GALE & UIMA

Hany Hassan
NCLT Seminar Series
Overview

- **GALE**
  - What is GALE?
  - GALE Tasks
  - GALE Linguistics Resources

- **UIMA Architecture**
GALE

- Global Autonomous Language Exploitation
- Largest ever DARPA funded project
- $50 million over 3 years (Maybe 2)
- 17 Participants (Companies and Universities)
- Three Teams (IBM, BBN, SRI)
- IBM Team:
  - Brown, CMU, Johns Hopkins, Stanford, University of Maryland, University of Pittsburgh
The GALE Project

“The goal of the GALE program is to improve timeliness and quality of information available to the war fighters and those that support them via a single, comprehensive system that supports transcription, translation, and distillation of multiple languages in text and audio formats.

The ultimate goal of GALE is to demonstrate transcription and translation capabilities at near-human levels, coupled with the ability to identify relevant information via distillation at levels exceeding human ability.”
GALE tasks

- Language Exploitation Environment (LEE)
- Transcription Engines
- Translation Engines
- Distillation Engines
- Linguistic Data
- Utility Evaluations
- Technology Insertion Projects into Military Applications
GALE tasks

- Transcription Engines
- Translation Engines
- Distillation Engines
- Linguistic Data
- Language Exploitation Environment (LEE)
Transcription Engines

- **Speech Transcription**
  - Arabic, English & Chinese
  - Broadcast news (radio, television)
  - Talk shows (studio, call-in)
  - Telephone conversations
- **State of-the-art**: 65% : 75%
- **Initial Target**: 65%
- **Final Target**: 95%
Machine Translation

• Machine Translation
  ➢ Arabic, English & Chinese
  ➢ Sources:
    • Broadcast news (radio, television)
    • Talk shows (studio, call-in)
    • Newswire
    • Newsgroups
    • Weblogs
    • Telephone conversations
Machine Translation: 2

- Translation engines should produce accurate, readable English.
- Evaluated by translation accuracy:
  - Post Editing: deletions, insertions and substitutions by a human editor necessary to make the machine translation readable, and to convey the proper meaning.
- First phase: 75%
- Final Target: 95%
Distillation Engine

• Integrates information of interest to the user
• Multiple sources and documents.
• Gather all documents relevant to a specific query
• Identify various kinds of new information
• Hyperlinked multimedia information but not necessarily text.
• Distillation engines will be evaluated differently depending on their outputs.
• Target: 95% recall with 90% precision.
Sample Query Types:

1. LIST FACTS ABOUT EVENTS DESCRIBED AS FOLLOWS: [event]
2. WHAT [people/organizations/countries] ARE RELATED TO [event] AND HOW?
3. PRODUCE A BIOGRAPHY OF [person]
4. PROVIDE INFORMATION ON [organization]
5. FIND STATEMENTS MADE BY OR ATTRIBUTED TO [person] ON [topic(s)]
6. DESCRIBE THE RELATIONSHIP OF [person/org] TO [person/org]
7. DESCRIBE [topic(s)] AND INVOLVEMENT OF [country]
8. DESCRIBE THE PROSECUTION OF [person] FOR [crime]
GALE Linguistic Data

- Transcription
  - Arabic & Chinese
- Translation
  - Arabic Broadcast
  - Chinese Broadcast
  - Weblog
  - Newswire
- Word Alignment
- XBanks
  - Arabic Treebank, POS, Propbank
  - English Treebank, POS, Propbank
GALE Evaluation

• One task consisting of
  – Distillation from foreign text
  – Distillation from foreign speech

• 2 Evaluation Schemes:
  – Go/NoGo evaluation
    • The teams have to pass a predefined level of performance in each core technology
  – Utility evaluation
    • The systems will be judged as for how well they help an analyst perform their task
Facilitate the interconnection of processing engines
Facilitate application development
Well-defined interfaces and control software for interconnecting the major processing modules in GALE
Overview

- GALE
  - GALE Tasks
  - GALE Linguistics Resources
- UIMA Architecture
UIMA is an open, industrial-strength, scaleable and extensible platform for creating, integrating and deploying unstructured information management solutions from powerful text or multi-modal analysis and search components. IBM is making the UIMA SDK available as free software to provide a common foundation for industry and academia to collaborate and accelerate the world-wide development of technologies critical for discovering vital knowledge present in the fastest growing sources of information today, source of information available to businesses and governments. The web is just the tip of the iceberg. Consider the mounds of information hosted in the enterprise and around the world and across different media including text, voice and video. The high-value content in these vast collections of unstructured information is, unfortunately, buried in lots of noise. Searching for what you need or doing sophisticated data mining over unstructured information sources presents new challenges. An unstructured information management (UIM) application may be generally characterized as a software system that analyzes large volumes of unstructured information (text, audio, video, images, etc.) to discover, organize and deliver relevant knowledge to the client or application end-user. An example is an application that processes millions of medical abstracts to discover critical drug interactions. Another example is an application that processes tens of millions of documents to discover key evidence indicating probable competitive threats. First and foremost, the unstructured data must be analyzed to interpret, detect and locate concepts of interest that are not explicitly tagged or annotated in the original artifact, for example, named entities like persons, organizations, locations, facilities, products etc. More challenging analytics may detect things like opinions, complaints, threats or facts. And then there are relations, for example, located in, finances, supports, purchases, repairs etc. The lists of concepts important for applications to detect and find in unstructured resources are large and often domain specific. Specialized component analytics must be combined and integrated to do the job. The result of analysis must, in turn, be put in structured forms so that powerful data mining and search technologies like search engines, database engines or OLAP (On-Line Analytical Processing, or Data Mining) engines may be leveraged to efficiently find the concepts you need, when you need them. In analyzing unstructured content, UIM applications make use of a variety of analysis technologies including: • Statistical and rule-based Natural Language Processing (NLP) • Information Retrieval (IR) • Machine learning • Ontologies • Automated reasoning and • Knowledge Sources (e.g., CYC, WordNet, FrameNet, etc.) These technologies are developed independently by highly specialized
UIMA is an open, industrial-strength, scalable and extensible platform for creating, integrating and deploying unstructured information management solutions. IBM is making the UIMA SDK available as free software to provide a common foundation for industry and academia to collaborate and accelerate the world-wide development of technologies critical for discovering vital knowledge present in the fastest growing sources of information today. Source of information available to businesses and governments. The web is just the tip of the iceberg. Consider the mounds of information hosted in the enterprise and around the world and across different media including text, voice and video. The high-value content in these vast collections of unstructured information is, unfortunately, buried in lots of noise. Searching for what you need or doing sophisticated data mining over unstructured information sources presents new challenges. An unstructured information management (UIM) application may be generally characterized as a software system that analyzes large volumes of unstructured information (text, audio, video, images, etc.) to discover, organize and deliver relevant knowledge to the client or application end-user. An example is an application that processes millions of medical abstracts to discover critical drug interactions. Another example is an application that processes tens of millions of documents to discover key evidence indicating probable competitive threats. First and foremost, the unstructured data must be analyzed to interpret, detect and locate concepts of interest that are not explicitly tagged or annotated in the original artifact, for example, named entities like persons, organizations, locations, facilities, products etc. More challenging analytics may detect things like opinions, complaints, threats or facts. And then there are relations, for example, located in, finances, supports, purchases, repairs etc. The lists of concepts important for applications to detect and find in unstructured resources are large and often domain specific. Specialized component analytics must be combined and integrated to do the job. The result of analysis must, in turn, be put in structured forms so that powerful data mining and search technologies like search engines, database engines or OLAP (On-Line Analytical Processing, or Data Mining) engines may be leveraged to efficiently find the concepts you need, when you need them. In analyzing unstructured content, UIMA applications make use of a variety of analysis technologies including: • Statistical and rule-based Natural Language Processing (NLP) • Information Retrieval (IR) • Machine learning • Ontologies • Automated reasoning and • Knowledge Sources (e.g., CYC, WordNet, FrameNet, etc.) These technologies are developed independently by highly specialized
From Unstructured to Structured

The Big Picture

Analytics Bridge the Unstructured & Structured worlds

Unstructured Information:
- Text
- Chat
- Email
- Audio
- Video

Identify Relevant Entities → Build Structure:
- Docs, Emails, Phone Calls, Reports
- Topics, Entities, Relationships
- People, Places, Org, Times, Events
- Customer Opinions, Products, Problems
- Chemicals, Drugs, Drug Interactions

Structured Information:
- Indices
- DBs
- KBs

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

High-Value
Most Current Content
Fastest Growing

BUT...
- Buried in Huge Volumes – Lots of Noise
- Implicit Semantics
- Inefficient Search
UIMA Architecture
UIIMA Design

Architecture for:
- Composing analytics that extract knowledge from unstructured sources
- Integrating results with structured information
- Providing Interfaces, Data Representation Schemes, Design Patterns

Principal Architectural Commitments
- Common representation scheme
- Common component engine interfaces (task and domain-independent)
- Common component metadat
- Plugable Workflow

Independent of but interoperable with:
- Specific data models
- Specific algorithms
- Specific Languages
- Specific Workflow
- Specific Back-end systems (DB, Search Engines, ..)
Fred Center is the CEO of Center Micros.

He is a graduate of State University and excels at golf.

Text Document: D102
UIIMA Annotation Viewer

- Analyzed by a combination of Analysis Engines
- Semantic Entities & Relations Represented
- Highlighted here in a GUI
Video Concept detection and indexing
UIMA Complaint Analysis Engines

- Deep and Shallow Parsing
- Categorization
- Summarization
- Semantic Class Detection
- POS Tagging
- Named Entity Identifier
- Machine Translation
- Video Analysis
- Audio Analysis & Speech Recognition
- Bio-informatics
- etc.
UIMA: Software

- 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
  - Open-Source
  - Includes Tutorial and Development-Level Utilities and Tooling
- Ships with a “Semantic Search” Engine and CAS Indexer
UIMA Open Source

- 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
- Rapid composition and integrated deployment of best-of-breed analytics
- Help build a World-Wide ecosystem of analysis and application developers
UIMA: Adoption

- Mayo Clinic – an early adopter
- UIMA Working Group driven by DARPA and IBM
  - Stanford, Carnegie Mellon, Columbia, UMASS
  - BBN, MITRE, SAIC (Object Sciences)
- DARPA/ITPO GALE Project
- TC-STARSpeech-to-Speech Project
- GATE Interoperability Layer (University of Sheffield)
- OpenNLP Components UIMAfied (Tokeniser, Parser, POS, NE, Sent Chunker)
- Components from UIMA working group members
- Endorsement by 16+ software companies
UIMA: links

- **UIMA Homepage**  IBM Research
  http://www.research.ibm.com/UIMA/

- **UIMA SourceForge:**
  http://sourceforge.net/projects/uima-framework
Thank you