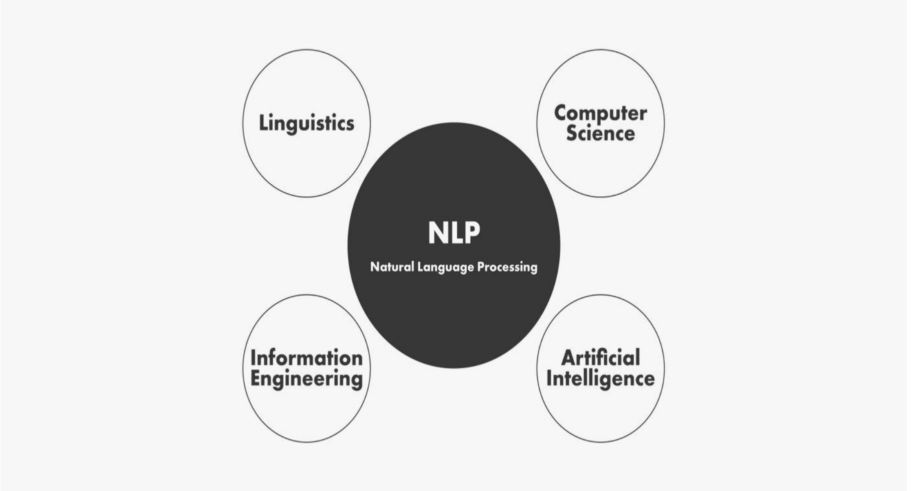
Basic NLP

Introduction To NLP

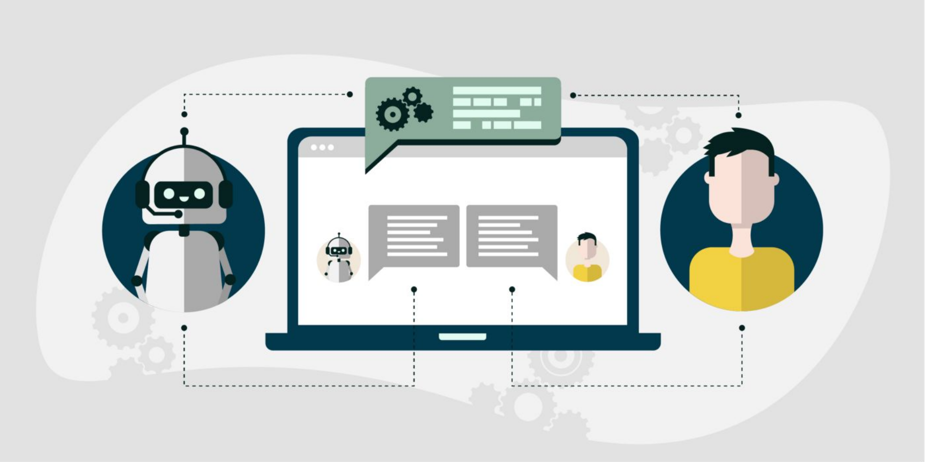


Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Wikipedia

Why NLP Useful?

NLP **enables computers to understand natural language as humans do**. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand.



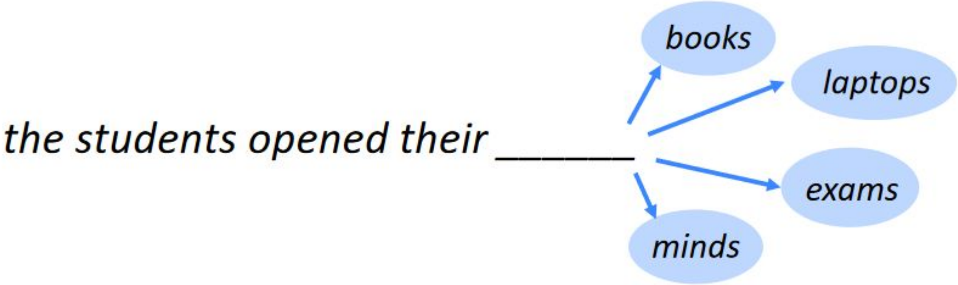
NLP in the Real World/Core applications:

- Email platforms, such as Gmail, Outlook, etc., use NLP extensively to provide a range of product features, such as spam classification, priority inbox, calendar event extraction, auto-complete, etc.
- Voice-based assistants, such as Apple Siri, Google Assistant, Microsoft Cortana, and Amazon Alexa rely on a range of NLP techniques to interact with the user, understand user commands, and respond accordingly.
- Modern search engines, such as Google and Bing, which are the cornerstone of today's internet, use NLP heavily for various subtasks, such as query understanding, query expansion, question answering, information retrieval, and ranking and grouping of the results so on.
- Machine translation services, such as Google Translate, Bing Microsoft Translator, and Amazon Translate are increasingly used in today's world to solve a wide range of scenarios and business use cases.
- Chatbots

Various NLP Tasks

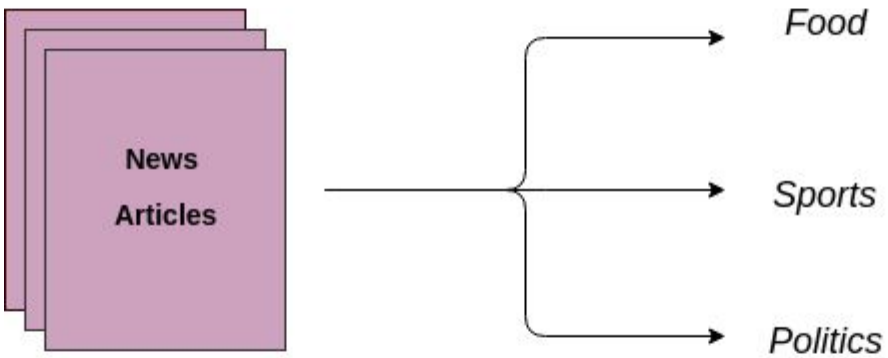
Language modeling

This is the task of predicting what the next word in a sentence will be based on the history of previous words. The goal of this task is to learn the probability of a sequence of words appearing in a given language. Language modeling is useful for building solutions for a wide variety of problems, such as speech recognition, optical character recognition, handwriting recognition, machine translation, and spelling correction.



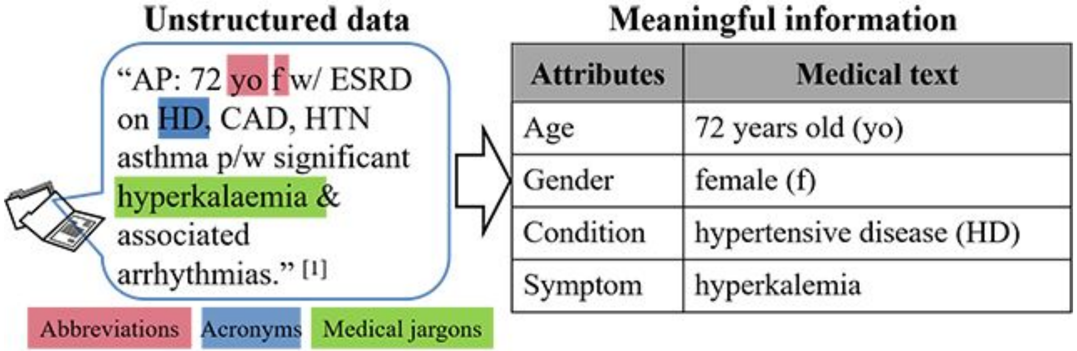
Text classification

This is the task of bucketing the text into a known set of categories based on its content. Text classification is by far the most popular task in NLP and is used in a variety of tools, from email spam identification to sentiment analysis.



Information extraction

As the name indicates, this is the task of extracting relevant information from text, such as calendar events from emails or the names of people mentioned in a social media post.



Unstructured data

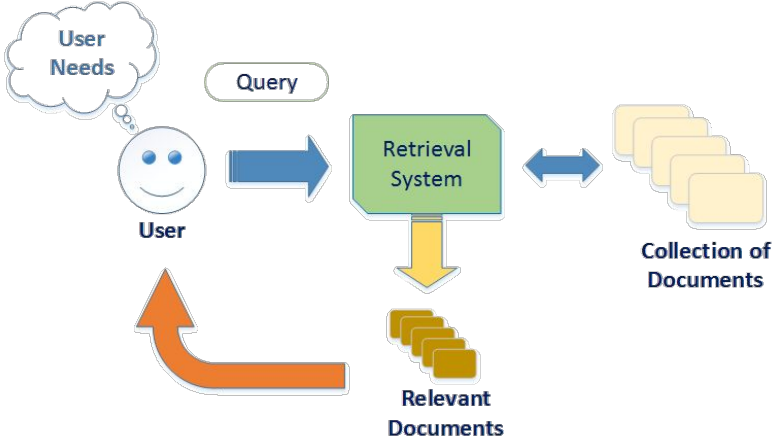
"AP: 72 yo f w/ ESRD on HD, CAD, HTN asthma p/w significant hyperkalaemia & associated arrhythmias." [1]

Meaningful information

Attributes	Medical text
Age	72 years old (yo)
Gender	female (f)
Condition	hypertensive disease (HD)
Symptom	hyperkalemia

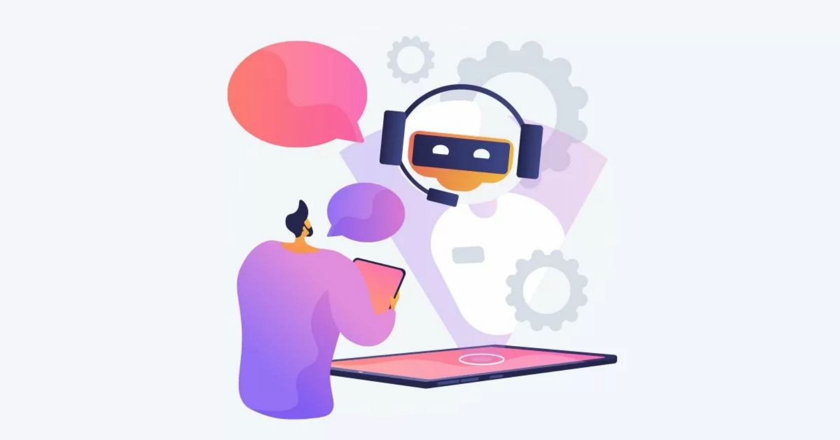
Information retrieval

This is the task of finding documents relevant to a user query from a large collection. Applications like Google Search are well-known use cases of information retrieval.



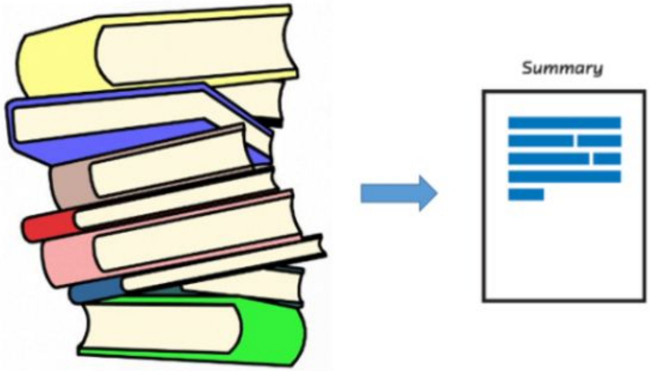
Conversational agent

This is the task of building dialogue systems that can converse in human languages. Alexa, Siri, etc., are some common applications of this task.



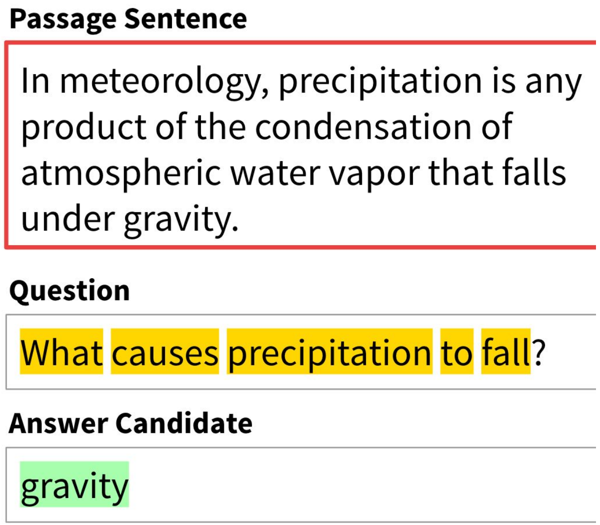
Text summarization

This task aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.



Question answering

This is the task of building a system that can automatically answer questions posed in natural language.



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

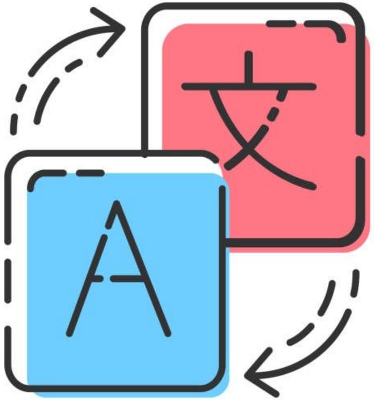
What causes precipitation to fall?

Answer Candidate

gravity

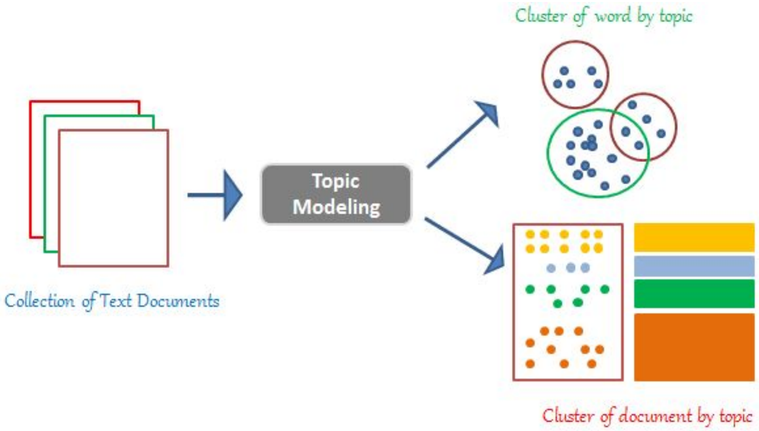
Machine translation

This is the task of converting a piece of text from one language to another. Tools like Google Translate are common applications of this task.

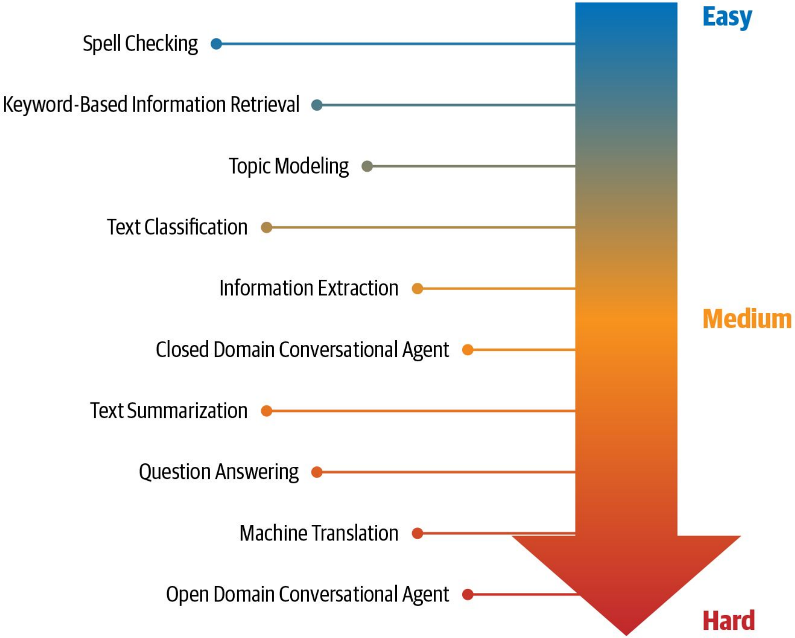


Topic modeling

This is the task of uncovering the topical structure of a large collection of documents. Topic modeling is a common text-mining tool and is used in a wide range of domains, from literature to bioinformatics.



Difficulty in terms of developing comprehensive solutions.



What Is Language?

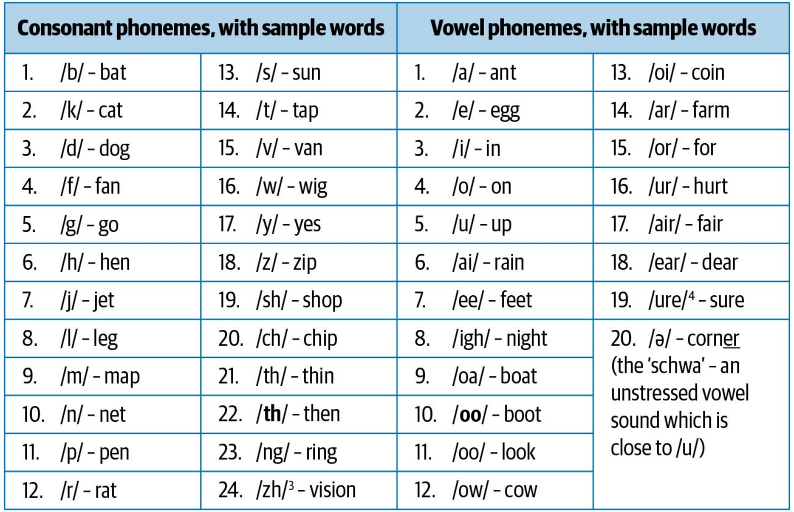
Language is a structured system of communication that involves complex combinations of its constituent components, such as characters, words, sentences, etc.

We can think of human language as composed of four major building blocks: **phonemes, morphemes and lexemes, syntax, and context**.

Building Blocks of Language

Phonemes

Phonemes are the smallest units of sound in a language. They may not have any meaning by themselves but can induce meanings when uttered in combination with other phonemes.



Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/[4] - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /**th**/ - then	10. /**oo**/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/[3] - vision	12. /ow/ - cow	

Morphemes and lexemes

A morpheme is the smallest unit of language that has a meaning. It is formed by a combination of phonemes. Not all morphemes are words, but all prefixes and suffixes are morphemes.

unbreakable
un + break + able

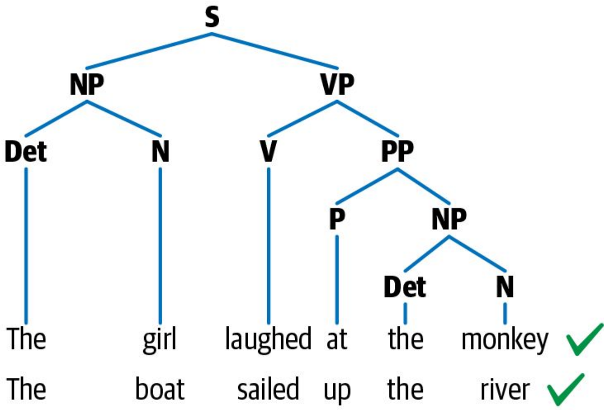
cats
cat + s

tumbling
tumble + ing

unreliability
un + rely + able + ity

Syntax

Syntax is a set of rules to construct grammatically correct sentences out of words and phrases in a language. Syntactic structure in linguistics is represented in many different ways. A common approach to representing sentences is a parse tree. In this representation, N stands for noun, V for verb, and P for preposition. Noun phrase is denoted by NP and verb phrase by VP.

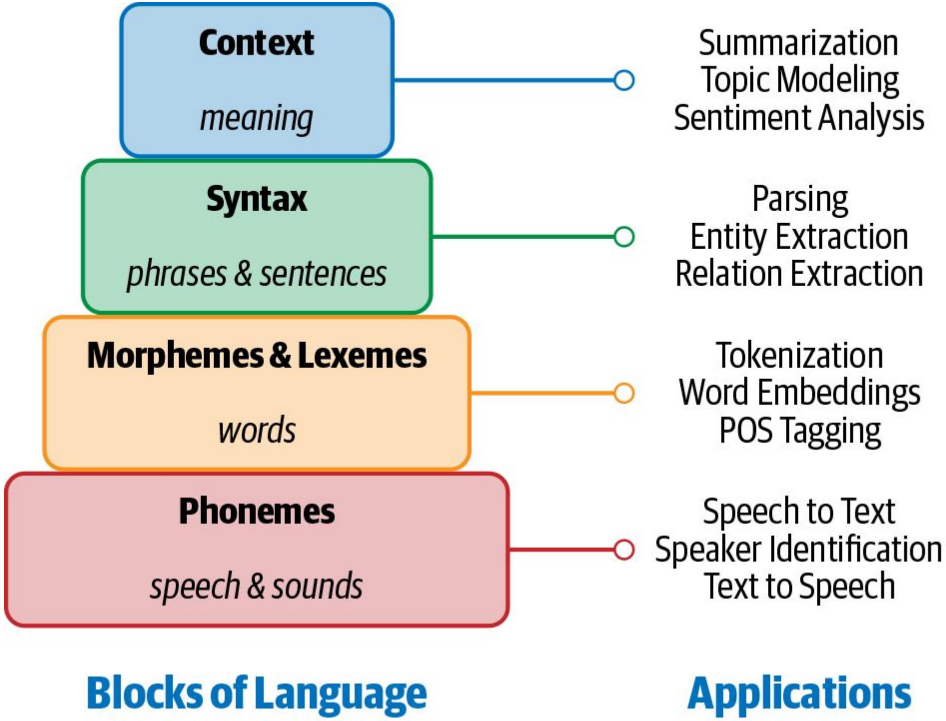

S
NP
VP
Det
N
V
PP
P
NP
Det
N
The girl laughed at the monkey ✓
The boat sailed up the river ✓

Context

Context is how various parts in a language come together to convey a particular meaning. Context includes long-term references, world knowledge, and common sense along with the literal meaning of words and phrases.

Complex NLP tasks such as sarcasm detection, summarization, and topic modeling are some of tasks that use context heavily.

Building blocks of language and their applications



Context

meaning

Summarization
Topic Modeling
Sentiment Analysis

Syntax

phrases & sentences

Parsing
Entity Extraction
Relation Extraction

Morphemes & Lexemes

words

Tokenization
Word Embeddings
POS Tagging

Phonemes

speech & sounds

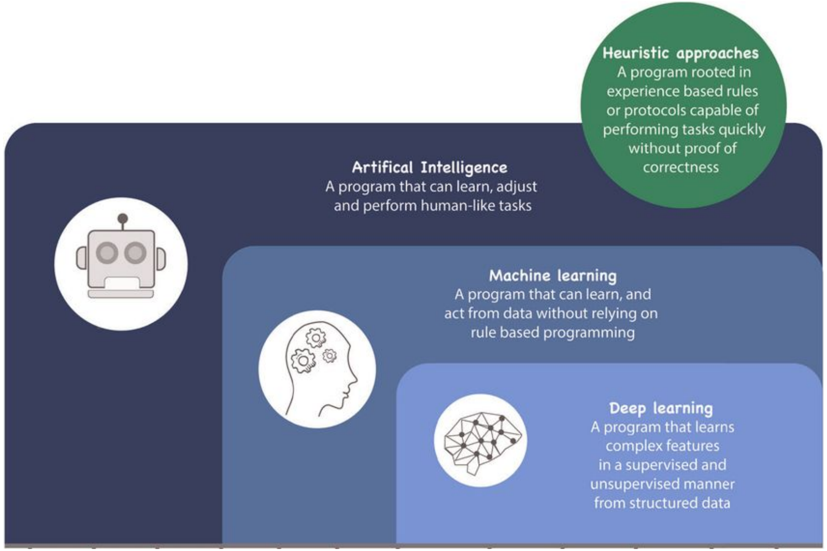
Speech to Text
Speaker Identification
Text to Speech

Blocks of Language

Applications

Approaches to NLP

The different approaches used to solve NLP problems commonly fall into three categories: **heuristics, machine learning, and deep learning.**



Heuristic approaches
A program rooted in experience based rules or protocols capable of performing tasks quickly without proof of correctness
Artifical Intelligence
A program that can learn, adjust and perform human-like tasks
Machine learning
A program that can learn, and act from data without relying on rule based programming
Deep learning
A program that learns complex features in a supervised and unsupervised manner from structured data

Heuristics-Based NLP

Similar to other early AI systems, early attempts at designing NLP systems were based on building rules for the task at hand.

Examples:

- Regular Expression
- Wordnet
- Open Mind Common Sense

Machine Learning for NLP

Machine learning techniques are applied to textual data just as they're used on other forms of data, such as images, speech, and structured data. Supervised machine learning techniques such as classification and regression methods are heavily used for various NLP tasks.

- **Naive Bayes**
- **Support vector machine**
- **Hidden Markov Model**

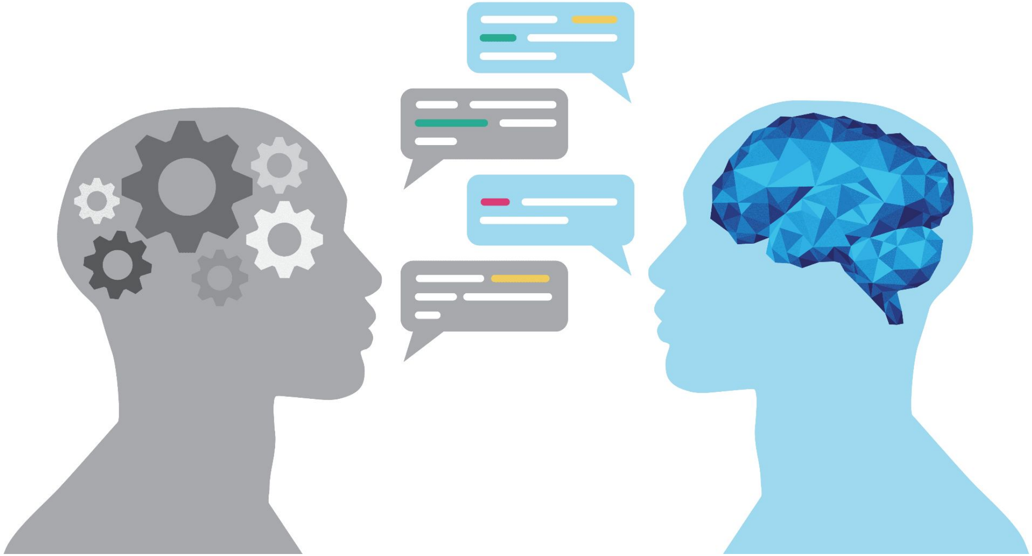
Deep Learning for NLP

Huge surge in using neural networks to deal with complex, unstructured data. Language is inherently complex and unstructured. herefore, we need models with better representation and learning capability to understand and solve language tasks. Here are a few popular deep neural network architectures that have become the status quo in NLP.

- Recurrent neural networks (RNN)
- Long short-term memory (LSTM)
- Convolutional neural networks (CNN)
- Transformers
- Autoencoders

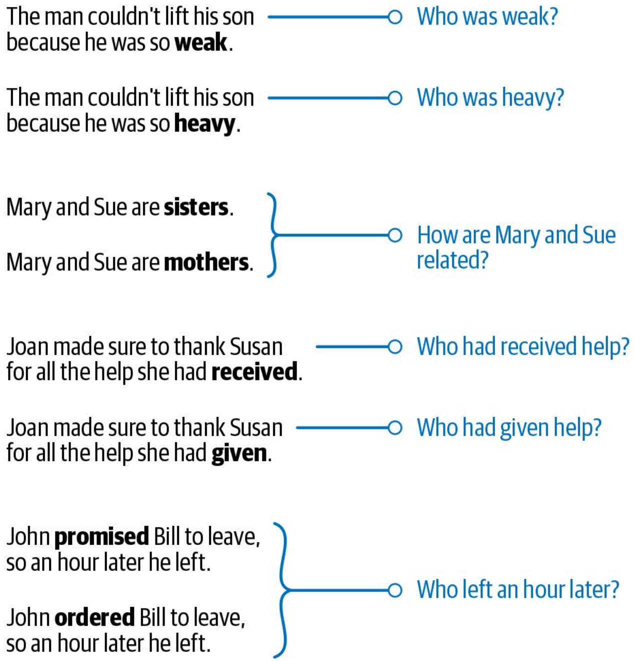
Why Is NLP Challenging?

The **ambiguity and creativity** of human language are just two of the characteristics that make NLP a demanding area to work in.



Ambiguity

Ambiguity means uncertainty of meaning!



The man couldn't lift his son because he was so **weak**. — Who was weak?

The man couldn't lift his son because he was so **heavy**. — Who was heavy?

Mary and Sue are **sisters**.

Mary and Sue are **mothers**.

How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**. — Who had received help?

Joan made sure to thank Susan for all the help she had **given**. — Who had given help?

John **promised** Bill to leave, so an hour later he left.

John **ordered** Bill to leave, so an hour later he left.

Who left an hour later?

Common knowledge

A key aspect of any human language is “common knowledge.” It is the set of all facts that most humans are aware of.

Example:

consider two sentences: **“man bit dog” and “dog bit man.”**

Creativity

Language is not just rule driven; there is also a creative aspect to it. Various **styles, dialects, genres, and variations** are used in any language. **Poems** are a great example of creativity in language. Making machines understand creativity is a hard problem not just in NLP, but in AI in general.

& Diversity: There are many languages

Wrapping Up

- Introduction To NLP
- Why NLP Useful?
- NLP in the Real World/Core applications:
- Various NLP Tasks
- Difficulty in terms of developing comprehensive solutions.
- What Is Language?
- Building Blocks of Language
- Approaches to NLP
- Why Is NLP Challenging?

ASSESSMENT

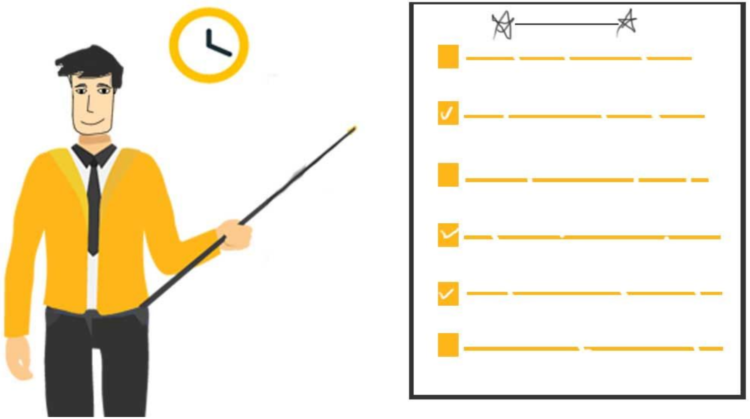
End-to-end NLP Pipeline

What is NLP Pipeline?

Break the problem down into several sub-problems, then try to develop a step-by-step procedure to solve them. Since language processing is involved, we would also list all the forms of text processing needed at each step. This step-by-step processing of text is known as a pipeline.

- **Data acquisition**
- **Text Preparation**
  - Text Cleanup
  - Basic Preprocessing
  - Advance Preprocessing
- **Feature engineering**
- **Modeling**
- **Evaluation**
- **Deployment**
- **Monitoring and model updating**

Points to Remember

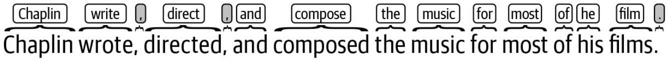


- It's not universal
- Deep Learning pipelines are slightly different
- Pipeline is non-linear

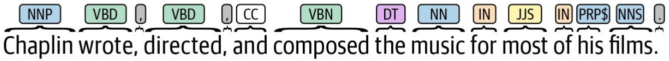
Input

Chaplin wrote, directed, and composed the music for most of his films.

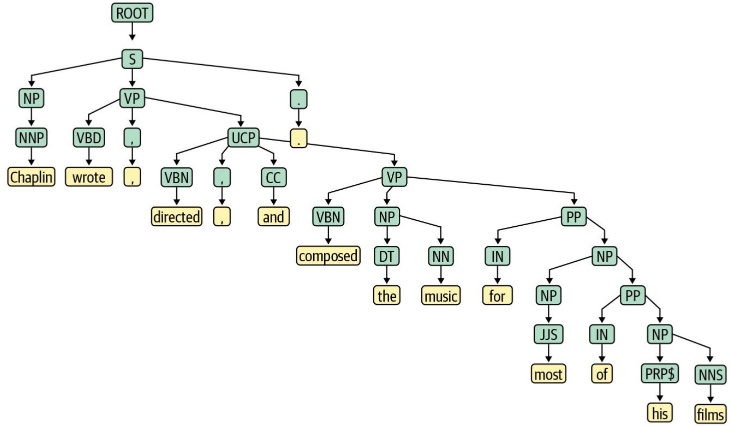
Tokenization with Lemmatization


Chaplin
write
,
direct
,
and
compose
the
music
for
most
of
he
film
.
Chaplin wrote, directed, and composed the music for most of his films.

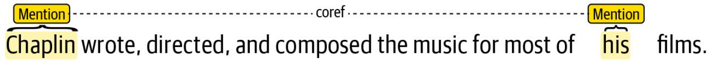
POS Tagging


NNP
VBD
,
VBD
,
CC
VBN
DT
NN
IN
JJS
IN
PRP$
NNS
.
Chaplin wrote, directed, and composed the music for most of his films.

Parse Tree



Coreference Resolution


Mention
coref
Mention
Chaplin wrote, directed, and composed the music for most of his films.

Lets get Hands on

Q
A

Some Resources to Follow

Books:

- Practical natural language processing
- Natural language processing with transformers