01 – INTRODUCTION TO NLP

MACHINE LEARNING FOR NATURAL LANGUAGE PROCESSING, AIMS 2024

Lecture 01

COURSE LOGISTICS

Dr. Elvis Ndah

Contact: elvis.ndah@gmail.com

Data Science consultant

Anju Software

European Commission

COURSE OBJECTIVE

Goal: Provide the student with the fundamental understanding of natural language processing methods and strategies.

- Introduce core ideas at the basis of mordern NLP algorithms
- Focus on Deep learning techniques applied to NLP.
- Understand the strenght and weaknesses of various NLP technologies and frmeworks.

PREREQUISITES

Programming language: Python

Recommended: intermediary

Acceptable: Basic

Probability and statistics

Machine learning (deep learning)

COURSE LOGISTICS

- 12 Lectures
- 2 hours per lecture
- 1.5-hour Lab (notebooks)

Materials at https://github.com/ndaheanalytics/natural-language-processing

COURSE EVALUATION

- Final assignment (80% of the final grade).
  - Solve a NLP problem and present your results .
    - Self-contained notebook.
    - Report/presentation.

- Quiz (20% of final grade).
  - During lectures.

COURSE OUTLINE

1. **Introduction to natural language processing**
2. Text representation
3. Text classification
4. Language modelling
5. Sequence generation
6. Machine translation
7. Conversational Dialogue Systems and Chatbots

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

- Why Natural Language processing?
- What is Natural Language Processing?
  - Why is NLP Hard
  - Modelling Framework
  - Tokenization and vocabulary
  - Overview of NLP task
- How to tackle NLP problems
- A brief history of NLP

WHY NATURAL LANGUAGE PROCESSING?

- What’s the purpose of natural langauge?
  - Communicate using language
  - We think (partly) with language
  - Develop scientific theories with language
  - Create relationships friends/business with

WHY NATURAL LANGUAGE PROCESSING?

What is language?

- **Wikipedia:** Language is a structured system of communication that consists of grammar and vocabulary.
- **Cambridge dictionary:** a system of communication by speaking, writing, or making signs in a way that can be understood.
- **Linguistic (Krach 2007):** A sign consists in a phonological structure, a morphological structure, a syntactic structure and a semantic structure.

WHY NATURAL LANGUAGE PROCESSING?

NLP: ability to automatically processes and undertand human or natural language.

- Allow humans to quickly access knowledge (Information extraction, text summariztion).
- Communicate across language barries (Machine translation).
- Analyse languages themselves (Linguistic and cognitive sciences)

WHY NATURAL LANGUAGE PROCESSING?

Build systems that help humans communicate

Help humans interact with each other and/or devices.

Useful systems

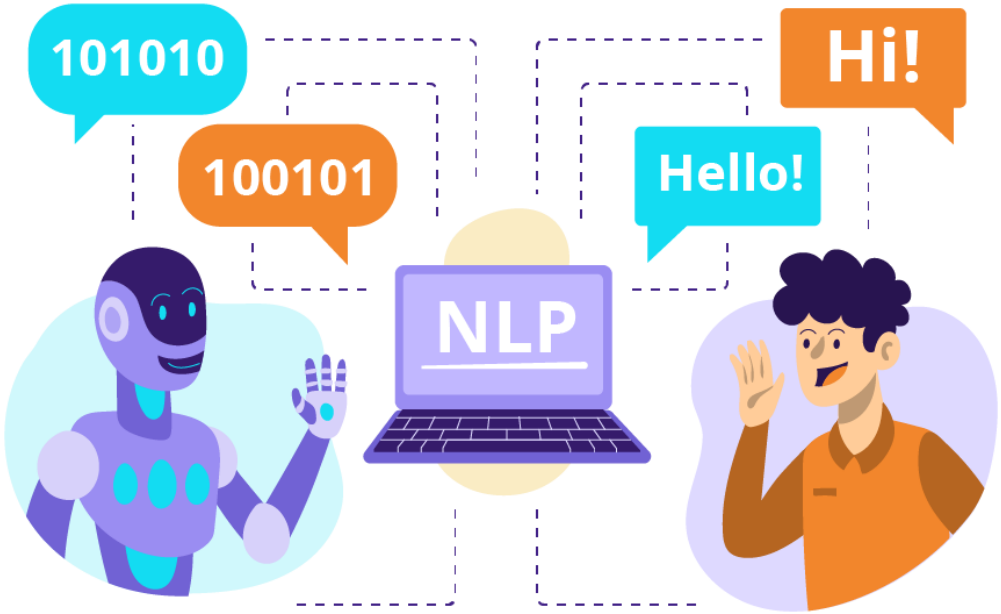
Automatic text summarization
Communicate without language barrier
Model and analyse properties of language
Speech recognition

WHY NATURAL LANGUAGE PROCESSING?

- Search: +2 billion Google users, 700+ million Baidu users
- Social Media: +3 billion users (Facebook, Instagram, twitter, WeChat).
- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)
- Machine Translation: 500 million users on google translate

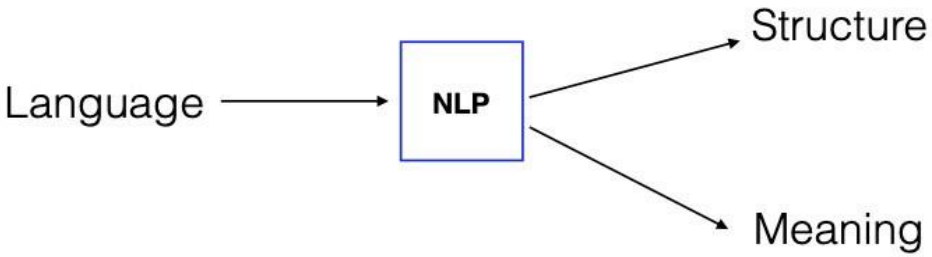
WHY NATURAL LANGUAGE PROCESSING?

Wiki: Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.



WHAT IS NATURAL LANGUAGE PROCESSING?

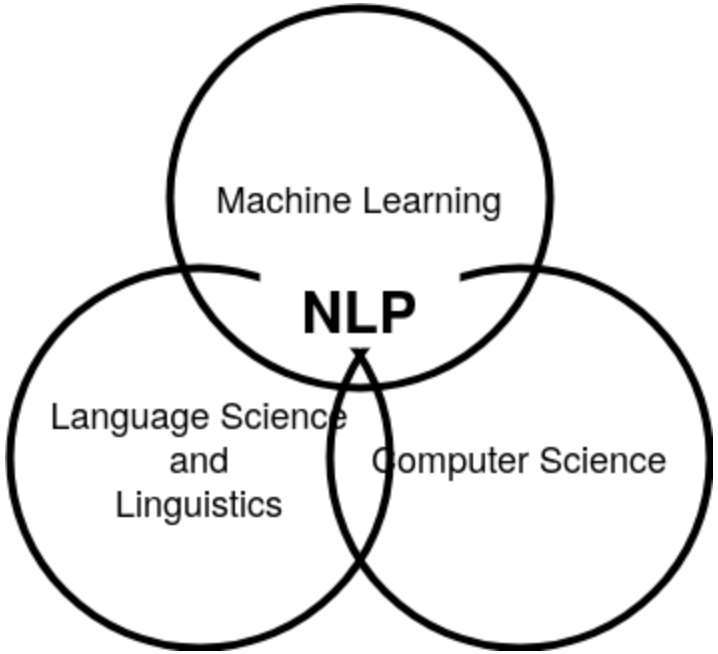
Develop methods for processing, analysing and understanding the structure and meaning of all natural or human languages.



It concerns with the interaction between natural languages and computing devices.

WHAT IS NATURAL LANGUAGE PROCESSING?

- Multidisciplinary field in the crossroad between Linguistics, Computer science and Machine learning



WHAT IS NATURAL LANGUAGE PROCESSING?

- Identify the structure and meaning of words, sentences, text and conversations.
- Deep understanding of broad language
- NLP is all around us

-- Unstructured text --
Add eggs and milk to my shopping list

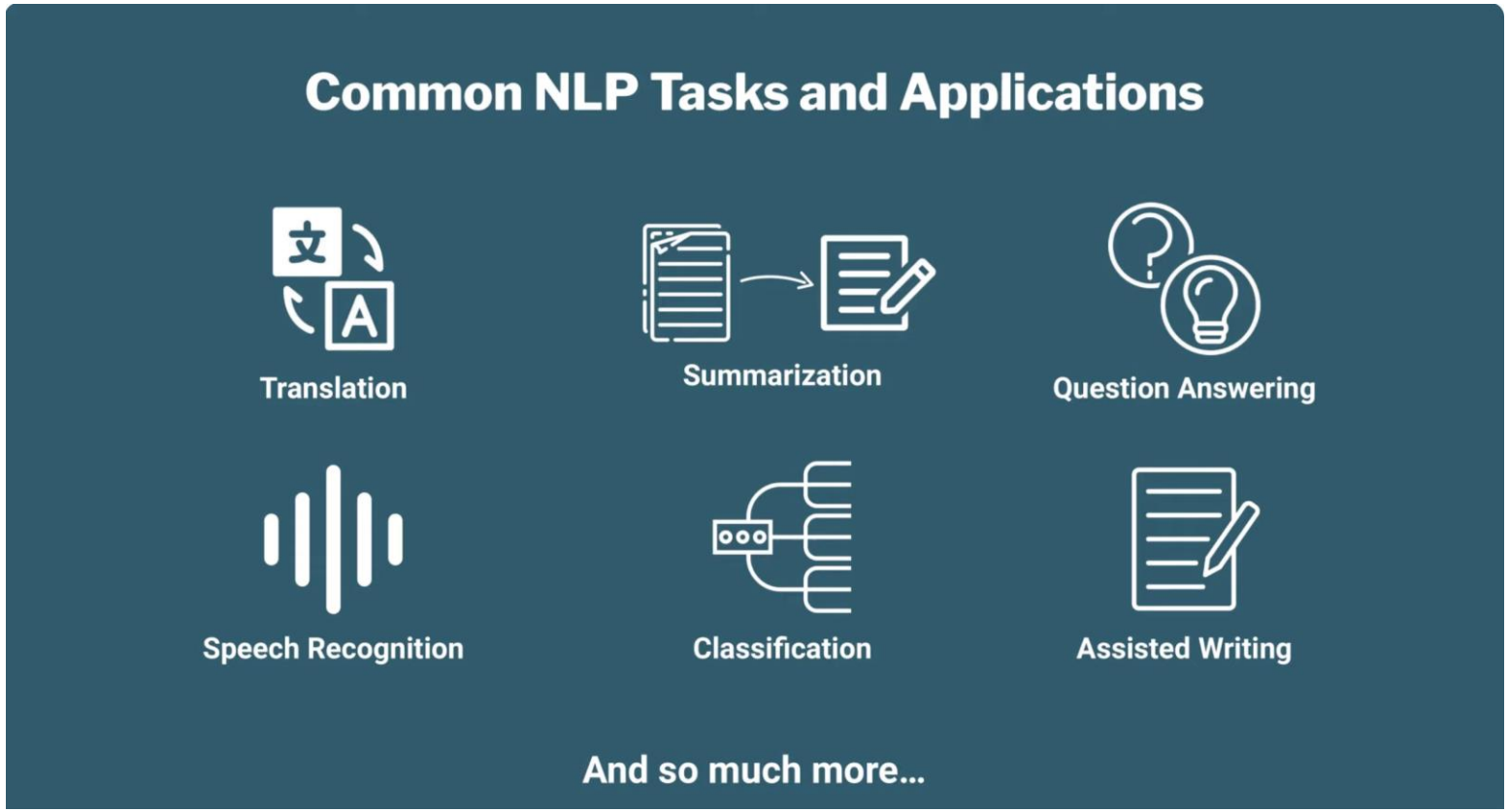
NLP

NLU →

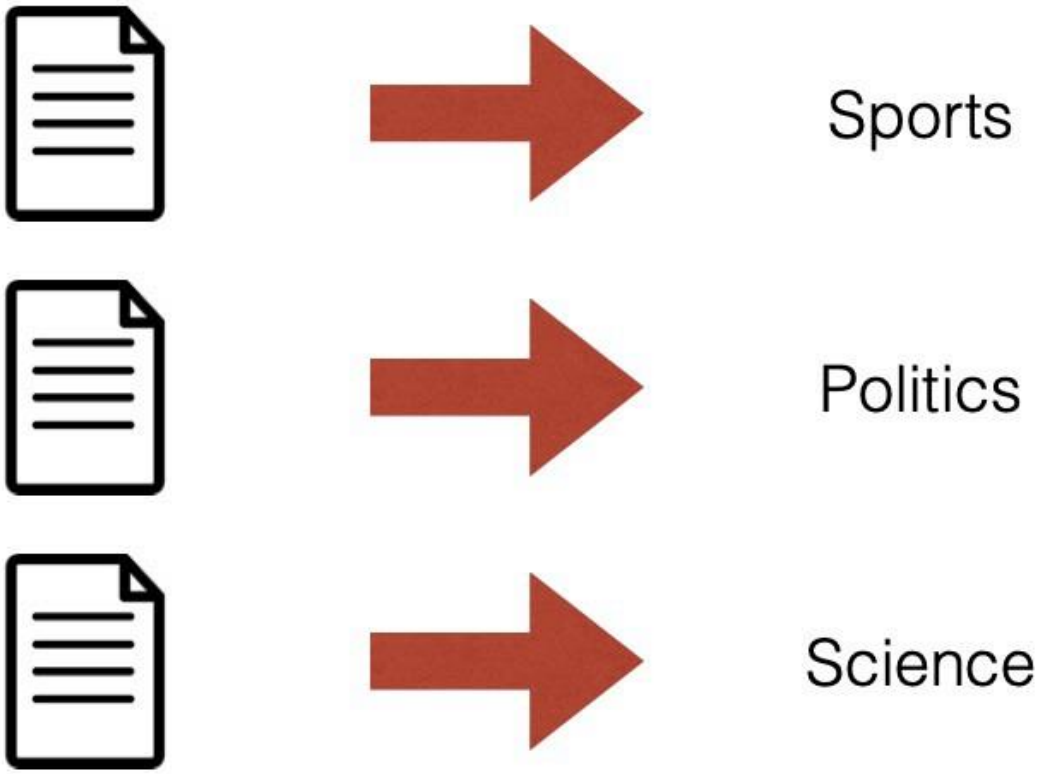
← **NLG**

-- Structured text –
<Shopping list>
<item>Eggs<\item>
<item>milk<\item>

MOST COMMON APPLICATIONS OF NLP



NLP APPLICATIONS – TEXT OR DOCUMENT CATEGORIZATION



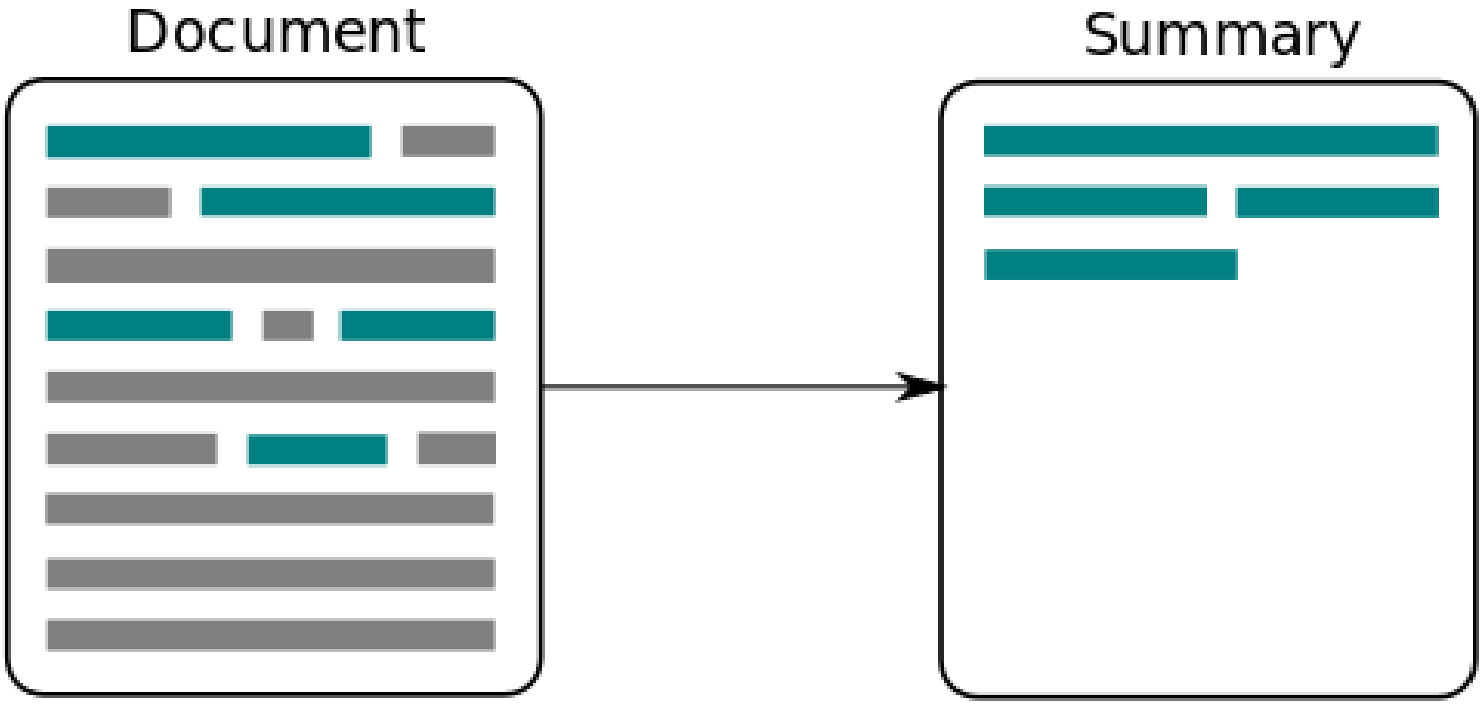
NLP APPLICATIONS – INFORMATION EXTRACTION

The task of **Information Extraction** involves extracting meaningful information from unstructured text

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

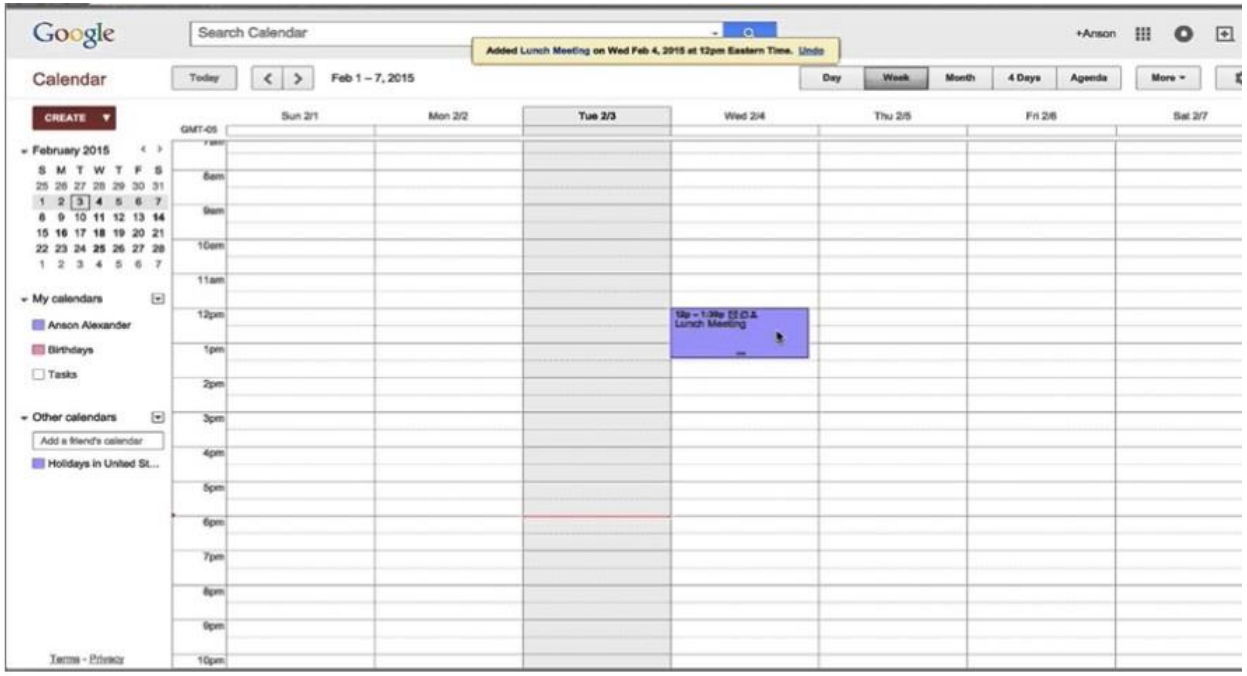
NLP APPLICATIONS – SUMMARIZATION



https://hackernoon.com/summarization-with-wine-reviews-using-spacy-b49f18399577

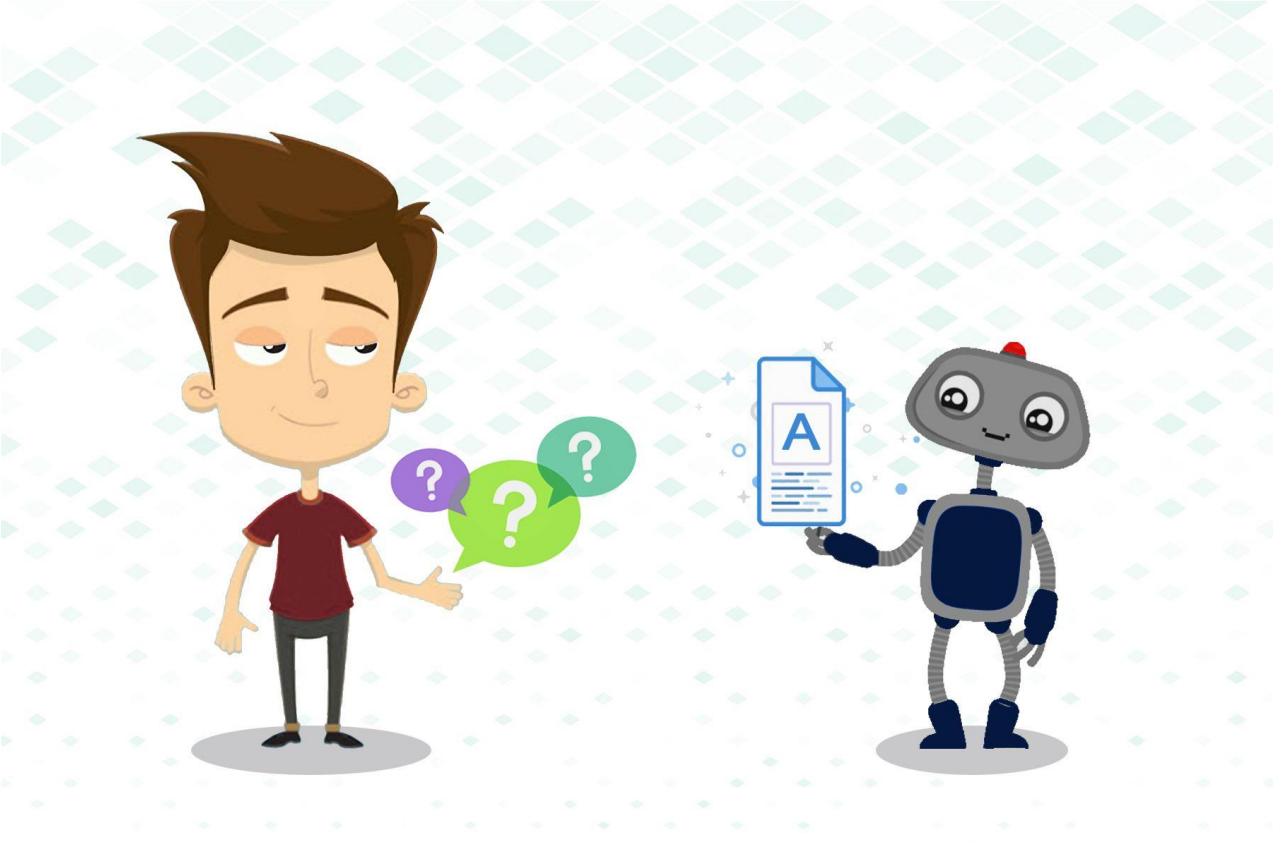
NLP APPLICATIONS – VIRTUAL ASSISTANTS

Move all my Wednesday meetings in April with John to 5pm



https://www.rogerpeele.com/how-to-choose-the-right-virtual-assistant-for-your-company-in-sydney/

NLP APPLICATIONS – QUESTION ANSWERING



https://blog.paperspace.com/question-answering-models-a-comparison/

NLP APPLICATIONS – READING COMPREHENSION

More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur.

Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field – a multi-millionaire.

- Who is Carl Lewis?
- Did Carl Lewis break any world records?
- Is Carl Lewis wealthy? What about Jesse Owens?

NLP APPLICATIONS – MACHINE TRANSLATION

WHAT IS SPECIAL ABOUT NATURAL LANGUAGE?

Lingustic analysis

- Phonology - sounds that make up language.
- Morphology – study of words and how they are formed.
- Syntax - structure of phrases, how words modify one another.
- Semantics - meaning of language in the world.
- Discourse: relations between clauses and sentences

WHAT ARE THE CHALLENGES OF NLP?

AMBIGUITY - LANGUAGES ARE AMBIGUOUS

VARIABILITY - LANGUAGES ARE COMPLEX

UNDERSTANDING REQUIRES VAST KNOWLEDGE AND EXPERIENCE

WHY IS NLP HARD – SYNTACTIC AMBIGUITY

Syntactic ambiguity: *two or more possible meanings within a single sentence.*

"Finally, a computer that understands you like your mother" (Ad, 1985)

- The computer understands you as well as your mother understands you.
- The computer understands that you like your mother.
- The computer understands you as well as it understands your mother.

WHY IS NLP HARD – SEMANTIC AMBIGUITY

Semantic ambiguity

- *occurs when a word, phrase or sentence, taken out of context, has more than one interpretation.*

"We saw her duck"

- The word *"her duck"* can can refer either to
  - the person's bird - the noun *"duck"* modified by the possessive pronoun *"her"*
  - a motion she made - the verb *"duck"*, subject of the objective pronoun *"her"*, object of the verb *"saw"*

WHY IS NLP HARD – LEXICAL AMBIGUITY

Lexical ambiguity: *two or more possible meanings within a single word*

Finally, a computer that understands your lie cured mother"

- The word *lie* can have multiple meanings in sentence the and will not change the context of the sentence.

- The ambiguity is on what cured mother
  - *lie:* an intentionally false statement
  - *lie:* spice or home-made remedy

WHY IS NLP HARD? – VARIABILITY

There are many ways to express the same meaning in language.

PWD ends up with 6 points.

PWD climbs by 6 points in the table.

PWD gains 6 points

Key computational challenge in NLP is to compute similarity of the above phrases.

WHY IS NLP HARD? – LANGUAGE IS NOT STATIC

New words added to dictionary

google, googling
laggy
greenwash

cyber lingo

#TBT => throwback Thursday
DM => direct message
LOL => laugh out loud
AMA => ask me anything
Troll => online troll
Epic fall => when some one fails

THE NLP PIPELINE

Tokenizer and segmentation	identify words and sentences boundaries Text normalization and vocabulary creation
Morphological analyzer	identify the structure of words
Word sense disambiguation	identify the meaning of words
Syntactic/semantic parser	obtain the structure and meaning of sentences
discourse analysis	keep track of the various entities and events mentioned

TEXT BOOKS

1. Speech and Language Processing 3rd ed, Jurafsky and Martin. https://web.stanford.edu/~jurafsky/slp3/
2. Natural Language Processing, Jacob Eisenstein. https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf