Infrastructure and Systems for Adaptive Speech and Text Analytics

Salim Roukos
roukos@us.ibm.com
Natural Language Technologies
IBM T. J. Watson Research Center
UIMA Pipe Line for Fact Search

Collection Processing Engine

Collection Reader
Collection of Text Docs

File Reader
Analysis Aggregate

Phrase Finder
Deep Parser
Relation Detector
SimpleToken AndSentence Annotator
EAnnotator (NE Detector)
Remote ACE (Remotely deployed NE Detector)
Relation Aggregate
Cross Annotator Coreference

Analysis Engines

CAS Consumers

Semantic Search Index

Semantic Search Indexer
XcasCas Consumer
Relation Coref
Generic Coref
EKDB Writer

Local File System

EKDB

Merge Lists
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.
The **Common Analysis Structure** (CAS) enables Analysis Writers access to the artifact being analyzed and the current analysis results.

Original artifacts may have multiple “subjects of analysis” (different views on the same artifact) and new ones may be created during and as a result of analysis.
The UIMA framework allows the developer to **compose & encapsulate** a set of engines as a single component, insulating applications or users from its implementation details/complexities – Promoting Reuse.
UIMA Component Architecture from “Source to Sink”
A common platform for development, composition and deployment of multi-modal analytics into different carriers.

Development Tools       Component Repository

UIMA Framework

UIMA Analysis Components

Large-Scale Distributed Deployment  Standalone  Product Embeddings

Web Fountain, ....  UIMA SDK  OmniFind, WPS, LWP...

Powered by UIMA-Compliant Run-Times