What is Natural Language Processing?

IBM Cloud Education

2 July 2020

1 Natural Language Processing (NLP)

Natural language processing strives to build machines that understand and respond to text or voice data—and respond with text or speech of their own—in much the same way humans do.

> natural language processing means machines that...
>
> - understand text or speech
> - respond with text or speech

2 What is natural language processing?

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

> $NLP \subset AI$

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

- elements of NLP...

  computational linguistics (rule-based modeling)
  statistical learning
  machine learning
  deep learning
  \+ natural language processing

- extract from spoken or written language...
  - meaning
  - intent
  - sentiment

NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

- functions
  - translate one natural language to another (e.g., English → Chinese)
  - respond to spoken commands
  - summarize text
- examples
  - voice-operated GPS
  - speech-to-text dictation
  - customer service chatbots
- goals / benefits
  - streamline business operations
  - increase employee productivity
  - simplify mission-critical processes

3 NLP tasks

Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions, variations in sentence structure—these just a few of the irregularities of human language that take humans years to learn, but that programmers must teach natural language-driven applications to recognize and understand accurately from the start, if those applications are going to be useful.

- challenges
  - ambiguity (multiple meanings)
  - homonyms, homophones, homographs
    - same sound / different meaning (to, too, two)
    - same spelling / different sound (bow of ship, bow of violin)
    - different spelling / same sound (enough, fluff)
  - sarcasm
  - idioms (miss the boat)
  - metaphors (he is the black sheep of the family)
  - variations in the structure of sentences
  - exceptions to rules of grammar, usage

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

Speech recognition , also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar.

Part of speech tagging , also called grammatical tagging, is the process of determining the part of speech of a particular word or piece of text based on its use and context. 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 is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps distinguish the meaning of the verb

'make' in 'make the grade' (achieve) vs. 'make a bet' (place).

Named entity recognition , or NEM, identifies words or phrases as useful entities. NEM identifies 'Kentucky' as a location or 'Fred' as a man's name.

Co-reference resolution 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 (e.g., 'she' = 'Mary'), but it can also involve identifying a metaphor or an idiom in the text (e.g., an instance in which 'bear' isn't an animal but a large hairy person).

Sentiment analysis attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.

Natural language generation is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language.

- recognize speech when the speaker...
  - speaks quickly
  - speaks with an accent
  - slurs words
  - breaks grammatical rules
  - varies intonation (speed and volume)
- determine meaning from context ('make the grade' vs. 'make a bet')
- categorize words ('Iowa' is a state, 'goldfinch' is a bird)
- match pronouns with nouns (to which person does the word 'he' or 'she' refer?)
- determine a speaker's or writer's mood (confused, skeptical, pleased, angry, etc.)
- generate speech

See the blog post "NLP vs. NLU vs. NLG: the differences between three natural language processing concepts" for a deeper look into how these concepts relate.

4 NLP tools and approaches

4.1 Python and the Natural Language Toolkit (NLTK)

The Python programing language provides a wide range of tools and libraries for attacking specific NLP tasks. Many of these are found in the Natural Language Toolkit, or NLTK, an open source collection of libraries, programs, and education resources for building NLP programs.

The NLTK includes libraries for many of the NLP tasks listed above, plus libraries for subtasks, such as sentence parsing, word segmentation, stemming and lemmatization (methods of trimming words down to their roots), and tokenization (for breaking phrases, sentences, paragraphs and passages into tokens that help the computer better understand the text). It also includes libraries for implementing capabilities such as semantic reasoning, the ability to reach logical conclusions based on facts extracted from text.

4.2 Statistical NLP, machine learning, and deep learning

The earliest NLP applications were hand-coded, rules-based systems that could perform certain NLP tasks, but couldn't easily scale to accommodate a seemingly endless stream of exceptions or the increasing volumes of text and voice data.

- programmers translated language rules into code
- but too many rules!
- and too many exceptions in our languages!
- clients want software that would work with lots of long, varied inputs

Enter statistical NLP, which combines computer algorithms with machine learning and deep learning models to automatically extract, classify, and label elements of text and voice data and then assign a statistical likelihood to each possible meaning of those elements. Today, deep learning models and learning techniques based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) enable NLP systems that 'learn' as they work and extract ever more accurate meaning from huge volumes of raw, unstructured, and unlabeled text and voice data sets.

- statistical NLP = algorithms + machine learning + deep learning
- extract, classify, label elements of spoken/written language
- assign likelihood to each possible meaning
- convolutional neural networks (CNNs)
- recurrent neural networks (RNNs)
- software handles large, unstructured, unlabeled inputs

For a deeper dive into the nuances between these technologies and their learning approaches, see "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?"

5 NLP use cases

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:

Spam detection: You may not think of spam detection as an NLP solution, but the best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more. Spam detection is one of a handful of NLP problems that experts consider 'mostly solved' (although you may argue that this doesn't match your email experience).

Machine translation: Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another. Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language. Machine translation tools are making good progress in terms of accuracy. A great way to test any machine translation tool is to translate text to one language and then back to the original. An oft-cited classic example: Not long ago, translating "The spirit is willing but the flesh is weak" from English to Russian and back yielded "The vodka is good but the meat is rotten." Today, the result is "The spirit desires, but the flesh is weak," which isn't perfect, but inspires much more confidence in the English-to-Russian translation.

Virtual agents and chatbots: Virtual agents such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice com-

mands and natural language generation to respond with appropriate action or helpful comments. Chatbots perform the same magic in response to typed text entries. The best of these also learn to recognize contextual clues about human requests and use them to provide even better responses or options over time. The next enhancement for these applications is question answering, the ability to respond to our questions—anticipated or not—with relevant and helpful answers in their own words.

Social media sentiment analysis: NLP has become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

Text summarization: Text summarization uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text. The best text summarization applications use semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

Use cases:

- detect spam (a mostly solved problem?)
- machine translation (it's getting better!)
- virtual agents ("Siri, play 'Haven't met you yet' by Michael Bublé")
- analyze sentiments (for example, look at blog posts to determine whether or not citizens favor a proposed new law)
- summarize text (for example, summarize newspaper articles)

6 Natural language processing and IBM Watson

IBM has innovated in the artificial intelligence space by pioneering NLP-driven tools and services that enable organizations to automate their complex business processes while gaining essential business insights. These tools include:

- Watson Discovery - Surface high-quality answers and rich insights from your complex enterprise documents - tables, PDFs, big data and more - with AI search. Enable your employees to make more informed decisions and save time with real-time search engine and text mining capabilities

that perform text extraction and analyze relationships and patterns buried in unstructured data. Watson Discovery leverages custom NLP models and machine learning methods to provide users with AI that understands the unique language of their industry and business. Explore Watson Discovery

- Watson Natural Language Understanding (NLU) - Analyze text in unstructured data formats including HTML, webpages, social media, and more. Increase your understanding of human language by leveraging this natural language tool kit to identify concepts, keywords, categories, semantics, and emotions, and to perform text classification, entity extraction, named entity recognition (NER), sentiment analysis, and summarization. Explore Watson Natural Language Understanding

- Watson Assistant - Improve the customer experience while reducing costs. Watson Assistant is an AI chatbot with an easy-to-use visual builder so you can deploy virtual agents across any channel, in minutes. Explore Watson Assistant

- Purpose-built for healthcare and life sciences domains, IBM Watson Annotator for Clinical Data extracts key clinical concepts from natural language text, like conditions, medications, allergies and procedures. Deep contextual insights and values for key clinical attributes develop more meaningful data. Potential data sources include clinical notes, discharge summaries, clinical trial protocols and literature data.

IBM's NLP products:

Watson Discovery finds relationships and patterns in a businesses documents (PDFs, spreadsheets), data mining and real-time search, helps business make more informed decisions

Watson Natural Language Understanding extracts meanings (keywords, categories, emotions, semantics) from webpages, social media

Watson Assistant chatbot to serve customers, reduce vendor's costs

Watson Annotator for Clinical Data is a medical application—looks for patients' conditions, allergies, medications and so on in doctor's notes, journal articles

For more information on how to get started with one of IBM Watson's natural language processing technologies, visit the IBM Watson Natural Language Processing page.