



# 01 – INTRODUCTION TO NLP

MACHINE LEARNING FOR NATURAL LANGUAGE PROCESSING, AIMS 2024

Lecture 05  
Dr. Elvis Ndah

# COURSE LOGISTICS

Dr. Elvis Ndah

Contact: [elvis.ndah@gmail.com](mailto:elvis.ndah@gmail.com)

Data science consultant

- Anju Software
- European Commission
- University of Ghent

# THE FOCUS OF THIS COURSE

01

Introduction to human language understanding.

02

Why natural language processing is difficult?

03

Understanding of the modern techniques for NLP

04

Learn to build systems for major problems in NLP

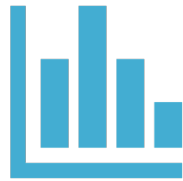
# PREREQUISITES



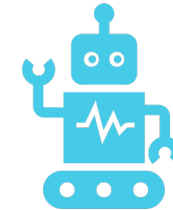
## **Programming language: Python**

Recommended: intermediary

Acceptable: Basic



## **Probability and statistics**



## **Machine learning (deep learning)**

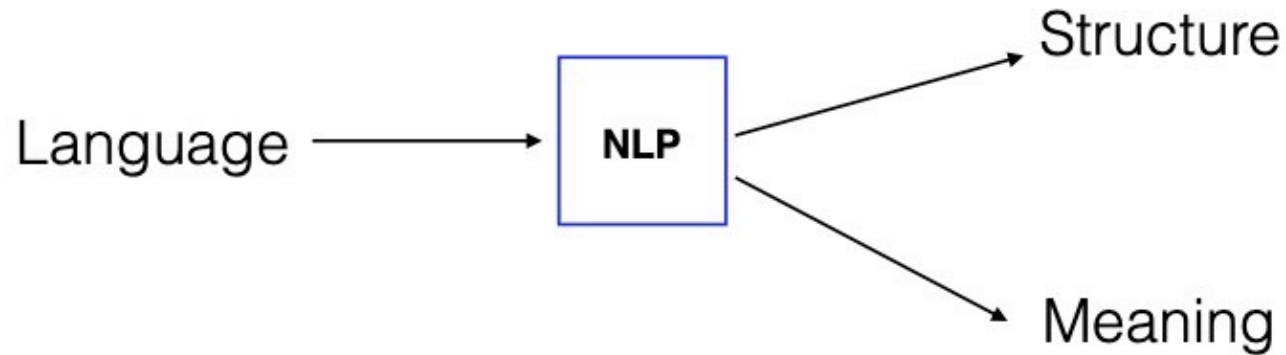
# LECTURE 1: INTRODUCTION TO NLP



- Course Overview
  - What is NLP? Why it is important?
  - Applications of NLP
  - What are the challenges?
  - What types of ML methods used in NLP?

# WHAT IS NLP?

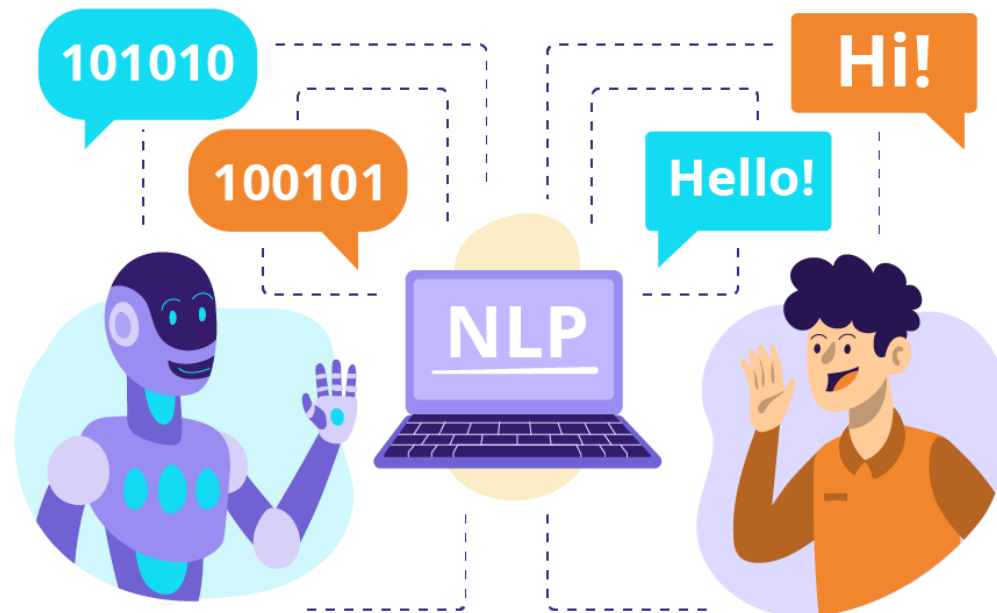
Develop methods for processing, analyzing and understanding the structure and meaning of all natural (human) language.



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

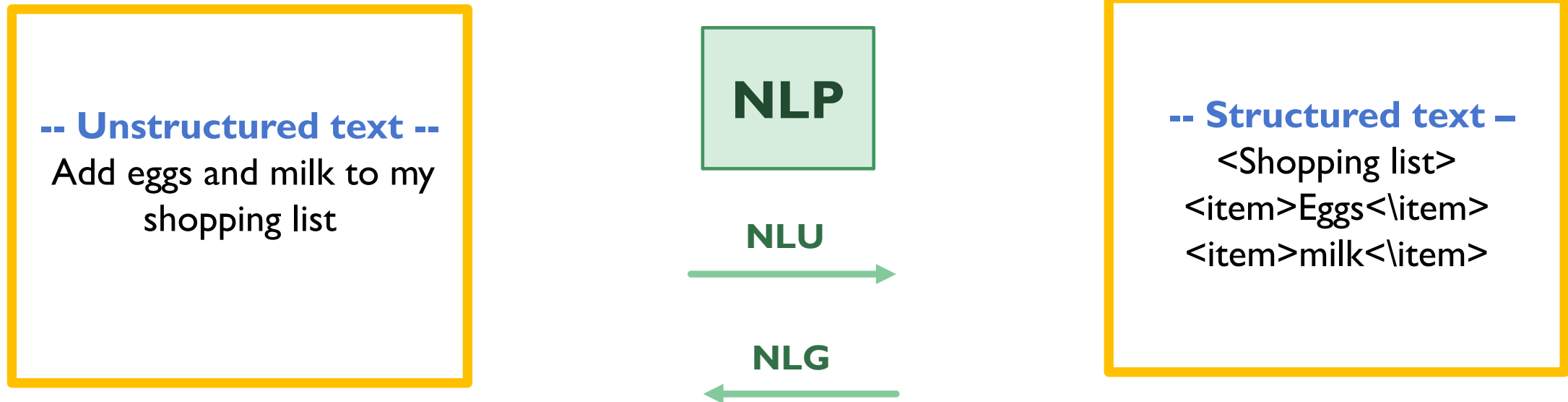
# WHAT IS NLP?

**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 NLP?

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

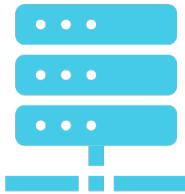




# WHY WORK ON NLP?



**Build systems that help humans communicate**



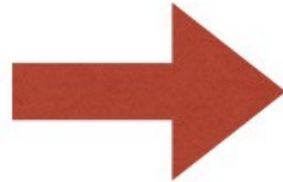
**Help humans interact with each other and/or devices.**



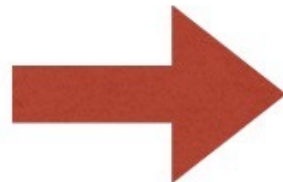
**Useful systems**

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

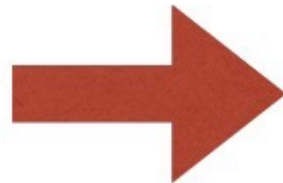
# NLP APPLICATIONS – TEXT OR DOCUMENT CATEGORIZATION



Sports



Politics



Science

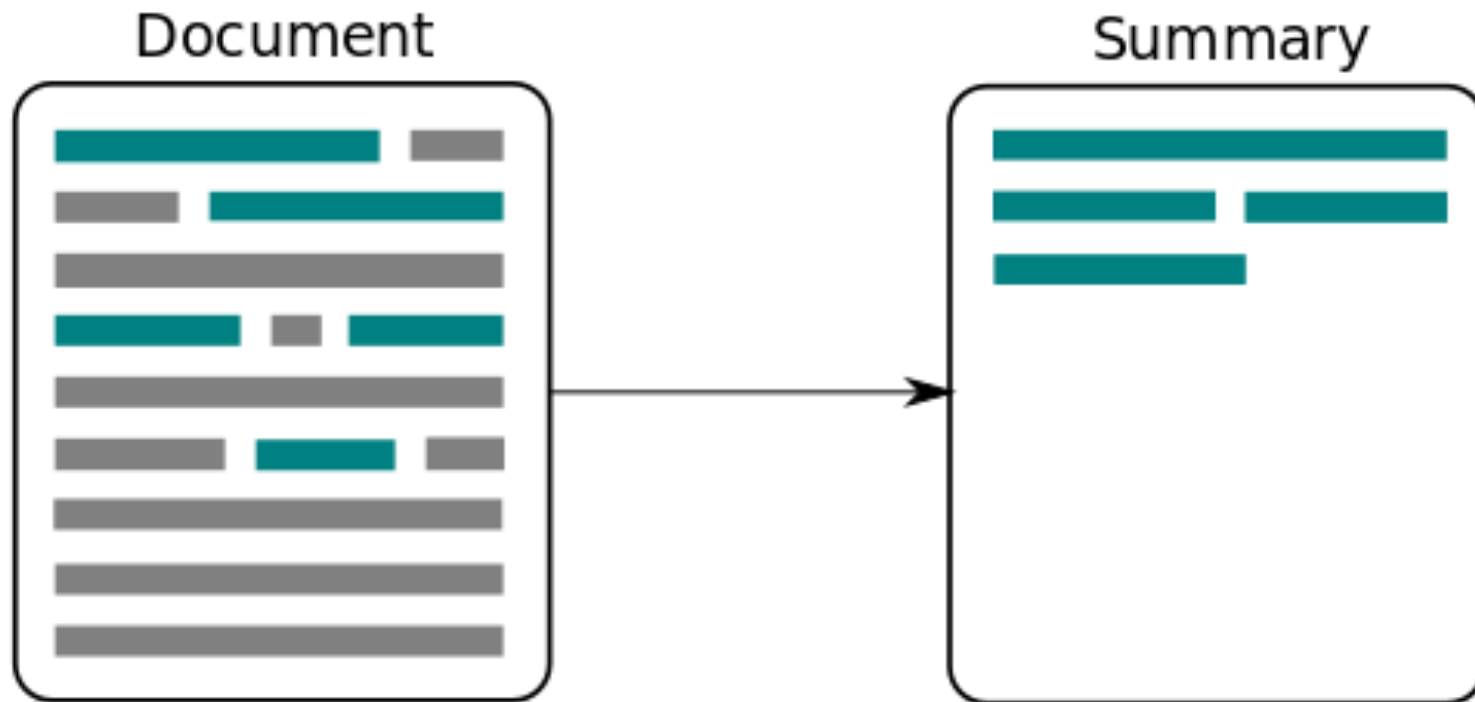
# NLP APPLICATIONS – INFORMATION EXTRACTION

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

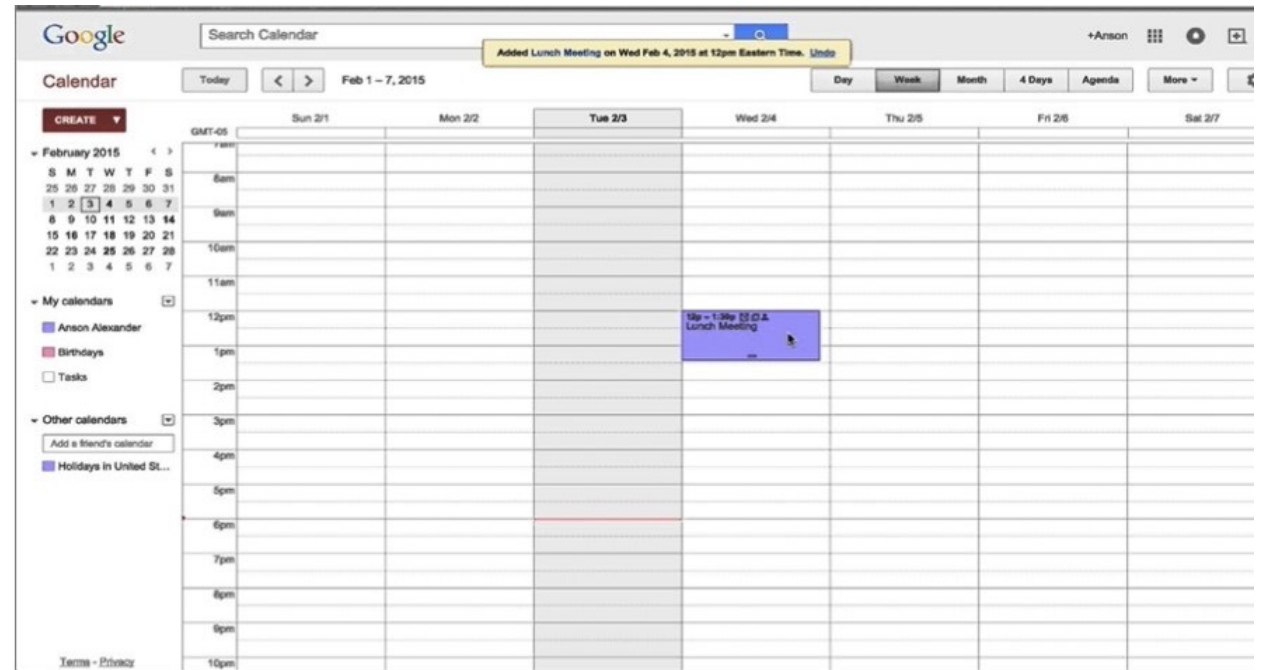
# NLP APPLICATIONS – SUMMARIZATION



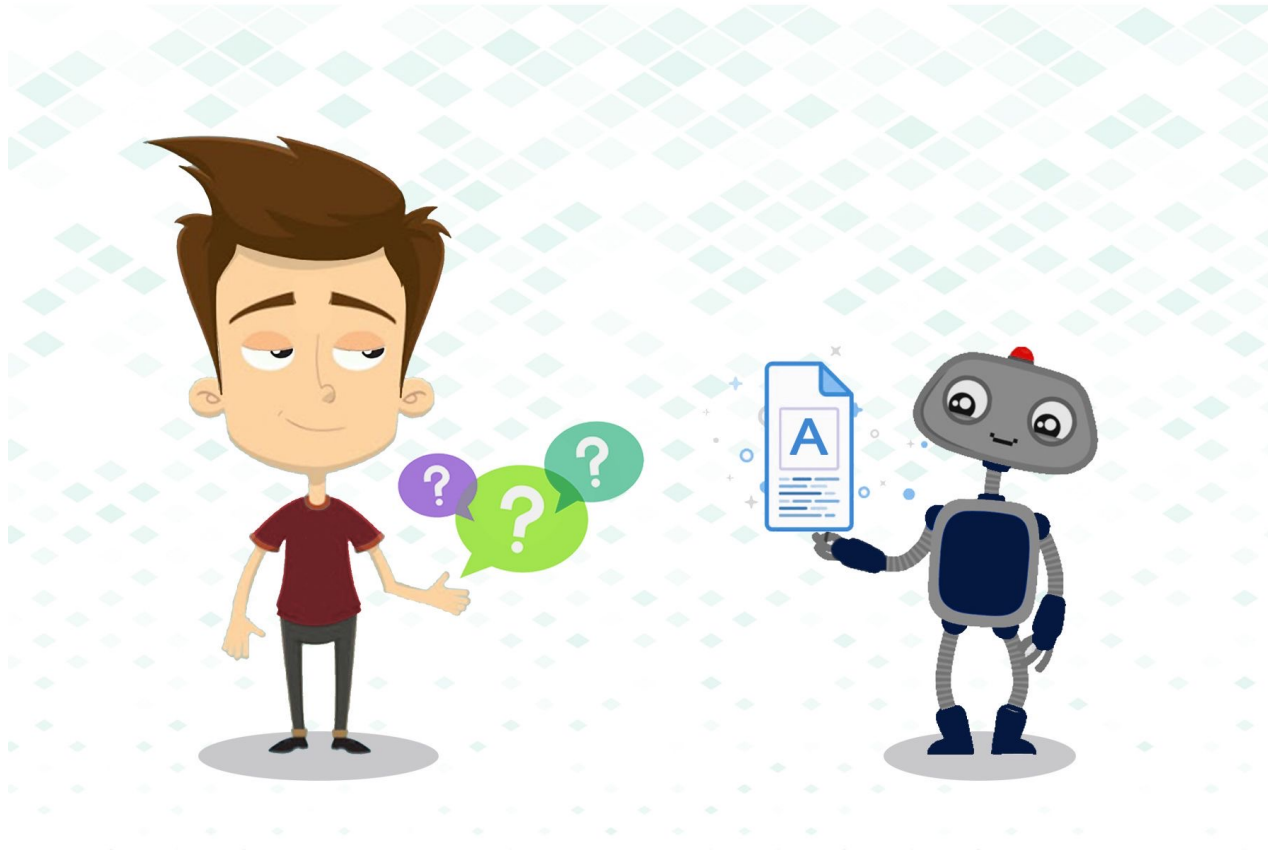
<https://hackernoon.com/summarization-with-wine-reviews-using-spacy-b49f18399577>

# NLP APPLICATIONS – VIRTUAL ASSISTANTS

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



# NLP APPLICATIONS – QUESTION ANSWERING



# 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

FRENCH - DETECTED

FRENCH

DUTCH

GERMAN

⌵

↔

ENGLISH

SPANISH

ARABIC



⌵


La grande illusion est un magnifique film de Jean Renoir, sorti en 1937. C'est aussi le titre d'un essai de Norman Angell, paru en 1910, dans lequel l'auteur anglais juge la guerre impossible du fait des liens économiques et financiers qui unissent les nations.


✕




La grande illusion is a magnificent film by Jean Renoir, released in 1937. It is also the title of an essay by Norman Angell, published in 1910, in which the English author considers war impossible because of economic and financial ties. that unite nations.

☆

261 / 5000 





# WHAT IS SPECIAL ABOUT NATURAL LANGUAGE?

## Linguistic 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 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 fail => 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>