

Most of the world's information comes in the form of unstructured natural language text. Accessing that text in a meaningful way (beyond the keyword level) and extracting relevant content is inherently difficult, in large part because relevant concepts may be expressed in a wide variety of ways. The goal of our work is to **make relevant concepts in natural-language text easily accessible** to analysts or higher-order automated systems by developing a **suite of efficient, precise, and domain-adaptable tools** that implement a range of information extraction, synthesis, and display applications. You can try many of our tools at our web site, <http://l2r.cs.uiuc.edu/~cogcomp/demos.php>.

## Core Information Technologies

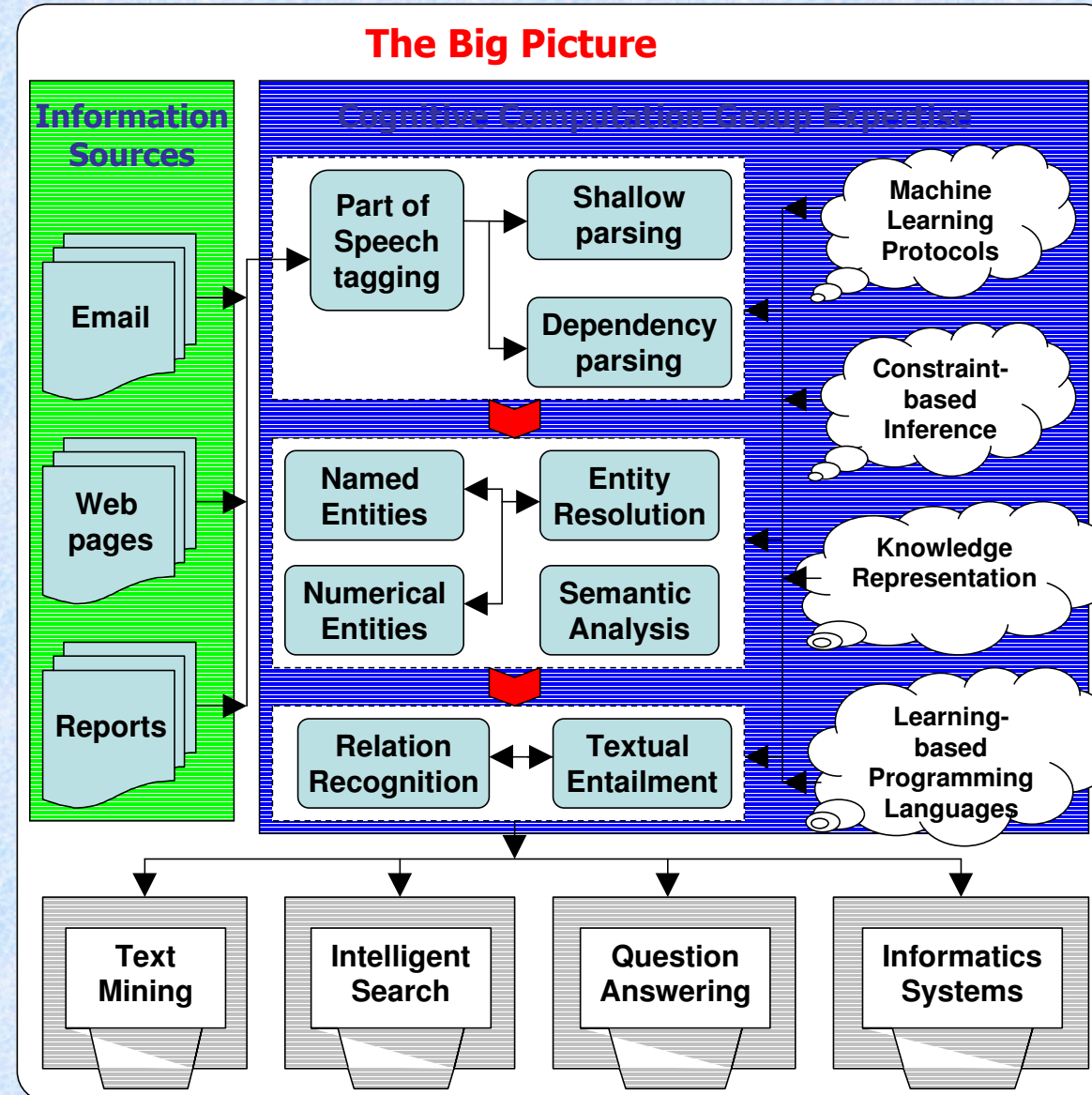
Although many systems tackle the problem of presenting large quantities of information efficiently and/or selectively, most assume their input comes from databases. However, **a huge percentage of data is not organized in databases, but instead written as free-form text.** The **Multimodal Information Access and Synthesis** center at the University of Illinois develops the machine learning theory, algorithms, and tools needed to **locate, access, organize and integrate this unstructured data.**

## Adaptable Technology

One-size-fits-all approaches seldom produce the best solutions for specialized applications, such as those required by government organizations. We base our work on **Machine Learning techniques** that allow our tools to be adapted to **new domains.** Members of our research team also develop new approaches to make such adaptation as **efficient and low-cost** as possible. To see an example, try our English-Russian Entity Discovery tool at the URL address above.

## Integrating Diverse Solutions

Many resources have been developed by the academic community that solve one sub-problem in isolation, such as syntactic parsing. However, **combining the output of tools that solve different problems** is extremely challenging. We have developed and implemented **principled, constraint-based approaches** to tackle this problem, such as our widely-used **Semantic Analysis system.**



## Concepts, not Keywords

A key problem that drives much of our current research is that even with the most sophisticated algorithms and powerful computers, simply using individual words (or even two-, three-, or five-word sequences) as the fundamental building blocks of understanding text cannot capture all relevant information in natural language text, because **meaning is highly dependent on context.** For example, "JFK" might refer to the former US president or the airport.

## Knowledge and Understanding

Working at the level of **concepts rather than words, while still using free-form text** as our input, allows us to directly represent knowledge and reasoning. One strand of our research we are investigating is a **hybrid natural language and logical approach** to text representation and understanding, to allow an automated system to recognize, for example, that "Smith gave Jones the means to destroy the bridge" implies that "Jones has dangerous ordnance equipment."

## Leveraging Human Expertise

We are investigating techniques that **allow human experts to efficiently train machine learning-based systems.** We are developing protocols that help the system to identify what it needs to know, so that it can ask for the minimal amount of guidance that achieves the maximum improvement in performance.

Try demonstrations of many of our tools at <http://l2r.cs.uiuc.edu/~cogcomp/demos.php>

We have a range of **established tools** that are **already widely used** in the Natural Language Processing research community, and many more under development. The selections presented below focus on concept-level analysis; one group focuses on recognizing entities, the other on relations between entities. You can try **real-time demonstrations** of these tools at our web site, <http://l2r.cs.uiuc.edu/~cogcomp/demos.php>, and download some of our software from <http://l2r.cs.uiuc.edu/~cogcomp/software.php>.

## Concept Recognition and Disambiguation

Concept recognition, a superset of Named Entity recognition, focuses on recognizing groups of words that represent a concept of interest. To integrate information from different sources, it is critical both to adapt the tools to the new domains, and to determine when different representations refer to the same underlying entity. This technology is critical to intelligent search, analyst support, and data integration.

### Named Entity Recognition

A critical first step in extracting meaning from text is to recognize groups of words that refer to entities of interest—for example, names of people, locations, and organizations, and numbers such as times, dates, and quantities. Two of our projects developed machine-learning-based tools for recognizing such entities: our **Named Entity Recognizer** and our **Number Quantization tool** directly annotate raw text to indicate boundaries and types of entities.

### Context-Sensitive Verb Paraphrasing

**Word sense disambiguation**, the problem of selecting the correct meaning of a word that has more than one interpretation, remains one of the more difficult text-analysis tasks. Our **Context-Sensitive Verb Paraphraser** takes a sentence containing an ambiguous verb or verb phrase, together with a candidate replacement verb, and determines whether this is a valid replacement. This tool uses a novel protocol that **minimizes the human annotation effort** needed to train the machine-learning algorithm by using text samples gathered from the World-wide Web.

### Multi-lingual Entity Discovery

Domain Adaptation is an extremely challenging task that is a major focus of the Natural Language Processing research community. If we train a named entity recognizer in one domain (e.g. English), can we adapt it with minimal effort to a new domain? Our **Multi-Lingual Named Entity Discovery** software uses automatically generated temporally aligned news articles in English and Russian to use our Named Entity Recognizer to learn the Russian names of entities of interest.

### Named Entity Resolution

Once you have identified entities of possible interest, you may need to collect information on that entity from a range of sources. But how can you tell whether different representations of the same name refer to the same person—for example, “George W. Bush” and “President Bush”? Our **Name Identification and Tracing tool** tackles this problem, retrieving relevant news items from a set of documents and indexes them with the entity originally specified, and with other entities that are related to the original individual.

## Relation Recognition and Reasoning

Relations are complex, abstract concepts that relate entities; for example, group membership, location, and possession are relations that can be expressed in a huge variety of ways, some of which are implicit and thus impossible to recognize from words alone (think of the location relation in “London’s Hard Rock Café”). Our Relation Recognition and Reasoning tools identify relations. Identifying relations is a cornerstone of Question Answering and Intelligent Search.

### Semantic Analysis

The need to work at the level of concepts and relations rather than individual words motivates the need for automated annotation tools that extract a level of information that is more abstract than the words themselves, but sufficiently generic and flexible that they are not restricted in scope to very limited domains.

Our **Semantic Analysis tool** (“**Semantic Role Labeler**”) detects verbs in sentences and, for each one, extracts the related entities and their respective roles (**Who did What to Whom, Where and When they did it**). For example, in the sentence “Mary was annoyed by the noise during her seminar,” it will detect that Mary was the entity annoyed, that the noise was what annoyed her, and that the time this occurred was during her seminar.

### Information Extraction

Our Information Extraction tools are trained to recognize specific relations, such as “Located In,” and concepts, such as “start time” and “title.” This demonstration extracts information from job listings and seminar announcements, and fills out pre-specified templates that indicate information of interest.

### Textual Entailment

The development of tools that work with concepts and relations is motivated by the need for **Intelligent Search—the ability to search based on meaning rather than on words used**. For example, if we wish to find documents that refer to, “The party that won the German election,” we may be interested in a document with the sentence, “The SPD garnered 32.6% of the popular vote, as opposed to the CDU’s 38.4%.” The task of Textual Entailment addresses the problem of recognizing when one text fragment implies another—in the example just given, that the CDU won the German Election, and hence is relevant to the query.

Our **Textual Entailment system** combines multiple levels of automated analysis, including that provided by our Semantic Analysis tool. It uses **hybridized Natural Language/Logical representation** with a suite of tools that can solve comparatively focused sentence simplification and comparison tasks. Used together, these tools can find a chain of reasoning that links two text fragments, if such a chain exists. It is designed to be extensible, so that as more resources are developed that can identify additional concepts and relations, they can be easily incorporated into the system.