

Information Extraction

● Entities

- ◆ Named entities such as People, Locations, and Organizations.
- ◆ Numerical Expressions such as Date, Year, Percentage, etc.
- ◆ Nominal entities such as *the truck*, etc. which get slotted into the proper category.

● Relationships and Attributes

- ◆ Entity relationship extraction, e.g., relationship associated with *person* include employer, associates, spouse, location, etc.
- ◆ Descriptive information such as age, job title, affiliation, etc.
- ◆ Contact information such as address, e-mail, URL, phone number, etc.
- ◆ Generic descriptor phrases used to describe entities, e.g., *the largest software manufacturer*, in reference to Microsoft.

● Entity Profiles

- ◆ Rich profiles for People, Organizations, Locations (including Geopolitical Entities), facilities, nominal entities (e.g., *the soldier*).
- ◆ Consolidated information for entity from single document, including event links.
- ◆ Cross-document merged entity profiles.
- ◆ Includes named, nominal, and pronominal entities.
- ◆ Consolidated local relationships.

● Events

- ◆ Captures *who did what (to whom)*.
- ◆ Template based domain-specific events.
- ◆ Support for user-defined event detection.

● Normalization

- ◆ Temporal normalization, support for TIMEX2 standard.
- ◆ Resolution of relative time and day mentions based on context.
- ◆ Location normalization, e.g., resolving ambiguous names based on context.

● Multilingual Support

- ◆ Native support for Unicode UTF-8 text.
- ◆ English, Simplified Chinese.

Architecture

● Hybrid model

- ◆ Flexible pipeline of processing modules.
- ◆ Statistical modules, e.g., entity tagging.
- ◆ Grammar modules, e.g., relationship detection.
- ◆ Lexicon look-up.

● Operating System Support

- ◆ Windows XP, Solaris 8 & 9, Linux RedHat EL 3 & 4.

● Modes of Operation

- ◆ Active Retrieval of documents via HTTP, FTP or local file system.
- ◆ Service Oriented Architecture, HTTP Post request and response for processing.

● Integration and Scalability

- ◆ Modules communicate via CORBA.
- ◆ Support for multiple CPUs for near-linear scalability of performance throughput.
- ◆ SOAP interface.

● Document Formats

- ◆ Support for Northern Light, Dialog, Factiva, LexisNexis, and other news sources.
- ◆ Support for a variety of XML and military message formats including USMTF.
- ◆ Extensible to support additional file formats and document sources.
- ◆ Multi-user utterance-level monitoring and alerts from XMPP and IRC chat rooms.

● Performance Throughput

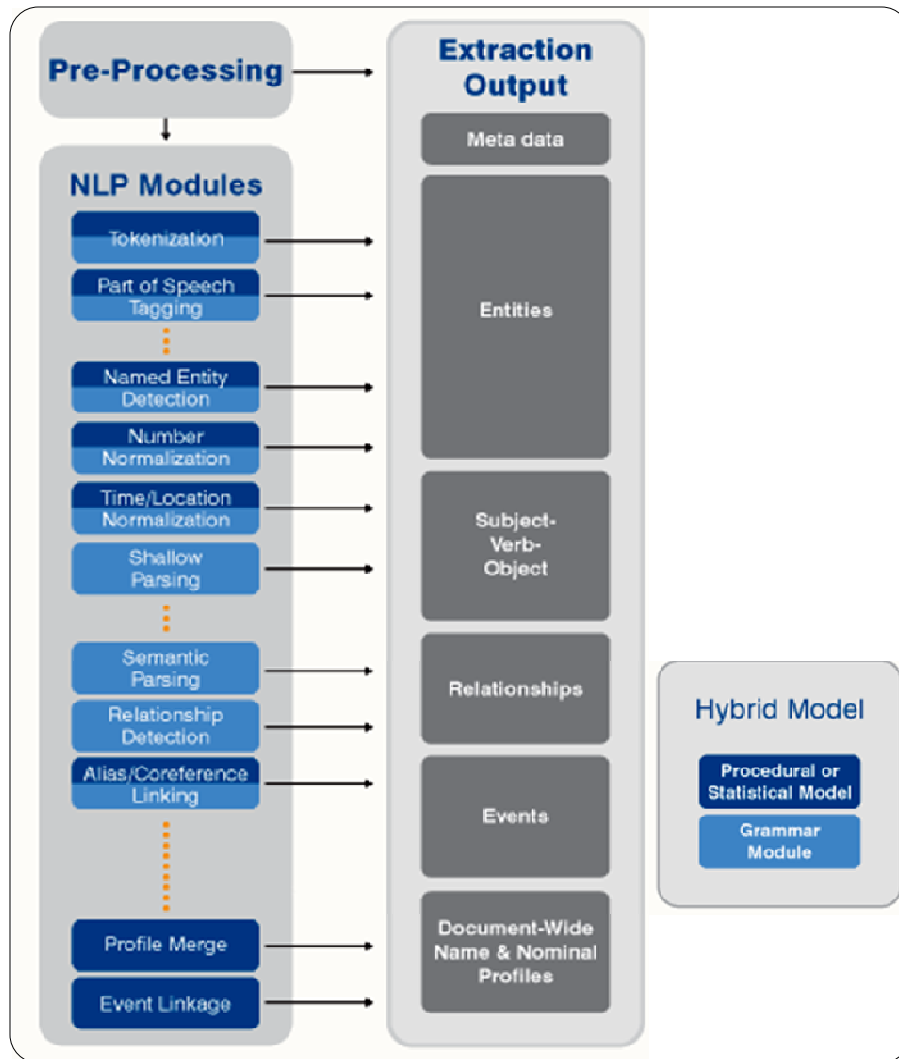
- ◆ 12-14 MB of documents processed per hour for complete extraction, per CPU.

● Extraction Output

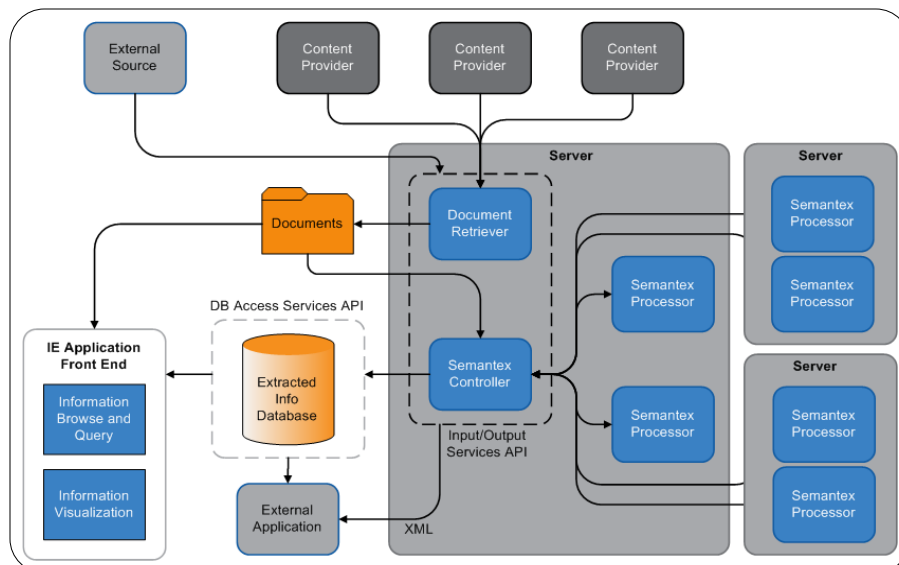
- ◆ In-line and stand-off XML markup.
- ◆ APF format.
- ◆ Relational Database support.

● Customization

- ◆ *Semantex Workbench*™ tool for user-enabled customization of grammars, lexicons, statistical models, entity profiles, and events.



Semantex™ Hybrid Processing Pipeline



Typical Semantex™ Deployment