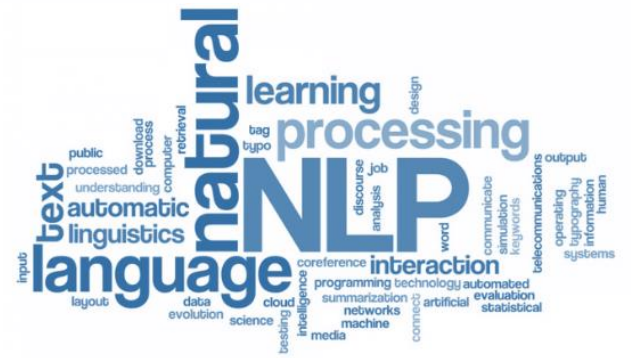
CENG 3526
Natural Language Processing

Lecture 1
Introduction to NLP - Fundamentals

Instructor
Bekir Taner Dinçer

Teaching Assistant
Selahattan Aksoy



MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
COMPUTER ENGINEERING

1

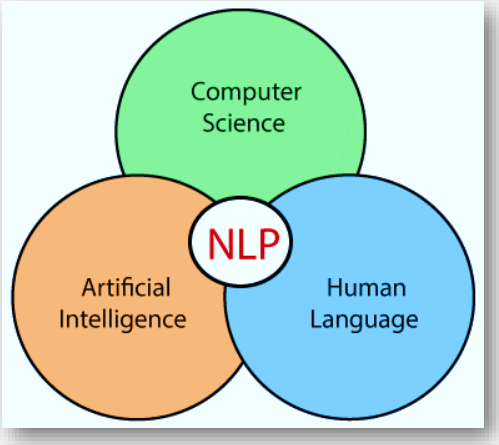
What is NLP?

Data, Information & Knowledge. Information Science. Artificial Intelligence. Data Science.

2

What NLP is : Functional Definition

Recap !

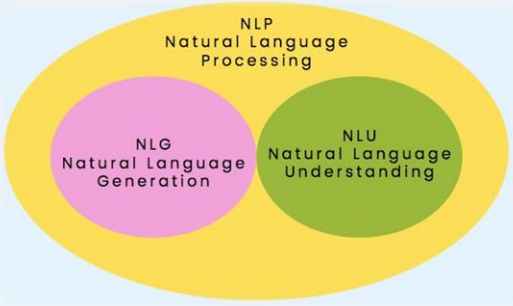


Natural language processing is **a subfield of**
linguistics, computer science, and artificial intelligence (AI)

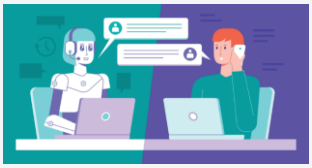
concerned with
the interactions between
computers and **human language**

What NLP is : The Goal

Recap !



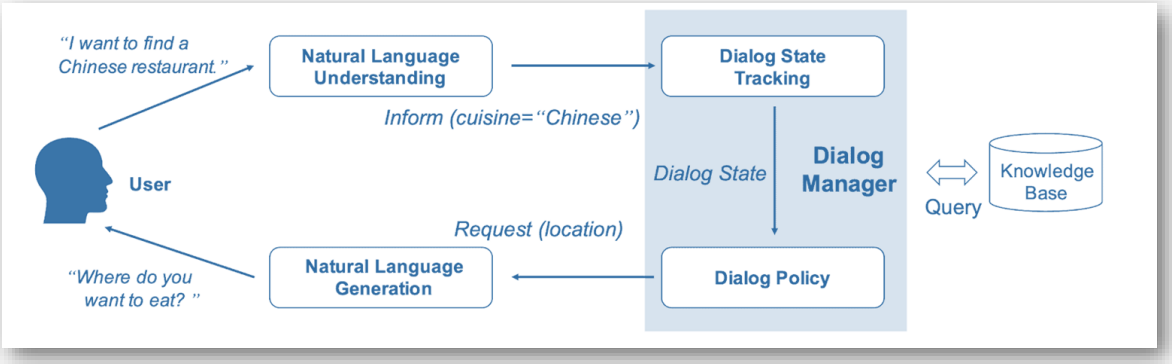
The goal
is
a computer capable of
"understanding" & "generating"
the contents of
written texts, and speech, ...



What NLP is : The Goal

Recap !

General Model of NLP – Conversation Agent



Natural Language Processing (NLP) : Use Cases in Industries

Education
- Machine Translation, Spell Checking and Grammar, etc.

Healthcare
- Speech Recog./Synthesis, Language Gener./Under., Question-Answering

Marketing/Advertising
- Machine Translation, Document Classification, Sentiment analysis, etc.

Pharmaceuticals/BioTech
- Document Classification, NER,Entity-Linking / Knowledge Graphs

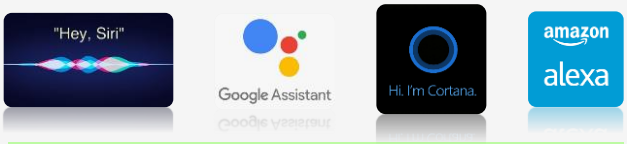
Banking/Finance
- Information Extraction, Text Summarization, NER etc.

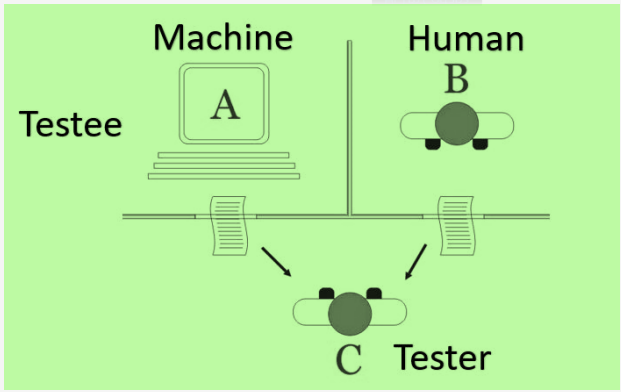
Miscellaneous
- Automotive, Defense & National Security, Food, Tourism, Law, etc.

Natural Language Processing (NLP) : Where are we now?

Turing Test

A method of inquiry in **artificial intelligence** (AI) for determining whether or not **a computer is capable** of **thinking like a human being**.

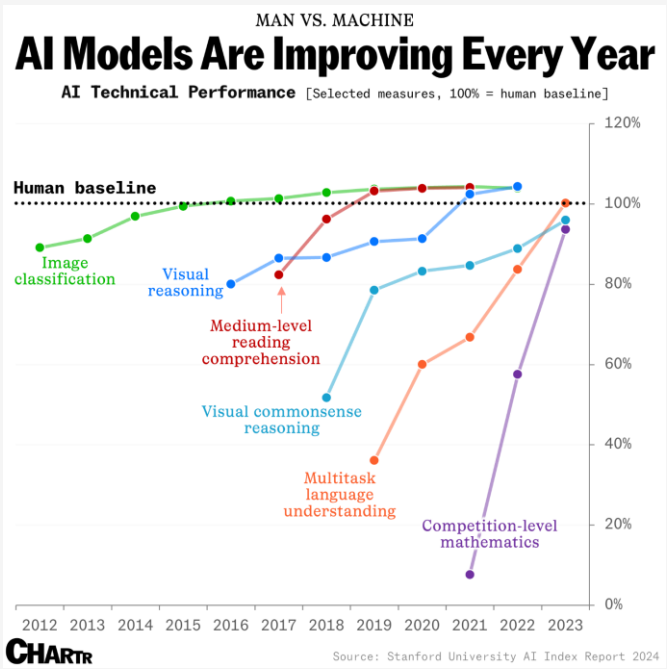




MSKU NLP

7

Where are we now?



MSKU CENG

8

Where are we now?

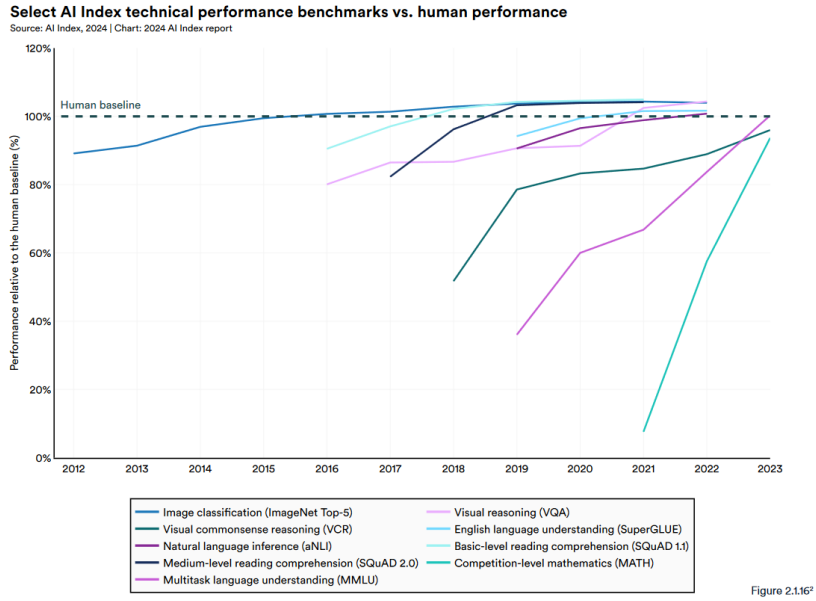


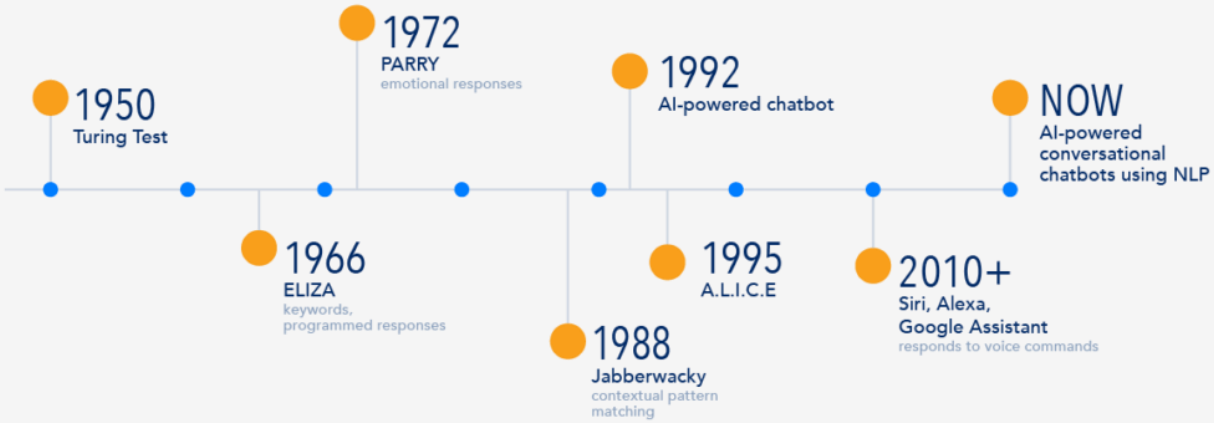
Figure 2.1.16[2]

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, "*The AI Index 2024 Annual Report*," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

MSKU CENG

9

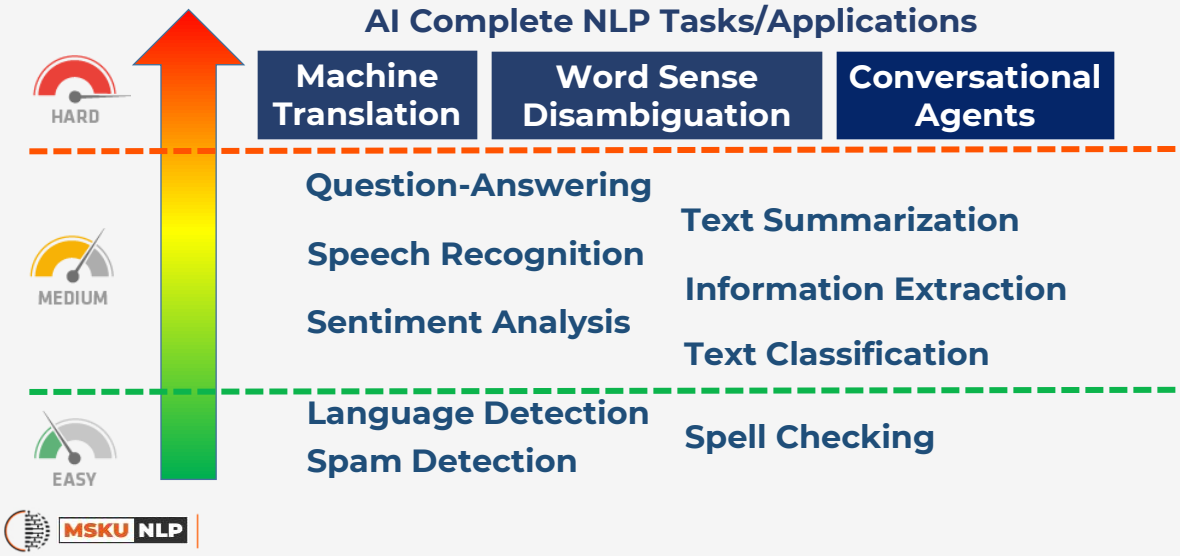
Natural Language Processing (NLP) : Where we are now?



MSKU NLP

11

Natural Language Processing (NLP) : NLP is Hard!



MSKU NLP

12

Natural Language Processing (NLP) : NLP is Hard, Why?

1. Ambiguity

Uncertainty in Meaning

Ayşe and Fatma are **sisters**.

Ayşe and Fatma are **mothers**.

Metaphors

My lawyer is a shark.

Idioms

He is as good as John Doe.

2. Common Sense/Knowledge

The facts that all humans are aware of

Dog bit man. ✓

Man bit dog. ✗

3. Creativity / Open Vocabulary

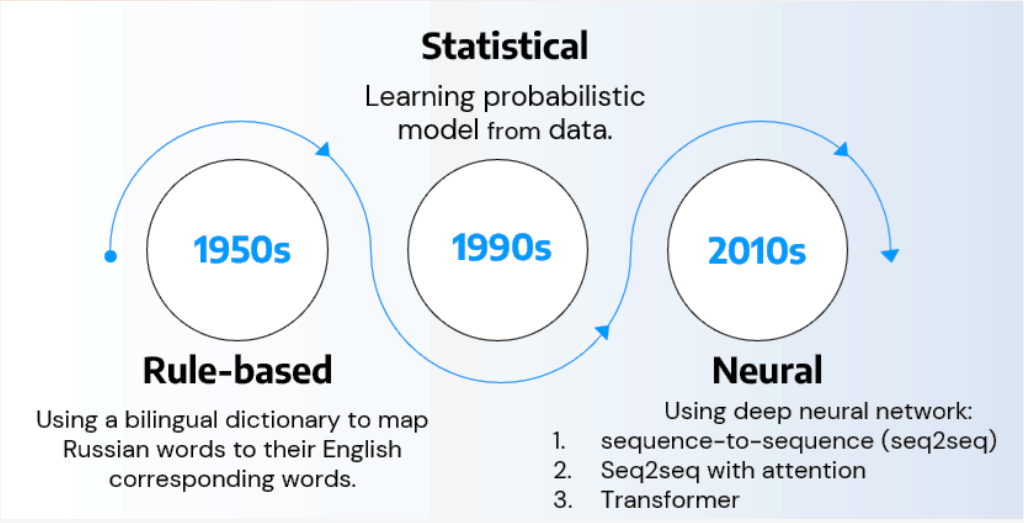
Poems, Genres (Literature) / Languages are generative.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. "The Winograd Schema Challenge." *The Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning* (2012)

MSKU NLP

13

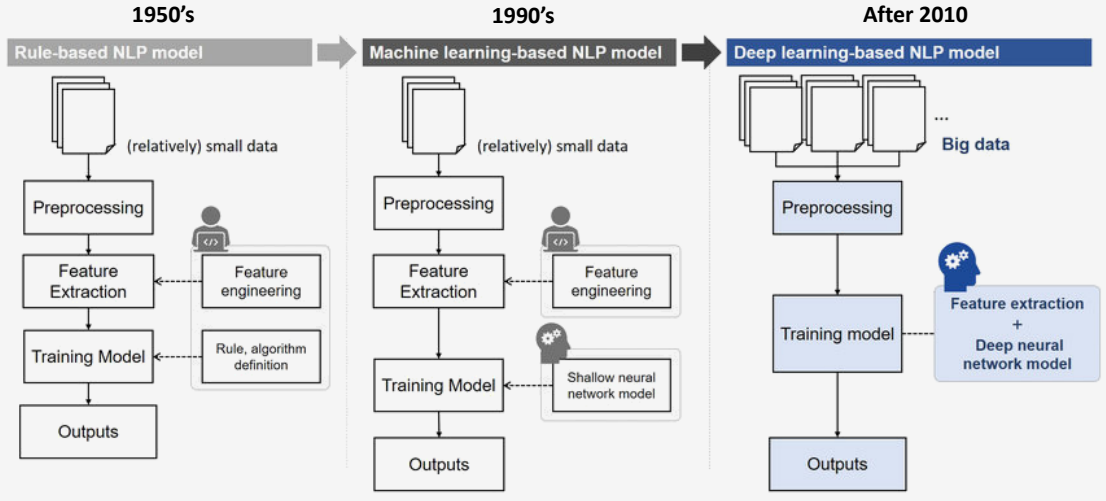
Natural Language Processing (NLP) : Methods



MSKU NLP

14

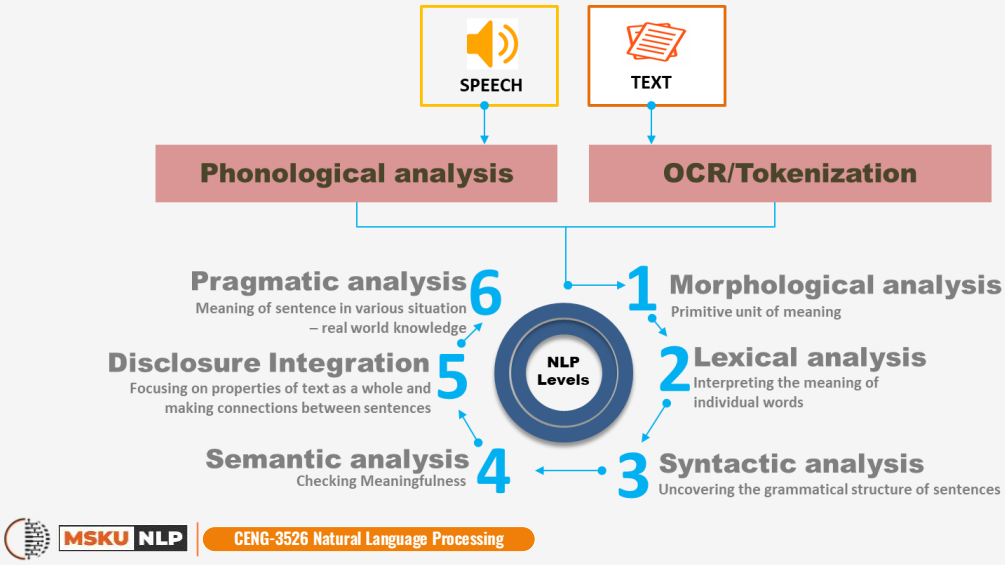
Natural Language Processing (NLP) : Methods



MSKU NLP

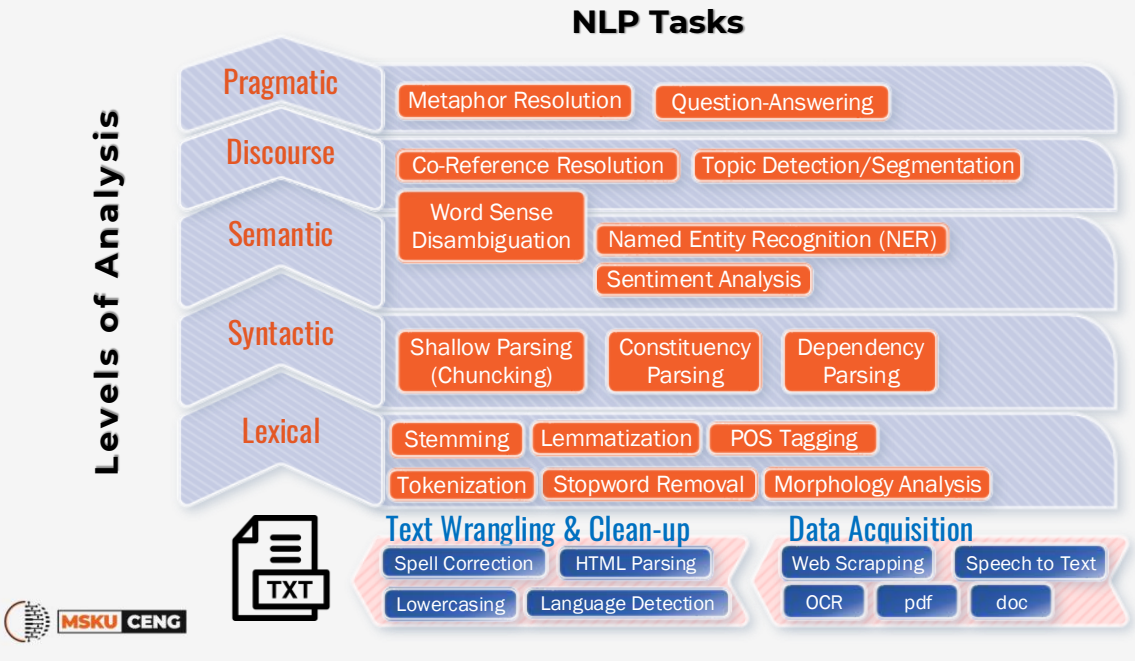
15

NLP : Levels of Analysis



16

NLP Tasks

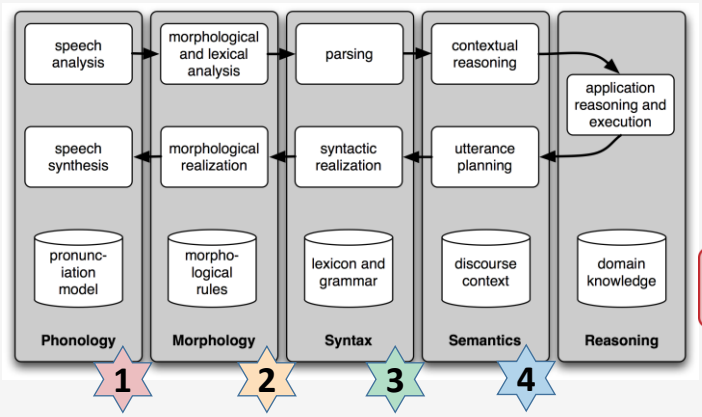


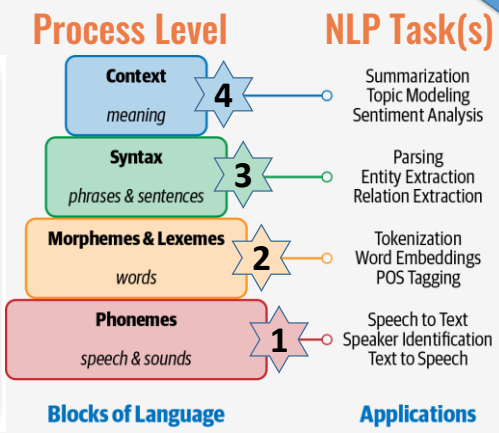
17

Natural Language Processing (NLP) : An NLP Walkthrough

More on Later !

NLP Task Pipeline for Conversation Agent

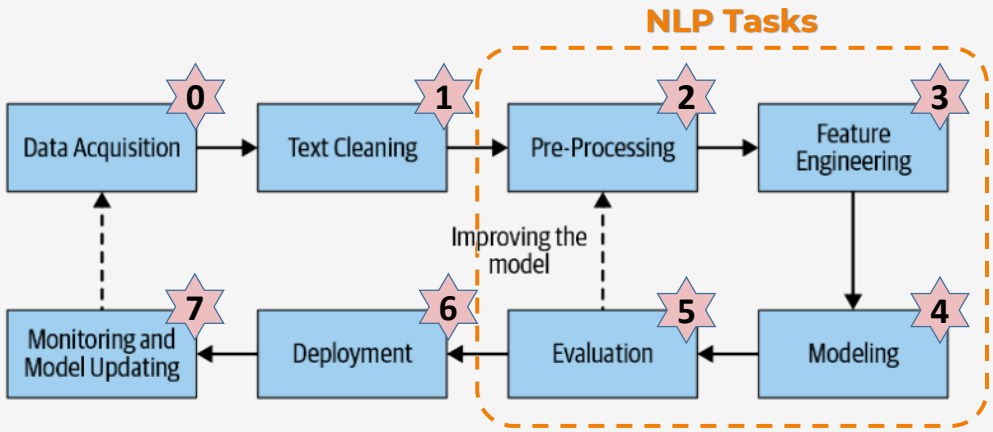




MSKU NLP

18

NLP : Pipelines for Real World NLP Applications



MSKU NLP | CENG-3526 Natural Language Processing

19

Text Representation

tokens, types and n-grams

20

Text Representation

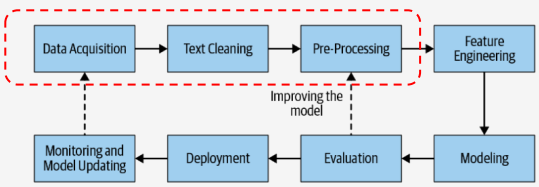
Text Representation in (NLP)

- is the process of **converting textual data into a numerical format** that can be understood and processed by machine learning algorithms.

Goal

- **transform raw text** into **a structured representation** so that the **semantic** and **syntactic information** contained within the text **is captured**.

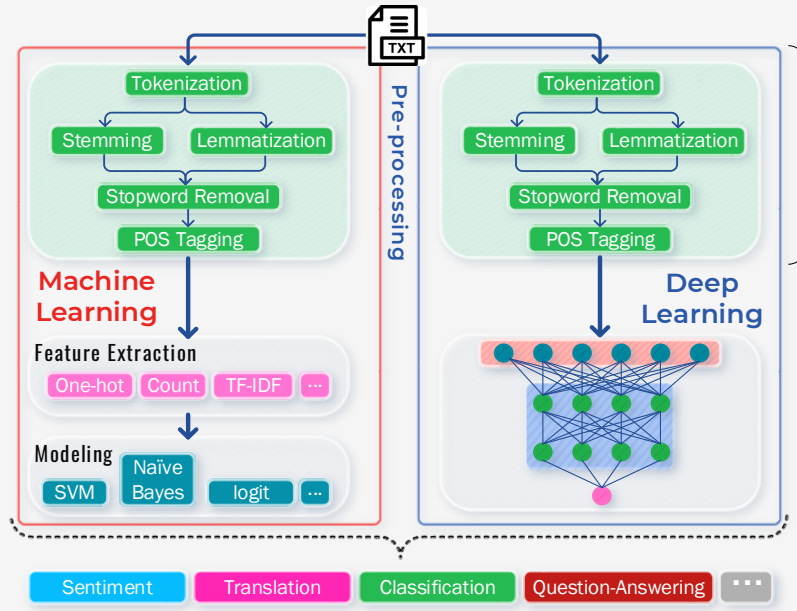
Text Representation Process



Common text representation techniques include:

- **Bag of Words** (BoW)
- **TF-IDF** (Term Frequency-Inverse Document Frequency):
- **Word Embeddings** (e.g., Word2Vec, GloVe)
- **Character-level Embeddings**
- **n-gram language models**
- **Document Embeddings** (Doc2Vec, or Deep Learning Models)

21



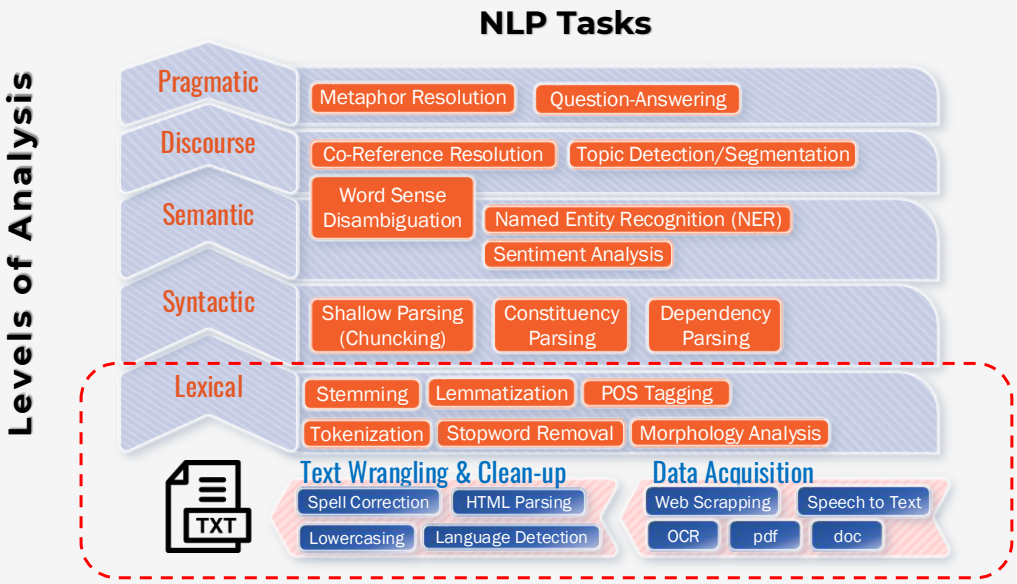
Common NLP Tasks

1. **Tokenization**
2. **Text Normalization**
3. **Text Cleaning**

=> Text Representation

MSKU CENG

22



MSKU CENG

23

Tokenization

CENG3526 - Week 1 - Text Representation and Pre-processing.ipynb

Breaks the text into individual words or **tokens**.

"The quick brown fox jumps over the lazy dog."

```
text = "The quick brown fox jumps over the lazy dog."

tokens = text.split()

print(tokens)
```

Split using whitespace as the separator

```
['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog.']
```

Tokens

Normalization

CENG3526 - Week 1 - Text Representation and Pre-processing.ipynb

1. **Tokenization**
2. **Text Normalization**
3. **Text Cleaning**

=> Text Representation

Converting text to a consistent format, such as lowercase or uppercase.

Before normalization:
"I can't believe this is happening! It's so awesome!“

After normalization:
“i can't believe this is happening! it's so awesome!"

Text Cleaning

CENG3526 - Week 1 - Text Representation and Pre-processing.ipynb

Text Cleaning is a crucial preprocessing step in NLP

- It involves **removing noise**, **inconsistencies**, and **irrelevant information** from text data, to improve the quality and consistency of the data, making it more suitable for downstream tasks like sentiment analysis, machine translation, and text summarization

1. **Tokenization**
2. **Text Normalization**
3. **Text Cleaning**

=> Text Representation

Text cleaning usually involves handling of the followings:

- Punctuations: commas, periods, question marks, etc.
- Contractions: Expanding contractions like "can't" to "cannot" and "don't" to "do not."
- Spelling errors: Identifying and correcting misspelled words.
- Special characters: Removing non-text characters like emojis, symbols, and control characters.
- HTML tags: Removing HTML tags from web pages.
- Dealing with noise: Removing noise like typos, OCR errors, or formatting inconsistencies.

Before cleaning:
"i can't believe this is happening! it's so awesome!"

After cleaning:
"i cannot believe this is happening it is so awesome"

Text Representation: Bag-of-Words Model (BOW)

Definition

- a Bag of Words (BoW) representation transforms a text document into a numerical vector, where each element corresponds to the frequency of a specific word in that document. This essentially treats the document as a collection of words, ignoring the order in which they appear.

Key Idea: Independence Assumption

The core assumption behind BoW is that the occurrence of one word in a document is independent of the occurrence of another word.

This means that the presence or absence of a word does not influence the probability of another word appearing.

Doc 1: "The quick brown fox jumps over the lazy dog."

Doc 2: "The lazy dog jumps over the quick brown fox."

Vocabulary:

- the
- quick
- brown
- fox
- jumps
- over
- lazy
- dog

BoW Vectors:

Word	Document 1	Document 2
the	2	2
quick	1	1
brown	1	1
fox	1	1
jumps	1	1
over	1	1
lazy	1	1
dog	1	1

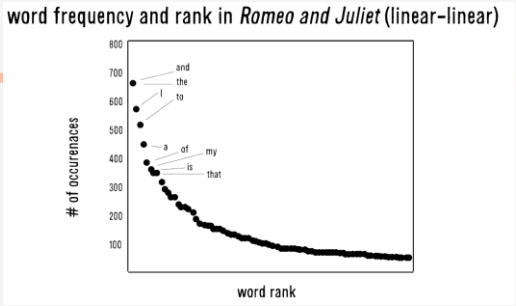
Zipf's Law

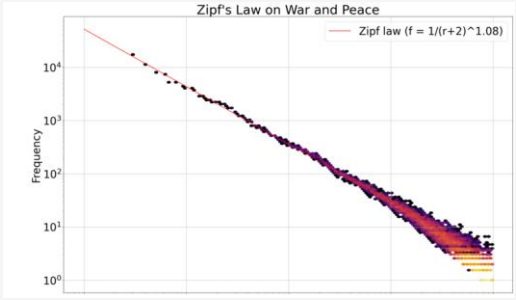
Zipf's Power Law

The product of the frequency of words (f) and their ranks (r) is approximately constant.

For English:

$$f = C \times \frac{1}{r}$$
$$C \cong N/10$$





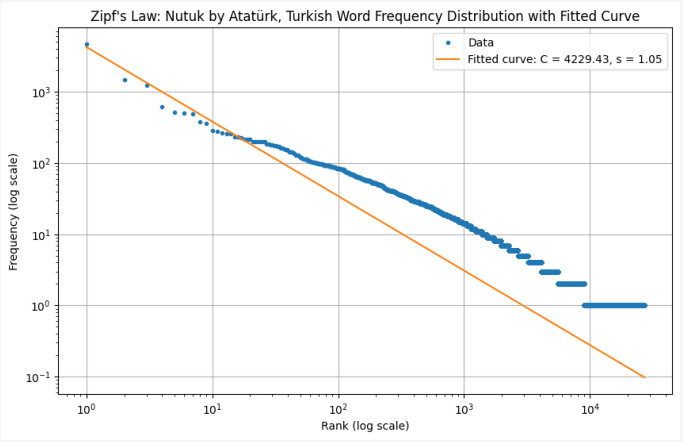
Zipf's Law: Turkish, Nutuk by Atatürk

Zipf's Power Law

The product of the frequency of words (f) and their ranks (r) is approximately constant.

For English:

$$f = C \times \frac{1}{r}$$
$$C \cong N/10$$



WordCloud: Turkish, Nutuk by Atatürk

What Nutuk is about?
Can you get the main theme/topic from the WordCloud?
Can you make it apparent?