telescope</td>
</tr>
<tr>
<td>Star</td>
<td>Microscope</td>
</tr>
<tr>
<td>Molecule</td>
<td>Binoculars</td>
</tr>
</tbody>
</table>

*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?
- Where is "Bob" located?
Making Sense [3]

"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.
Ontology Example -- Computer Intrusion

Ontology Example -- Beer
Ontology Example -- W3C Datatypes

Rick Jellife's interactive datatype hierarchy -- each type is formally related to those around it by restriction relationships

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

Hyponymy/Hypernymy
Meronymy/Holonymy

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 (Wells Fargo)
- 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 Homonymy

Two kinds of homoyms:

A HOMOGRAPH 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 homography may be polysemy from a historical perspective...and native speakers sometimes disagree about whether two senses are polysemous or homonymous.

HOMOPHONES are two words with the same pronunciation but different spellings and meanings.
**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
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
- 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

\[
\text{hyponym} = \{\text{adjective+}\} \ \text{hypernym} \ \{\text{distinguishing clause+}\}
\]

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/Holonymy

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"

Indexes

A "map" to the knowledge contained in a text or collection of texts

Consists of a list of (names of) topics and references to occurrences of those topics

Topics can be arranged / decomposed hierarchically

Topics can be associated with other topics

Topics and references can be "typed"

Topics and references can be "aliased"

Typographic conventions can be used to reinforce these distinctions and relationships
Index - Opera (from TAO of Topic Maps)

La Bohème, 10, 70, 197-198, 326
Cavalleria Rusticana, 71, 203-204
The Girl of the Golden West, see La fanciulla del West
Leoncavallo, Ruggiero
    I Pagliacci, 71-72, 122, 247-249, 326
Madama Butterfly, 70-71, 234-236, 326
Manon Lescaut, 294
Mascagni, Pietro
    Cavalleria Rusticana, 71, 203-204
Puccini, Giacomo, 69-71
    La Bohème, 10, 70, 197-198, 326
    La fanciulla del West, 291
Madama Butterfly, 70-71, 234-236, 326
Manon Lescaut, 294
Tosca, 26, 70, 274-276, 326
Turandot, 70, 282-284, 326
Rustic Chivalry, see Cavalleria Rusticana
singers, 39-52,
    See also individual names
    baritone, 46
    bass, 46-47
    soprano, 41-42, 337
    tenor, 44-45

Index - Document Engineering

Forecasts:
    type of document 10.5.3;
foreign key: 14.4.2 (sidebar);
formal ontology: 6.5.2;
    model based applications 15.0;
    naming components 7.7.2; 12.1.11;
    semantic rules 8.3.3;
    semantic web 15.3.7;
Forwarding Instruction: 1.1;
    entry point 14.4.3;
    incremental information trail 9.8.6;
franchising:
    pattern 10.4.4;
Frege, Gottlob:
    intension and extension 12.1.9 (sidebar);
fulfilment: see distribution;
functional dependency: 6.5.1; 7.8.1;
    normalization 13.4.1;
    rules 13.4.4;
functional organization: 9.2;
G2B: See e-government;
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 "narrower term" or "see also")

USE in a thesaurus refers the reader from a variant term to a preferred term; the inverse of UF.
Thesaurus Example (Graphical Format)

![Diagram](image)

Thesaurus Example (Textual Format)

<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>
Thesaurus Example - ERIC [1]

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]
Synonym Rings

"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 (Initialisms, acronyms, apocopeations, nicknames)
Topic Maps

A recent invention... designed to support the distributed management of information and "knowledge"

Motivation is "merging the indexes" of printed and digital information collections

Two-layer model: an "information layer" consisting of "topics" and "associations" and a "knowledge layer" that are linked together by "occurrences"

(Try the "Omnigator" - generic topic map browser - at http://www.ontopia.net/omnigator)
Assignment 4

Designing a faceted classification system to organize household "tools" using 10 instances provided to you

Test the scope and robustness of your system with additional instances provided to you by someone else

Use Facetmap (http://facetmap.com/)

Turn in your facetmap and a report about your experiences by October 6

Readings for INFO Lecture #10

R. Glushko, "Modeling Methods and Artifacts for Crossing the Data/Document Divide"

K. Thomas, "XML in the Pharmaceutical Industry: Structured Product Labeling"

Tim Bray, "On Language Creation"