NLP Introduction

講者：Isaac

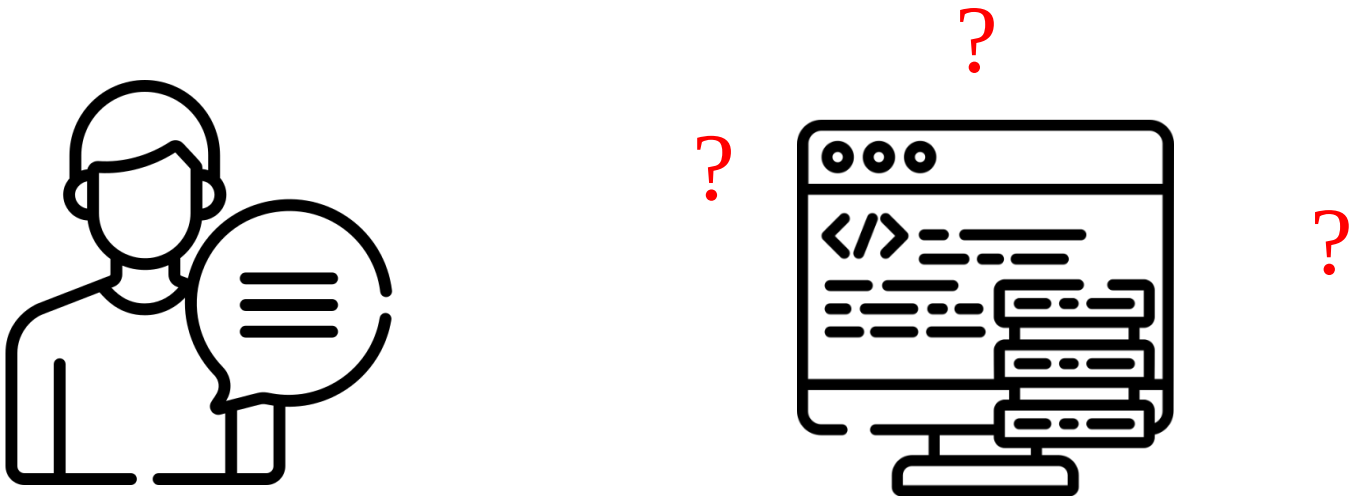
Outline

- Why NLP
- What is Natural Language Processing
- NLP Applications
- NLTK Introduction

Why NLP?

Why NLP?

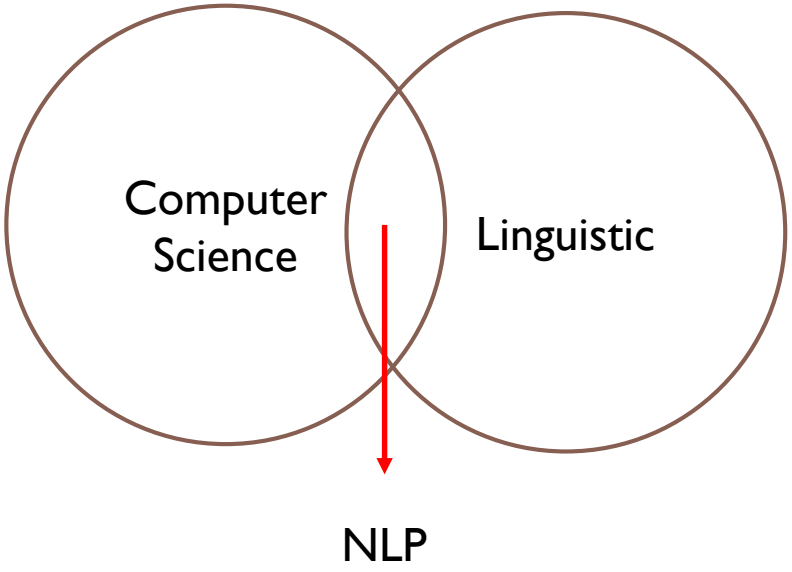
- goal of Natural Language Processing(NLP) is to the fill the gap how the humans communicate(natural language) and what the computer understands(machine language)



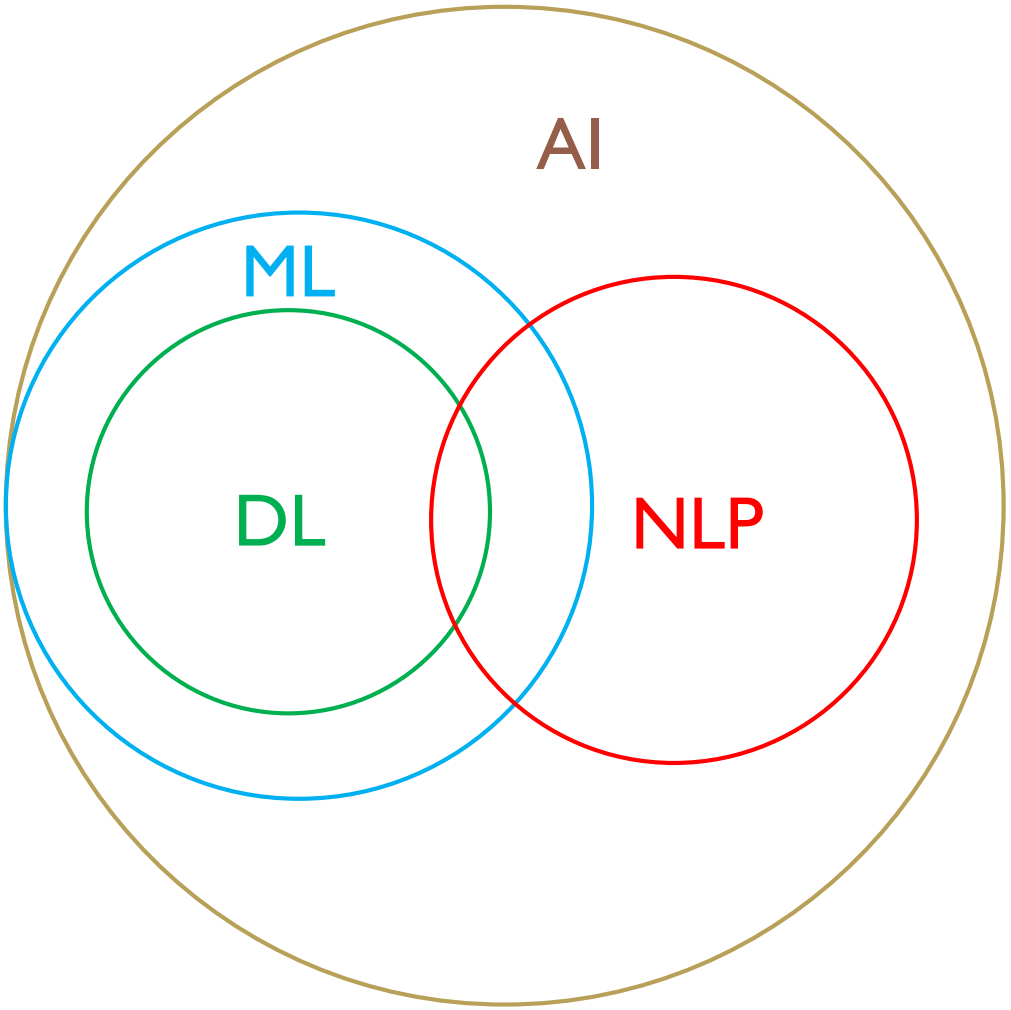
What is Natural Language Processing

What's NLP

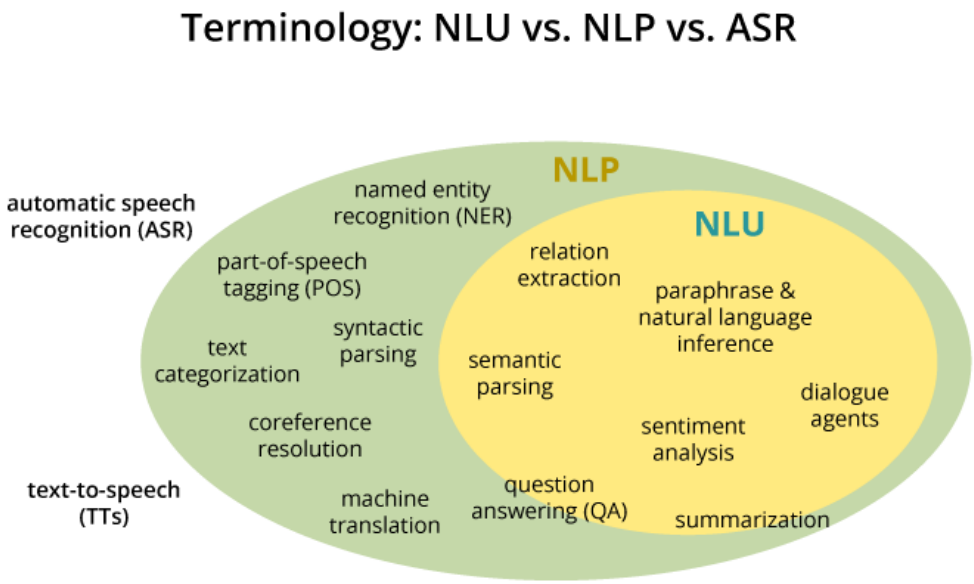
- Natural Language Processing(NLP) is the technology used to aid computers to understand the human's natural language
  - hard to teach computers to understand how we communicate



Relationship between AI, ML, DL, and NLP



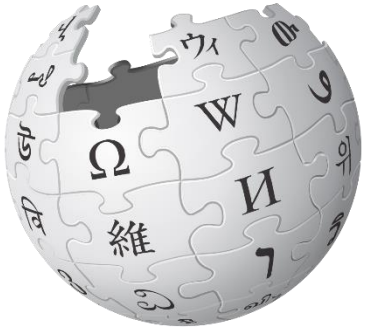
NLU V.S. NLP V.S. ASR



https://www.kdnuggets.com/2019/07/nlp-vs-nlu-understanding-language-processing.html

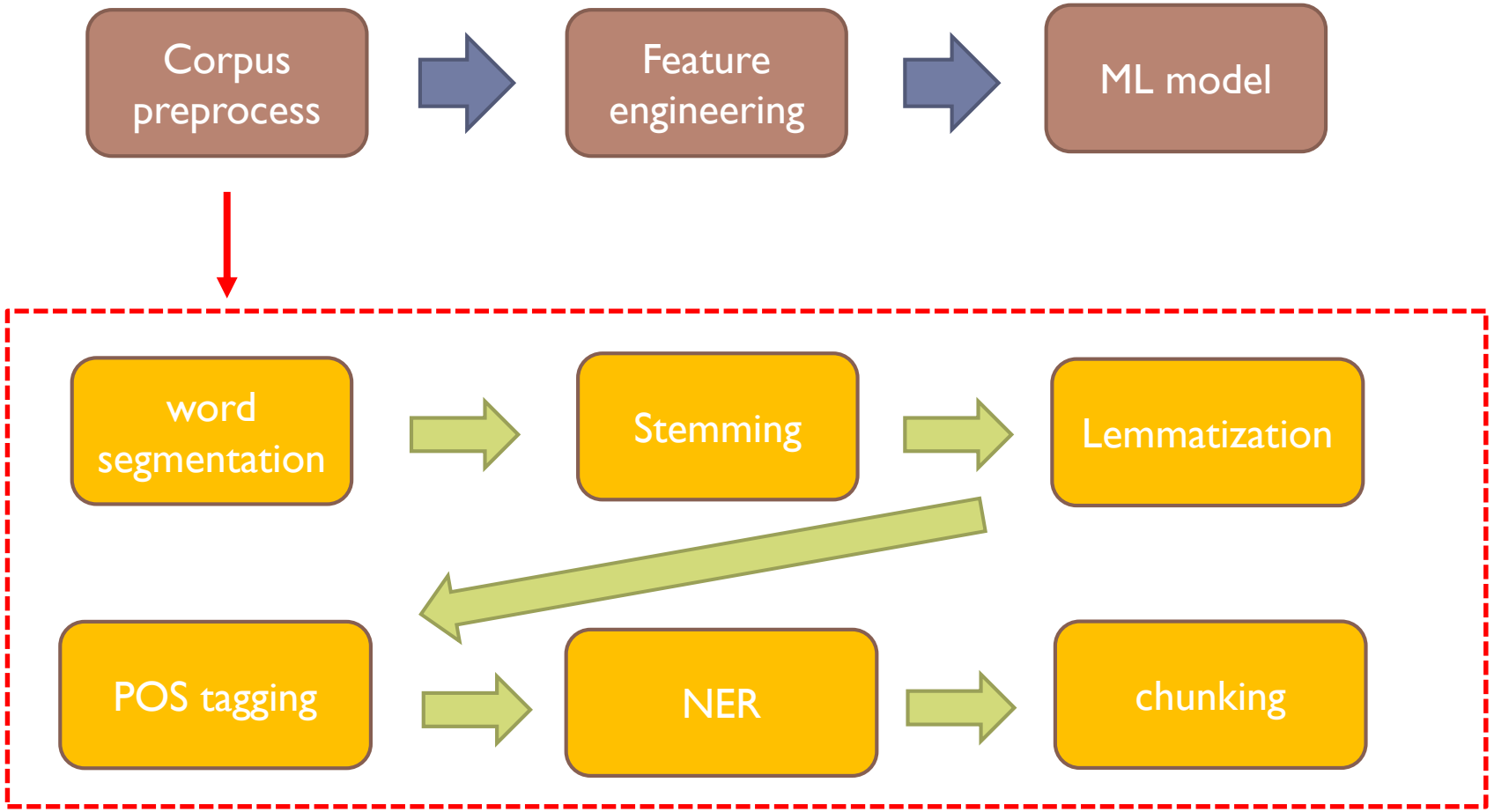
What's Corpus

- corpus refers to a collection of texts
  - may be single language or multiple languages
  - common corpus including Google Books Ngram, Brown, American National



WIKIPEDIA
The Free Encyclopedia

Flows in NLP


Corpus preprocess
Feature engineering
ML model
word segmentation
Stemming
Lemmatization
POS tagging
NER
chunking

Techniques used in NLP?

- Syntax
  - If arrangement of words in a sentence makes grammatical sense
- Semantics
  - the meaning that is conveyed by a text

syntax techniques

- Syntax techniques
  - Word segmentation
  - Stemming
  - Lemmatization
  - Part-of-speech tagging
  - Parsing
  - Sentence breaking

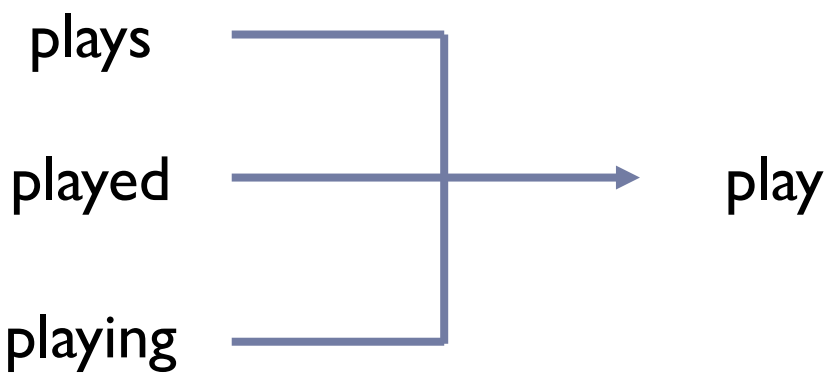
syntax - word segmentation

- task of splitting text into words
  - Like Chinese/Japanese characters
- Famous open source
  - **Jieba (**https://github.com/fxsjy/jieba**)**

我|電腦|當機|了

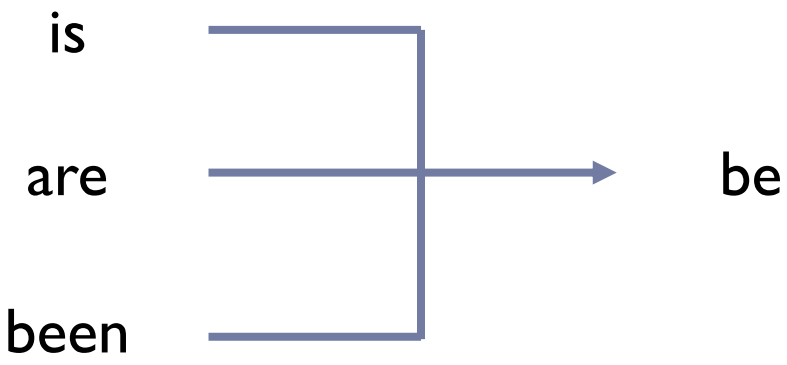
syntax - Stemming

- cutting the inflected words to their root form
- famous method including Porter, Snowball, and Lancaster



syntax -Lemmatization

- reducing the inflected forms of a word into a single form for easy analysis



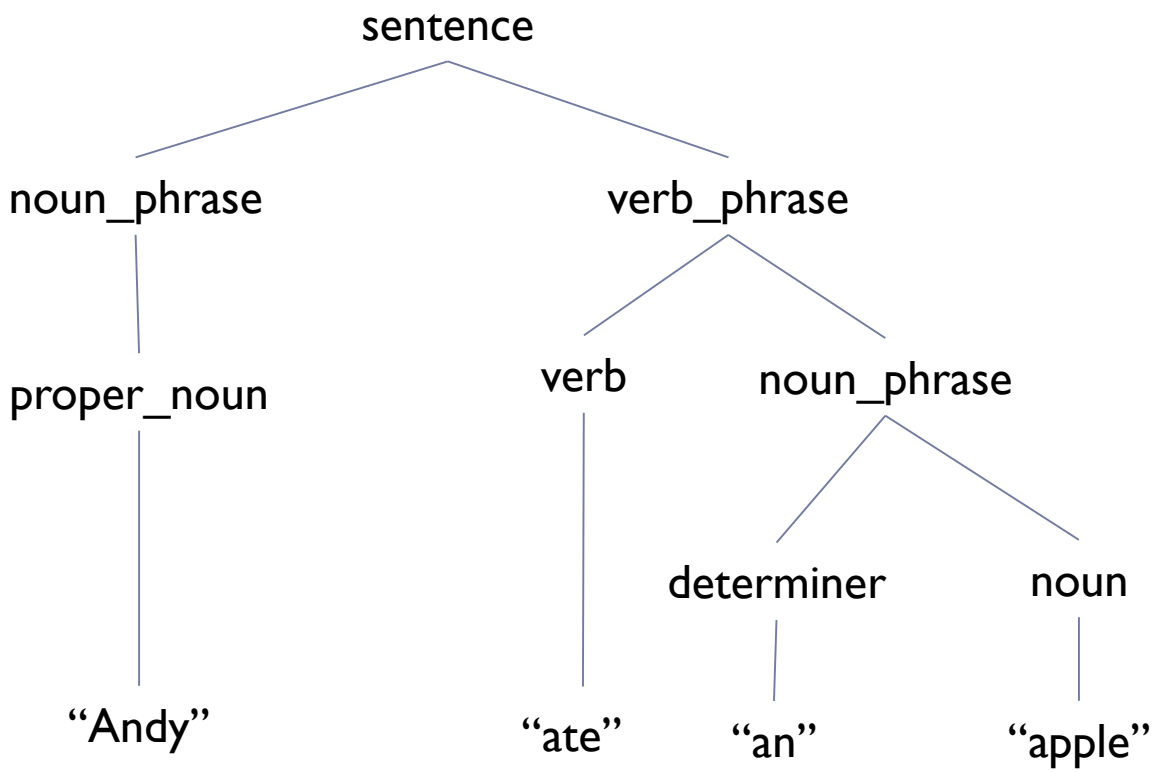
syntax - Part-of-speech tagging

- Identifying the part of speech for every word

She sells seashells on the seashore

syntax - Parsing

- Grammatical analysis of a sentence

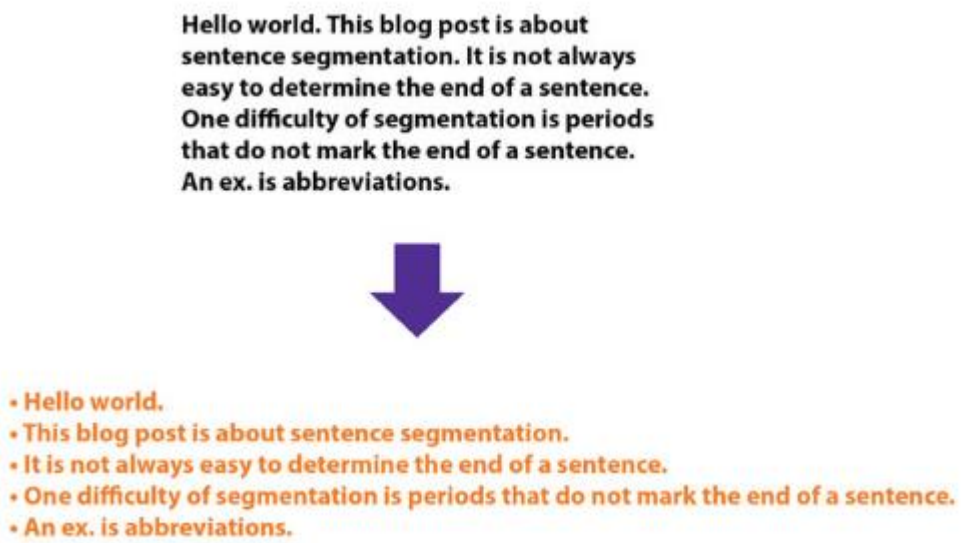


syntax - Parsing

I saw a girl with a telescope

syntax - Sentence breaking

- Placing sentence boundaries on a continuous text
- Also called sentence tokenize, sentence boundary disambiguation, sentence segmentation

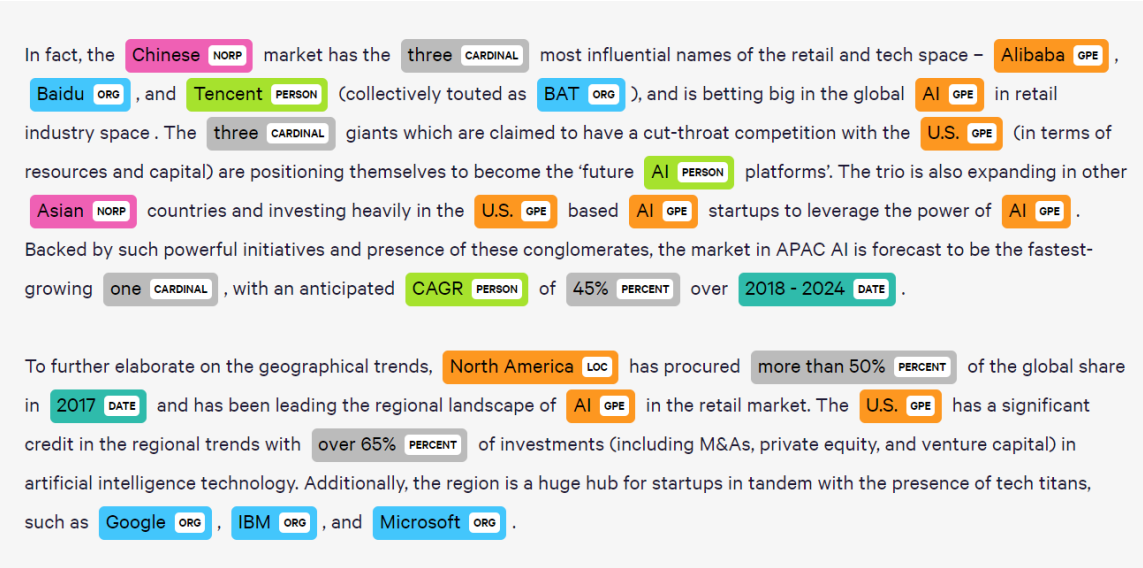


Semantics techniques

- Syntax techniques
  - Named entity recognition (NER)
  - Word sense disambiguation
  - Natural language generation

Semantics - Named Entity Recognition (NER)

- Finding references to entities in text and labeling them with their location and type



Semantics - Word sense disambiguation

- Assign the most appropriate meaning to a word within a given context

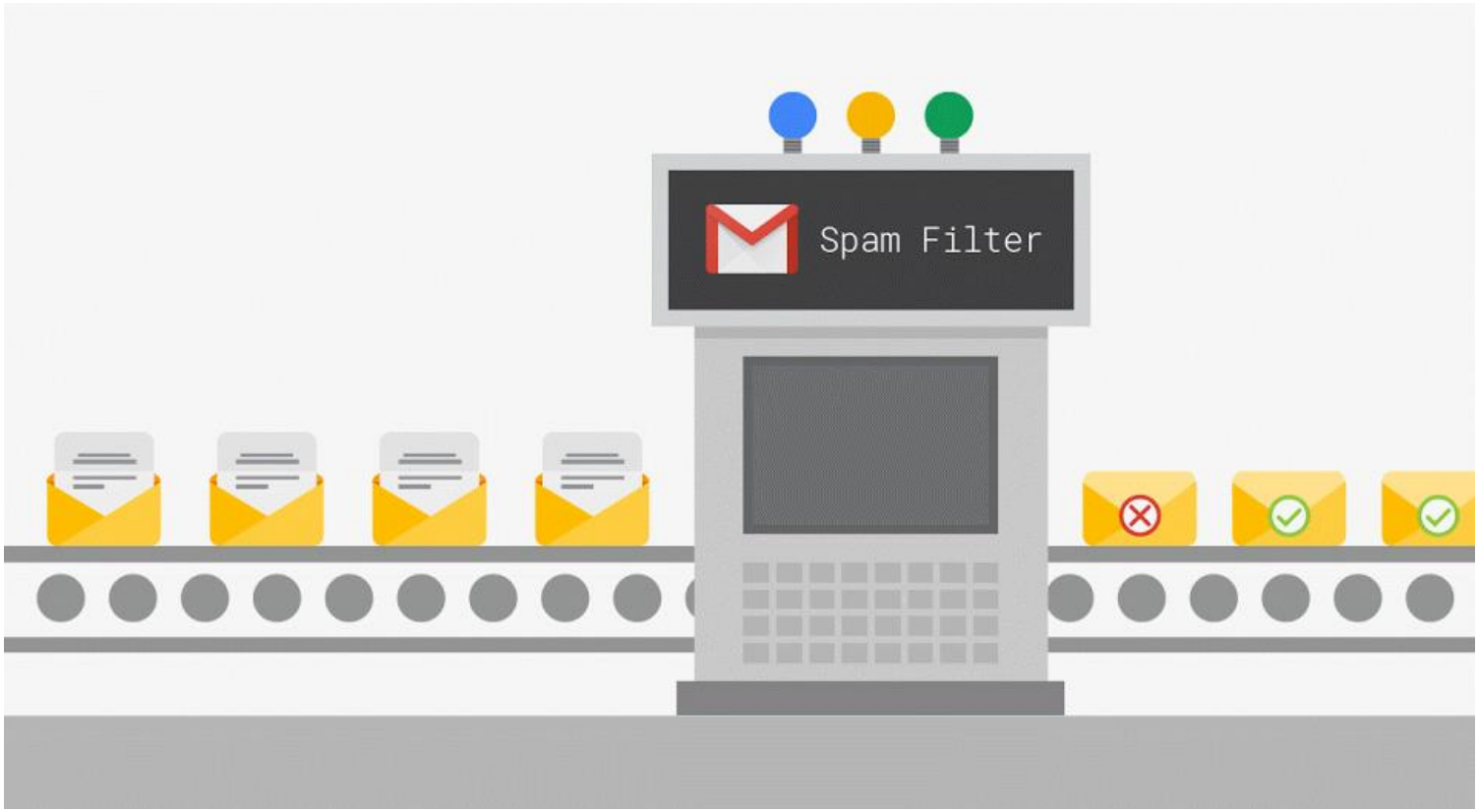
"The ***bank*** will not be accepting cash on Saturdays."

"The river overflowed the ***bank.***"

Semantics - Natural language generation

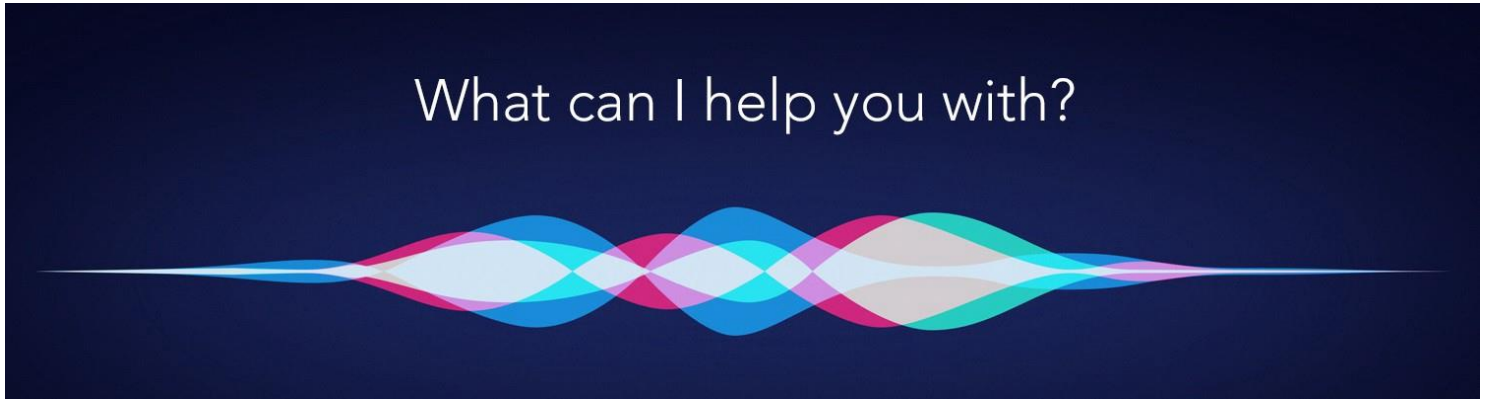
NLP Applications

Spam detection



Source: https://i.gifer.com/Ou1t.gif

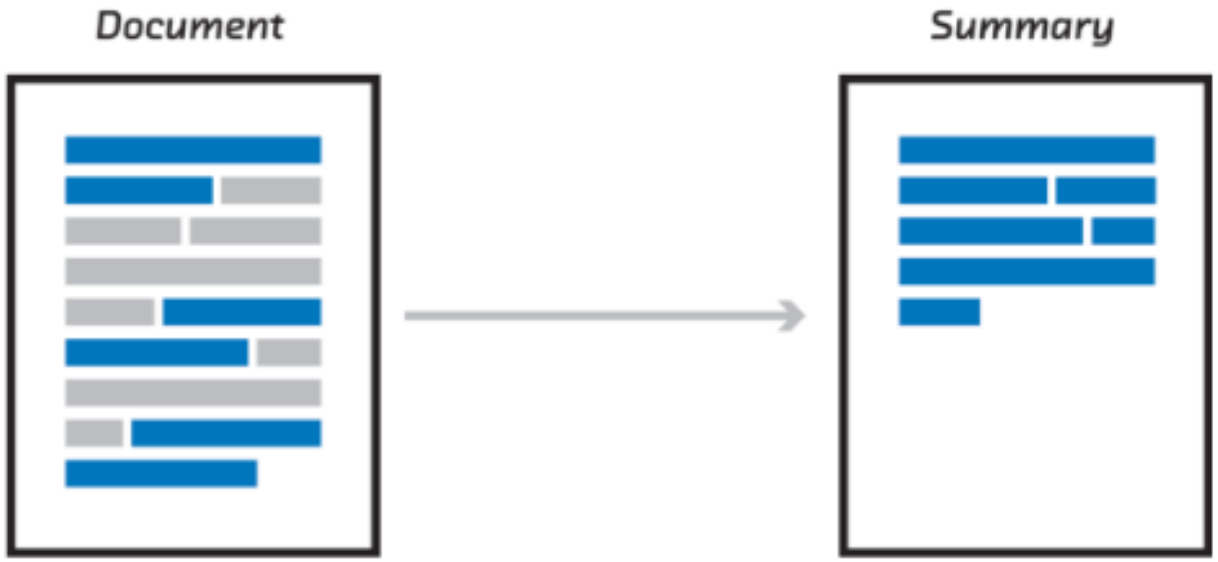
Siri



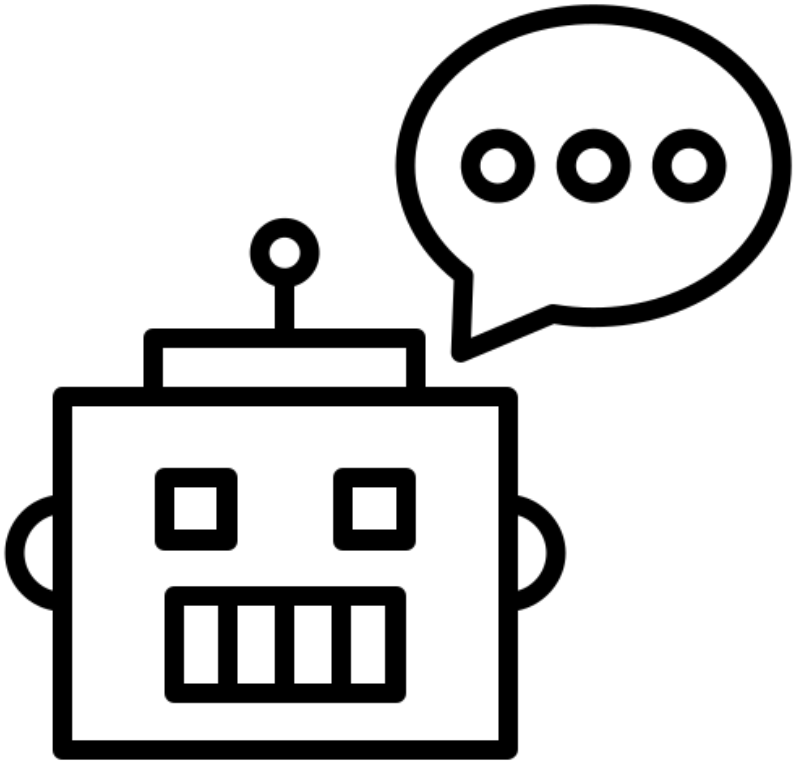
Source: https://www.analyticsindiamag.com/behind-hello-siri-how-apples-ai-powered-personal-assistant-uses-dnn/

Spell checking

Automatic Summarization



Chatbots



NLTK Introduction

What's NLTK?

- NLTK (Natural Language Toolkit) is a open source for work with human language data
  - provide many corpora and lexical resources
  - https://www.nltk.org/

Four kinds of corpus in NLTK

- isolate corpus
  - collection of natural language text
- categorized corpus
  - collection of different kinds of categorized text
- overlapping corpus
  - collection of different kinds of categorized text but some of category are overlapped
- temporal corpus
  - collection of text in a period of time

Common open source in NLP