Semantic Research using Natural Language Processing at Scale; A continued look behind the scenes of Semantic Insights *Research Assistant™* and *Research Librarian™*

Presented to

The Federal Big Data Working Group Meetup

On 07 June 2014

By Chuck Rehberg, CTO

Semantic Insights™ a Division of Trigent Software

Setting Context for this Exploration

- This brief exploration is a continuation (by request) of my previous introduction of two web-enabled Semantic-based products for "Large Document Corpus[1] Research";
  - Research Assistant™
  - Research Librarian™
- These products are just two of many under development based on the **SIRA** (Semantic Insights Research Assistant) ***Technology***
- Note: These two products are currently in limited Beta Test

[1] by "Document Corpus" we mean any discrete [evolving] online collection of documents

Key Concepts (from SIRA's POV)

1. "World View" represents what you hold to be true or possible. (think "Ontology")
2. "Understanding" is the mapping of an experience to your "World View".
   - This "meaning map" is stored in SIRA as a "memory" of the experience.
3. "Word Sense Disambiguation"
   - In Natural Language text, identifying the appropriate senses of a term, requires a combination of "context analysis" and previous experience…and may be further constrained by one's World View. (e.g. "She is _running_ for the office.").
4. "Semantic Search"
   - SIRA does not require "Semantic Tags". Since what something means can change with the reader, SIRA determines the meaning by "reading" the text each time.
   - SIRA expands the concept of search by identifying overlapping semantics rather than how the information is presented.
5. SIRA "reads" multiple documents very quickly.
   - SIRA turns each research investigation into a speed reading application that implements all the ways the semantics of the investigation can be represented in a given Natural Language.
   - This requires a Dictionary, a Language Specification (**_created by using our Language Lab_**) and a high speed inference engine.
6. "Concept Clusters"
   - Natural Language Text can be thought of as an encryption of concepts and their relationships. A given sentence can encode many Concept Clusters. These Concept Clusters can be represented in many ways. (e.g. using Verbs, Clauses, Phrases…)

Forms of World Knowledge used by SIRA

1. **Language**
   - An encryption of Concepts and Relationships (decrypted using *Meaning Maps*)
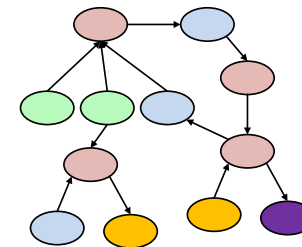   - Equivalent ways of expressing the same meaning are handled using *Equivalent Pattern Sets*
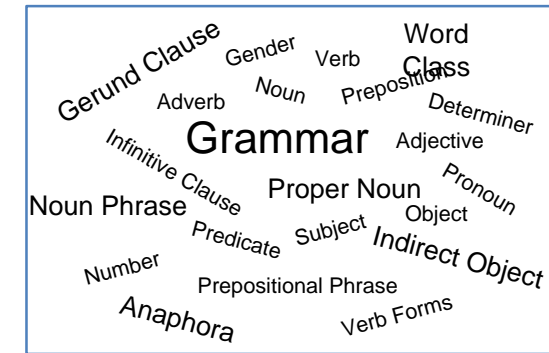
2. **Dictionary**
   - Terms and Senses (+ *Linguistic Metadata*) and Synonymy…
   - Domains (knowledge disciplines)

3. **Ontology**
   - A "World View" in terms of Concepts, Relationships,…

4. **Implication Patterns**
   - Allow for "Do what I want, not what I said"
   - Apply knowledge from experience and expertise to improve results

Gerund Clause   Gender   Verb   Word Class
Adverb   Noun   Preposition   Determiner
Infinitive Clause   **Grammar**   Adjective
Noun Phrase   Proper Noun   Pronoun
Predicate   Subject   Object
Number   Indirect Object
Prepositional Phrase
Anaphora   Verb Forms

DICTIONARY

Sample requirement for implication pattern: For companies to be similar, they must have nearly (+/- 10%) the same; market cap, age, and annual revenue.

Forms of Knowledge acquired by SIRA over time

1. Memory of Experiences
   - SIRA (like people?), stores experiences (e.g. reading) together with what that experience meant at the time of the experience.
   - This memory is not automatically integrated into the ontology. A process of reflection can be initiated to update the Ontology.

2. "Emergent Categories" (EC)
   - Each time SIRA experiences a "Concept Cluster", the EC database may be updated to reflect its similarity with previous experiences. This can then be used to improve "Word Sense Disambiguation" and meaning analysis. [not available in current products]

Who we are

- We are:

    SEMANTIC INSIGHTS™

    – Semantic Insights™ is the R&D division of Trigent Software, Inc. www.trigent.com

    – We focus on developing semantics-based information products that produce high-value results serving the needs of general users requiring little or no special training.

    – Visit us at www.semanticinsights.com

Chuck Rehberg

As CTO at Trigent Software and Chief Scientist at Semantic Insights, Chuck Rehberg has developed patented high performance rules engine technology and advanced natural language processing technologies that empower a new generation of semantic research solutions.

Chuck has more than thirty years in the high-tech industry, developing leading-edge solutions in the areas of Artificial Intelligence, Semantic Technologies, analysis and large –scale configuration software.