Semantic Acceleration or "The Practical Web"

Dr. Alfred Spector
Chief Technology Officer
IBM Software Group

(Due credit for this talk to Chris Welty and the UIMA team)
Outline

- Innovation and Semantics
- Semantic Web
- The Challenge
- The Opportunity
  - Unstructured Information Management Architecture: UIMA (!)
- Connections to the Semantic Web
- Successes to date
- Opportunities and the IBM Innovation Grants
- Conclusions and Summary
Semantic Acceleration or “The Practical Web”

Abstract

The Semantic Web envisions a future where applications (computer programs) can make sense and therefore more productive use of all the information on the web by assigning common "meaning" to the millions of terms and phrases used in billions of documents. AI and knowledge representation must rise to the occasion and work with decentralized representations, imprecision and incompleteness. Standard web-based representations are an essential enabler and we have made good progress in their design. But we still rely on humans to assign semantics and here there is a big leap of faith: The World Wide Web has grown at startling rates because humans are prolific at producing enormous volumes of unstructured information, that is, information without explicit semantics; on the other hand navigating this mass of information has proven to be both possible and profitable to the point that there is a $6 B search advertising industry. It's is not practical to expect the same will automatically happen for semantically enriched content. And yet we need semantics to better leverage the huge value on the web.

The Practical Web is about confronting this challenge. Its about realizing that we will need to automate the assignment of semantics to unstructured content to ultimately realize the vision of the Semantic Web. If well done the results will be synergistic with the motors of web expansion: user value and commercial value.
Society Moves Forward Because of Innovation

Innovation is the intersection of invention and insight, leading to the creation of social and economic value.

US National Innovation Initiative
Information Semantics will Drive Greatly Increased Value …in Virtually Every Domain.

- Incremental value leveling off
- Semantics
- No semantics

1990s 2000s
Semantic Web

The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

From the W3C Semantic Web home: [http://www.w3.org/2001/sw/](http://www.w3.org/2001/sw/)
Structured Information:

Semantics of information captured in DB schema

<table>
<thead>
<tr>
<th>Name</th>
<th>Occupation</th>
<th>Organization</th>
<th>Age</th>
<th>Office Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>Engineer</td>
<td>IBM</td>
<td>29</td>
<td>San Francisco</td>
</tr>
<tr>
<td>Carbonell</td>
<td>Professor</td>
<td>CMU CSD</td>
<td>39</td>
<td>The Burgh.</td>
</tr>
<tr>
<td>Brown</td>
<td>CEO</td>
<td>Textract</td>
<td>42</td>
<td>New York</td>
</tr>
</tbody>
</table>

Unstructured Information:

Semantics inherent in usage and context

The associated press reported today that, Ms Jones, an Engineer at IBM has been recently spotted at the Summit meeting in Zurich,............At, 29, Ms. Brown, Is the youngest CEO at the Summit,
Wherefrom the Semantics?

- Some will be manually created
- Some web content generated from existing databases
  - Structured, but semantics often hidden
  - Still may requires efforts to harmonize, extend, declare, & expose
- However, most web and enterprise data contains only latent structure
  - Manual markup hard –perhaps even impossible– to scale
  - Therefore, automated and semi-automated methods required
The associated press reported today that, Ms Brown, CEO of Textract Corp., had been recently spotted at the Summit meeting in Zurich. At 42, Ms. Brown, is the youngest CEO at the Summit,…

Using Analysis Engines (pattern recognition techniques),
to do Automatic Tagging,….

imposes structure (Reveals Semantics), …..

allowing us to process text as we would structured data.
Analytics: The Promise and the Challenge

- Independently developed
- From an increasing # of sources

Analysis Capabilities:
- Language, Speaker Identifiers
- Tokenizers
- Part of Speech Detectors
- Document Structure Detectors
- Parsers, Translators
- Named-Entity Detectors
- Face Recognizers
- Relationship Detectors
- Classifiers...

Capability Specializations:
- Modality
- Human Language
- Domain of Interest
- Source: Style and Format
- Input/Output Semantics
- Privacy/Security
- Precision/Recall Tradeoffs
- Performance/Precision Tradeoffs...

The right analysis for the job will likely be a best-of-breed combination integrating across many dimensions.
Key point: The combination hypothesis: If intimately integrated, various KM technologies will provide higher quality results (accuracy, recall, etc.)
UIMA: The Project

- **Start**
  - IBM Research, Watson and Worldwide beginning 2001
  - An internal project to accelerate Research and Technology Transfer
  - *And to bring order out of our own chaos 😊*

- **Focus**
  - Text and multi-modal analysis integration and component reuse in support of Information and Knowledge Management products and solutions

- **Requirements**
  - Text, video and speech analysis
  - Advanced (concept/semantic) search and knowledge representation and reasoning

- **Architecture**
  - Informed by TIPSTER, Catalyst, Atlas, GATE, TAF, Talent, WebFountain
  - Modern software engineering approaches

- **Individuals involved**
  - David Ferrucci, Arthur Ciccolo, Andrei Broder, and many more
UIMA Objectives

- Enable Separation Of Roles
- Facilitate Combination Hypothesis
- Enable Third Party Contribution
- Speed Technology Transfer
- Support Ease Deployment
- Promote Technology Sharing
- Capitalize On Web Services
- Support Multiple Data Modalities
- Support Wide Breadth of Software Prod’s
- Support Large Scale Applications
- Support Breadth Of UIM Research
- Accelerate Performance Of Core Technologies

UIMA
UIMA Defined

- 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
  - Pluggable Transports
  - Embeddable
- **Independent of but interoperable with**
  - Specific data models
  - Specific algorithms
  - Specific Language-level or domain-level concepts or tools
  - Specific workflows or workflow engines
  - Specific Back-end Systems (DB, Search Engine, KB Interfaces)
UIMA: The Software

- Supports UIMA-compliant development, composition & deployment
- Java and C++ framework implementations
  - Analytics in other languages possible through service-based interfaces
- Support for co-located and service-oriented deployments
- Support for specialized APIs to common data representation
- UIMA SDK (Software Development Kit)
  - Stand-alone Java Install
  - Freely Available from IBM alphaWorks
  - Includes Tutorial and Development-Level Utilities and Tooling
  - Ships with a “Semantic Search” Engine and CAS Indexer
  - Core framework goes open-source by end of 2005
The Open Framework Approach

**Key Assumptions**
- Businesses will realize increasing value from discovering knowledge in unstructured information using component analytics
- Analytics improve differentiate search & knowledge management solutions
- **There will be as many providers of analytics as there are application software creators/vendors**

**Key Enabler for UIM Solutions in many industries and application domains**
- Rapid composition and integrated deployment of best-of-breed analytics

**IBM’s Approach**
- Provide **UIMA** as an Open, Plug-n-Play Integrating Framework
- Enable IBM and other products and services with UIMA
- Help build a World-Wide ecosystem of analysis and application developers
- Differentiate IBM in a number of dimensions
Text and Multi-Modal Analytics

Unstructured Information

- High-Value
- Most Current
- Fastest Growing (80% of Corporate Data)

...BUT ...
- Buried in Huge Volumes (Noise)
- Implicit Semantics
- Inefficient Search

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....

Structured Information

- Explicit Semantics
- Efficient Search
- Focused Content

...BUT...
- Slow Growing
- Narrow Coverage
- Less Current/Relevant
Report Date 10 March 2003. Slick business dealings keep local olive oil importer out of the pits. Robert Crane was recognized by local business leaders for his skill at leading the Gorman Food Importers Inc. to strong profits while others are struggling. Mr. Crane, owner of Gorman Food Importers Inc., has consistently been able to produce exceptional results, while still keeping a focus on his employees. Gorman Food Importers Inc. has been in business since 1970 and specializes in food imports from the Middle East, including olive oil and figs. Gorman Food Importers Inc. is headquartered in NYC, and their warehouse is located in Paramus, NJ. The company employs 659 people in the two locations. Robert Crane can be reached at 608-703-2317.
UIIMA Pipeline for Video Concept Detection & Indexing

Video segments about Basketball, skiing, vehicles

**Collection Processing Engine**
- Video Segment Parser
- Video Annotation Aggregate
- Analysis Engine

**CAS Consumers**
- xCAS Writer
- Semantic Search Indexer
- MARVEL Indexer

**Local File System**
- Semantic Search Index
- MARVEL Index

**Query**
- Hit List

**MARVEL Search Engine**
UIMA Pipeline for Video Concept Detection & Indexing

Video Collection

Keyframe And Icon Extractor

Video Segment Parser

UIMA Pipeline for Video Concept Detection & Indexing

Video segments about Basketball,...skiing, vehicles

Media Metadata Annotator

Collection Reader

Analysis Engine

CAS Consumers

Collection Processing Engine

Visual Feature Extractor

Semantic Concept Detector

Semantic Search Indexer

MARVEL XCAS Writer

Local File System

Semantic Search Index

MARVEL Index

Query Hit List

Local Intranet
An example annotator that searches for room numbers in the IBM Watson research buildings.
More on Relationship to Semantic Web: The Return of the Combination Hypothesis

*If intimately integrated, various KM technologies will provide higher quality results (accuracy, recall, etc.)*

- Can this be generalized to combination of UIMA & Semantic Web?
  - Can we combine annotators and formal ontologies and reasoners to accelerate the population of the semantic web?
  - What would “higher quality” mean in this context? How would it be measured?
From UIMA Analytics to the Semantic Web

“13 delegates from Turkey arrived today.”

<country rdf:id="Turkey" />

Easy!!!
UIMA and Semantic Web Technologies

Text Analytics

Semantic Web

- Relations
- Entities vs. Mentions
- Precision
- Recall
- Explainability
- Brittleness
- Scale

Entities vs. Mentions
Traditional analysis just focuses on types
Not much use to SW technology
• End up with a graph of just nodes, no arcs
Need to have relations
State of the art for relation extraction not so good
• 30% Precision & 40% Recall

Must improve relation identification & extraction
Entities vs. Mentions

- End up with a graph of disconnected triples
- Need to identify nodes that refer to the same entities – coreference analysis
UIMA and Semantic Web Technologies

Text Analytics vs. Semantic Web

Relations
Entities vs. Mentions
Precision
Recall
Explainability
Brittleness
Scale

Very interesting and fruitful work to be done!
Adoption: UIMA within IBM

- IBM Research Labs developing UIMA compliant Analysis Engines
  - Deep and Shallow Parsing
  - Categorization
  - Summarization
  - Semantic Class Detection
  - POS, English/Chinese/Japanese NE
  - Classifier Trainers
  - Machine Translation
  - Video and Speech Analytics
  - BioInformatics, etc.

- Some products:
  - Portal
  - Omnifind

- IBM Internal Component Repository
  - 80+ Analysis Components and 23+ UIMA-based systems/solutions
Adoption: Outside of IBM

- First Version of UIMA SDK Released on AlphaWorks Dec 2004
  - 2,600+ Downloads as of September 2005
  - Open Source Announcement met with broad industry interest
- Mayo Clinic – an early adopter
- UIMA Working Group driven by DARPA and IBM
  - Small initial group of academics & researchers to evaluate & provide feedback
    - Stanford, Carnegie Mellon, Columbia, UMASS
    - BBN, MITRE, SAIC (Object Sciences)
- DHS/National Labs - Threat Assessment Project
- DARPA/ITPO GALE Project (Speech-to-text, Translation, Distillation)
- TC-STAR Speech-to-Speech Project
- Third party development of UIMA compliant analytics
  - GATE Interoperability Layer (University of Sheffield)
  - OpenNLP Components UIAMfied (Tokeniser, Parser, POS, NE, Sent Chunker)
  - Components from UIMA working group members
  - Endorsement by 16+ software companies
DARPA GALE Project

- Awarded mid-2005
- Aimed at breakthroughs in Recognition, Language Processing, and Distillation
- 3 Teams led by BBN, SRI International, and IBM
- IBM Team:
  - Brown, Carnegie Mellon, Johns Hopkins, Stanford, University of Maryland, University Of Pittsburgh
- UIMA selected as the architecture for analytical engine integration for all teams
Enterprise Search Middleware - Omnifind

Unstructured data in the Enterprise forces Innovation in Search Engine
Differentiated value based on Unstructured Information Mgmt. Architecture (UIMA)

- **Delivers the best results with sub-second response**
  - Sophisticated relevancy algorithms for corporate content

- **Scales for large collections or enterprises**
  - 500K documents and above
  - 1000s of concurrent users

- **Fits easily into enterprise applications**
  - Java APIs
  - Document level security

- **Eases administration and maintenance**
  - Analysis features all under-the-covers

HTTP/HTTPS, News groups (NNTP), File systems, Domino databases, MS Exchange public folders, DB2
Content Mgr, DB UDB, Informix, Oracle Documentum & FileNet via integrated WebSphere II Content Edition
Where does UIMA fit in to the Business World?

- **Computing Infrastructure**
  - (machines, networks, grids)

- **System Middleware**
  - (WebSphere, DB2, Web services, MQ, SOAP, Globus etc.)

- **UIM Middleware (UIMA)**
  - (Search Engines, Docs & Meta-Data, Collection Processing, Analysis Engine Frameworks, Ontology Integration etc., )

- **UIM & KM Application Frameworks**
  - (e.g, Life Sciences, National Intelligence)

- **UIM & KM Applications**
## ISVs Supporting UIMA and OmniFind

<table>
<thead>
<tr>
<th>Deliver content to platform for analysis</th>
<th>Provide components text analytics</th>
<th>Provide applications that leverage text analysis and semantic search</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.factiva.com">Factiva</a></td>
<td><a href="https://www.attensity.com">Attensity</a></td>
<td><a href="https://www.inquira.com">Inquira</a></td>
</tr>
<tr>
<td><a href="https://www.ql2software.com">QL2 Software</a></td>
<td><a href="https://www.clearforest.com">ClearForest</a></td>
<td><a href="https://www.cognos.com">Cognos</a></td>
</tr>
<tr>
<td><a href="https://www.inxight.com">Inxight</a></td>
<td><a href="https://www.ipharse.com">Iphrase</a></td>
<td><a href="https://www.endeca.com">Endeca</a></td>
</tr>
<tr>
<td><a href="https://www.nstein.com">nstein</a></td>
<td><a href="https://www.sas.com">Sas</a></td>
<td><a href="https://www.kana.com">Kana</a></td>
</tr>
<tr>
<td><a href="https://www.schemaLogic.com">SchemaLogic</a></td>
<td><a href="https://www.spss.com">Spss</a></td>
<td><a href="https://www.siebel.com">Siebel</a></td>
</tr>
<tr>
<td><a href="https://www.emi.com">EMI</a></td>
<td><a href="https://www.semagix.com">Semagix</a></td>
<td></td>
</tr>
</tbody>
</table>
Some UIMA Links

- [UIMA Homepage](http://www.research.ibm.com/uima) at IBM Research
- Download [UIMA SDK](http://www.alphaworks.ibm.com) from IBM alphaWorks site
- IBM Systems Journal Article: [Building an example application with the Unstructured Information Management Architecture](http://www.research.ibm.com/uima)
  - Related Press
    - [Volume Analytics: IBM's UIMA - and Why You Should Care](http://www.research.ibm.com/uima)
      - [DMReview](http://www.research.ibm.com/uima)
  - In addition for folks in Ireland, contact [Elaine Stephen@ie.ibm.com](mailto:Elaine.Stephen@ie.ibm.com), Director of IBM Dublin lab to find out about our text analytics in Dublin and for career opportunities there!
IBM Innovation Awards

- Unstructured Information Management Architecture (UIMA) Innovation Award for 2006
- Background: The UIMA framework separates the hard work of advancing the state-of-the-art in Natural Language Processing and more generally algorithms for unstructured information (text, audio, video) analysis...
- Curriculum and Research.
- Grant size: $10,000. - $30,000.
- Objective: Proposals are sought in this area, in the porting of significant analysis algorithms to the UIMA framework, and in the use of UIMA to support knowledge acquisition for the semantic web.
- Online submission will open on January 26, 2006
- Information will be posted at: www.ibm.com/university
- Key dates:
  - January 26, 2006 Online submission opens.
  - February 17, 2006 Evaluation begins for proposals rec’d by this date.
  - February 28, 2006 Deadline for submitting a proposal.
  - April 28, 2006 Award winners notified via email & postal mail.
The early 90’s Web was elegantly simple. However, today’s high aspirations for the Web require so many additional capabilities that we may fail to meet our information handling quality goals, and our systems may become enormously expensive and brittle. This presentation postulates the need to develop a coherent Architecture for Knowledge Middleware.

This software platform would enable the integration of the diverse technologies required to process accurately diverse forms of information – from natural language and image to highly structured databases. Using technologies that work together cleanly would benefit us by permitting reuse of commonly needed components and improving information processing quality via simplifying the combination of multiple approaches.

This presentation will make a case for (1) the economic and technical imperative, (2) the major challenges to be overcome, (3) the practicality, and (4) both near and long-term benefits of this coherent approach to Knowledge Middleware.
Conclusions

- Semantic processing of unstructured information seems exceedingly useful
- The semantic processing will be based on many forms of analytics, developed by many – yet operating together.
- The combination of these analytics will result in higher accuracy analytics: a.k.a. the Combination Hypothesis is true.
- UIMA provides very valuable engineering support for this
- IBM intends to Open Source UIMA shortly to facilitate adoption
- We think UIMA will be of value to the semantic web
- It seems to us there is valuable research to be done here
- IBM will make available Innovation Grants available to catalyze efforts in this important area
Thank You