Natural Language Processing (NLP)

For other uses, see NLP. This article is about natural language processing done by computers. For the natural language processing done by the human brain, see Language processing in the brain.

Natural language processing (NLP) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e., statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

History

Natural language processing has its roots in the 1940s. Already in 1940, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence, though at the time that was not articulated as a problem separate from artificial intelligence. The proposed test includes a task that involves the automated interpretation and generation of natural language.

Symbolic NLP (1950s – early 1990s)

The premise of symbolic NLP is well-summarized by John Searle's Chinese room experiment: Given a collection of rules (e.g., a Chinese phrasebook, with questions and matching answers), the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it confronts.

- **1950s:** The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem. However, real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted in America (though some research continued elsewhere, such as Japan and Europe) until the late 1980s when the first statistical machine translation systems were developed.
- **1960s:** Some notably successful natural language processing systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction.
- **1970s:** During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data.
- **1980s:** The 1980s and early 1990s mark the heyday of symbolic methods in NLP. Focus areas of the time included research on rule-based parsing, morphology, semantics, reference, and other areas of natural language understanding.

Statistical NLP (1990s–2010s)

Up until the 1980s, most natural language processing systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing.

- **1990s:** Many of the notable early successes on statistical methods in NLP occurred in the field of machine translation, due especially to work at IBM Research, such as IBM alignment models. These systems were able

to take advantage of existing multilingual textual corpora that had been produced by the Parliament of Canada and the European Union as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding systems of government.

- **2000s:** With the growth of the web, increasing amounts of raw language data has become available since the mid-1990s. Research has thus increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data.

Neural NLP (present)

In 2003, the word n-gram model, at the time the best statistical algorithm, was overperformed by a multi-layer perceptron with a single hidden layer and context length of several words trained on up to 14 million of words with a CPU cluster in language modeling.

In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing. This popularity was due partly to a flurry of results showing that such techniques can achieve state-of-the-art results in many natural language tasks.

Common NLP tasks

The following is a list of some of the most commonly researched tasks in natural language processing. Some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks.

- **Text and speech processing**
    - Optical character recognition (OCR)
    - Speech recognition
    - Speech segmentation
    - Text-to-speech
    - Word segmentation (Tokenization)
    - Morphological analysis
- **Syntactic analysis**
    - Grammar induction
    - Sentence breaking
    - Parsing
    - Lexical semantics (of individual words in context)
- **Relational semantics (semantics of individual sentences)**
    - Relationship extraction
    - Semantic parsing
    - Semantic role labeling
    - Coreference resolution
- **Discourse (semantics beyond individual sentences)**
    - Coreference resolution
    - Discourse analysis

    - Implicit semantic role labeling
    - Recognizing textual entailment
- **Higher-level NLP applications**
    - Automatic summarization (text summarization)
    - Grammatical error correction
    - Logic translation
    - Machine translation (MT)
    - Natural-language understanding (NLU)
    - Natural-language generation (NLG)

General tendencies and (possible) future directions

Based on long-standing trends in the field, it is possible to extrapolate future directions of NLP. As of 2020, three trends among the topics of the long-standing series of CoNLL Shared Tasks can be observed:

- Interest on increasingly abstract, "cognitive" aspects of natural language
- Increasing focus on multilinguality, including, to some extent, cross-linguality
- The recent shift from "shallow" to deep architectures and the accompanying large increase in quantitative evaluation of different methods using corpora such as the Penn Treebank and similar, although still small by comparison with corpora acquired from the web and through unsupervised or semi-supervised methods.

The subfield of NLP devoted to learning approaches in NLP is known as Natural Language Learning (NLL) and its conference CoNLL and peak body SIGNLL are sponsored by ACL, recognizing also their links with computational linguistics and language acquisition. The Conference on Computational Natural Language Learning (CoNLL) and Empirical Methods in Natural Language Processing (EMNLP) are closely related conferences which both have a significant learning emphasis.