The Goal of Language Understanding

Chapter 7: Learning by reading

John F. Sowa

VivoMind Research, LLC

1 March 2014

The Goal of Language Understanding

Outline:

1. Problems and challenges (goal.pdf)
2. Psycholinguistics and neuroscience (goal2.pdf)
3. Semantics of natural languages (goal3.pdf)
4. Wittgenstein's early and later philosophy (goal4.pdf)
5. Dynamics of language and reasoning (goal5.pdf)
6. Analogy and case-based reasoning (goal 6.pdf)
7. Learning by reading

Each chapter is in a separate file. Later chapters make occasional references to earlier chapters, but they can be read independently.

7. Learning by Reading

Perfect understanding of natural language is an elusive goal:

- **Even native speakers don't understand every text in their language.**
- **Without human bodies and feelings, computer models will always be imperfect approximations to human thought.**

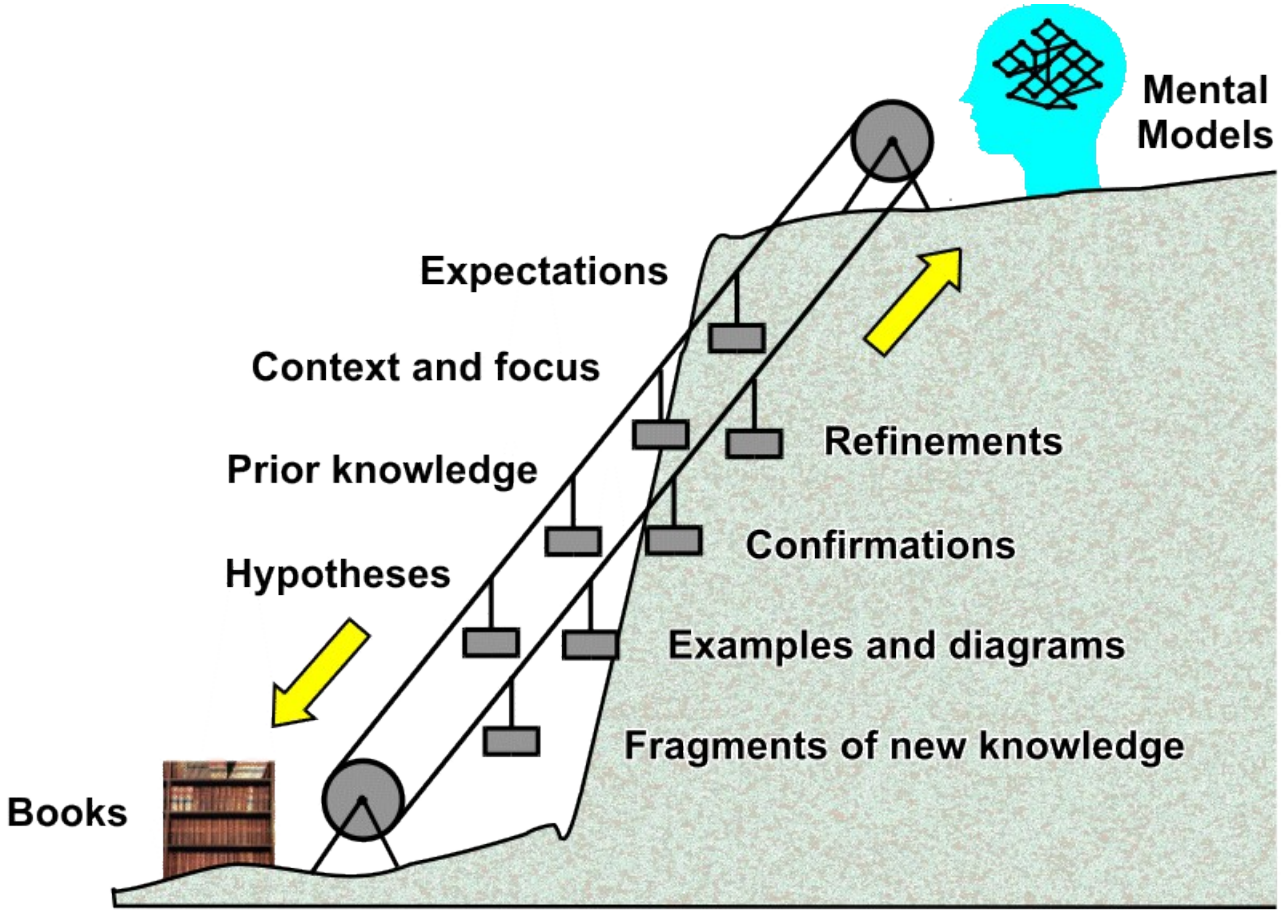
For technical subjects, computer models can be quite good:

- **Subjects that are already formalized, such as mathematics and computer programs, are ideal for computer sytems.**
- **Physics is harder, because the applications require visualization.**
- **Poetry and jokes are the hardest to understand.**

But NLP systems can learn background knowledge by reading:

- **Start with a small, underspecified ontology of the subject.**
- **Use some lexical semantics, especially for the verbs.**
- **Read texts to improve the ontology and the lexical semantics.**
- **The primary role for human tutors is to detect and correct errors.**

Reading is More Than a Translation to Logic



Reading a book is an interactive process that constructs mental models from information in the text, prior knowledge, and the evolving context.

A system for machine reading should implement equivalent processes.

Diagram adapted from Peter Clark, et al., Reading to learn, in *Proc AAAI 2007 Spring Symposium on Machine Reading*, 2007. Slides: http://www.cs.utexas.edu/users/pclark/papers/ss07.ppt

Application to Information Extraction

The next slide shows an excerpt of the data generated as part of a contract with the US Department of Energy.

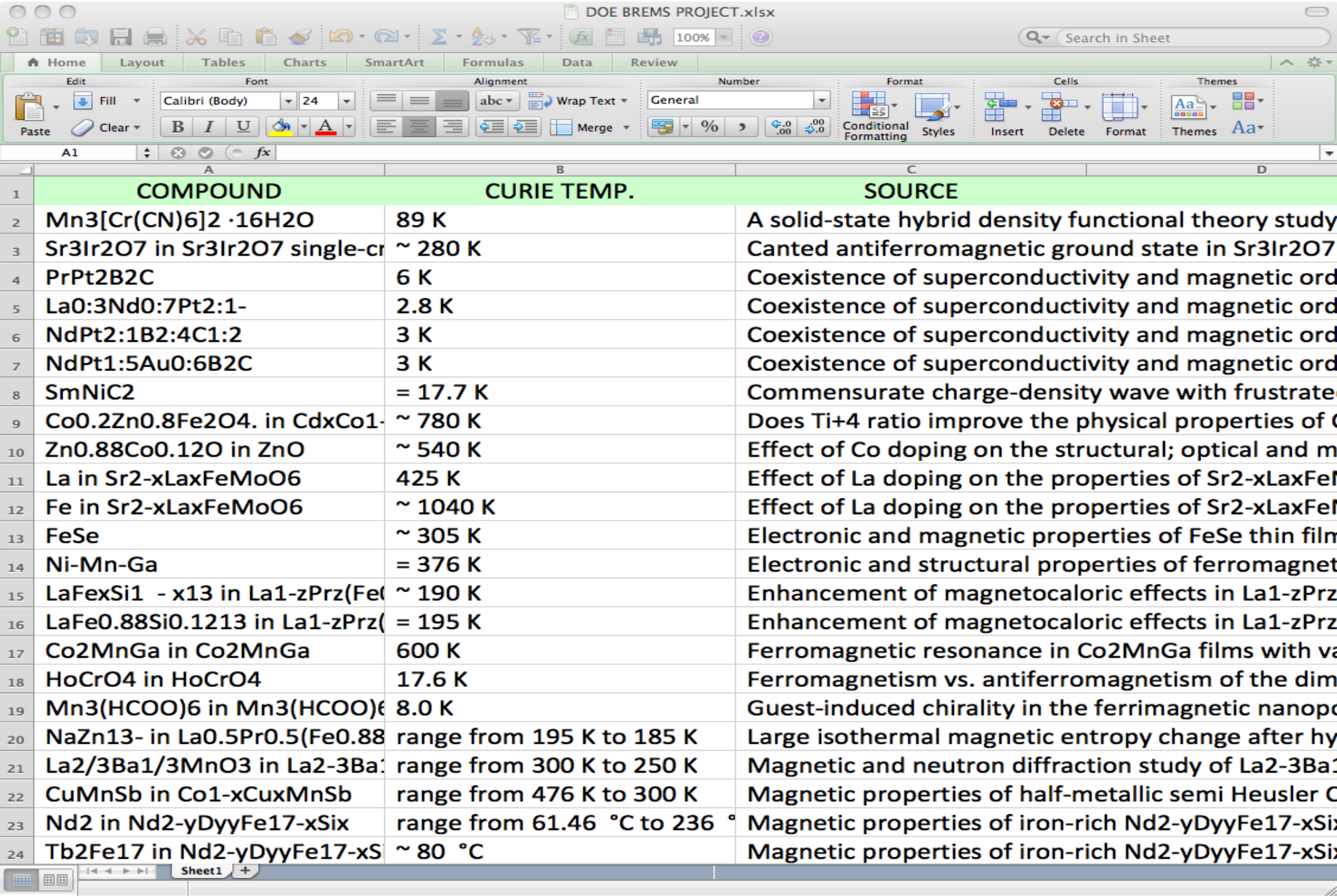
Analysis performed on a collection of reports written in English:

- **Map every sentence of each document to a conceptual graph.**
- **Find coreference links between concept nodes.**
- **The result is a large CG that represents all the data in the document.**
- **For each column in the table, such as source or Curie temperature, match a query CG that asks for the data that belongs in that slot.**
- **Move all matching data from the document to the appropriate slots.**

VivoMind Research, LLC, won that contract in a competition among a dozen groups from universities and corporations.

- **On the trial set of documents, VivoMind got 96% of the entries correct.**
- **The second best score was 73%. Most scores were below 50%.**
- **The Vivomind method of information extraction is summarized in the following article: http://www.jfsowa.com/pubs/template.htm**

Information Extracted from Published Documents


DOE BREMS PROJECT.xlsx

	COMPOUND	CURIE TEMP.	SOURCE
2	Mn3[Cr(CN)6]2 ·16H2O	89 K	A solid-state hybrid density functional theory study
3	Sr3Ir2O7 in Sr3Ir2O7 single-cr	~ 280 K	Canted antiferromagnetic ground state in Sr3Ir2O7
4	PrPt2B2C	6 K	Coexistence of superconductivity and magnetic ord
5	La0:3Nd0:7Pt2:1-	2.8 K	Coexistence of superconductivity and magnetic ord
6	NdPt2:1B2:4C1:2	3 K	Coexistence of superconductivity and magnetic ord
7	NdPt1:5Au0:6B2C	3 K	Coexistence of superconductivity and magnetic ord
8	SmNiC2	= 17.7 K	Commensurate charge-density wave with frustrate
9	Co0.2Zn0.8Fe2O4. in CdxCo1-	~ 780 K	Does Ti+4 ratio improve the physical properties of
10	Zn0.88Co0.12O in ZnO	~ 540 K	Effect of Co doping on the structural; optical and m
11	La in Sr2-xLaxFeMoO6	425 K	Effect of La doping on the properties of Sr2-xLaxFe
12	Fe in Sr2-xLaxFeMoO6	~ 1040 K	Effect of La doping on the properties of Sr2-xLaxFe
13	FeSe	~ 305 K	Electronic and magnetic properties of FeSe thin film
14	Ni-Mn-Ga	= 376 K	Electronic and structural properties of ferromagnet
15	LaFexSi1 - x13 in La1-zPrz(Fe	~ 190 K	Enhancement of magnetocaloric effects in La1-zPrz
16	LaFe0.88Si0.1213 in La1-zPrz(	= 195 K	Enhancement of magnetocaloric effects in La1-zPrz
17	Co2MnGa in Co2MnGa	600 K	Ferromagnetic resonance in Co2MnGa films with va
18	HoCrO4 in HoCrO4	17.6 K	Ferromagnetism vs. antiferromagnetism of the dim
19	Mn3(HCOO)6 in Mn3(HCOO)6	8.0 K	Guest-induced chirality in the ferrimagnetic nanopo
20	NaZn13- in La0.5Pr0.5(Fe0.88	range from 195 K to 185 K	Large isothermal magnetic entropy change after hy
21	La2/3Ba1/3MnO3 in La2-3Ba1	range from 300 K to 250 K	Magnetic and neutron diffraction study of La2-3Ba1
22	CuMnSb in Co1-xCuxMnSb	range from 476 K to 300 K	Magnetic properties of half-metallic semi Heusler C
23	Nd2 in Nd2-yDyyFe17-xSix	range from 61.46 °C to 236 °	Magnetic properties of iron-rich Nd2-yDyyFe17-xSix
24	Tb2Fe17 in Nd2-yDyyFe17-xSi	~ 80 °C	Magnetic properties of iron-rich Nd2-yDyyFe17-xSix

Application to Legacy Re-engineering

Analyze the software and documentation of a large corporation.

Programs in daily use, some of which were up to 40 years old.

- **1.5 million lines of COBOL programs.**
- **100 megabytes of English documentation — reports, manuals, e-mails, Lotus Notes, HTML, and program comments.**

Goal:

- **Analyze the COBOL programs.**
- **Analyze the English documentation.**
- **Compare the two to generate:**
  English glossary of all terms with index to the software,
  Structure diagrams of the programs, files, and data,
  List of discrepancies between the programs and documentation.

An Important Simplification

An extremely difficult and still unsolved problem:

- **Translate English specifications to executable programs.**

Much easier task:

- **Translate the COBOL programs to conceptual graphs.**
- **Those CGs provide the ontology and background knowledge.**
- **The CGs derived from English may have ambiguous options.**
- **VAE matches the CGs from English to CGs from COBOL.**
- **The COBOL CGs show the most likely options.**
- **They can also insert missing information or detect errors.**

The CGs derived from COBOL provide a formal semantics for the informal English texts.

Excerpt from the Documentation

The input file that is used to create this piece of the Billing Interface for the General Ledger is an extract from the 61 byte file that is created by the COBOL program BILLCRUA in the Billing History production run. This file is used instead of the history file for time efficiency. This file contains the billing transaction codes (types of records) that are to be interfaced to General Ledger for the given month.

For this process the following transaction codes are used: 32 — loss on unbilled, 72 — gain on uncollected, and 85 — loss on uncollected. Any of these records that are actually taxes are bypassed. Only client types 01 — Mar, 05 — Internal Non/Billable, 06 — Internal Billable, and 08 — BAS are selected. This is determined by a GETBDATA call to the client file.

Note that none of the files or COBOL variables are named.

By matching the graphs derived from English to the graphs derived from COBOL, VAE identified all the file names and COBOL variables involved.

Interpreting Novel Patterns

Many texts contain unusual or ungrammatical patterns.

They may be elliptical forms that could be stored in tables.

But some authors write them as phrases in a sentence:

- ***32 — loss on unbilled***
- ***72 — gain on uncollected***
- ***85 — loss on uncollected***

Intellitex generated a CG with a default relation (Link):

[Number: 32]→(Link)→[Punctuation: "–"]→(Link)→[Loss]→(On)→[Unbilled]

The value 32 was stored as a constant in a COBOL program.

The phrase "loss on unbilled" was written as a comment.

The CGs derived from the COBOL data and comments matched the CGs derived from the English documentation.

Results

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

- **Four weeks for customization:**

  Design, ontology, and additional programming for I/O formats.

- **Three weeks to run VLP + VAE + extensions:**

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

  Weak matches were confirmed or corrected by Majumdar and LeClerc.

- **One week to produce a CD-ROM with the desired results:**

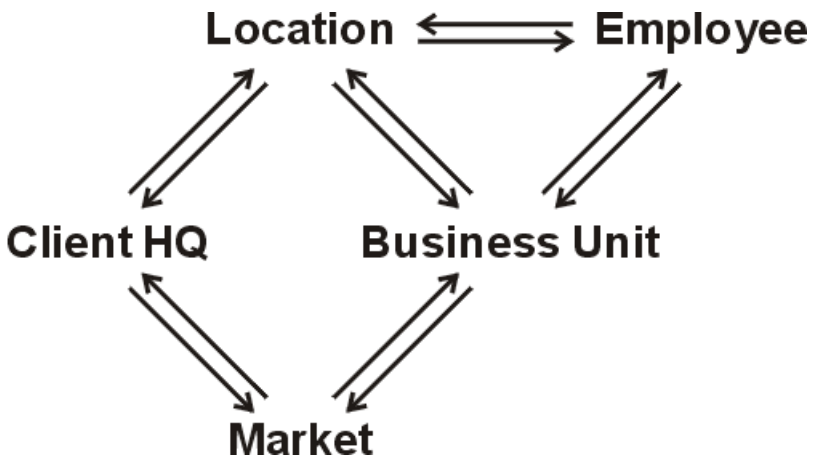
  Glossary, data dictionary, data flow diagrams, process architecture, system context diagrams.

A major consulting firm estimated that the job would take 40 people two years to analyze the documentation and generate the cross references.

With VivoMind software, it was completed in 15 person weeks.

Mismatch Found by VAE

A diagram of relationships among data types in the database:



Question: Which location determines the market?

According to the documentation: Business unit.

According to the COBOL programs: Client HQ.

Management had been making decisions based on incorrect assumptions.

Contradiction Found by VAE

From the ontology used for interpreting English:

- **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.”

Relating Formal and Informal CGs

The legacy-reengineering task required two kinds of processing.

Precise reasoning:

- **Analyzing the COBOL programs and translating them to CGs.**
- **Detecting discrepancies between different programs.**
- **Detecting discrepancies between programs and documentation.**

Indexing and cross references:

- **Creating an index of English terms and names of programs.**
- **Mapping English documents to the files and programs they mention.**

Conceptual graphs derived from COBOL are precise.

But the CGs derived from English are informal and unreliable.

Informal CGs are adequate for cross-references between the English documents and the COBOL programs.

All precise reasoning was performed on CGs from COBOL or on CGs from English that were *corrected* by CGs from COBOL.

Application to Oil and Gas Exploration

Source material:

- **79 documents, ranging in length from 1 page to 50 pages.**
- **Some are reports about oil or gas fields, and others are chapters from a textbook on geology used as background information.**
- **English, as written for human readers (no semantic tagging).**
- **Additional data from relational DBs and other structured sources.**
- **Lexical resources derived from WordNet, CoreLex, IBM-CSLI Verb Ontology, Roget's Thesaurus, and other sources.**
- **An ontology for the oil and gas domain written in controlled English by geologists from the University of Utah.**

Queries:

- **A paragraph that describes a potential oil or gas field.**
- **Analogies compare the query to the documents.**

Answering Questions with VAE

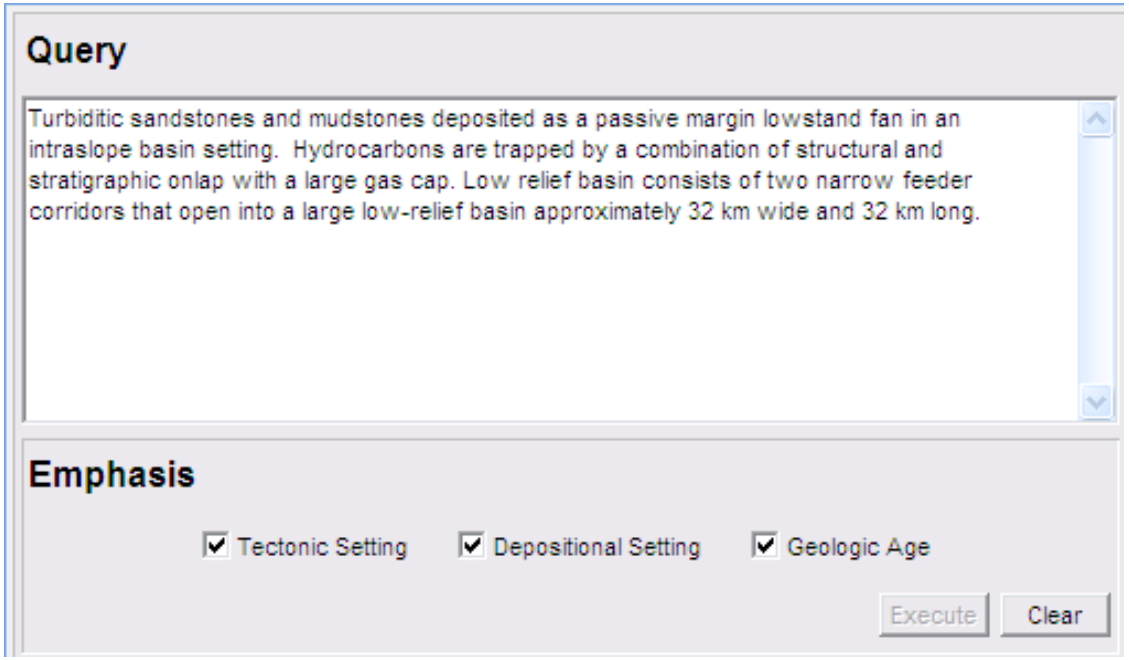
For the sources, either NL documents or structured data:

- **Translate the text or data to conceptual graphs.**
- **Translate all CGs to Cognitive Signatures™ in time proportional to (N log N), where N is the total number of CGs.**
- **Store each Cognitive Signature in Cognitive Memory™ with a pointer to the original CG and the source from which that CG was derived.**
- **Use previously translated CGs to help interpret new sentences.**

For a query stated as an English sentence or paragraph,

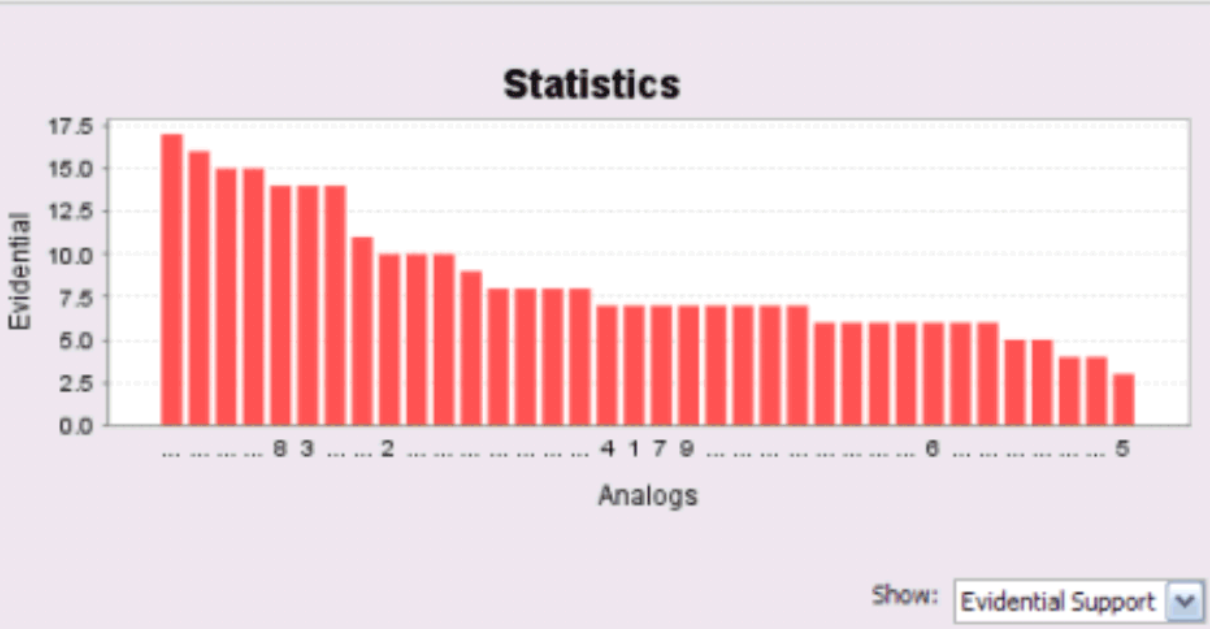
- **Translate the query to conceptual graphs.**
- **Find matching patterns in the source data and rank them in order of semantic distance.**
- **For each match within a given threshold, use structure mapping to verify which parts of the query CG match the source CG.**
- **As answer, return the English sentences or paragraphs in the source document that had the closest match to the query.**

A Query Written by a Geologist



Turbiditic sandstones and mudstones deposited as a passive margin lowstand fan in an intraslope basin setting. Hydrocarbons are trapped by a combination of structural and stratigraphic onlap with a large gas cap. Low relief basin consists of two narrow feeder corridors that open into a large low-relief basin approximately 32 km wide and 32 km long.

GeoMind Results Interface



Show: Evidential Support

Query Results

10) 17%- Vautreuil
23) 16%- Hogsnyta Type II Shelf Margin
25) 15%- Tanqua Karoo Subbasin
36) 15%- des
3) 14%- Espy Ranch, Spine 1, and Rattlesnake Ridge
8) 14%- Songpan-Ganzi Complex
19) 14%- Pukearuhe Beach
31) 11%- Waikiekie South Beach and Inland
2) 10%- Brushy Canyon Outcrop Belt
16) 10%- Atlapexco Road Cut
35) 10%- depocenter
22) 9%- Storvola Type 1 Shelf Margin
21) 8%- Punta Naranio

Sort By: Evidential Support

Query Results: Analog Summary

Source Visualization

```
NAME : Vautreuil
COUNTRY : France
FORMATION : Gres d_Annot Formation (Annot Sandstones)
AGE : Eocene-Oligocene
```

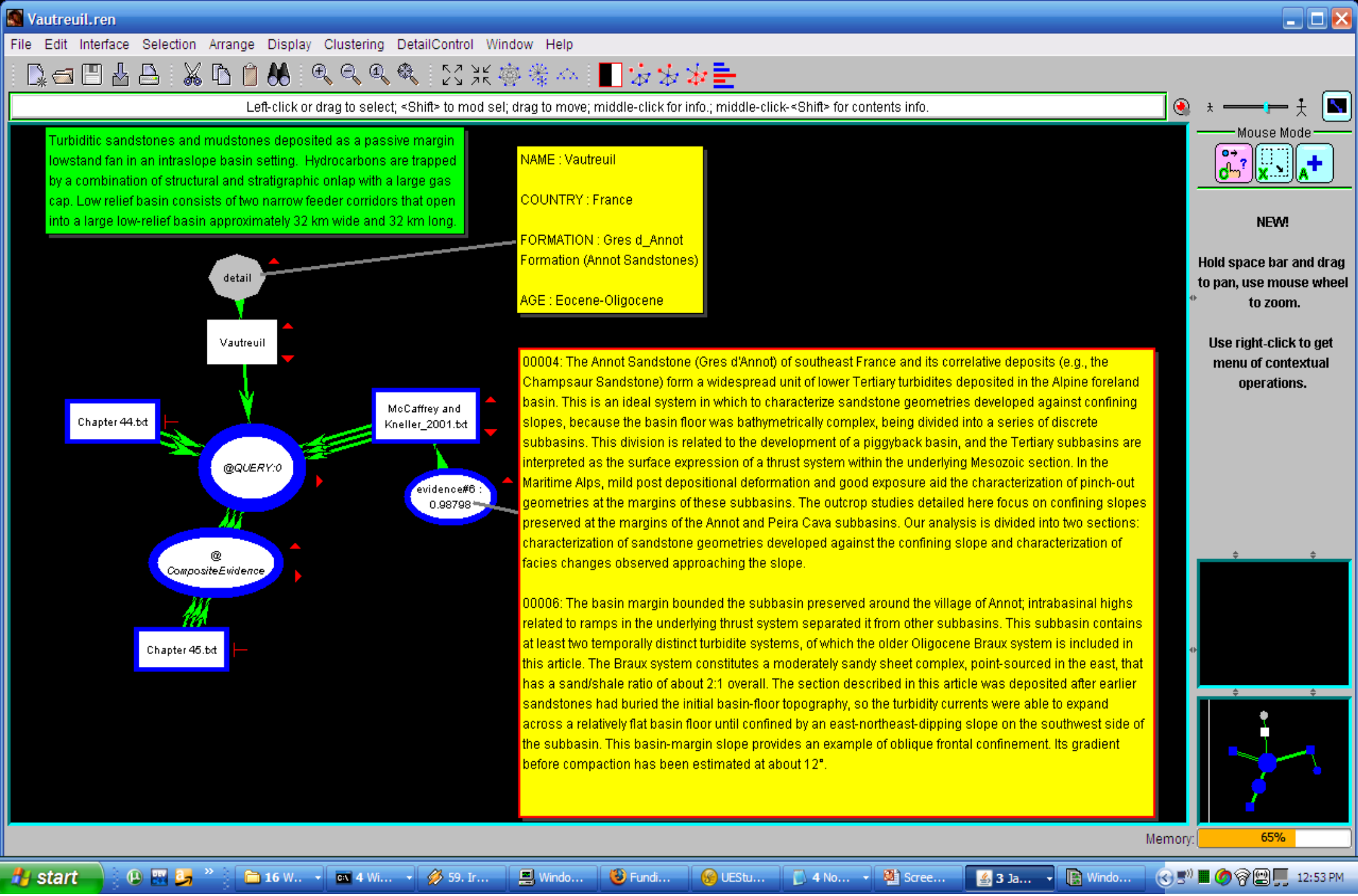
Query Results: Evidence

```
                  vautreuil  chapter 44 lomas, et. al. onlapping
sheet sandstones in the gres d_annot, vautreuil, france   cliffs
forming the east side of the vautreuil de laverq (44?18-n;
valley, west of the foret domaniale 6?29-e) region:
provence-alpes-cote d_azur, departement: alpes-de-haute-provence
france overview montage: 2700 m (8850 ft), detailed panel: 800 m
```

Preferences

```
Emphasis:
          On:  Tectonic Setting
          On:  Depositional Setting
          On:  Geologic Age
Weights:
          On:  Provenance
          On:  Profile
Sources:
          Corperate
                On:  Exploration
                On:  Production
                On:  Financial
          Vendor
                On:  AAPG
                On:  Wood
          External
```

Details of the closest matching hydrocarbon fields



Linking the query to the paragraphs that contain the answer

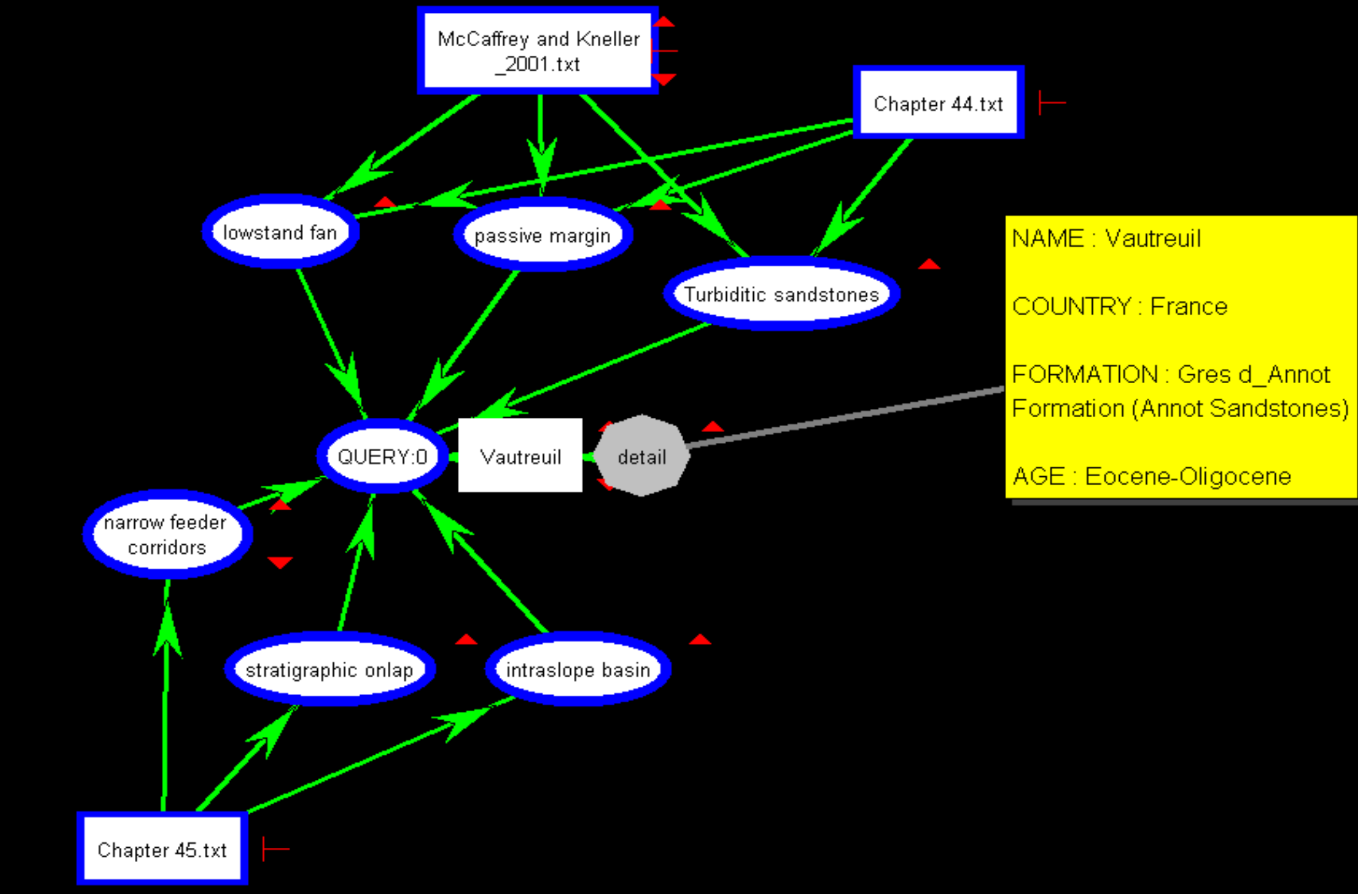
What the Screen Shots Show

Information shown in the previous screen shot:

- The query in the green box describes some oil or gas field.
- The data in the small yellow box describes the Vautreuil field.
- The large yellow box shows the paragraphs in a report by McCarthy and Kneller from which that data was extracted.

The next screen shot shows how the answer was found:

- Many terms in the query were not defined in the ontology: *lowstand fan, passive margin, turbiditiic sandstones, narrow feeder cables, stratigraphic onlap, intraslope basin*.
- VLP generated tentative CGs for these phrases and looked in Cogntive Memory to find similar CGs derived from other sources.
- Chapters 44 and 45 of the textbook on geology contained those CGs as subgraphs of larger graphs that had related information.
- Patterns found in the larger graphs helped relate the CGs derived from the query to CGs derived from the report that had the answer.



Uing background knowledge from a textbook to find the answer

Emergent Knowledge

When reading the 79 documents,

- **VLP translates the sentences and paragraphs to CGs.**
- **But it does not do any further analysis of the documents.**

When a geologist asks a question,

- **The VivoMind system may find related phrases in many sources.**
- **To connect those phrases, it may need to do further searches.**
- **Some sources can be textbooks with background knowledge that helps VLP interpret the research reports.**
- **The result consists of conceptual graphs that relate the query to paragraphs in research reports that contain the answer.**
- **The new CGs can be added to Cognitive Memory for future use.**

By a "Socratic" dialog, the geologist can lead the system to explore novel paths and discover unexpected patterns.

Specializing a Basic Ontology

For VivoMind applications, the basic KR ontology is related to verbal patterns by means of conceptual graphs. *

Additions required for information extraction:

- **Grammar and ontology for chemical compounds.**
- **CGs that specify the patterns to be found in the documents.**

For legacy re-engineering:

- **An ontology that defines a subset of COBOL semantics.**
- **For each COBOL program, a model that consists of the CGs translated from the COBOL code.**

For oil and gas exploration,

- **A simple ontology written in controlled English by geologists.**
- **General CGs derived from a textbook on geology.**
- **More specialized CGs derived from research reports.**

*** For a summary of the KR ontology, see http://www.jfsowa.com/ontology**

The Process of Language Understanding

People relate patterns in language to patterns in mental models.

Simulating exactly what people do is impossible today:

- **Nobody knows the details of how the brain works.**
- **Even with a good theory of the brain, the total amount of detail would overwhelm the fastest supercomputers.**
- **A faithful simulation would also require a detailed model of the body with all its mechanisms of perception, feelings, and action.**

But efficient approximations to human patterns are possible:

- **Graphs can specify good approximations to continuous models.**
- **They can serve as the logical notation for a dynamic model theory.**
- **And they can support a high-speed associative memory.**

This engineering approach is influenced by, but is not identical to the cognitive organization and processing in the human brain.

What is Language Understanding?

Understanding a text in some language does not require a translation to a language of thought or logical form.

Instead, it requires an interpreter, human or robot, to relate the text to his, her, or its context, knowledge, and goals:

- **That process changes the interpreter's background knowledge.**
- **But the kind of change depends critically on the context, goals, and available knowledge.**
- **No two interpreters understand a text in exactly the same way.**
- **With different contexts, goals, or knowledge, the same interpreter may understand a text in different ways.**

The evidence of understanding is an appropriate response to a text by an interpreter in a given situation.

If a robot responds appropriately to a command, does it understand? What if it explains how and why it responded?

Related Readings

Future directions for semantic systems,
http://www.jfsowa.com/pubs/futures.pdf

From existential graphs to conceptual graphs,
http://www.jfsowa.com/pubs/eg2cg.pdf

Role of Logic and Ontology in Language and Reasoning,
http://www.jfsowa.com/pubs/rolelog.pdf

Fads and Fallacies About Logic,
http://www.jfsowa.com/pubs/fflogic.pdf

Conceptual Graphs for Representing Conceptual Structures,
http://www.jfsowa.com/pubs/cg4cs.pdf

Peirce's tutorial on existential graphs,
http://www.jfsowa.com/pubs/egtut.pdf

ISO/IEC standard 24707 for Common Logic,
http://standards.iso.org/ittf/PubliclyAvailableStandards/c039175_ISO_IEC_24707_2007(E).zip

References

For more information about the VivoMind software:

Majumdar, Arun K., John F. Sowa, & John Stewart (2008) Pursuing the goal of language understanding, http://www.jfsowa.com/pubs/pursuing.pdf

Majumdar, Arun K., & John F. Sowa (2009) Two paradigms are better than one and multiple paradigms are even better, http://www.jfsowa.com/pubs/paradigm.pdf

Sowa, John F. (2002) Architectures for intelligent systems, http://www.jfsowa.com/pubs/arch.htm

Sowa, John F., & Arun K. Majumdar (2003) Analogical reasoning, http://www.jfsowa.com/pubs/analog.htm

Sowa, John F. (2003) Laws, facts, and contexts, http://www.jfsowa.com/pubs/laws.htm

Sowa, John F. (2005) The challenge of knowledge soup, http://www.jfsowa.com/pubs/challenge.pdf

Sowa, John F. (2006) Worlds, models, and descriptions, http://www.jfsowa.com/pubs/worlds.pdf

Sowa, John F. (2011) Cognitive architectures for conceptual structures, http://www.jfsowa.com/pubs/ca4cs.pdf

Related references:

Johnson-Laird, Philip N. (2002) Peirce, logic diagrams, and the elementary processes of reasoning, *Thinking and Reasoning* 8:2, 69-95. http://mentalmodels.princeton.edu/papers/2002peirce.pdf

Lamb, Sydney M. (2011) Neurolinguistics, Class Notes for Linguistics 411, Rice University. http://www.owlnet.rice.edu/~ling411

Harrison, Colin James (2000) *PureNet: A modeling program for neurocognitive linguistics*, http://scholarship.rice.edu/bitstream/handle/1911/19501/9969261.PDF

For other references, see the combined bibliography for this site:

http://www.jfsowa.com/bib.htm