Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language.

NLP enables computers and digital devices to recognize, understand and generate text and speech by combining computational linguistics, the rule-based modeling of human language together with statistical modeling, machine learning and deep learning.

NLP research has helped enable the era of generative AI, from the communication skills of large language models (LLMs) to the ability of image generation models to understand requests. NLP is already part of everyday life for many, powering search engines, prompting chatbots for customer service with spoken commands, voice-operated GPS systems and question-answering digital assistants on smartphones such as Amazon's Alexa, Apple's Siri and Microsoft's Cortana.

NLP also plays a growing role in enterprise solutions that help streamline and automate business operations, increase employee productivity and simplify business processes.

AI Training Ground Natural Language Processing (2:33 min)

Benefits of NLP

NLP makes it easier for humans to communicate and collaborate with machines, by allowing them to do so in the natural human language they use every day. This offers benefits across many industries and applications.

- Automation of repetitive tasks
- Improved data analysis and insights
- Enhanced search
- Content generation

Automation of repetitive tasks

NLP is especially useful in fully or partially automating tasks like customer support, data entry and document handling. For example, NLP-powered chatbots can handle routine customer queries, freeing up human agents for more complex issues. In document processing, NLP tools can automatically classify, extract key information and summarize content, reducing the time and errors associated with manual data handling. NLP facilitates language translation, converting text from one language to another while preserving meaning, context and nuances.

Improved data analysis

NLP enhances data analysis by enabling the extraction of insights from unstructured text data, such as customer reviews, social media posts and news articles. By using text mining techniques, NLP can identify patterns, trends and sentiments that are not immediately obvious in large datasets. Sentiment analysis enables the extraction of subjective qualities, attitudes, emotions, sarcasm, confusion or suspicion from text. This is often used for routing communications to the system or the person most likely to make the next response.

This allows businesses to better understand customer preferences, market conditions and public opinion. NLP tools can also perform categorization and summarization of vast amounts of text, making it easier for analysts to identify key information and make data-driven decisions more efficiently.

Enhanced search

NLP benefits search by enabling systems to understand the intent behind user queries, providing more accurate and contextually relevant results. Instead of relying solely on keyword matching, NLP-powered search engines analyze the meaning of words and phrases, making it easier to find information even when queries are vague or complex. This improves user experience, whether in web searches, document retrieval or enterprise data systems.

Powerful content generation

NLP powers advanced language models to create human-like text for various purposes. Pre-trained models, such as GPT-4, can generate articles, reports, marketing copy, product descriptions and even creative writing based on prompts provided by users. NLP-powered tools can also assist in automating tasks like drafting emails, writing social media posts or legal documentation. By understanding context, tone and style, NLP sees to it that the generated content is coherent, relevant and aligned with the intended message, saving time and effort in content creation while maintaining quality.

Approaches to NLP

NLP combines the power of computational linguistics together with machine learning algorithms and deep learning. Computational linguistics uses data science to analyze language and speech. It includes two main types of analysis: syntactical analysis and semantical analysis. Syntactical analysis determines the meaning of a word, phrase or sentence by parsing the syntax of the words and applying preprogrammed rules of grammar. Semantical analysis uses the syntactic output to draw meaning from the words and interpret their meaning within the sentence structure.

The parsing of words can take one of two forms. Dependency parsing looks at the relationships between words, such as identifying nouns and verbs, while constituency parsing then builds a parse tree (or syntax tree): a rooted and ordered representation of the syntactic structure of the sentence or string of words. The resulting parse trees underly the functions of language translators and speech recognition. Ideally, this analysis makes the output either text or speech understandable to both NLP models and people.

Self-supervised learning (SSL) in particular is useful for supporting NLP because NLP requires large amounts of labeled data to train AI models. Because these labeled datasets require time-consuming annotation, a process involving manual labeling by humans, gathering sufficient data can be prohibitively difficult. Self-supervised approaches can be more time-effective and cost-effective, as they replace some or all manually labeled training data.

Three different approaches to NLP include:

Rules-based NLP

The earliest NLP applications were simple if-then decision trees, requiring preprogrammed rules. They are only able to provide answers in response to specific prompts, such as the original version of Moviefone, which had rudimentary natural language generation (NLG) capabilities. Because there is no machine learning or AI capability in rules-based NLP, this function is highly limited and not scalable.

Statistical NLP

Developed later, statistical NLP automatically extracts, classifies and labels elements of text and voice data and then assigns a statistical likelihood to each possible meaning of those elements. This relies on machine learning, enabling a sophisticated breakdown of linguistics such as part-of-speech tagging.

Statistical NLP introduced the essential technique of mapping language elements, such as words and grammatical rules to a vector representation so that language can be modeled by using mathematical (statistical) methods, including regression or Markov models. This informed early NLP developments such as spellcheckers and T9 texting (Text on 9 keys, to be used on Touch-Tone telephones).

Deep learning NLP

Recently, deep learning models have become the dominant mode of NLP, by using huge volumes of raw, unstructured data both text and voice to become ever more accurate. Deep learning can be viewed as a further evolution of statistical NLP, with the difference that it uses neural network models. There are several subcategories of models:

- *Sequence-to-Sequence* (seq2seq) models: Based on recurrent neural networks (RNN), they have mostly been used for machine translation by converting a phrase from one domain (such as the German language) into the phrase of another domain (such as English).
- *Transformer models*: They use tokenization of language (the position of each token words or subwords) and self-attention (capturing dependencies and relationships) to calculate the relation of different language parts to one another. Transformer models can be efficiently trained by using self-supervised learning on massive text databases. A landmark in transformer models was

Google's bidirectional encoder representations from transformers (BERT), which became and remains the basis of how Google's search engine works.

- *Autoregressive models*: This type of transformer model is trained specifically to predict the next word in a sequence, which represents a huge leap forward in the ability to generate text. Examples of autoregressive LLMs include GPT, Llama, Claude and the open-source Mistral.
- *Foundation models*: Prebuilt and curated foundation models can speed the launching of an NLP effort and boost trust in its operation. For example, the IBM® Granite™ foundation models are widely applicable across industries. They support NLP tasks including content generation and insight extraction. Additionally, they facilitate retrieval-augmented generation, a framework for improving the quality of response by linking the model to external sources of knowledge. The models also perform named entity recognition which involves identifying and extracting key information in a text.

NLP Tasks

Several NLP tasks typically help process human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include:

- Coreference resolution
- Named entity recognition
- Part-of-speech tagging
- Word sense disambiguation

Coreference resolution

This is the task of identifying if and when two words refer to the same entity. The most common example is determining the person or object to which a certain pronoun refers (such as "she" = "Mary"). But it can also identify a metaphor or an idiom in the text (such as an instance in which "bear" isn't an animal, but a large and hairy person).

Named entity recognition (NER)

NER identifies words or phrases as useful entities. NER identifies "London" as a location or "Maria" as a person's name.

Part-of-speech tagging

Also called grammatical tagging, this is the process of determining which part of speech a word or piece of text is, based on its use and context. For example, part-of-speech identifies "make" as a verb in "I can make a paper plane," and as a noun in "What make of car do you own?"

Word sense disambiguation

This is the selection of a word meaning for a word with multiple possible meanings. This uses a process of semantic analysis to examine the word in context. For example, word sense disambiguation helps distinguish the meaning of the verb “make” in “make the grade” (to achieve) versus “make a bet” (to place). Sorting out “I will be merry when I marry Mary” requires a sophisticated NLP system.

How NLP works

NLP works by combining various computational techniques to analyze, understand and generate human language in a way that machines can process. Here is an overview of a typical NLP pipeline and its steps:

Text preprocessing

NLP text preprocessing prepares raw text for analysis by transforming it into a format that machines can more easily understand. It begins with tokenization, which involves splitting the text into smaller units like words, sentences or phrases. This helps break down complex text into manageable parts. Next, lowercasing is applied to standardize the text by converting all characters to lowercase, ensuring that words like "Apple" and "apple" are treated the same. Stop word removal is another common step, where frequently used words like "is" or "the" are filtered out because they don't add significant meaning to the text. Stemming or lemmatization reduces words to their root form (e.g., "running" becomes "run"), making it easier to analyze language by grouping different forms of the same word. Additionally, text cleaning removes unwanted elements such as punctuation, special characters and numbers that may clutter the analysis.

After preprocessing, the text is clean, standardized and ready for machine learning models to interpret effectively.

Feature extraction

Feature extraction is the process of converting raw text into numerical representations that machines can analyze and interpret. This involves transforming text into structured data by using NLP techniques like Bag of Words and TF-IDF, which quantify the presence and importance of words in a document. More advanced methods include word embeddings like Word2Vec or GloVe, which represent words as dense vectors in a continuous space, capturing semantic relationships between words. Contextual embeddings further enhance this by considering the context in which words appear, allowing for richer, more nuanced representations.

Text analysis

Text analysis involves interpreting and extracting meaningful information from text data through various computational techniques. This process includes tasks such as part-of-speech (POS) tagging, which identifies grammatical roles of words and named entity recognition (NER), which detects specific entities like names, locations and dates. Dependency parsing analyzes grammatical relationships between words to understand sentence structure, while sentiment analysis determines the emotional tone of the text, assessing whether it is positive, negative or neutral. Topic modeling identifies underlying themes or topics within a text or across a corpus of documents. Natural language understanding (NLU) is a subset of NLP that focuses on analyzing the meaning behind sentences. NLU enables software to find similar meanings in different sentences or to process words that have different meanings. Through these techniques, NLP text analysis transforms unstructured text into insights.

Model training

Processed data is then used to train machine learning models, which learn patterns and relationships within the data. During training, the model adjusts its parameters to minimize errors and improve its performance. Once trained, the model can be used to make predictions or generate outputs on new, unseen data. The effectiveness of NLP modeling is continually refined through evaluation, validation and fine-tuning to enhance accuracy and relevance in real-world applications.

Different software environments are useful throughout the said processes. For example, the Natural Language Toolkit (NLTK) is a suite of libraries and programs for English that is written in the Python programming language. It supports text classification, tokenization, stemming, tagging, parsing and semantic reasoning functionalities. TensorFlow is a free and open-source software library for machine learning and AI that can be used to train models for NLP applications. Tutorials and certifications abound for those interested in familiarizing themselves with such tools.

Challenges of NLP

Even state-of-the-art NLP models are not perfect, just as human speech is prone to error. As with any AI technology, NLP comes with potential pitfalls. Human language is filled with ambiguities that make it difficult for programmers to write software that accurately determines the intended meaning of text or voice data. Human language might take years for humans to learn and many never stop learning. But then programmers must teach natural language-powered applications to recognize and understand irregularities so their applications can be accurate and useful. Associated risks might include:

Biased training

As with any AI function, biased data used in training will skew the answers. The more diverse the users of an NLP function, the more significant this risk becomes, such as in government services, healthcare and HR interactions. Training datasets scraped from the web, for example, are prone to bias.

Misinterpretation

As in programming, there is a risk of garbage in, garbage out (GIGO). Speech recognition, also known as speech-to-text, is the task of reliably converting voice data into text data. But NLP solutions can become confused if spoken input is in an obscure dialect, mumbled, too full of slang, homonyms, incorrect grammar, idioms, fragments, mispronunciations, contractions or recorded with too much background noise.

New vocabulary

New words are continually being invented or imported. The conventions of grammar can evolve or be intentionally broken. In these cases, NLP can either make a best guess or admit it’s unsure and either way, this creates a complication.

Tone of voice

When people speak, their verbal delivery or even body language can give an entirely different meaning than the words alone. Exaggeration for effect, stressing words for importance or sarcasm can be confused by NLP, making the semantic analysis more difficult and less reliable.

NLP use cases by industry

NLP applications can now be found across virtually every industry.

Finance

In financial dealings, nanoseconds might make the difference between success and failure when accessing data, or making trades or deals. NLP can speed the mining of information from financial statements, annual and regulatory reports, news releases or even social media.

Healthcare

New medical insights and breakthroughs can arrive faster than many healthcare professionals can keep up. NLP and AI-based tools can help speed the analysis of health records and medical research papers, making better-informed medical decisions possible, or assisting in the detection or even prevention of medical conditions.

Insurance

NLP can analyze claims to look for patterns that can identify areas of concern and find inefficiencies in claims processing, leading to greater optimization of processing and employee efforts.

Legal

Almost any legal case might require reviewing mounds of paperwork, background information and legal precedent. NLP can help automate legal discovery, assisting in the organization of information, speeding review and making sure that all relevant details are captured for consideration.