Putting the Semantic In the Semantic Web

An overview of UIMA and its role in Accelerating the Semantic Revolution

Ontolog Forum, May 11, 2006

David A. Ferrucci
Senior Manager, Semantic Analysis and Integration
Chief Architect, UIMA
IBM T.J. Watson Research Center
Opening Observations

- **Semantics are key to Knowledge Discovery in Unstructured Sources**
  - Unstructured sources is where the action is
  - Keywords and Known-Item Search don’t cut it
  - Need Richer Semantics

- **Massive manual semantic annotation is not likely**
  - A challenge for the Semantic Web vision
  - User View vs. Authors view
  - Automated Annotation is Essential
  - Many Component Annotators must emerge and combine

- **The right enabler and the right expectations**
  - Must enable rapid annotator development and integration
  - Applications can add value while tolerating less that perfect precision
  - Automation is necessary to tag and help humans reduce/focus content
Overview

- Semantics Matters – A Motivating Example
- UIMA: Facilitating Automatic Semantic Discovery
- Using Semantics to Raise the Search Bar
  - Finding the Sweet-Spot
  - Semantic Search and the SAW
  - Knowledge Extraction
Semantic Match vs. Key Word Match

SEARCH: Going rate for leasing a billboard near Triborough Bridge

Top hits from popular search engines miss the mark...

*Keywords may Match*

*But Wrong Content returned*

*And right content MISSED*
SEARCH: Going rate for leasing a billboard near Triborough Bridge

Wired New York Forum - Long Island City Development
"it's Right over The 59th Street Bridge." ... it's going to become our SoHo. The Powerhouse is most likely going to become a performing Arts facility for ...

Long Island City Development [Archive] - Wired New York Forum
... Creek to the Triborough Bridge - would become an attractive destination. ... Several Long Island City sites are now in play - Junger himself's going ...
www.wirednewyork.com/forum/archive/index.php/t-5160.html - 255k - Cached - Similar pages
[ More results from www.wirednewyork.com ]

[PDF] Austin's Wireless Future
File Format: PDF/Adobe Acrobat - View as HTML
In Section 2 we explore the near term future of wireless -- ubiquitous mobile ... applications such as the Triborough Bridge installation, wireless sensors ...
www.ic2.org/publications/AustinsWirelessFuture.pdf - Similar pages

[PDF] Funding of the High-Speed Rail System
File Format: PDF/Adobe Acrobat - View as HTML
Triborough Bridge and Tunnel Authority** Oklahoma, Turnpike Authority ... management fees, reserves for going T's, and leasing commissions. ...
www.ca.highspeedrail.ca.gov/plans/pdf/Financial_Plan.pdf - Similar pages

The rail magazine for The computer age ...
Surprisingly, thanks to 9600 baud modems, lower rates, and automated access ... his previous positions included President, Triborough Bridge and tunnel ...
www.railfan.net/railpix/techbro_1-5.htm - Supplemental Result - Similar pages

Checklist 1989
Triborough Bridge and Tunnel Authority, selected procurement and contract ... Distribution of high school graduates and college going rate, New York State. ...
www.nysl.nysed.gov/eds/documents/education/check89.htm - 513k - Cached - Similar pages

[PDF] www.icc.utexas.edu/publications/AustinsWirelessFuture...
File Format: PDF/Adobe Acrobat - View as HTML
Supplemental Result - Similar pages
Going rate for leasing a billboard near Triborough Bridge

* Were you looking for *Going rate for leaving a billboard near Triborough Bridge*

**Web Results**

Page 1 of 4 results containing *Going rate for leasing a billboard near Triborough Bridge* (0.70 seconds)

**Wired New York Forum - Long Island City Development**

… on Vernon Boulevard *near* the Queensboro Bridge, the founders of Silvercup … Newtown Creek, the Triborough Bridge - would become an attractive … Pearson St. - where he’s *going* to do a 20-story …


**FINANCIAL PLAN**

*Triborough Bridge* and Tunnel Authority** Oklahoma Turnpike Authority (Average Annual Growth Rate) 1979-1982 10.9% 1980-1990 5 … Leveraged *Leasing* of Rail Rolling Stock When rail rolling …

www.ca.highspeedrail.ca.gov/plane/pdf/Financial_Plan.pct  Cached page  PDF file

**NEW YORK: Stadiums/Arenas (NETS, JETS, YANKS, and others) [Archive ...**

… backward from the adjacent Triborough Bridge elevated roadway that leads to … I was *going* to the name of one of my companies on it to get … The area *near* the Continental Arena has become construction …

forum.skyscraperpage.com/archive/index.php/t-85433.html  Cached page

http://odur.let.rug.nl/ftp/pub/prolog-app/DutchBrillTagging/TaggerSoftware/Bin_and_Data/LEXICON.WSJ

Billboard NNP Billboarding NNP Billed VBN Billerica NNP Billheimer NNP Billiards NNP NN Billie NNP Billing NN Billings NNS NNP Billion CD Billionaire NN Billions NNS Billmeyer NNP Billock NNP Bills NNS …

odur.let.rug.nl/ftp/pub/prolog-app/DutchBrillTagging/TaggerSoftware/Bin_and_Data/LEXICON.W...  Cached page

**Another Search Engine**

*Some Different* Hits  but  *Still All Misses*

**Makes you wonder...**

**Could be out there?**
Semantic Match Improves Recall

SEARCH: Going rate for leasing a billboard near Triborough Bridge

No Keywords in Common
But a good “hit”

“…We were offered $250,000/year in 2001 for an outdoor sign in Hunts Point overlooking the Bruckner expressway. …”
Semantic Match Improves Precision

SEARCH: Going rate for leasing a billboard near Triborough Bridge

Common Keywords
Bad Semantic Match

“…Simon and Garfunkel's "The 59th Street Bridge Song" was rated highly by the Billboard magazine in the 60's…”
UIMA Overview
Unstructured Information Management Architecture

A open-source, pluggable framework for accelerating the science and business of semantic analysis
Analytics Bridge the Unstructured & Structured Worlds

Unstructured Information

- Text, Chat, Email, Audio, Video
- High-Value
- Most Current
- Fastest Growing (80% of Corporate Data)
...BUT...
- Buried in Huge Volumes (Noise)
- Implicit Semantics
- Inefficient Search

Structured Information

- Indices
- DBs
- KBs
- Explicit Semantics
- Efficient Search
- Focused Content
...BUT...
- Slow Growing
- Narrow Coverage
- Less Current/Relevant

Text and Multi-Modal Analytics

Discover Relevant Semantics → Build into Structure

- Docs, Emails, Phone Calls, Reports
- Topics, Entities, Relationships
- People, Places, Org, Times, Events
- Customer Opinions, Products, Problems
- Threats, Chemicals, Drugs, Drug Interactions....
Report Date 14 April, 2003. From an interrogation of a cooperative detainee in Guantanamo, detainee says he studied regularly with a man named William Davis at the Harvard Business School, Cambridge, MA in 1994. From a captured laptop computer in Bermuda it is learned that William Davis holds a Canadian passport in the name David Miller. INS check reveals that a David Miller, from Canada, entered the USA on a travel visa in January of 2003 stating that he would be visiting a person named Clark Webster in Richmond, VA. The contact address given by Miller was 1631 Capitol Ave., Richmond VA phone number: 804-759-6302.
Semantic Analytics: The Promise and the Challenge

- Independently developed
- From an increasing # of sources

Analysis Capabilities
- Language, Speaker Identifiers
- Tokenizers
- Part of Speech Detectors
- Modality
- Human Language
- Domain of Interest
- Relationship Detectors
- Classifiers ...

- Different technologies & interfaces
- Highly specialized & fine grained

Capability Specializations
- Performance/Precision Tradeoffs...

The right analysis for the job will likely be a **best-of-breed combination** integrating across many dimensions.
Over 200 IBM researchers across 6 WW labs with UIM as a Strategic Focus

Almaden

Watson

Zurich

Beijing

Austin

Haifa

Delhi

Tokyo
UIMA: Unstructured Information Management Architecture

The Architecture

- **Architecture** for composing analytics that extract knowledge from unstructured sources & integrate results with structured information
  - Interfaces, Data Representation Schemes, Design Patterns

- **Principal Architectural Commitments**
  - Common representation scheme
  - Common component engine interfaces (task and domain-independent)
  - Common component metadata
  - Pluggable Workflow and Transports
  - Embeddable (can be layered on top of system middleware)

- **Independent of but interoperable with application specific**
  - Data models
  - Algorithms
  - Language-level or domain-level concepts or tools
  - Workflows or workflow engines
  - Transports (e.g., SOAP, Vinci)
  - Back-end Systems (DB, Search Engine, KB Interfaces)
UIMA: The Software

- **Supports UIMA-compliant development, composition & deployment**
- **Java Framework Implementation**
  - Components in other languages (e.g., C++, PERL, PYTHON)
  - Support for co-located and service-oriented deployments
    - Just change XML Descriptor
    - Balances performance and flexibility
    - Support for C++ co-located Aggregates for high-performance communication
  - High-Performances APIs to common data representation (CAS)
- **Where to Get It**
  - Software Development Kit (SDK) Freely Available from IBM alphaWorks
    - Stand-alone Java Install
    - C++ Enablement Layer
    - Tutorial and Development-Level Utilities and Tooling
    - “Semantic Search” Engine and Indexer
  - Core Software Framework **Open-Source on Source Forge**
  - Emerging Component Community at CMU
    - [http://uima.lti.cs.cmu.edu/index.html](http://uima.lti.cs.cmu.edu/index.html)
Common Infrastructure
Development, Composition, and Deployment

Development Tools
Component Repository
UIMA Framework

UIMA Analysis Components

Large-Scale Distributed Deployment
Standalone
Product Embeddings

Across various types of deployments
Architecture Highlights
Common Representations, Interfaces and Design Patterns

How it works
UIMA’s Basic Building Blocks are **Annotators**. They iterate over an artifact to discover new types based on existing ones and update the **Common Analysis Structure (CAS)** for upstream processing.

**Common Analysis Structure (CAS)**

- **Artifact (e.g., Document)**
- **Parser**
- **Named Entity**
- **Relationship**
- **Person**
- **CeoOf**
- **Arg1:Person**
- **Arg2:Org**
- **Organization**
- **NP**
- **VP**
- **PP**

**Analysis Results (i.e., Metadata)**

Fred is the CEO of Center of Center to Micros
Sample Type System

- Top
  - String
  - Int
  - Annotation Begin: int
    - End: int
  - ... (placeholder)

- Entity Annotation
  - Begin: int
  - End: int
  - Person
    - Gov Official
    - Location

- Relation Annotation
  - Arg1: Entity Ant.
  - Arg2: Entity Ant.

- Gov Title
  - Located In
    - Arg1: Entity Ant.
    - Arg2: Location
  - ... (placeholder)

- Token

- Gram. Struc.
  - NP
  - VP
  - PP
Partial HUTT Type System (254 concepts in total)
Analyzed by different Analysis Engines
Semantic Entities & Relations detected
Highlighted in GUI

Analyzed by a combination of Analysis Engines
Semantic Entities & Relations Represented
Highlighted here in a GUI
Fred Center is the CEO of Center Micros. He is a graduate of State University.
The Basic UIMA Component Interfaces

Component Descriptor

- Declarative Description (XML)
- Identification
- Type System, Capabilities
- Configuration Parameters etc.

CAS

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Analysis (Metadata)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type System</td>
<td>Index Repository</td>
</tr>
</tbody>
</table>

- Provides access to Artifact and the current analysis (metadata)
- JCAS Interface presents native Java model of CAS

Annotator (e.g. Name-Entity Detector)

Analysis Algorithm

- Analysis Code goes Here
- Interface: CAS in/CAS Out

UIIMA Context

- Manages access to shared external resources
- Allowing local APIs to access resources shared by cooperating engines
The UIMA framework allows the developer to compose & encapsulate a set of engines as a single aggregate component, insulating applications or users from its implementation details/complexities – Promoting Reuse.

Aggregate Analysis Engine: Named-Entity Detector

Aggregate Analysis Engine: Relation Detector
Collection Processing Engines: Aggregate Analytic Components From “Source to Sink”
User Definable Workflow (3Q 2006)
CPE for Keyword Search

Collection Processing Engine

SimpleToken AndSentence Annotator

Collection Reader
File Reader
Analysis Aggregate
Analysis Engines
KW Search Indexer
KW Search Index
XcasCas Writer
Local File System
CAS Consumers

Collection of Text Docs
CPE for Semantic Search

Collection Processing Engine

Collection of Text Docs

SimpleToken AndSentence Annotator

Named-Entity Annotator

File Reader

Collection Reader

Analysis Aggregate

Analysis Engines

XcasCas Writer

CAS Consumers

Semantic Search Indexer

Semantic Search Index

Local File System
We index and search over tokens AND the semantics annotations from UIMA

“first” is an ambiguous term so is “center”

We are looking for these terms with particular semantics possibly detected by the UIMA analytics

The JuruXML Query Language Exploits the results of Analysis:

KeyWord Query: “first”
Semantic Search Query: <organization> first </organization>
Semantic Search Query: <ceo_of> <person> Center </person> </ceo_of>

“first” as it appears in the name of an organization

“Center” appearing in a name of a person who is the ceo of something
Raising the Search *Bar*

From matching keywords to gathering knowledge
From Finding Documents to Gathering Knowledge

**Known-Item Search**
- **Problem**
  - Find the items (e.g., documents) that individually contain some combination of keywords
- **Operating Assumptions**
  - User assumed that exactly what he/she wants is there in a document
  - There is one specific item out there that best solves information need
  - Native artifact boundaries are rigid
  - Contents = Used Words
  - User provides contained key words or phrases
- **Magic: Ranking**

**Knowledge Gathering and Synthesis**
- **Problem**
  - Produce an *efficient* collection of knowledge that is relevant to information need
- **Operating Assumptions**
  - Multiple items at varying grains together will solve information need
  - Artifact boundaries are permeable
  - Contents ~ Knowledge
  - User provides a conceptual space of interest (possibly using words or phrases)
- **Magic: Semantic Matching**
Raising the Search Bar

- **Exploit Semantic Analysis**
  - Tolerate less-than-perfect analytics
  - Use to Eliminate Noise and Focus Content
  - Degrade gracefully toward key-word search

- **The more in the query the better the results should be**
  - Web-Search conditions users to ~ 2 key words
  - Recall drops off dramatically
  - The more users type the less they know about the desired artifact & we punish them for it
  - Encourage the user to fully describe their information need in their terms

- **Predictive Query Interpretation and Refinement**
  - Should be clear on how to modify a query to direct results
  - Throttle Precision and Recall

- **Cross Artifact Boundaries**
  - The best solution is not always the single-based document

- **Identify the best Grain (relates to Density)**
  - Document, Passage, Entity, Fact, Set of...
Finding the Sweet-Spot

Using an Ontology and an interactive dialog to elicit priorities and select or focus the content
The QA-Cube

- Produces conceptual dimensions present in a scenario
  - Each represented by a concept taxonomy

- Interacts with user to elicit relative priorities
  - Where the user needs to be most specific and
  - Where he is willing to relax the requirements

- Evaluate or Generate Focused Corpora
  - Less Noise, More Relevant
  - Worthy of Deeper Analysis/Exploration
Example Base Ontology

Location
  is-a Geo
    is-a Continent
      part-of Country
        part-of State/Province
          part-of City

Agent
  is-a Org
    is-a Terrorist Org
      performs Activity

Plant
Scenario Ontology Expansion

“...explore the prevention of opium growing in Afghanistan...”

1. NLP on scenario text (e.g., Topic, NE Detection)

2. Identify and query ontologies & KBs
   - Find relevant concepts/relations
   - Integrate with Scenario Ontology accordingly
Selects key dimensions in QA Cube

- **Location**
  - Geo
    - Continent
      - Country
        - State/Province
          - Afghanistan
          - City
  - Continent
  - Geo
- **Agent**
  - Org
    - Terrorist Org
- **Activity**
- **Plant**
  - Drug
    - Narcotic
- **Opium**

- Immediate feedback on effectiveness of conceptual query

- Hits: 0
- Clusters: 0
Find the Sweet Spot

Location
  is-a Geo
  is-a Continent
  part-of Country
  part-of State/Province
  part-of City

Agent
  is-a Org
  part-of Terrorist Org

Activity

Plant

Drug

Terrorist Org

Opium

Narcotic

Immediate feedback on effectiveness of conceptual query

Hits: 300
Clusters: 10
Semantic Search and Analysis Workbench

Exploiting Task Ontologies & Automatically Discovered Semantics to Improve Search Quality
Semantic Search

Broadly Defined…

An extension of search that exploits semantic analysis to improve precision, recall, and density.
The SAW: A Tool for Knowledge Gathering and Synthesis

- Beyond Keyword Search
- Ontologies
- Automatically Discovered Semantics
- Throttle Precision and Recall
- Toward Knowledge Gathering and Synthesis
The Semantic Search 4-Step Program

1. **Corpus**
   - Detect the Semantic Content in the Corpus
   - Build the *Semantic Signatures*

2. **UIMA Corpus-Analysis**
   - Index the text and the *Semantic Signatures*

3. **Semantic Signatures in Query**
   - Detect Semantic Signatures in Query
   - Automatically generate Semantic Search queries For back-end

4. **User Query**
   - UIMA Query-Analysis
   - Semantic Search Engine (e.g., OmniFind)
   - SIAPI: Efficiently retrieve documents matching Semantic Signature

**The Semantic Search 4-Step Program**
Intelligent Passage Selection
- Based on semantic processing
- Hones in on relevant sections
- Redefining the right grain

Task Ontology
- Used as basis for interpreting query
- And for viewing/selecting relevant knowledge

Query Analysis
- Generates Semantic Search Query
- Based on UIMA analytics

User Types whatever
- Keywords, Phrases, Questions

500+ documents | 500+ passages

Intelligent Passage Selection
- Based on semantic processing
- Hones in on relevant sections
- Redefining the right grain
• Document detail
• Reveals task-relevant concepts and relations
• Easy to zoom in on key content based on semantic analysis
• Can highlight answers, concepts, passages or facts
• Can also track source of annotation

49 documents | 88 passages
• To further focus
• Query analysis exploits relations
• 49 results (rather than 500) use highly relevant relation
• Can combine Precision-Oriented and Recall-Oriented
• User want to get more specific
• Based task-model gets options for “restricting” selected concepts
• Goes from Weapon to Chemical Weapon
• Zoomed in on single highly relevant document
What weapons does Iraq produce?

Internet Query

- ProducesWeapon
  - ChemicalWeapon
    - Weapon

Germany Says It Seeks To Curb Its Arms

Exporters

Abstract: Germany has decided to seek new
sanctions against illegal arms exporters and plans to create in the Office of Trade and Industry and export control office. Foreign Minister Hans-Dietrich Genscher called exporters “merchants of death,” while Economics Minister Jergen Kohlmann says he is determined to close the “last loophole” in German export laws. As the fifth largest arms exporter, Germany earns about 6% of its export earnings from the sale of weapons abroad. In fact, according to Germany's "arms of small scale" in its export policies after being struck with Iraqi arms licenses which, it is alleged, had been improved and their range increased by German companies. It is also alleged that German companies built "poison gas plants in Iraq as well as sold sophisticated equipment such as precision timers to the Iraqi military. Under the proposed sanctions, violations of arms embargos on trade with Iran will be implemented for a minimum of one year, an increase from six months, with 10-year maximum sentences. At least, 10 German companies are under investigation on suspicion of violating the export laws.

Legend

- HolOnHeld
- At
- BaseOnAt
- MuslimOnNation
- Nation
- NationOnNation
- NationOnLocation
- Location
- GeogOnEntity
- Facility
- MilitaryOnOrganizational
- Organizational
- Physical
- PersonOnPhysical
- Social
- PersonOnSocial
- PersonalOnOrganizational
- SocialOnPerson
- PersonOnOrganizational
- PersonalOnSocial
- PersonalOnOrganizational
- OrganizationalOnSocial
- OrganizationalOnPerson
- SocialOnPerson

Select All  Deselect All  Hide Unselected
Knowledge Extraction and Integration

Assisting the Transformation from Text to Formal Knowledge
The Knowledge-Level Type (KLT) System imposes essential ontological distinctions on linguistic analysis results

The transformation of analysis results into formal knowledge…

Requires analysis engines to distinguish between entity annotations, relation annotations, mentions of individuals and the individuals themselves (referents).

Fred Center is the CEO of Center Micros. Center is a graduate of state University

Fred Center is the CEO of Center Micros.
EKDB (Extracted Entities And Relations)

Target Ontology (OWL)

KITE (Semantic Integration)

OWL (RDF: Instances of entities & relations)

OWL Tools
Formal Reasoners (e.g., Protégé, Pellet)
Knowledge Integration and Transformation Engine

The transformation of analysis results into formal knowledge requires the explicit mapping of types in the type system to classes and properties in the ontology.

**CAS**

**Analysis Type System**

**Entity (Person):** Fred Center

**Entity (Organization):** Center Micros

**Relation (ManagerOf)**

**KITE Mapping Plugins**

Person(?x) ^ ManagerOf(?x, ?y) → Executive(?x)

ManagerOf(?x, ?y) → hasManager(?y, ?x)

Organization(?x) → SocialAggregate(?x)

**Target Ontology**

**Executive: Fred Center**

**hasManager**

**SocialAggregate: Center Micros**

**Target KB**
Automated Knowledge Acquisition

...Not Gonna Happen...

Documents or other media

Automation + User Interaction Tools and process tolerates less than perfect precision to produce candidate task-relevant knowledge

OWL KB

Task-Relevant Knowledge
Acceptable to Humans & Automated Reasoners

• Perfect Knowledge
• Precise & Complete Rendering
• Ready to Reason

• Unstructured
• Implicit Semantics
• Ambiguous
• Imprecise
• Incomplete
• Inconsistent
• Partially relevant
Thank You
Query Analysis: Simple *Answer Type Insertion*

**Want User to Type these….**

- **What is Webster’s Phone Number**
  - Webster’s Phone Number
  - ...

- **Where is Smith?**
  - Smith’s Place
  - ...

**The System to Generate These….**

- +*Webster phone number*
- +<PhoneNumber> </PhoneNumber>

- +*Smith*
- <Location> </Location>

- +*Smith place*
- <Location> </Location>
Query Analysis: *Exploiting Relation Detection*

**Want User to Type these....**

- **Who owns Select Gourmet Foods?**
  - <Owner>
  - +<> <Person> </Person> <Organization> </Organization><>
  - +<Organization> Select Gourmet Foods </Organization>
  - </Owner>

- **Select Gourmet Foods owner?**
  - owner
  - <Organization> +Select +Gourmet +Foods </Organization>
  - +<>
  - <Organization> </Organization>
  - <Person> </Person>
  - </>

**The System to Generate These....**

- Greater Precision

Still targeted but **Greater Recall**
UIMA Data Representation Alignment with Standards

- **CAS (Instances)**
  - Semantics: Object Graph
    - A *stand-off* representation leaves original artifact untouched
    - Expressivity allows multiple interpretations of artifacts, overlapping annotations
  - Inline XML may be generated
  - OWL RDF may be generated

- **CAS Type System (Schema)**
  - Semantics: Ecore/EMOF
  - XML Syntax: XMI
  - XML Schema for partial instance validation can be generated
  - Java Classes for CAS Types can be generated

- **Component Metadata**
  - For Discovery, Reuse and Application-Level Plug-in-Play
  - XML document declaratively capture identification, configuration, behavior and deployment information about UIMA components.
  - Tools to create standard OSGi Bundles for painless application integration from servers to cell phones [http://www.aqute.biz/osgi.html](http://www.aqute.biz/osgi.html)
Leverage open-source Eclipse/EMF Tooling

- **Standard Syntax & Semantics (OMG)**
- **Potential for Increased Expressivity**

**Standard EMF Tooling**

**Standard EMF Tools +**

**Annotated Java Interfaces**

**UML**

**XML Schema**

**UIMA Type System (eCore/eMOF)**

**EMF Tooling +**

**XML Schema**

**Generated Java Classes (UIMA JCAS)**

**CAS Instances (XMI)**

**Partial Validation Of instance data**

**Standard XML For an Object-Graph**