Information about SIRA for TIBCO

Chuck Rehberg

CTO, Trigent Software

Chief Scientist, Semantic Insights

07Dec11

Context

- ***Semantic Insights Research Assistant*** (SIRA) is tool suite and technology automating knowledge-intensive research tasks
- It does this by:
    1. Understanding natural language,
    2. Knowing your subject,
    3. Quickly sifting through a vast amount of information, and
    4. Producing reports containing only the relevant information including hyperlinks and bibliography

SIRA goes beyond Search to Research!

SIRA deals in both Natural Language and Meaning

- Simply put; Natural Language is a powerful system for representing concepts (meaning)
- These concepts can be expressed in many ways:
    1. Single or multi-word terms,
    2. Phrases or groupings of phrases,
    3. The position of lexical items,
    4. The choice of combinations of terms, and
    5. What is missing or what is implied within a given context.
    6. References to other text

Reference Concepts and Concept Clusters

- Concepts can be abstract or physical.
- Concepts can be “simple” (i.e. a Reference Concept) or can be built up by relationships between a number of concepts (i.e. a Concept Cluster).
- ***Reference Concepts*** are usually expressed as individual words/terms.
- ***Concept Clusters*** are usually expressed using relationships between concepts.
- Many kinds of relationships bind concepts in a Concept Cluster:
  1. generalization/specialization (kinds)
  2. part-whole (composition)
  3. conceptual relationships (e.g. “is married to”)
  4. categorization (e.g. “songs I like”, “events that happened to me on Tuesday”,…).

More than a collection of terms…

- *Understanding what is going on in natural language text requires more than recognizing the meaning of individual words and more than just identifying the presence of Reference Concepts. You need robust handling of interconnected Concept Clusters.*
- ***SIRA focuses on Concept Clusters***
- But, why hasn't Google done this?
- There is the technical challenge: Many ways of expressing the same Concept Cluster using natural language:
  1. Some of ways bare little or no resemblance to each other.
  2. They may not share the same terms or same structure.
  3. They may not even use the same synonyms of a term.
  4. A given Concept Cluster may be spread across a number of sentences or even documents.
- ***But more importantly; Google sells eyes! SIRA sells time and expertise!***

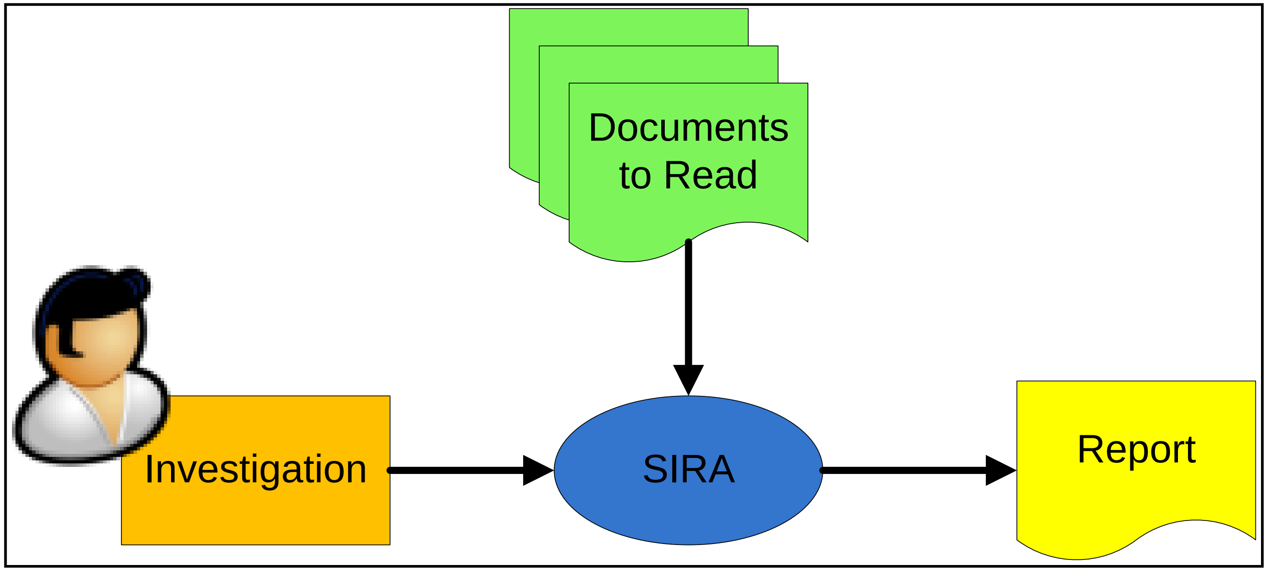
Technologies Applied

- Natural Language Processing (NLP)
- Word Sense Disambiguation (WSD)
- Ontologies and Dictionaries
- Rules Based Systems
- Machine Learning

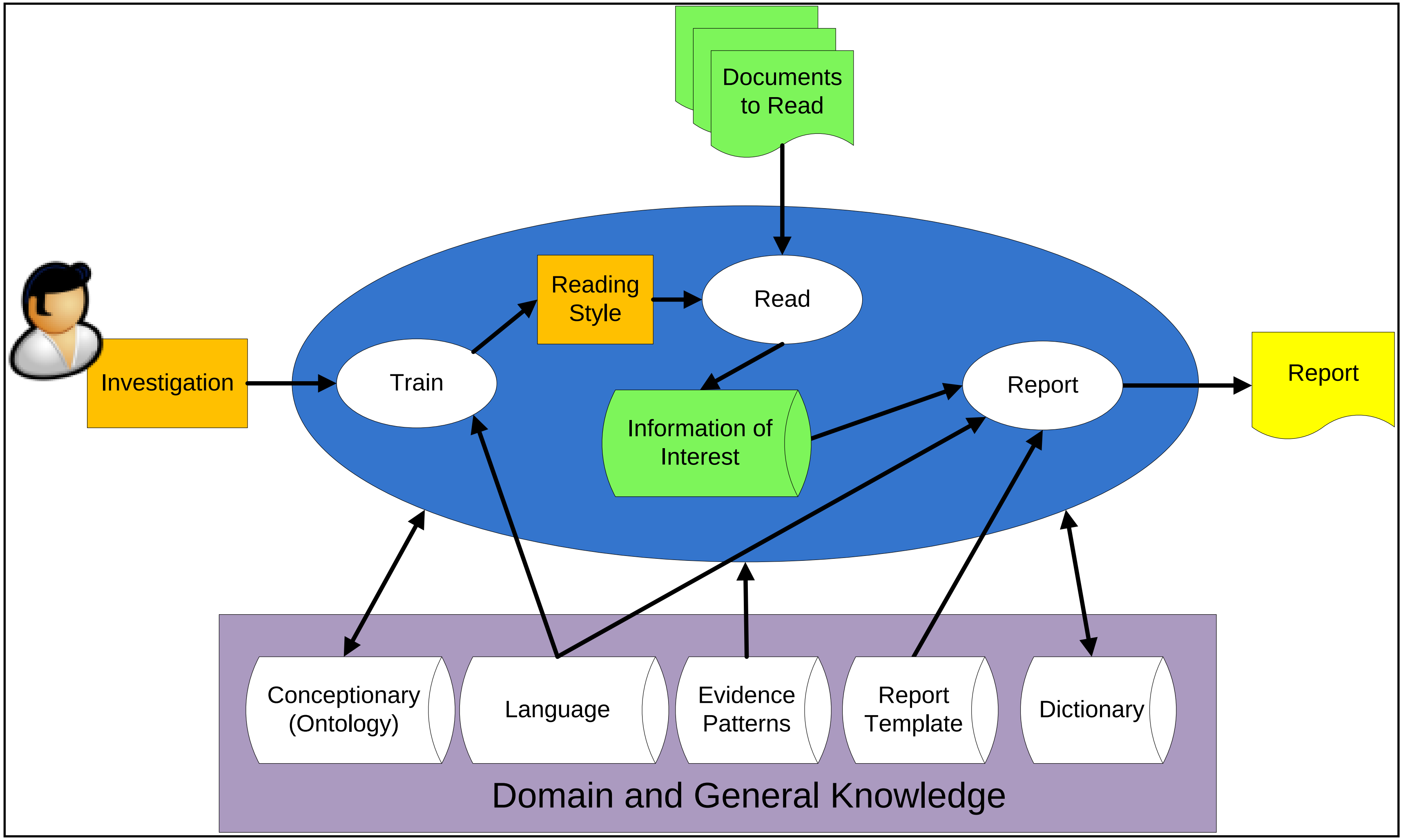
Behind the Scenes: The SIRA approach step-by-step

1. **Given a statement of investigation (questions, statements, or whole documents) in a given language:**
   1. SIRA derives one or more "valid" Concept Clusters. *This process uses patent pending NLP and WSD algorithms. [includes WSD]*
   2. SIRA then generates all the ways the Concept Cluster can be expressed in a given language. *This process is patent pending.*
   3. SIRA then generates a high-speed "Reader" application capable of identifying each sentence containing any fragment of the Concept Cluster. *This process is a trade secret.*
2. **Select a corpus of documents**
   1. SIRA executes the Reader application to process each document in the corpus [optionally including the hyperlinks] to identify each sentence containing any fragment of the Concept Cluster. *This uses our patented rules engine and trade secrets.*
   2. SIRA verifies each identified sentence both linguistically and semantically for valid Concept Cluster match. *This process is patent pending.*
3. **Select a report template**
   1. SIRA generates a written report presenting only the information of interest culled from the documents read. *This process is patented.*
4. **Examine the Research Report**

The basics (normal users)

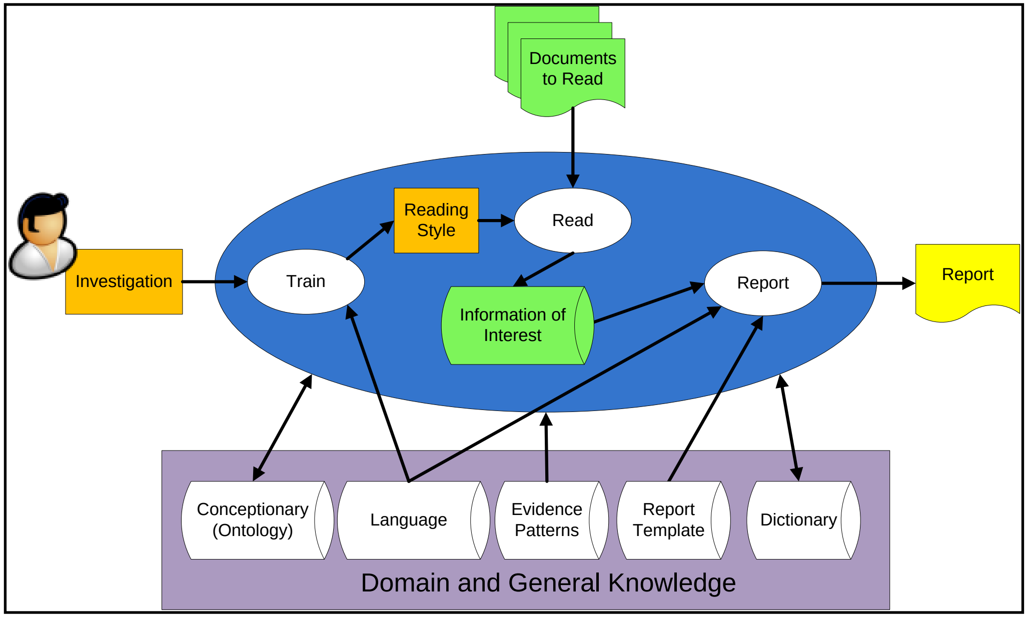


Behind the scenes: Information flow



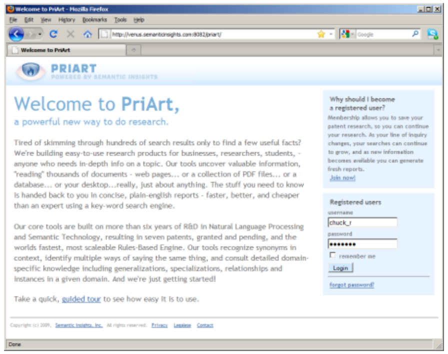
Deployment Cost-Time-Effort Considerations

1. The [domain] knowledge required
   a. The ontology
   b. The dictionary
   c. Evidence patterns
2. The Corpus researched
3. The interface you will use
4. The end user usage profile
5. Any special language requirements



SIRA Tool Suite

- PriArt: A Semantic Research Assistant (Web)
  - Examples: Autism report, high school assignment, homeland security, patent infringement
- SIRA Development Center* (Desktop)
  - Used to develop and manage World Knowledge
    - Ontologies, Dictionaries, Testing and Training
  - Live Demonstration of Ontology and Dictionary creation and curation
- Language Lab* (Web)
  - Used to define Language and Genre
  - Syntax, Grammar and Meaning Maps


Welcome to PriArt - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://venus.semanticinsights.com:8082/priart/
Google
Welcome to PriArt
PRIART
POWERED BY SEMANTIC INSIGHTS
Done

Welcome to PriArt,
a powerful new way to do research.

Tired of skimming through hundreds of search results only to find a few useful facts? We're building easy-to-use research products for businesses, researchers, students, - anyone who needs in-depth info on a topic. Our tools uncover valuable information, "reading" thousands of documents - web pages... or a collection of PDF files... or a database... or your desktop...really, just about anything. The stuff you need to know is handed back to you in concise, plain-english reports - faster, better, and cheaper than an expert using a key-word search engine.

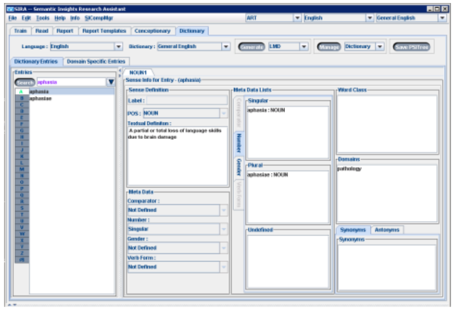
Our core tools are built on more than six years of R&D in Natural Language Processing and Semantic Technology, resulting in seven patents, granted and pending, and the worlds fastest, most scaleable Rules-Based Engine. Our tools recognize synonyms in context, identify multiple ways of saying the same thing, and consult detailed domain-specific knowledge including generalizations, specializations, relationships and instances in a given domain. And we're just getting started!

Take a quick, guided tour to see how easy it is to use.

Why should I become a registered user?
Membership allows you to save your patent research, so you can continue your research. As your line of inquiry changes, your searches can continue to grow, and as new information becomes available you can generate fresh reports.
Join now!

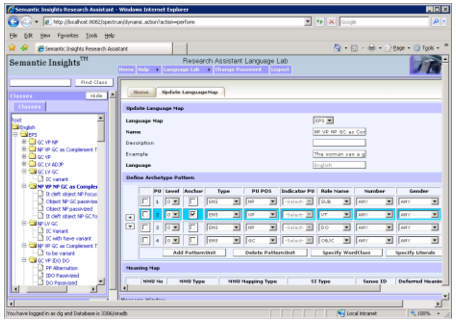
Registered users
username: chuck_r
password: •••••••
□ remember me
Login
forgot password?

Copyright (c) 2009, Semantic Insights, Inc. All rights reserved. Privacy Legalese Contact



Language : English
Dictionary : General English
Search: aphasia
aphasia
aphasiae

Sense Info for Entry - (aphasia)
Label :
POS : NOUN
Textual Definition : A partial or total loss of language skills due to brain damage
Meta Data
Comparator : Not Defined
Number : Singular
Gender : Not Defined
Verb Form : Not Defined
Singular: aphasia : NOUN
Plural: aphasiae : NOUN
Undefined:
Domains: pathology
Synonyms:


Semantic Insights Research Assistant - Windows Internet Explorer
http://localhost:8082/spectrum/dynamic.action?action=perform
Google
File Edit View Favorites Tools Help
Semantic Insights Research Assistant
Page
Tools
Semantic Insights™
Research Assistant Language Lab
Home
Help
Language Lab
Change Password
Logout
Find Class
Classes
Hide
Classes
Home
Update LanguageMap
Add PatternUnit
Delete PatternUnit
Specify WordClass
Specify Literals
Local intranet
100%

Root
English
EPS
GC VP NP
NP VP GC as Complement T
GC VP
GC LV ADJP
GC LV GC
IC variant
NP VP NP GC as Complem
It cleft object NP focus
Object NP GC passivize
Object NP passivized
It cleft object NP GC fo
NP LV GC
IC Variant
IC with have variant
NP VP GC as Complement T
to be variant
GC VP IDO DO
PP Alternation
IDO Passivized
DO Passivized

Update Language Map
Language Map: EPS
Name: NP VP NP GC as Cor
Description:
Example: The woman saw a g
Language: English

Define Archetype Pattern

	PU	Level	Anchor	Type	PU POS	Indicator PU	Role Name	Number	Gender
□	1	0	□	ENS	NP	-Select-	SUB	ANY	ANY
□	2	0	☑	ENS	VP	-Select-	VT	ANY	ANY
□	3	0	□	ENS	NP	-Select-	DO	ANY	ANY
□	4	0	□	ENS	GC	-Select-	OBJC	ANY	ANY

Meaning Map

	NMU No	NMU Type	NMU Mapping Type	SI Type	Sense ID	Deferred Meanin

You have logged in as dg and Database is 3306/siradb

Word Sense Disambiguation (WSD)

Included here to normalize the definition…

- In natural language processing, word sense disambiguation (WSD) is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people.
- WSD is a natural classification problem: Given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into one or more of its sense classes.
- The features of the context (such as neighboring words) provide the evidence for classification.

http://www.scholarpedia.org/article/Word_sense_disambiguation

About SIRA’s Approach to WSD

Patent Application No. 20100063796 “Word Sense Disambiguation Using Emergent Categories” – Charles Rehberg, et al., Published March 11, 2010

1. Based on the dynamic evolution of relationship categories
   - An example of machine learning; improved over time; can be pre-trained
2. Relies on linguistic metadata in Dictionary
   - Note: dynamically grows the Dictionary when unknowns are encountered
3. Does not use a statistical Parts-of-speech tagger
   - Avoids the determinant error in trained POS taggers
4. Does not use mathematical similarity algorithms

SEMANTIC INSIGHTS™

- Who we are:
  - 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 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 twenty five years in the high-tech industry, developing leading-edge solutions in the areas of Artificial Intelligence, Semantic Technologies , analysis and large –scale configuration software.
- chuck_r@trigent.com or chuck_r@semanticinsights.com