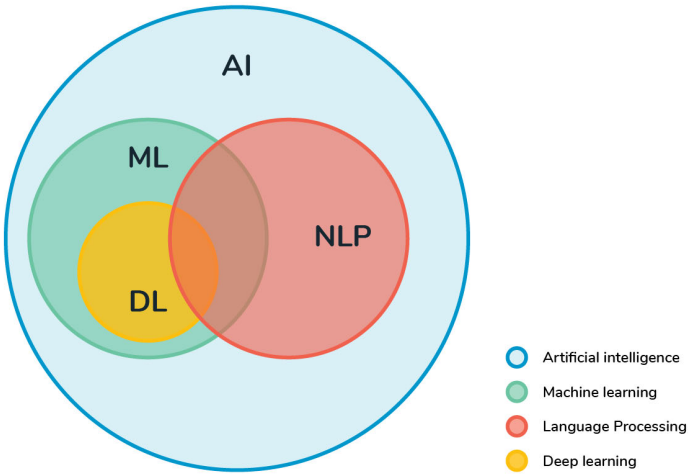
Customise appearance

Natural Language Processing



NLP is a subset of AI and uses ML/DL techniques. Source: Sathiyakugan 2018.

In computer science, languages that humans use to communicate are called "natural languages". Examples include English, French, and Spanish. Early computers were designed to solve equations and process numbers. They were not meant to understand natural languages. Computers have their own programming languages (C, Java, Python) and communication protocols (TCP/IP, HTTP, MQTT).

To instruct computers to perform tasks, we traditionally use a keyboard or a mouse. Why not speak to the computer and let it respond in a natural language? This is one of the aims of Natural Language Processing (NLP). NLP is an essential component of artificial intelligence.

NLP is rooted in the theory of linguistics. Techniques from machine learning and deep neural networks have also been successfully applied to NLP problems. While many practical applications of NLP already exist, NLP has many unsolved problems.

Discussion

- Why do computers have difficulty with NLP?

NLP has to parse unstructured textual content to extract useful information. Source: Waldron 2015.

Computers have mostly been dealing with **structured data**. This is data that's organized, indexed and referenced, often in databases. In NLP, we often deal with **unstructured data**. Social media posts, news articles, emails, and product reviews are examples of text-based unstructured data. To process such text, NLP has to learn the structure and grammar of the natural language. Importantly, 80% of enterprise data is unstructured.

Human languages are quite unlike the precise and unambiguous nature of computer languages. Human languages have plenty of complexities such as ambiguous phrases, colloquialisms, metaphors, puns, or sarcasms. The same word or text can have multiple meanings depending on the context. Language evolves with time. Worse still, we communicate imperfectly (spelling, grammar or punctuation errors) but still manage to be understood. These variations, so natural to human communication, are complex for computers.

Ambiguities in natural languages can be classified as lexical, syntactic or referential.

When the source of information is speech, more challenges arise: accent, tone, loudness, background noise or context, pronunciation, emotional content, pauses, and so on.

- Could you share some examples of the complexities of English? Consider the sentence, "One morning I shot an elephant in my pajamas". The man was in his pajamas but grammatically it's also

correct to think that the elephant was wearing his pajamas. Likewise, a person may say, "Listening to loud music slowly gives me a headache". Was she listening to music slowly or does the headache develop slowly?

A more confusing example is, "The complex houses married and single soldiers and their families". Confusion arises because we may initially interpret "complex houses" as an adjective-noun combination. The sentence makes sense only when we see that "complex" is a noun and "houses" is a verb. NLP addresses this via part-of-speech tagging.

Consider this one, "John had a card for Helga, but couldn't deliver it because he was in her way". Was John was in Helga's way? In fact, "he" refers to an earlier reference to a third person. NLP calls this coreference resolution.

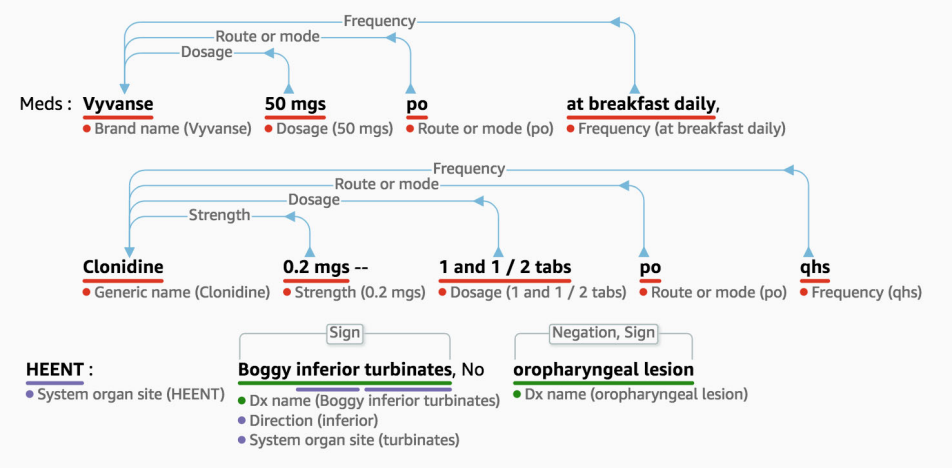
"The Kiwis won the match" is an example that requires context to make sense. New Zealand nationals are referred to as "Kiwis", after their national bird. Natural language is full of metaphors like this.

- What are some example problems that NLP can solve?
  From the number of problems that NLP solves, we describe a few:
  - **Sentiment Analysis**: From product reviews or social media messages, the task is to figure out if the sentiment is positive, neutral or negative. This is useful for customer support, engineering and marketing departments.
  - **Machine Translation**: Suppose original content is published only in one language, machine translation can deliver this content to a wider readership. Tourists can use machine translation to communicate in a foreign country.
  - **Question Answering**: Given a question, an NLP engine leveraging a vast body of knowledge, can provide answers. This can help researchers and journalists. Whitepapers and reports can be written faster.
  - **Text Summarization**: NLP can be tasked to summarize a long essay or an entire book. It can provide a balanced summary of a

story published at different websites with different points of view.

- **Text Classification**: NLP can classify news stories by domain or detect email spam.
- **Text-to-Speech**: This is an essential aspect of voice assistants. Audiobooks can be created for the visually impaired. Public announcements can be made.
- **Speech Recognition**: Opposite of text-to-speech, this creates a textual representation of speech.

- Who's been using NLP in the real world, and for what purpose?



Amazon Comprehend Medical is a service for healthcare. Source: Simon 2018.

Facebook uses machine translation to automatically translate posts and comments. Google Translate processes 100 billion words a day. To connect sellers and buyers across language barriers, eBay is using machine translation.

Using speech recognition and text-to-speech synthesis, voice assistants such as Amazon Alexa, Apple Siri, Facebook M, Google Assistant, and Microsoft Cortana are enabling human-to-device interaction using natural speech.

Amazon Comprehend offers an NLP API to perform many common NLP tasks. This has been extended by Amazon Comprehend Medical for healthcare domain.

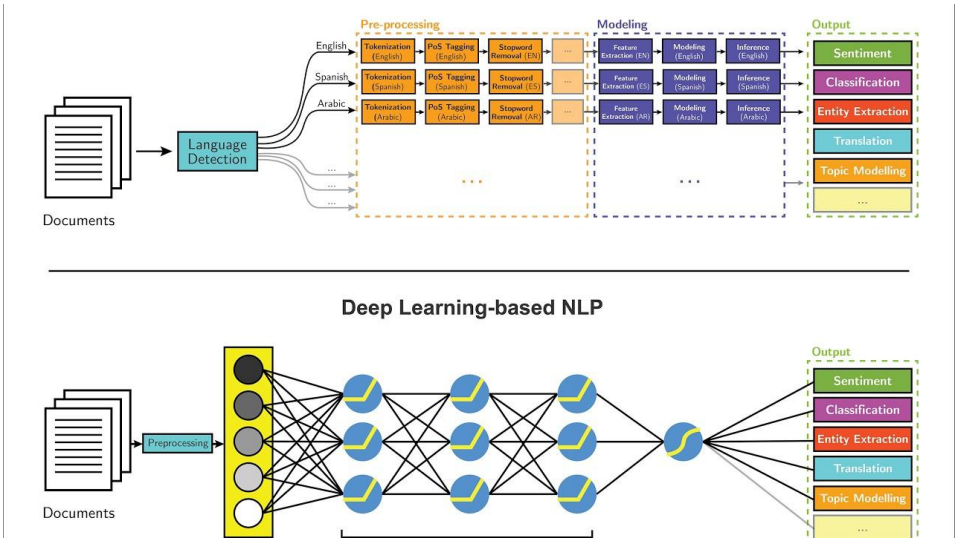
Uber uses NLP for better customer support. Human agents are involved but they are assisted by NLP models that suggest top

three solutions. This has reduced ticket resolution time by over 10%.

Perception offers an NLP-based product to do theme clustering and sentiment analysis. This helps with performance reviews and employee retention while minimizing bias.

For aircraft maintenance, NLP is used for information retrieval, troubleshooting, writing summary reports, or even directing a mechanic via a voice interface. It's been observed that NLP can classify defects better than humans.

- What are the main approaches adopted by NLP?



Classical NLP has given way to Deep Learning NLP. Source: Le 2018.

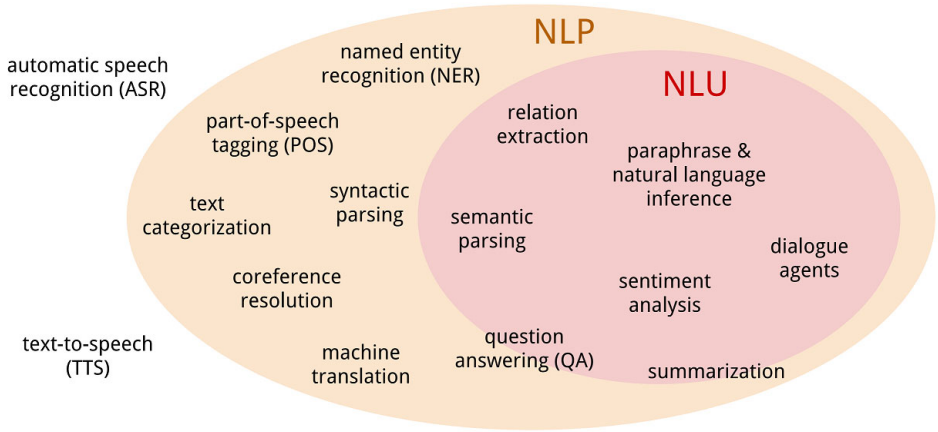
Classical NLP from the 1950s took the **symbolic approach** rooted in linguistics. Given the rules of syntax and grammar, we could obtain the structure of text. Using logic, we could obtain the meaning. But rules had to be hand-crafted and were often numerous. They didn't handle colloquial text well. Rules worked well for specific use cases but couldn't be generalized.

In practice, better accuracy was achieved by using a **statistical approach** that began in the 1980s. Rules were learned and they had associated probabilities. Machine Learning (ML) models came in with support vector machines and logistic regression. More recently, Deep Learning (DL) models that employ a neural network of many layers have brought better accuracy. This success is partly

due to the more efficient representations given by *word embeddings*.

NLP involves different levels or scope of analysis. **Low-level** analysis is about word tokens and structure. **Mid-level** analysis is about identifying entities, topics, and themes. **High-level** analysis leads to meaning and understanding. Alternatively, some classify text processing into two parts: **shallow parsing or chunking** and **deep parsing**.

- How is NLP related to NLU and NLG?



NLU is a subset of NLP. Source: MacCartney 2014, slide 8.

NLP is broadly made of two parts:

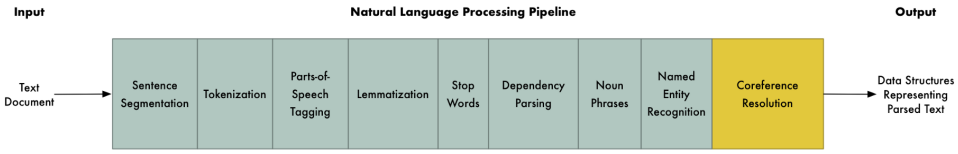
  - **Natural Language Understanding (NLU)**: This involves converting speech or text into useful representations on which analysis can be performed. The goal is to resolve ambiguities, obtain context and understand the meaning of what's being said. Some say NLP is about text parsing and syntactic processing while NLU is about semantic relationships and meaning. NLU tackles the complexities of language beyond the basic sentence structure.
  - **Natural Language Generation (NLG)**: Given an internal representation, this involves selecting the right words, forming phrases and sentences. Sentences need to ordered so that information is conveyed correctly.

NLU is about analysis. NLG is about synthesis. An NLP application may involve one or both. Sentiment analysis and semantic search are examples of NLU. Captioning an image or video is mainly an

NLG task since input is not textual. Text summarization and chatbot are applications that involve NLU and NLG.

There's also **Natural Language Interaction (NLI)** of which Amazon Alexa and Siri are examples.

- What's the typical data processing pipeline in NLP?



A typical text processing pipeline with optional coreference resolution. Source: Geitgey 2018.

A typical NLP pipeline consists of text processing, feature extraction and decision making. All these steps could apply classical NLP techniques, machine learning or neural networks. Where ML and NN are used, we would have to train a model from sufficient volume of data before it can be used for prediction and decision making.

In text processing, the input is just text and the output is a structured representation. This is done by identifying words, phrases, parts of speech, and so on. Since words have variations (go, going, went), it's common to reduce them to a root form with techniques such as **stemming** and **lemmatization**. Common words that don't add value to analysis (the, to, and, etc.) are called *stop words* and these are removed. Punctuations are also removed to simplify analysis. **Named Entity Recognition (NER)** involves identifying entities such as places, names, objects, and so on. **Coreference resolution** tries to resolve pronouns (he, they, it, etc.) to the correct entities.

More formally, text processing involves analysis of three types: syntax (structure), semantics (meaning), pragmatics (meaning in context).

- What are some challenges that NLP needs to solve?
NLU is still an unsolved problem. Systems are as yet incapable of understanding the way humans do. Until then, progress will be

limited to better pattern matching. Where NLU is lacking, it affects the success of NLG.

In the area of chatbots, there's a need to model common sense. It's also not clear if models should begin with some understanding or should everything be learned using the technique of reinforcement learning. Computing infrastructure needed to build a full-fledged agent that can learn from its environment is also tremendous.

Not much has been done for low-resource languages where the need for NLP is greater. Africa alone has about 2100 languages. We need to find a way to solve this even if training data is limited.

Current systems are unable to reason with large contexts, such as entire books or movie scripts. Supervision with large documents is scarce and expensive. Unsupervised learning has the problem of sample inefficiency.

Just measuring progress is a challenge. We need datasets and evaluation procedures tuned to concrete goals.

- Could you mention some of the tools used in NLP?
  In Python, two popular NLP tools are **Natural Language Toolkit (NLTK)** and **SpaCy**. NLTK is supposedly slower and therefore not the best choice for production. TextBlob extends NLTK. Textacy is based on SpaCy and handles pre-processing and post-processing tasks. There's also PyTorch-NLP suited for prototyping and production. AllenNLP and Flair are built on top of PyTorch for developing deep learning NLP models. Intel NLP Architect is an alternative. Gensim is a library that targets topic modelling, document indexing and similarity retrieval.

  There are also tools in other programming languages. In Node.js, we have Retext, Compromise, Natural and Nlp.js. In Java, we have OpenNLP, Stanford CoreNLP and CogCompNLP. The last two have Python bindings as well. There are libraries in R and Scala as well but these haven't been updated for over a year.

  For execution, **Jupyter Notebook** provides an interactive environment. If you don't want to install Jupyter, it's also available

as web services. Azure Notebook Service is an example. Via subscriptions, these services allow you to use powerful cloud computing resources.

Milestones

1948

In the area of automated translation, a dictionary look-up system developed at Birkbeck College, London can be seen as the first NLP application. In the years following World War II, researchers attempt translating German text to English. Later during the era of Cold War, it's about translating Russian to English.

1957

American linguist Noam Chomsky publishes *Syntactic Structures*. Chomsky revolutionizes the theory of linguistics and goes on to influence NLP a great deal. The invention of Backus-Naur Form notation in 1963 for representing programming language syntax is influenced by Chomsky's work. Another example is the invention of Regular Expressions in 1956 for specifying text search patterns.

1966

In the U.S., the Automatic Language Processing Advisory Committee (ALPAC) Report is published. It highlights the limited success of machine translation. This results in a lack of funding right up to 1980. Nonetheless, NLP advances in some areas including case grammar and semantic representations. Much of the work till late 1960s is about syntax though some addressed semantic challenges.

1970

In this decade, NLP is influenced by AI with focus on world knowledge and meaningful representations. Thus, semantics becomes more important. SHRDLU (1973) and LUNAR (1978) are two systems of this period. Into the 1980s, these lead to the adoption of logic for knowledge representation and reasoning. **Prolog** programming language is also invented in 1970 for NLP applications.

1980

This decade sees the growing adoption of Machine Learning and thereby signalling the birth of **statistical NLP**. Annotated bodies of text called *corpora* are used to train ML models to provide the gold standard for evaluation. ML approaches to NLP become prominent through the 1990s, partly inspired by the successful application of Hidden Markov Models to speech recognition. The fact that statistics has brought more success than linguistics is echoed by Fred Jelinek,

> Every time I fire a linguist, the performance of our speech recognition system goes up.

1982

Project Jabberwacky is launched to simulate natural human conversations in the hope of passing the Turing Test. This heralds the beginning of **chatbots**. In October 2003, Jabberwacky wins third place in the Loebner Prize.

1998

The *FrameNet* project is introduced. This is related to **semantic role modelling**, a form of shallow semantic parsing that's continues to be researched even in 2018.

2001

For language modelling, the classical N-Gram Model has been used in the past. In 2001, researchers propose the use of a **feed-forward neural network** with vector inputs, now called *word embeddings*. In later years, this leads to the use of RNNs (2010) and LSTMs (2013) for language modelling.

2003

Latent Dirichlet Allocation (LDA) is invented and becomes widely used in machine learning. It's now the standard way to do topic modelling.

2013

Improvements to word embeddings along with an efficient implementation in *Word2vec* enable greater adoption of neural networks for NLP. RNNs and LSTMs become obvious choices since they deal with dynamic input sequences so common in NLP. CNNs from computer vision get repurposed for NLP since CNNs are more

parallelizable. Recursive Neural Networks attempt to exploit the hierarchical nature of language.

Mar
2016

Microsoft launches *Tay*, a chatbot on Twitter that would interact with users and get better in conversing. However, Tay is shut down within 16 hours after it learns to talk in racist and abusive language. A few months later Microsoft launches *Zo* chatbot.

Sep
2016

Google replaces its phrase-based translation system with **Neural Machine Translation (NMT)** that uses a deep LSTM network with 8 encoder and 8 decoder layers. This reduces translation errors by 60%. This work is based on **sequence-to-sequence learning** proposed in 2014, which later becomes a preferred technique for NLG.