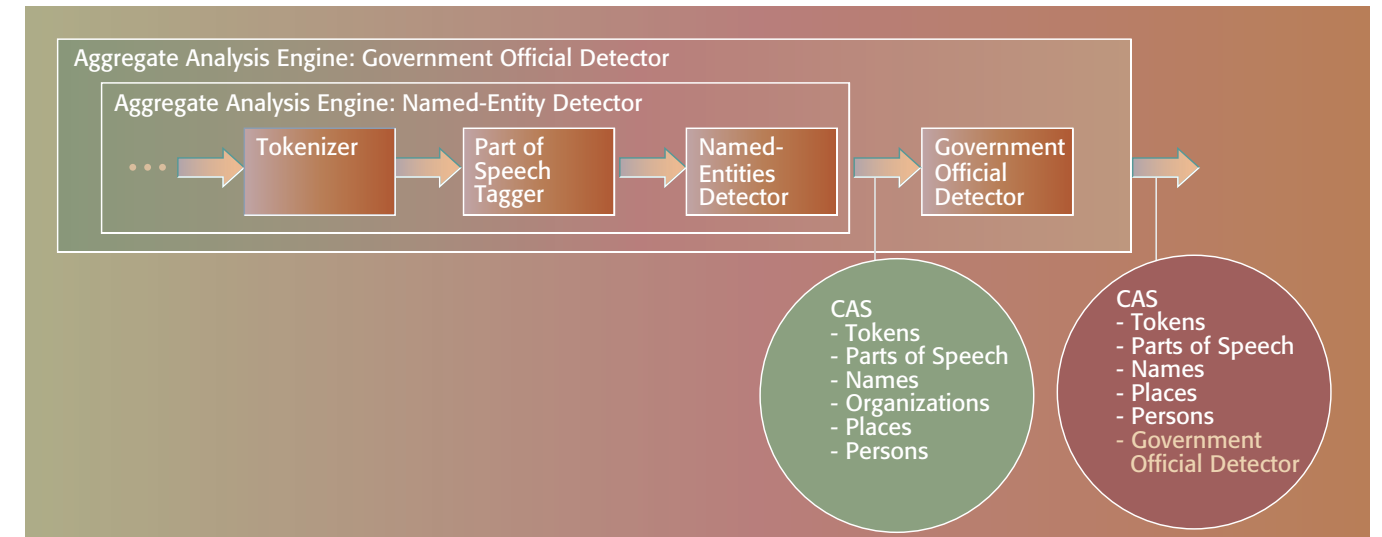


IBM Research has made many seminal contributions in Natural Language Processing (NLP). Notable among these are three paradigm shifts: the widely used conditional maximum entropy modeling framework; the first grammar-free, data-driven statistical parser; and statistical machine translation techniques. Today, our broad array of projects, built on several decades of NLP research, range from natural language understanding and generation to automatic translation, multimodal interaction, Web data mining, and next generation search. IBM Research's multi-disciplinary teams are addressing the challenges of helping users uncover exact but hard-to-find information in increasingly large, multilingual, and multimodal content collections. The impact of our NLP work can be seen in several IBM software products and services offerings that deliver Web-scale multilingual search and information discovery applications.

**INFORMATION ANALYTICS**

Historically, information management has focused on structured, tabular data suitable for relational databases. However, with increased availability of unstructured information modalities, such as varied forms of text, images, and video, the content, not immediately machine-accessible, must be extracted via a sequence of analytics. To facilitate the creation, composition, and deployment of a broad range of multimodal and multilingual analysis capabilities along with their integration into search, Research has developed the Unstructured Information Management Architecture (UIMA). This framework, already embedded in some IBM products, is now available through open source.



An encapsulation and composition of Text Analysis Engines. Adapted from *IBM Systems Journal*, Volume 43, No. 3, (2004), p. 459.

**LANGUAGE ANALYSIS AND INFORMATION EXTRACTION**

The extraction and structured organization of information from unstructured text is an important NLP research area. In addition to language processing modules with hand-coded symbolic knowledge, machine learning as well as statistical and finite-state techniques are now essential components of a hybrid toolkit for language analysis. Machine learning techniques provide scalable methods for shallow parsing, word sense disambiguation, named entity and topic detection and tracking, and relation extraction from unstructured data. Applications developed at Research include automatic content extraction systems for news stories in English, Chinese, and Arabic, and the Biological Text Knowledge Services system, a UIMA application for automatic analysis of medical abstracts, records, and patents.

**QUESTION-ANSWERING AND MULTIMODAL INTERACTION**

Question-answering and multimedia/multimodal conversational systems allow users to interact with information systems using natural language, graphical user interfaces, and gestures. Question-answering enables precise search through support of natural language questions or phrases as input. Multimodal systems can generate synchronized natural language and graphics presentations in response to a user's request. By taking advantage of complementary modalities, these systems are effective for information browsing and searching. IBM's research in these areas focuses on multi-agent approaches with multiple knowledge sources and answering strategies, robust and adaptive natural language understanding, and adaptation-based natural language generation.

**MACHINE TRANSLATION AND MULTILINGUAL PROCESSING**

With economic globalization and growth of multilingual information on the Internet, Machine Translation (MT) technology is an important business tool. IBM's MT research pursues two principal approaches: symbolic translation algorithms that make use of detailed linguistic information, including slot grammar deep parsing, and statistical machine translation algorithms that discover translation correspondences from large multilingual corpora and apply them by means of statistical decoding methods. Other applications related to MT technology include controlled language compliance checking and cross-lingual information retrieval, where, for example, queries in one language can be made against documents in a second language. In recent years, IBM's Bleu metric has emerged as the *de facto* standard for evaluating MT system performance.

