parc®
Palo Alto Research Center

PARC Natural Language Processing

What is "natural language processing", and why does it matter?

A method of human-computer interaction, natural language processing enables computers to extract meaning from the words and phrases that people use – and respond in kind when presenting information back to them.

Natural language technologies convert human language into formal semantic representations which computer applications can interpret, act on, *and* respond with easily understood grammatical sentences. The result: people can interact more naturally with computer-based information – using normal, familiar expressions – instead of using carefully constructed computer jargon.

Combining the scalability and processing power of computers with natural language understanding leads to many highly useful applications: ranging from spell checking and fact checking, to automated question answering and knowledge-based reasoning systems.

Why has natural language processing been so difficult to achieve?

Resolving ambiguity is one of the most difficult tasks in natural language processing because computers lack "common sense".

Natural language processing requires analyzing underlying linguistic structures and relationships, grammatical rules, explicit concepts, implicit meanings, logic, discourse context, and more. Because individual words and sentences often have multiple meanings, and a single concept can be expressed in many different forms, a significant challenge in natural language processing is how to handle the ambiguity that arises when interpreting a single sentence. For example,

Seemingly similar sentences may differ radically in meaning:

- "The CEO was fired up about his new role."
- "The CEO was fired from his new role."

Seemingly different sentences can have the same meaning:

- "IBM's PC division was acquired by Lenovo."
- "Lenovo bought the PC division of IBM."

People who read these sentences can figure out the intended meanings by inferring from context or by drawing on personal knowledge and understanding. But computers don't benefit from the subtleties of human experience and learning, which is why determining intended meaning is such a difficult task in natural language processing.

What are the common approaches to natural language technologies?

Most natural language technologies trade off between broad-but-shallow OR deep-but-narrow approaches.

Many statistical and machine-learning approaches focus on breadth instead of depth. For example, standard word-occurrence methods use the words in a query to retrieve popular passages with as many possible matching words, and rank responses by different quality metrics. These systems find matching information by trading off deep understanding for frequent good guesses. Since they usually ignore things like word order and functional stop words (e.g., "of", "by", "this"), statistical approaches are shallow in their ability to process meaning or nuances. So people who ask this kind of system "who bought IBM" have to wade through false responses about *what IBM bought* – even though they only want to find responses about companies who bought IBM divisions or equipment. A language processing system based on word occurrence doesn't distinguish between these kinds of responses.

Other approaches focus on depth instead of breadth. These systems go for deeper understanding rather than being right "on-the-average". For example, by interpreting texts within built-in models and ontologies – which formally capture a specific knowledge domain – these approaches can map sentences into models in that domain and obtain responses by consulting the model. So in the query "I want to fly from Los Angeles to the Silicon Valley area", the system can consult a travel company's table for filling in a web page to transform the phrase "the Silicon Valley area" into the airport "SJC". Such model-based systems can represent the difference between IBM as a buyer, and IBM as seller. However, natural language systems built like this are narrow, because they're effective only in domains where models and ontologies have been created.

What differentiates PARC's approach to natural language?

PARC's hybrid approach provides broad and *deep natural language processing and retrieval – spanning syntax, semantics, context, knowledge.*

Unlike other approaches that work well on either keywords *or* specialized grammars, PARC's natural language platform uses a broad coverage grammar and general ontology to capture the subtleties of language use. PARC's approach combines:

- deep linguistic capabilities, engineered for efficiency;
- a highly tuned parser and interpretation framework with a unique ambiguity-management method;
- a knowledge representation designed to capture the essential content of text; and
- carefully crafted database retrieval and indexing that uses the content representation to locate relevant answers to questions.

Ambiguity management poses a significant problem in natural language systems, because processing all possible natural language meanings leads to exponential, time-consuming computations. PARC's method postpones ambiguity resolution while dramatically reducing the resources (e.g., time and space) required to resolve it. The system therefore isn't forced to pick the "right" solution up front, when it has less information and is less likely to be accurate. Since this approach preserves processing power *without* prematurely narrowing possible solutions, PARC's natural language systems don't trade off efficiency and accuracy.

In contrast, statistical parsers try to get around the ambiguity problem by using the highest-scoring or "best parse" as early as possible. Though the statistical methods are often correct, they don't take into account information inherent in a sentence or embedded in its context. In a head-to-head comparison with the best statistical parser on the field-standard data set, PARC's parser showed comparable speed, accuracy, and coverage [1].

The benefit: PARC's systems incorporate the precision of natural language understanding with equivalent efficiency, portability, and robustness of purely statistical approaches – and do much more.

PARC's approach normalizes the various ways of expressing the same meaning, so our natural language system:

- encodes synonyms (e.g., "buying" is the same as "purchasing");
- determines word classes (e.g., "buying" and "purchasing" both mean "acquiring");
- identifies claims for existence or non-existence of entities (events, people, places, dates, named concepts); and
- identifies *relationships* among these entities (who did what to whom).

How does PARC's approach work?

Building upon decades of scientific research and technology refinement, PARC has amassed several highly scalable and *high-performing natural language capabilities.*

The science of natural language processing involves artificial intelligence, linguistics, logic, computer science, statistics, and human-computer interaction – so PARC's unique, interdisciplinary approach is ideally suited for this work.

Furthermore, since PARC has long been engaged in making scientific breakthroughs that have significant commercial applications, our natural language researchers have uniquely combined work on foundational theories, computational frameworks, *and* robust technology implementations.

---

[1] Kaplan, R. M.; Riezler, S; King, T. H.; Maxwell, J. T.; Vasserman, A. 2004. Speed and accuracy in shallow and deep stochastic parsing. In Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04); May 2-7; Boston, MA.

PARC's contributions to the field of natural language processing – developed over three decades and continually enhanced and advanced – include:

- *Finite-state morphology (FSM).* PARC pioneered the theoretical concepts, algorithms, and engineering platforms that enabled finite-state machines to be used for linguistic descriptions. Today FSM is a standard technology used for spellchecking, identifying and classifying named entities, OCR language modeling, and information extraction.

- *Lexical functional grammar (LFG) and XLE parser and generator.* PARC designed the formal theory of LFG (in collaboration with Joan Bresnan at MIT/Stanford University). PARC created XLE, a robust implementation of LFG that consists of algorithms for efficiently parsing, generating, debugging, and rewriting broad-coverage grammars. Used by research institutions around the world, XLE is also the basis of the Parallel Grammar (ParGram) Project. The ParGram Project is developing industrial-strength grammars for English, as well as Arabic, Chinese, French, German, Japanese, Norwegian, Urdu, and other languages.

- *Disjunctive Constraint Satisfaction ambiguity-management method.* As pointed out earlier, resolving ambiguity is a major challenge. PARC's unique method maintains the set of possible interpretations *without* sacrificing accuracy and efficiency. The system can therefore utilize more information before making a choice, and even postpone choices until presented with a query – using the query itself as help in choosing the right interpretation. Implemented in XLE, this method enables the system to produce a tightly packed representation of all possible solutions as its output.

- *Knowledge representation and entailment-and-contradiction detection.* As part of a recent Intelligence Community-funded initiative, PARC researchers worked to bridge the gap between text-based information retrieval systems and logical formula-based knowledge representation and reasoning systems. Information retrieval systems can find relevant natural language *passages*, but not answers to questions. Knowledge representation and reasoning systems can provide answers, but only with *formal* input. So PARC researchers built a system that *automatically maps text to abstract knowledge representations.* PARC's entailment and contradiction engine maps a query into this abstract knowledge representation, retrieves texts with similar representation from a database, and tries to determine if the text answers the query.

Together, these highly scalable and high-performing core capabilities let PARC operate at a much deeper level where computers systematically process not just language, but *meaning*.

Where does PARC focus its natural language research and development?

By working on a pipeline of technologies – bridging surface language analysis to deep knowledge representation – PARC researchers have provided robust solutions today while enabling new applications for tomorrow.

Semantic search and question-answering focuses on understanding queries expressed in natural language and retrieving information to answer them. While other approaches approximate answers by matching isolated passages and calculating the *most likely* information, PARC's approach retrieves the *most accurate and most relevant* search results or answers by focusing on underlying meaning. This means that people don't have to waste their time scanning through a long list of popularity-skewed guesses and vague snippets: they can more efficiently locate the precise information they need. Instead of being limited to "good enough" results, people can expect the best results.

Large-scale information understanding focuses on understanding embedded concepts, overall sentiment, and relationships in massive document collections. For example, by semantically indexing their specialized document collections, corporations can more efficiently and productively mine text, identify document clusters, locate redundancies or contradictions, and understand trends.

Deep content analysis and sensemaking focuses on understanding concepts, identifying relationships among entities, creating more meaningful text summarizations, enabling inference beyond the text, and differentiating facts versus beliefs. An important element of "making sense" of massive, diverse, unstructured information collections involves presenting information in a way that can be adapted for specific interests. Since PARC's approach formalizes underlying semantic representations, people can name and select concepts of interest (i.e., what they're trying to make sense of) – and thus receive information specifically tailored to them. Some of this work was developed for government intelligence analysts, and includes cutting-edge information visualization and indexing technologies from PARC's user-interface/human-information interaction research groups.

Intelligent redaction semi-automatically highlights text segments of interest, and also identifies segments that need to be protected. By applying natural language techniques to detect entities and entity relationships in texts, people can find important or sensitive text segments more efficiently. When combined with the work of PARC's security and user-interface researchers, this technology allows different people to access different portions of the document based on their specific information needs and roles.

Understanding dialogue could enable better human-computer interaction and natural language processing in specialized domains.

Where are PARC's natural language technologies in commercial use today, and what are the commercialization opportunities?

PARC's natural language processing technologies are implemented in many products. PARC continues to create new intellectual property and to commercialize applications in new fields of use.

Our intellectual property portfolio holds more than 60 patents in natural language processing research. Various spin-offs and licensees have commercialized more than 100 other PARC patents.

PARC natural language technologies have been embedded in a wide range of commercial products offered by Xerox; Scansoft (now Nuance); and Microlytics (which licensed to Microsoft, Symantec, Apple, WordPerfect, Fuji Xerox, and others). In 1997, PARC spun out Inxight Software, Inc. By transforming text into actionable information, Inxight provides scalable, multilingual, information-discovery solutions to customers such as Boeing, Charles Schwab, Deutsche Telekom, Eli Lilly, Factiva, GlaxoSmithKline, Korean Telecom, LexisNexis, MCI, Merrill Lynch, Microsoft, Oracle, PricewaterhouseCoopers, Reuters, SAP, Yahoo, and various U.S. government agencies.

Recently, PARC signed a licensing and collaboration agreement with Powerset to develop and commercialize natural language technology for a *consumer search* engine. The agreement between Powerset and PARC includes technology and patent licenses, as well as long-term research collaboration.

ABOUT PARC:
Transforming the ways in which advanced research creates business value

PARC has tackled some of the most difficult natural language problems and developed cutting-edge technologies to address them. With a proven ability to *instantiate theories, build industrial-strength prototypes,* and *anticipate market needs,* PARC continues to attract businesses seeking to obtain or extend their natural language capabilities and deliver innovative applications.

PARC partners with both global organizations and with promising start-ups to discover and commercialize breakthrough technology and business concepts that solve real needs, and transform how enterprises deliver value to customers. PARC's physical, computer, biological, and social scientists take an agile, cross-disciplinary approach to innovation, with the vision, expertise, and instinct to convert scientific findings into industrial-strength prototypes. Founded in 1970 as part of Xerox Research, PARC was incorporated in 2002 as an independent, advanced R&D organization. PARC is a wholly owned subsidiary of Xerox Corporation.

Interested in learning more?

Media contact: Linda Jacobson, Manager, Marketing & Communications, ljacobson@parc.com or 650-812-4035

Business contact: Lawrence Lee, Director of Business Development, Intelligent Systems Laboratory, lawrence.lee@parc.com or 650-812-4756

www.parc.com/nlp