The Goal of Language Understanding

John F. Sowa

VivoMind Intelligence, Inc.

6 March 2008

The Ultimate Learning System

Remark by a three-year-old child named Laura:

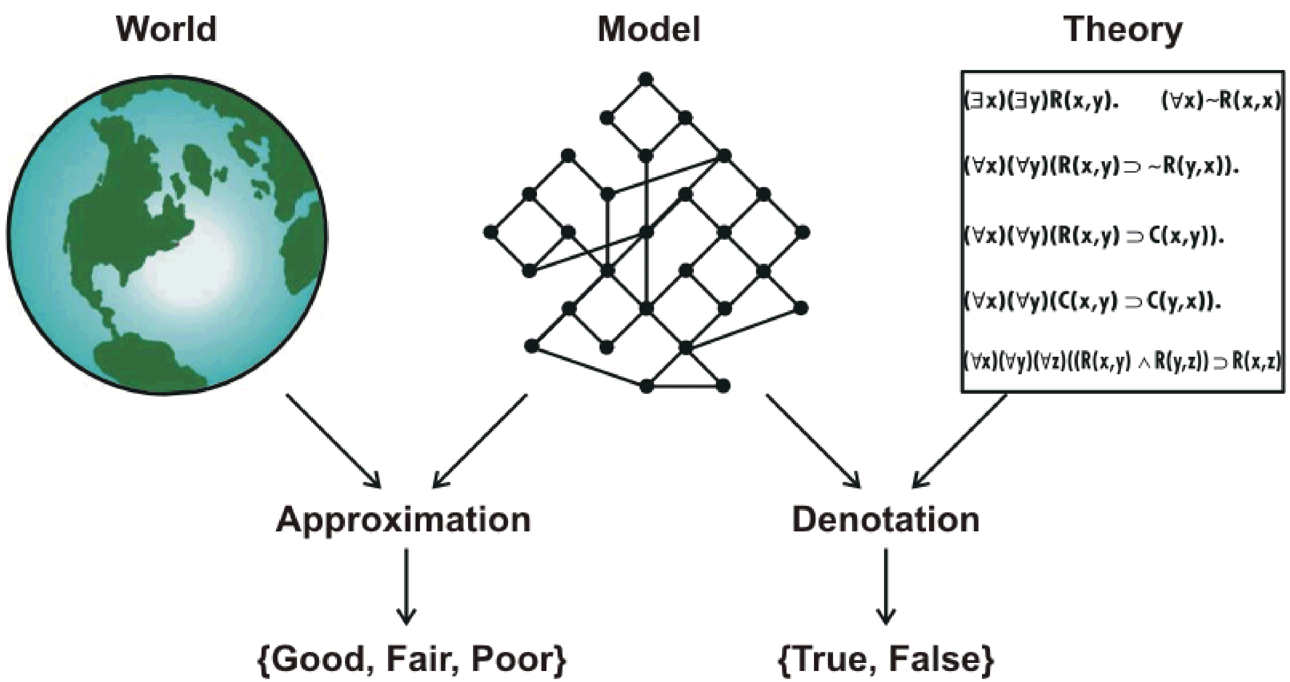
When I was a little girl, I could go “geek, geek” like that;
but now I can go “This is a chair.”

Enormous logical and grammatical complexity:

* Subordinate and coordinate clauses
* Tenses: Earlier time contrasted with “now”
* Modal auxiliaries: can and could
* Quotations: “geek, geek” and “This is a chair"
* Metalanguage about her own linguistic abilities
* Contrast shown by but
* Parallel stylistic structure

Could we ever simulate Laura’s ability to learn language?

Models and Reality



George Box: "All models are wrong, but some are useful."

Evaluating Truth in Terms of a Model

Truth is not fuzzy — it is context dependent.

Truth is determined by the correspondence of a proposition to reality, as mediated by some model of reality for some purpose.

Examples:

Is the earth spherical?

Is a ball bearing spherical?

Is a meatball spherical?

The answer is yes or no, depending on the chosen model.

And the choice of model depends on the purpose.

Aristotle’s Levels of the Psyche

A hierarchy with each function built on top of the earlier ones:

1. Nutrition: level of plants.
2. Perception: level of sedentary animals.
3. Locomotion: level of worms.
4. Imagery: level of animals with sight and hearing.
5. Thought: human level with speech and reason (logos).

Before the age of two, children acquire an enormous amount of knowledge at the prelinguistic levels.

Robot Levels of Competence

For mobile robots, Rodney Brooks made some finer subdivisions:

1. Avoiding: Avoid contact with other objects.
2. Wandering: Wander around without hitting things.
3. Exploring: Head for places that seem reachable.
4. Mapping: Build a map of the environment.
5. Noticing: Recognize changes in the environment.
6. Reasoning: Reason about and act upon objects.
7. Planning: Formulate and execute plans.
8. Anticipating: Reason about the behavior of other agents.

Language and logic are helpful for levels 6, 7, and 8.

But chimpanzees behave at those levels even without language.

Image-like Mental Models

Modeling hypothesis by Kenneth Craik:

> ***If the organism carries a small-scale model of external reality and of its own possible actions within its head, it is able to carry out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way react in a fuller, safer, and more competent manner to the emergencies which face it.***

The amount of information represented in an image is much larger than any description in language or logic.

And it is rarely expressed in words, even by adults.

Mental models could be simulated as "virtual reality."

Conceptual Graph Models

Models that bridge the gap between images, logic, and language.

They are not as detailed as the image-like models.

They represent only one perspective on an infinity of options.

But they are manageable with our current technology.

Each model represents one conception with one ontology.

But the same reality could be represented with different ontologies.

And different ontologies would result in different conceptual graphs that represent the same reality.

Need to Align Ontologies

Any topic can be described at many levels of detail with different choices of labels for the concept and relation types.

There is no standard upper-level ontology, but there are many important ontologies that must be accommodated:

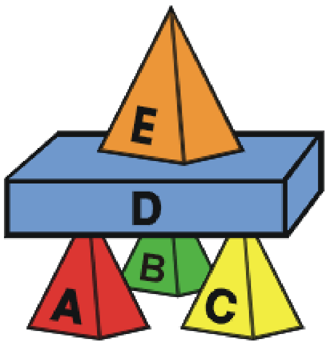
- Some widely used ontologies, such as Cyc, OpenCyc, SUMO, Dolce, BFO, etc.
- Many ontologies required by governments and large organizations.
- Example: The Amazon.com ontology.
- An enormous number of legacy systems with no explicit ontologies.

Critical requirement:

Enable heterogeneous systems to interoperate despite differences in their underlying ontologies, whether implicit or explicit.

Example of Different Ontologies

- The structure on the right may be described in different ways:



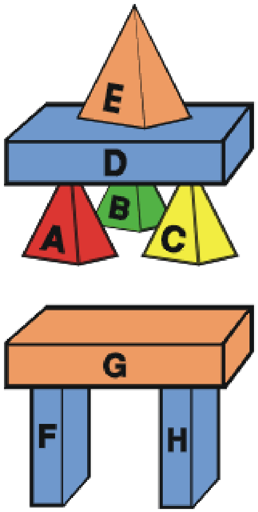
- English sentence: "A red pyramid A, a green pyramid B, and a yellow pyramid C support a blue block D, which supports an orange pyramid E."

- A relational database would use tables.

- But there are many different options for organizing the tables and choosing labels for the tables and columns.

- These choices lead to different ontologies, which may be structurally very different from one another.

Representation in a Relational DB



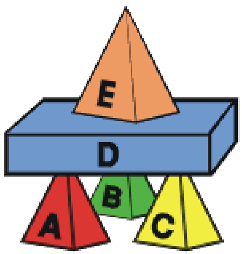
Objects

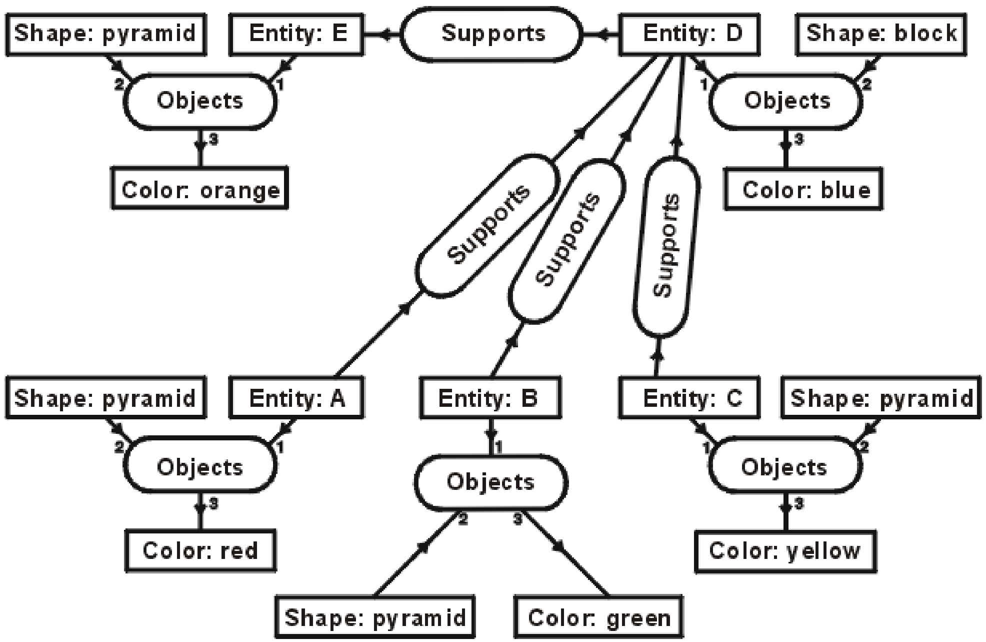
ID	Shape	Color
A	pyramid	red
B	pyramid	green
C	pyramid	yellow
D	block	blue
E	pyramid	orange
F	block	blue
G	block	orange
H	block	blue

Supports

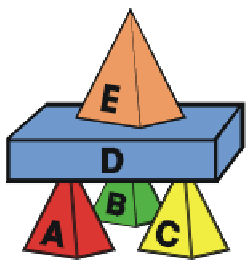
Supporter	Supportee
A	D
B	D
C	D
D	E
F	G
H	G

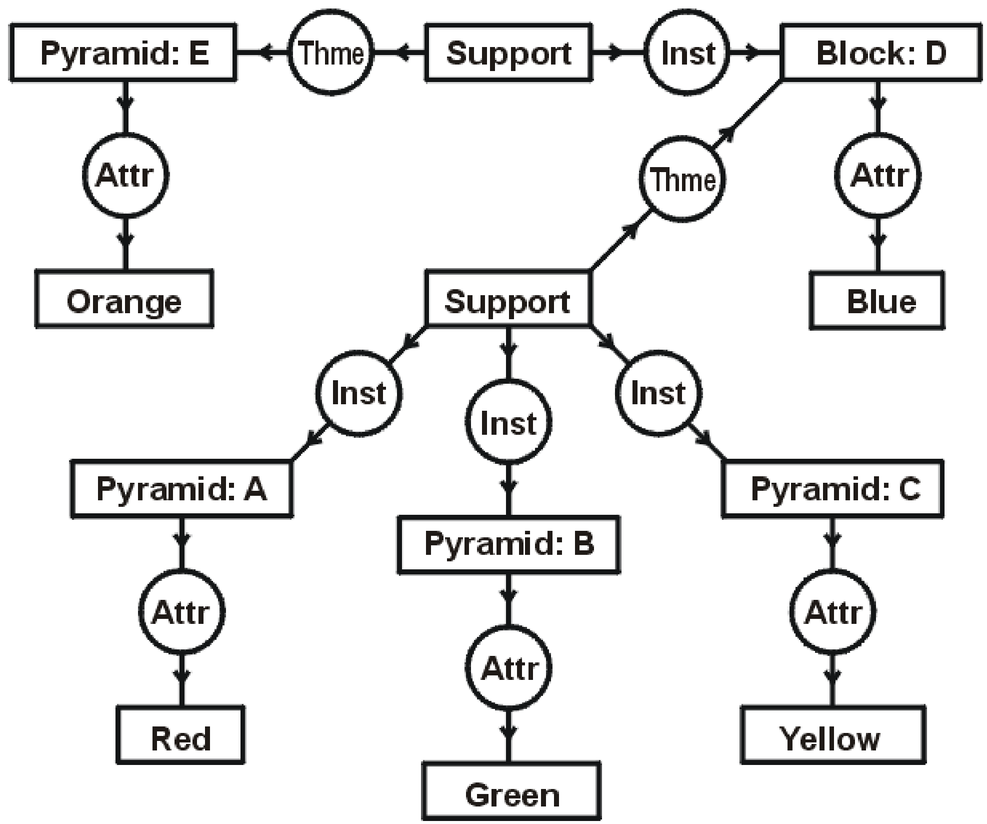
Conceptual Graph from Relational DB





Conceptual Graph from English Sentence



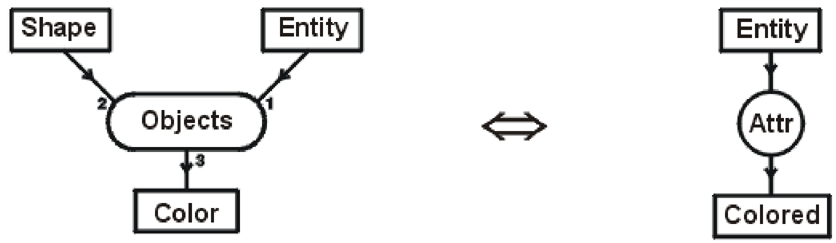


“A red pyramid A, a green pyramid B, and a yellow pyramid C support a blue block D, which supports an orange pyramid E.”

The Two CGs Look Very Different

* CG from RDB has 15 concept nodes and 8 relation nodes.
* CG from English has 12 concept nodes and 11 relation nodes.
* All the type labels of concept and relation nodes are different.
* But there are some structural similarities.
* VAE uses method #3 to find them.

Transformations Found by VAE



Top transformation was applied to 5 subgraphs.

Bottom one was applied to 4 subgraphs.

One application of any given transformation could be due to chance, but 4 or 5 applications are strong evidence for its significance.

The fact that these two transformations completely map each graph to the other is convincing.

Exact or Approximate Mappings

* This example illustrates exact mappings between graphs.

* For some applications, approximate mappings are acceptable.

* Approximate maps may merge nodes or violate type constraints.

* The basic algorithms are the same for both kinds of mappings.

* With strict constraints, only exact mappings are permitted.

* More relaxed constraints permit approximate mappings.

Legacy Re-engineering

Version of Intellitex applied to three languages — English, COBOL, and JCL:

* **1.5 million lines of COBOL.**
* **Several hundred JCL scripts.**
* **100 megabytes of English documentation — reports, manuals, e-mails, Lotus Notes, HTML, and transcriptions of oral communications.**

Goal:

* **Analyze the COBOL and JCL to determine:**

  Data dictionary, data flow diagrams, process architecture diagrams, system context diagrams.

* **Analyze the English documentation to determine:**

  Discrepancies between the documentation and the implementation.

  English glossary of all terms and their changes over the years.

  English captions on the diagrams derived from COBOL and JCL.

Estimate by a major consulting firm:

* **The project would require 40 people for two years = 80 person years.**

An Important Simplification

Extremely difficult problem:

* Trying to understand English specifications.

* Mapping the results to an executable COBOL program.

Much, much easier problem:

* Mapping COBOL to conceptual graphs.

* Using CGs derived from COBOL to analyze English documentation.

* The CGs derived from COBOL represent the necessary background knowledge for interpreting the English text.

* Any discrepancies that VAE finds between the CGs derived from English and the CGs derived from COBOL represent possible errors.

Results

Job finished in 8 weeks by two programmers, Arun Majumdar and André LeClerc.

* Four weeks for customization:

  Design and logistics.

  Additional programming for I/O formats.

* Three weeks to run Intellitex + VAE + extensions:

  24 hours a day on a 750 MHz Pentium III.

  VAE handled matches with strong evidence (close semantic distance).

  Matches with weak evidence were confirmed or corrected by Majumdar and LeClerc.

* One week to produce a CD-ROM with integrated views of the results:

  Glossary, data dictionary, data flow diagrams, process architecture, system context diagrams.

Total time of 16 person weeks instead of 80 person years.

Contradiction Found by VAE

From analyzing English documentation:

* Every employee is a human being.

* No human being is a computer.

From analyzing COBOL programs:

* Some employees are computers.

What is the reason for this contradiction?

Quick Patch in 1979

A COBOL programmer made a quick patch:

* Two computers were used to assist human consultants.

* But there was no provision to bill for computer time.

* Therefore, the programmer named the computers Bob and Sally, and assigned them employee ids.

For more than 20 years:

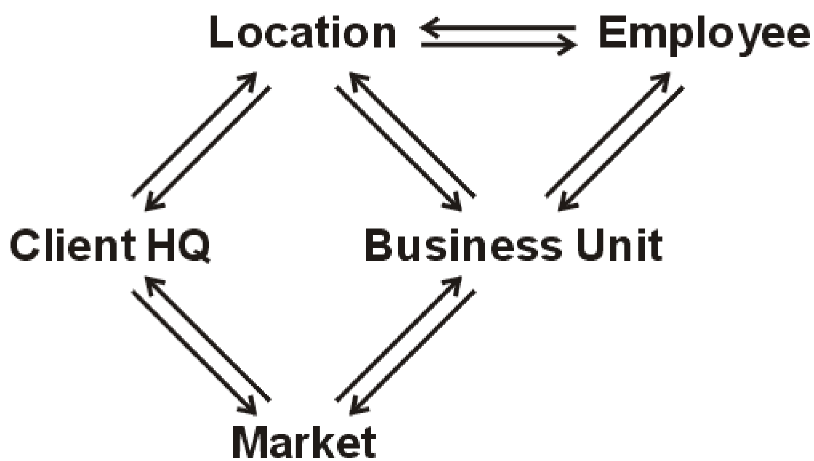
* Bob and Sally were issued payroll checks.

* But they never cashed them.

VAE discovered the two computer “employees.”

Mismatch Found by VAE

A diagram of relationships among data types in the database:



Question: Which location determines the market?

According to documentation: Business unit.

According to COBOL programs: Client HQ.

Management had been making decisions based on incorrect assumptions.

Conclusions

Deductive methods are good when there are widely applicable theories, as in physics, engineering, and established accounting procedures.

When there are no reliable theories, analogical reasoning is necessary.

Even when good theories are available, analogical reasoning can be a valuable supplement for handling exceptions.

Analogical reasoning can also be used at the metalevel to find mappings between different theories and ontologies.

But we are still very far from representing the level of language and learning of a three-year-old child.

Much more work is needed, especially in representing and processing image-like mental models.

Related Readings

A summary of conceptual graphs and their relationship to Peirce's existential graphs, the ISO standard for Common Logic, and other logic-based notations, such as RDF(S), OWL, SQL, Prolog, and Z. "Conceptual Graphs," in F. van Harmelen, V. Lifschitz, and B. Porter, eds., *Handbook of Knowledge Representation*, Elsevier, 2008, pp. 213-237. http://www.jfsowa.com/cg/cg_hbook.pdf

A survey of the psychological, philosophical, and linguistic issues related to artificial intelligence. "Categorization in Cognitive Computer Science," in H. Cohen & C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*, Elsevier, 2006, pp. 141-163. http://www.jfsowa.com/pubs/cogcat.htm

A discussion of analogical reasoning and applications of the VivoMind Analogy Engine (VAE). "Analogical reasoning," co-authored with Arun K. Majumdar, in de Moor, Lex, Ganter, eds., *Conceptual Structures for Knowledge Creation and Communication*, Proceedings of ICCS 2003, LNAI 2746, Springer-Verlag, Berlin, 2003, pp. 16-36. http://www.jfsowa.com/pubs/analog.htm

The complexity of what people have in their heads and the challenge of implementing systems that can handle anything remotely similar. "The Challenge of Knowledge Soup," in J. Ramadas & S. Chunawala, *Research Trends in Science, Technology and Mathematics Education*, Homi Bhabha Centre, Mumbai, 2005, pp. 55-90. http://www.jfsowa.com/pubs/challenge.pdf

The problems of making ontologies more flexible and dynamically adaptable to the changing aspects of the world. "A Dynamic Theory of Ontology," in B. Bennett & C. Fellbaum, eds., *Formal Ontology in Information Systems*, IOS Press, Amsterdam, 2006, pp. 204-213. http://www.jfsowa.com/pubs/dynonto.htm

A summary of Peirce's work on logic, semeiotic, and related topics and their relevance to modern work on semantic technology. "Peirce's contributions to the 21st Century," in H. Schärfe, P. Hitzler, & P. Øhrstrøm, eds., *Conceptual Structures: Inspiration and Application*, LNAI 4068, Springer, Berlin, 2006, pp. 54-69. http://www.jfsowa.com/pubs/csp21st.pdf

Issues in modal logic, a comparison of the approaches to semantics by Kripke and Dunn, and their application to conceptual graphs. "Worlds, Models, and Descriptions," *Studia Logica*, Special Issue Ways of Worlds II, 84:2, 2006, pp. 323-360. http://www.jfsowa.com/pubs/worlds.pdf